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

Ward Whitt
Industrial Engineering and Operations Research, Columbia University, ww2040@columbia.edu

Wei You
Industrial Engineering and Operations Research, Columbia University, wy2225@columbia.edu

Queueing applications are often complicated by dependence among interarrival times and service times, e.g., when there are multiple customer classes with class-dependent service-time distributions, or when arrivals are departures or overflows from other queues or superpositions of such complicated processes. We show that the robust queueing approach for single-server queues proposed by Bandi, Bertsimas and Youssef (2015) can be extended to describe the impact of dependence among interarrival times and service times on customer waiting times and the remaining workload in service time as a function of the traffic intensity in the queue. Thus, robust queueing can be useful to develop performance approximations for queueing networks and other complex queueing systems.

Key words: robust queueing theory, queueing approximations, queueing network analyzer, dependence among interarrival times and service times, indices of dispersion

History: May 2, 2016

1. Introduction

Robust optimization is proving to be a useful approach to optimization problems for complex stochastic models; e.g., see Bertsimas et al. (2011), Ben-Tal et al. (2009) and Beyer and Sendhoff (2007). Bandi et al. (2015) have applied this approach to create a robust queueing (RQ) theory, which can be used to generate approximations for performance measures in complex queueing systems, including networks of queues as well as single queues. They show that this approach provides an alternative way to develop relatively simple performance approximations like those in the queueing network analyzer in Whitt (1983b).

The starting point of their RQ approach is the representation of the customer waiting times in a general stable stationary $G/G/1$ single-server queue with unlimited waiting space and the first-come first-served (FCFS) service discipline as the maximum of a sequence of partial sums, using the Loynes (1962) reverse-time construction. As we explain in §2,

$$W_n = M_n \equiv \max\{S_k : 0 \le k \le n\}, \quad n \ge 1, \tag{1}$$

where S_k is the k^{th} partial sum with $S_0 \equiv 0$ and $\equiv$ denotes equality by definition. Instead of a detailed stochastic model, they place deterministic constraints on the possible interarrival times and service times thorough the partial sums S_k. Then the RQ optimization problem is solved to yield an upper bound on the waiting time, which can be a basis for approximations of the mean steady-state waiting time.

In any robust optimization problem, a critical role is played by the deterministic constraints representing the stochastic elements. Given the representation of the waiting times in terms of partial sums, Bandi et al. (2015) base their constraints on the central limit theorem (CLT) for partial sums S_k. Treating the partial sums S_k^a of the interarrival times U_k and the partial sums S_k^s of the service times V_k separately leads to the two uncertainty sets

$$\begin{aligned} \mathcal{U}^a &\equiv \{(U_1,\ldots,U_n) : S_k^a \ge k m_a - \sqrt{k} b_a,\ 1 \le k \le n\}, \quad \text{and} \\ \mathcal{U}^s &\equiv \{(V_1,\ldots,V_n) : S_k^s \le k m_s + \sqrt{k} b_s,\ 1 \le k \le n\}, \end{aligned} \tag{2}$$

where $m_a \equiv E[U_k]$ and $m_s \equiv E[V_k]$, while b_a and b_s are parameters to be specified. Thinking of the $GI/GI/1$ model in which the interarrival times U_k and service times V_k come from independent sequences of independent and identically distributed (i.i.d.) random variables with finite variances σ_a^2 and σ_s^2, the CLT suggests that $b_a = \beta\sigma_a$ and $b_s = \beta\sigma_s$ for an appropriate constant β, which measures the number of standard deviations away from the mean in a Gaussian approximation.

The RQ optimization with objective function (1) subject to the constraints in (2) has a simple solution, which depends on only a few parameters; see Theorem 1 in §2.

In this paper we develop new RQ formulations for the general stationary $G/G/1$ model that provide ways to expose the impact of dependence among the interarrival times and service times on the mean steady-state waiting time and its continuous-time analog, the virtual waiting time or workload. This dependence commonly arises in queueing networks and multi-class settings, as illustrated by Fendick et al. (1989, 1991). In turn, multi-class queues and associated queueing network models are applied widely, e.g., to analyze the performance of communication, healthcare and production systems, as in Badidi et al. (2005), Cochran and Roche (2009), Gayon et al. (2009), Hall (2006) and Hall (2012).

We show that the new RQ is intimately connected to previous performance approximations for queues based on indices of dispersion (scaled variance-time curves) in Fendick and Whitt (1989). There it was shown that the impact of the dependence in the offered traffic upon the mean steady-state workload $E[Z_\rho]$ as a function of the traffic intensity ρ can be approximated characterized by the *index of dispersion for work* (IDW), $\{I_w(t) : t \geq 0\}$. The IDW $I_w(t)$ is a scaled version of the variance of the total input of work over the interval $[0, t]$; see §5 and §7.4. The main idea is that the time interval over which dependence has impact on the steady-state performance should increase as the traffic intensity ρ increases; i.e., (3) and (9) in Fendick and Whitt (1989) suggest the approximation

$$E[Z_\rho] \approx \frac{\tau \rho I_w(t(\rho))}{2(1-\rho)}, \quad 0 < \rho < 1, \tag{3}$$

where the mean service time is τ and $t(\rho)$ is an increasing function of ρ on the interval $(0, 1)$ with $t(\rho) \to \infty$ as $\rho \to 1$ and $t(\rho) \to 0$ as $\rho \to 0$. Heavy-traffic and light-traffic limits show that approximation (3) is asymptotically correct as $\rho \uparrow 1$ and as $\rho \downarrow 0$; see (45). The new RQ provides ways to define the function $t(\rho)$ in the challenging intermediate cases; see Theorem 5 and the examples in §8.

Here is how the rest of this paper is organized: After reviewing the basic RQ approach from Bandi et al. (2015) in §2, we relate it to established theory for the $GI/GI/1$ queue and associated heavy-traffic (HT) limits in §3. We then present a version of RQ involving one uncertainty set instead of two in §4. We introduce a new RQ formulation for the continuous-time workload process in §5.

In §6 we propose the new uncertainty sets leading to RQ formulations exposing the impact of dependence among the interarrival times and service times. We also provide background on the supporting functional CLT (FCLT) for the arrival and service processes and the HT FCLT for the waiting time and workload processes. In §7 we establish theory for the RQ with dependence, showing the connections to the indices of dispersion and associated approximations. In §8 we illustrate with numerical examples. In §9 we draw conclusions. In §9.2 we discuss how the new RQ can be applied.

2. Robust Queueing for the Single Server Queue

We now review the robust queueing (RQ) approach developed in Bandi et al. (2015), elaborating upon (1) and (2). We consider customer waiting times (before receiving service) in the single-server queue with unlimited waiting space and the FCFS service discipline; e.g., as in Chapter X of Asmussen (2003) and Chapter 6 of Sigman (1995). The waiting time of customer n satisfies the recursion

$$W_n = (W_{n-1} + V_{n-1} - U_{n-1})^+ \equiv \max\{W_{n-1} + V_{n-1} - U_{n-1}, 0\}, \tag{4}$$

where V_{n-1} is the service time of customer $n-1$ and U_{n-1} is the interarrival time between the arrival times of customers $n-1$ and n. If we initialize the system by having a customer 0 arrive to find an empty system, then W_n can be represented in (1) using reverse-time indexing with $S_k \equiv X_1 + \cdots + X_k$ and $X_k \equiv V_{n-k} - U_{n-k}$, $1 \le k \le n$.

If we extend the reverse-time construction indefinitely into the past from a fixed present state, then $W_n \uparrow W$ with probability 1 as $n \to \infty$, allowing for the possibility that W might be infinite. For the stable stationary $G/G/1$ case with $E[U_k] < \infty$, $E[V_k] < \infty$ and $\rho \equiv E[V_k]/E[U_k] < 1$, $P(W < \infty) = 1$; e.g., see Loynes (1962) or §6.2 of Sigman (1995).

Bandi et al. (2015) developed approximations for W_n in (1) and the limit W for the $G/G/1$ model by performing the maximization in (1) subject to the constraints in (2). They also provided an extension to cover the heavy-tailed case, where finite variances might not exist; then $\sqrt{k}$ in (2) is replaced by $k^{1/\alpha}$ for $0 < \alpha \le 2$.

The RQ formulation in (1) and (2) is attractive because the optimization has a simple solution in which all constraints are satisfied as equalities.

Theorem 1. (*worst-case waiting time, Theorem 2 of Bandi et al. (2015)*) *For the stationary $G/G/1$ single-server queue, the solution of the RQ optimization* (1) *with uncertainty sets in* (2), *where $m = E[V_k] - E[U_k] < 0$ and $b \equiv b_s + b_a > 0$, is*

$$\begin{aligned} W_n^* &= \max\{mk + b\sqrt{k} : 0 \le k \le n\} \\ &\le \max\{mx + b\sqrt{x} : x \ge 0\} = mx^* + b\sqrt{x^*} = \frac{b^2}{4|m|} \quad \textit{and} \quad x^* = \frac{b^2}{4m^2}, \end{aligned} \tag{5}$$

In addition, W_n^* is maximized at one of the integers immediately above or below x^* for all $n \ge x^*$.

Remark 1. (when steady-state is reached) Unlike the stochastic $G/G/1$ model, where steady state is approached over time, W_n^* in (5) is actually constant for $n \ge x^*$. The deterministic time x^* in RQ is analogous to the relaxation time for the stochastic single server queue, as discussed in Cohen (1982), and can serve as an approximation of it. The scaling by $1-\rho$ in W_n^* for $n \ge x^*$ in (5) is consistent with the spatial scaling in the heavy-traffic limit, while the scaling by $(1-\rho)^2$ in x^* is consistent with the time scaling in the heavy-traffic limit and thus the relaxation time; e.g., see §3 and (28) and (29) in §6.1.

3. The $GI/GI/1$ Queue and Heavy Traffic

It turns out that RQ is intimately connected to heavy-traffic theory for the single-server queue, as in Ch. 9 of Whitt (2002). Hence, we provide a quick review of the $GI/GI/1$ special case and associated heavy-traffic approximations.

Let the model be specified by (i) the traffic intensity ρ, $0 < \rho < 1$, and (ii) two independent sequences of i.i.d. nonnegative random variables $\{V_k : k \ge 1\}$ and $\{U_k : k \ge 1\}$ with $E[V_k] = E[U_k] =$

1, $Var(V_k) = \sigma_s^2 < \infty$ and $Var(U_k) = \sigma_a^2 < \infty$. In the model with traffic intensity ρ, we use the scaled interarrival times be $\rho^{-1} U_k$, $k \ge 1$.

The associated squared coefficients of variation (scv's, variance divided by the square of the mean) are then $c_s^2 = \sigma_s^2$ and $c_a^2 = \sigma_a^2$ (ρ cancels out). Then for $X_k \equiv X_k(\rho)$, $E[X_k] \equiv m \equiv m(\rho) = (1 - \rho^{-1}) = -(1-\rho)/\rho < 0$ and $Var(X_k) \equiv \sigma_x^2 \equiv \sigma_x^2(\rho) = \sigma_s^2 + \sigma_a^2(\rho) = c_s^2 + \rho^{-2} c_s^2 < \infty$. The reverse-time construction in (1) is not needed in this setting because the random variables X_k are i.i.d.

An exact expression for the distribution of the steady-state waiting time W is known, but complicated in general. As reviewed in Abate et al. (1993), where computation is discussed, the Laplace transform is given by the Spitzer formula

$$E[e^{-sW}] = \exp\{\sum_{k=1}^{\infty} k^{-1} E[e^{-s(S_k)^+} - 1]\},$$

which implies that

$$E[W] = \sum_{k=1}^{\infty} E[(S_k)^+]/k \quad \text{and} \quad Var(W) = \sum_{k=1}^{\infty} E[((S_k)^+)^2]/k$$

but the distribution of S_k^+ is complicated in general. For the $M/GI/1$ special case with a Poisson arrival process, we have $c_a^2 = 1$ and the classic Pollaczek-Khintchine formula (for $E[V_k] = 1$)

$$E[W(\rho)] = \frac{\rho(c_s^2 + 1)}{2(1-\rho)} = \frac{\sigma_x^2(1)}{2|m(\rho)|}.$$

The standard heavy-traffic (HT) approximation is obtained by letting $\rho \uparrow 1$, e.g., see Chapters 5 and 9 of Whitt (2002). In that limit, the final PK formula has the limit

$$\lim_{\rho \uparrow 1} \{(1-\rho) E[W(\rho)]\} \to \frac{c_s^2 + 1}{2} = \frac{\sigma_x^2(1)}{2}.$$

The HT logic leads to approximating the sequence of partial sums $\{S_k : k \ge 0\}$ by Brownian motion (BM) with drift, $\{\sigma B(t) + mt : t \ge 0\}$, where $m \equiv m(\rho)$, $\sigma \equiv \sigma_x(1) = \sqrt{c_a^2 + c_s^2}$ and the associated sequence of waiting times by reflected Brownian motion (RBM). That leads to the associated HT approximation for the steady-state waiting time

$$W_{HT} \stackrel{\mathrm{d}}{=} M_{HT} \equiv \sup\{\sigma B(t) + mt : t \ge 0\},$$

for $m = m(\rho) = 1 - \rho^{-1} < 0$ and $\sigma^2 = \sigma^2(1) = (c_a^2 + c_s^2) > 0$, which has an exponential distribution, i.e.,

$$P(W_{HT} > x) = e^{2xm/\sigma^2}, \quad x \ge 0, \quad \text{and} \quad E[W_{HT}] = \frac{\sigma^2(1)}{|2m(\rho)|} = \frac{\rho(c_s^2 + c_a^2)}{|2(1-\rho)|}. \tag{6}$$

Insightful derivations of (6) for RBM as well as the transient distributions of RBM, exploiting martingales and stochastic calculus, are given in Harrison (1985); see §1.8, §1.9, §3.6 and §5.6 plus the background material.

Just as for the RQ formulas, the HT formulas can benefit from tuning. For example, the HT approximation could be taken to be $E[W_{HT}] = \sigma^2(\rho)/|2m(\rho)| = \rho(c_s^2 + \rho^{-2}c_a^2)/|2(1-\rho)|$, which corresponds to the Kingman (1962) upper bound. The ratio of these two mean formulas goes to 1 as $\rho \to 1$.

4. One Uncertainty Set Instead of Two

From §2, it is evident that the waiting times depend on the service times and interarrival times only through their difference X_n. Thus, instead of the two uncertainty sets in (2), we propose the single uncertainty set

$$\mathcal{U}^x \equiv \{\tilde{X}_n : S_k^x \le (k \vee k_L)E[X_k] + b_x\sqrt{k \vee k_L}, 1 \le k \le n\}, \tag{7}$$

where $a \vee b \equiv \max\{a, b\}$, $\tilde{X}_n \equiv (X_1, \ldots X_n) \in \mathbb{R}^n$, for $S_k^x = S_k^s - S_k^a$ and $X_k \equiv V_{n-k} - \rho^{-1}U_{n-k}$. To avoid excessively strong constraints for small values of k, not justified by the CLT, we replace k by $k \vee k_L$ on the right in (7), but the lower bound k_L has no impact if chosen appropriately.

The conclusions of Theorem 1 remain unchanged with the uncertainty set changed from (2) to (7), but there is a significant difference in the interpretation of the constant b.

Corollary 1. (*worst-case waiting time with a single uncertainty set*) *For the stationary $G/G/1$ single-server queue with $m = E[V_k] - \rho^{-1}E[U_k] < 0$, the solution of the RQ optimization* (1) *with uncertainty sets in* (7) *is*

$$\begin{aligned} W_n^* &\equiv \max\{W_n : \tilde{X}_n \in \mathcal{U}^x\} = \max\{m(k \vee k_L) + b_x\sqrt{k \vee k_L} : 1 \le k \le n\} \\ &\le \max\{my + b_y\sqrt{y} : y \ge k_L\} = my^* + b_x\sqrt{y^*} = \frac{b_x^2}{4|m|} \quad \textit{and} \quad y^* = \frac{b_x^2}{4m^2}, \end{aligned} \tag{8}$$

provided that $k_L < y^*$. In addition, W_n^* is maximized at one of the integers immediately above or below y^* for all $n \geq y^*$.

COROLLARY 2. (*the $GI/GI/1$ queue with a single uncertainty set*) *If, in addition to the assumptions of Corollary* 1, *the model is $GI/GI/1$, where the service times are independent of the interarrival times, and if we let $b = \beta\sqrt{Var(X_1)}$, then*

$$b_x = \beta\sqrt{Var(V_1) + \rho^{-2}Var(U_1)}, \tag{9}$$

If we furthermore let $\beta \equiv \sqrt{2}$, then formula (8) *with* (9) *agrees with the Kingman (1962) upper bound for the mean wait EW, which is asymptotically correct in the HT limit as $\rho \to 1$. In contrast, the RQ in Theorem 1 is not asymptotically correct in HT if we let $b_s = \beta(\sqrt{Var(V_1)}$ and $b_a = \beta(\sqrt{Var(U_1)}$; then we would have $b = b_s + b_a = \beta(\sigma_s + \rho^{-1}\sigma_a)$ instead of the asymptotically correct form of $b = \sqrt{b_s^2 + b_a^2} = \beta(\sqrt{\sigma_s^2 + \rho^{-2}\sigma_a^2})$ in* (9).

Bandi et al. (2015) used the two constants b_a and b_s in (2) as tuning control parameters to develop approximations, e.g., by doing statistical fitting with data. Corollaries 1 and 2 suggest doing the same with (9) and the single parameter β in (9).

5. The Continuous-Time Workload

We now extend the RQ approach to the continuous-time workload in the $G/G/1$ single-server queue. The workload at time t is the amount of unfinished work in the system at time t; it is also called the virtual waiting time because it represents the waiting time a hypothetical arrival would experience at time t. The workload is more general than the virtual waiting time because it applies to any work-conserving service discipline. We consider the workload primarily because it can serve as a convenient more tractable alternative to the waiting time, as shown in Fendick and Whitt (1989).

Given a sequence $\{(U_k, V_k)\}$ of interarrival times and service times, the arrival counting process can be defined by

$$A(t) \equiv \max\{k \geq 1 : U_1 + \cdots + U_k \leq t\} \quad \text{for} \quad t \geq U_1 \tag{10}$$

and $A(t) \equiv 0$ for $0 \le t < U_1$, while the total input of work is

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \ge 0, \tag{11}$$

and the remaining workload at time t, starting empty at time 0, is

$$Z(t) \equiv Y(t) - t - \inf\{Y(s) - s : 0 \le s \le t\}, \quad t \ge 0. \tag{12}$$

We start with the stationary sequence of mean-1 variables $\{(U_k, V_k)\}$ and insert the traffic intensity by letting the interarrival times be $\rho^{-1}U_k$. That makes $A_\rho(t) = A(\rho t)$, $Y_\rho(t) = Y(\rho t)$ and $Z_\rho(t) = Z(\rho t)$, where A, Y and Z are defined in terms of the mean-1 variables as in (10)-(12). Finally, we let these continuous-time processes be time-stationary versions; see Sigman (1995) for the technical details about stationary random marked point processes.

5.1. Another Reverse-Time Construction

As in §6.3 of Sigman (1995), we again use a reverse-time construction to represent the workload in a single-server queue as a supremum, so that the RQ optimization problem becomes a maximization over constraints expressed in an uncertainty set, just as before, but now it is a continuous optimization problem. Let $Z_\rho(t)$ be the workload at time 0 of a system that started empty at time $-t$. Then $Z_\rho(t)$ can be represented as

$$Z_\rho(t) \equiv \sup\{Y(\rho s) - s : 0 \le s \le t\}, \quad t \ge 0, \tag{13}$$

where $Y(s)$ is defined as (11), but is interpreted as the total work in service time to enter over the interval $[-s, 0]$. That is achieved by letting V_k be the k^{th} service time indexed going backwards from time 0 and $A(s)$ counting the number of arrivals in the interval $[-s, 0]$. Then $N_\rho(s) \equiv Y(\rho s) - s$ is the net input over the interval $[-s, 0]$ with traffic intensity ρ. Paralleling the waiting time in §2, $Z_\rho(t)$ increases monotonically to Z_ρ as $t \to \infty$. In an appropriate stationary framework, Z_ρ corresponds to the steady-state workload with traffic intensity $\rho < 1$ and satisfies $P(Z_\rho < \infty) = 1$; see §6.3 of Sigman (1995).

5.2. Robust Queueing in Continuous Time

Paralleling (7), (13) specifies an RQ optimization problem with the uncertainty set

$$\mathcal{U}^z \equiv \{\tilde{N}_\rho(t) : N_\rho(s) \le -(1-\rho)(s \vee t_L) + b_z\sqrt{s \vee t_L},\ 0 \le s \le t\}, \tag{14}$$

where we regard $\tilde{N}_\rho(t) \equiv \{N_\rho(s) : 0 \le s \le t\}$ as an arbitrary real-valued function on the interval $[0,t]$. We again include the lower bound, t_L here, but again it does not affect the RQ optimization if chosen appropriately.

As in §2, our formulation is motivated by a CLT. In §6.1 we will show that $Y(s)$ in (11) obeys a CLT, which supports (14). The same reasoning as before yields the following analog of Theorem 1 and Corollary 1.

Corollary 3. (*worst-case workload for the single-server queue*) *For the stationary* $G/G/1$ *single-server queue, the solution of the RQ optimization* (13) *with uncertainty sets in* (14) *is*

$$\begin{aligned} Z_\rho^*(t) &= \max\{-(1-\rho)(s \vee t_L) + b_z\sqrt{s \vee t_L} : 0 \le s \le t\} \\ &= -(1-\rho)x^* + b_z\sqrt{x^*} = \frac{b_z^2}{4|1-\rho|} \quad \textit{and} \quad x^* \equiv x^*(\rho) = \frac{b_z^2}{4(1-\rho)^2} \end{aligned} \tag{15}$$

for all $t \ge x^*$ *provided that* $t_L \le x^*$.

Formula (15) coincides with (8) if $b_z = \rho b_x$, so that the two RQ frameworks are essentially equivalent. That should not be surprising, because the steady-state workload is the same as the steady-state waiting time in the $M/GI/1$ queue and the HT limit is the same as for the waiting time in the $GI/GI/1$ queue. Hence, (15) also can be compared to the formulas in §3.

6. Stochastic Dependence

We can extend the RQ formulations for the general $G/G/1$ model in §4 and §5 to allow time dependence and stochastic dependence in the interarrival times and service times by expressing the uncertainty sets in (7) and (14) directly in terms of the means and variances, in particular, respectively as

$$\mathcal{U}^{x'} \equiv \{\tilde{X}_n : S_k^x \le E[S_k^x] + b_x'\sqrt{Var(S_k^x)}, \quad 1 \le k \le n\}. \tag{16}$$

and

$$\mathcal{U}^{z'} \equiv \{\tilde{N}_{\rho}(t) : N_{\rho}(s) \leq E[N_{\rho}(s)] + b'_z\sqrt{Var(N_{\rho}(s))},\, 0 \leq s \leq t\}. \tag{17}$$

For simplicity, we omit the lower bounds on the indices k and t in (16) and (17), but they could be included. For the $GI/GI/1$ model with its i.i.d. assumptions, in the case of finite variances, the previous case in (7) emerges as a special cases with and $b'_x = b_x/\sqrt{Var(X_k)}$. These are the approximations we would use if we simply chose to ignore the dependence; i.e., they correspond to using the stationary-interval method for approximating the arrival and service processes in Whitt (1982).

Remark 2. (dependence between interarrival times and service times) For the general $G/G/1$ model, the RQ uncertainty set (16) is more general than the RQ uncertainty set (2) because it can capture the impact of correlations between the interarrival times and service times, as can be seen from the variance formula

$$Var(S_k^x) = Var(S_k^s - S_k^a) = Var(S_k^s) + Var(S_k^a) - 2Cov(S_k^s, S_k^a), \quad k \geq 1.$$

These correlations can have an impact, as illustrated for queues with multiple customer classes having class-dependent service times by Fendick et al. (1989), which is reviewed in Example 9.6.1 of Whitt (2002).

Remark 3. (justification of uncertainty set (14)) Even for the $GI/GI/1$ model, the justification for the continuous-time workload uncertainty set (14) is more complicated than the previous discrete-time uncertainty sets, because there are constants c_1 and c_2 such that $Var(A_{\rho}(s)) = c_1 s$ and $Var(Y_{\rho}(s)) = c_2 s$ for all s if and only if the arrival process is a Poisson process, in which case $c_1 = \rho$ and $c_2 = \rho E[V^2]$; see §2.5 of Ross (1996). Nevertheless, the uncertainty set (14) can be justified for all $GI/GI/1$ queues and more general models by the CLT for $Y(t)$, as we explain in the next section; see Corollary 5.

6.1. The FCLT with Weak Dependence

Henceforth we focus on the general stationary $G/G/1$ model, allowing stochastic dependence among the interarrival times and service times. In this context we can still apply the CLT to motivate the uncertainty sets, but now we apply the CLT and its generalization to a functional CLT (FCLT) for weakly dependent stationary sequences, as in Theorems 19.1-19.3 of Billingsley (1999) and Theorem 4.4.1 of Whitt (2002).

To state the basic FCLT underlying the RQ approach to the waiting time and workload processes, we consider a sequence of models indexed by n with stationary sequence of interarrival times and service times. As in §3, we assume that the models are generated by a fixed sequence of mean-1 random variables $\{(U_k, V_k)\}$, with the interarrival times in model n being $U_{n,k} \equiv \rho_n^{-1} U_k$. For each n, let the sequence of pairs of partial sums be $\{(S^a_{n,k}, S^s_{n,k} : k \geq 1\}$. Let $\lambda_n = \rho_n$ and $\mu_n = 1$ denote the arrival rate and service rate in model n. Let $\lfloor x \rfloor$ denote the greatest integer less than or equal to the real number x. Let D^2 be the two-fold product space of the function space D and let $\Rightarrow$ denote convergence in distribution. For this initial FCLT, we let $\rho_n \to \rho$ as $n \to \infty$ for arbitrary $\rho > 0$. Let random elements in the function apace D^2 be defined by

$$\left(\hat{\mathbf{S}}^a_n(t), \hat{\mathbf{S}}^s_n(t)\right) \equiv n^{-1/2}\left(\left[S^a_{n,\lfloor nt \rfloor} - \rho_n^{-1} nt\right], \left[S^s_{n,\lfloor nt \rfloor} - nt\right]\right), \quad t \geq 0.$$

Theorem 2. (*FCLT for partial sums of interarrival times and service times*) *Let $\{(U_k, V_k) : k \geq 1\}$ be a weakly dependent stationary sequence with $E[U_k] = E[V_k] = 1$. Let $U_{n,k} = \rho_n^{-1} U_k$ and $V_{n,k} = V_k$, $n \geq 1$, and assume that the variances and covariances satisfy*

$$0 < \rho^{-2}\sigma_A^2 \equiv \lim_{n\to\infty} \{n^{-1} Var(S^a_n)\} < \infty, \quad 0 < \sigma_S^2 \equiv \lim_{n\to\infty} \{n^{-1} Var(S^s_n)\} < \infty$$
$$\text{and} \quad \rho^{-1}\sigma^2_{A,S} \equiv \lim_{n\to\infty} \{n^{-1} Cov(S^a_n, S^s_n)\}. \tag{18}$$

Then (under additional regularity conditions assumed, but not stated here)

$$\left(\hat{\mathbf{S}}^a_n, \hat{\mathbf{S}}^s_n\right) \Rightarrow \left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s\right) \quad in \quad D^2 \quad as \quad n \to \infty, \tag{19}$$

where $\left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s\right)$ *is distributed as zero-drift two-dimensional Brownian motion (BM) with covariance matrix*

$$\Sigma = \begin{pmatrix} \rho^{-2}\sigma_A^2 & \rho^{-1}\sigma_{A,S}^2 \\ \rho^{-1}\sigma_{A,S}^2 & \sigma_S^2 \end{pmatrix}.$$

Proof. The one-dimensional FCLT's for weakly dependent stationary sequences in D can be used to prove the two-dimensional version in Theorem 2. First, the limits for the individual processes $\hat{\mathbf{S}}_n^a$ and $\hat{\mathbf{S}}_n^s$ imply tightness of these processes in D, which in turn implies joint tightness in D^2. Second, the Cramer-Wold device in Theorem 4.3.3 of Whitt (2002) implies that limits for the finite-dimensional distributions for all linear combinations (which should be implied by the unstated regularity condition) implies the joint limit for the finite-dimensional distributions (fidi's). Finally, tightness plus convergence of the fidi's implies the desired weak convergence by Corollary 11.6.2 of Whitt (2002). ■

As a consequence of Theorem 2, we also have an associated FCLT for scaled random elements associated with $S_{n,k}^x \equiv S_{n,k}^s - S_{a,k}^a$ and $Y_n(s) \equiv \sum_{i=1}^{A_n(s)} V_{n,i} = \sum_{i=1}^{A(\rho_n s)} V_i = Y(\rho_n s)$, $s \ge 0$, for A and Y in (10) and (11). Let $\mathbf{B}(t)$ be standard (zero drift and unit variance) one-dimensional BM and let $\mathbf{e}$ be the identity function in D, i.e., $\mathbf{e}(t) = t$. Let $\overset{\mathrm{d}}{=}$ mean equal in distribution, as processes if used for stochastic processes.

Corollary 4. (*joint FCLT for basic processes*) *Under the conditions of Theorem 2,*

$$\left(\hat{\mathbf{S}}_n^a, \hat{\mathbf{S}}_n^s, \hat{\mathbf{S}}_n^x, \hat{\mathbf{Y}}_n\right) \Rightarrow \left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s, \hat{\mathbf{S}}^x, \hat{\mathbf{Y}}\right) \quad in \quad D^4 \quad as \quad n \to \infty, \tag{20}$$

where

$$\left(\hat{\mathbf{S}}_n^x(t), \hat{\mathbf{Y}}_n(t)\right) \equiv n^{-1/2}\left(\left[S_{n,\lfloor nt \rfloor}^x - (1-\rho_n^{-1})nt\right], \left[Y_{n,\lfloor nt \rfloor} - \rho_n nt\right]\right), \quad t \ge 0, \tag{21}$$

and $\hat{\mathbf{S}}^x = \hat{\mathbf{S}}^s - \hat{\mathbf{S}}^a \overset{\mathrm{d}}{=} \sigma_X \mathbf{B}$, *with variance function*

$$\sigma_X^2 \equiv \sigma_X^2(\rho) = \rho^{-2}\sigma_A^2 + \sigma_S^2 - 2\rho^{-1}\sigma_{A,S}^2, \quad 0 < \sigma_X^2 < \infty, \tag{22}$$

for $\rho^{-2}\sigma_A^2$, σ_S^2 *and* $\rho^{-1}\sigma_{A,S}^2$ *in* (18), *while*

$$\hat{\mathbf{Y}} = \hat{\mathbf{S}}^s \circ \rho\mathbf{e} - \rho\hat{\mathbf{S}}^a \circ \rho\mathbf{e} \overset{\mathrm{d}}{=} \sigma_Y \mathbf{B} \circ \rho\mathbf{e} \overset{\mathrm{d}}{=} \sqrt{\rho}\sigma_Y \mathbf{B}, \tag{23}$$

where

$$\sigma_Y^2 \equiv \sigma_Y^2(\rho) = \sigma_A^2 + \sigma_S^2 - 2\sigma_{A,S}^2, \quad 0 < \sigma_Y^2 < \infty, \quad \textit{for all} \quad \rho. \tag{24}$$

Hence, $\hat{\mathbf{Y}} = \hat{\mathbf{S}}^{\mathbf{x}}$ *for* $\rho = 1$*, but not otherwise.*

Proof. We apply the continuous mapping theorem (CMT) using several theorems from Whitt (2002). The CMT itself is Theorem 3.4.4. We treat the process $S_{n,k}^x$ using addition. We treat the random sum Y_n in two steps. We first apply the inverse map to go from the FCLT for $S_{n,k}^a$ to the FCLT for the associated scaled counting processes, applying Theorems 7.3.2 and 13.7, which yields limit $-\rho\hat{\mathbf{S}}^a \circ \rho\mathbf{e}$. Then we apply composition with centering in Corollary 13.3.2 of Whitt (2002) to get (23). ■

Condition (18) implies that $k^{-1}Var(S_k^x) \rightarrow \sigma_X^2$ as $k \rightarrow \infty$ for σ_X^2 in (22). In addition to the conclusions of Theorem 19 and Corollary 4, we assume that the appropriate uniform integrability holds, so that we also have the continuous-time analog

$$s^{-1}Var(Y(s)) \rightarrow \sigma_Y^2 \quad \text{as} \quad s \rightarrow \infty \tag{25}$$

for σ_Y^2 in (24).

Theorem 2 and Corollary 4 imply ordinary CLT's for the processes S_n^x and $Y_n(s)$, which we discuss in the next subsection.

6.2. Alternative Scaling in the CLT

We now explain how the new uncertainty sets in (16) and (17) lead to interesting new RQ optimization problems. To do so, we apply the ordinary CLT's that follow from §6.1, illustrating by focusing on S_n^x. As usual, the ordinary CLT follows immediately by applying the CMT with the projection map $\pi : D \rightarrow \mathbb{R}$ with $\pi(x) \equiv x(1)$.

Under the assumptions of Theorem 2, the CLT for the partial sums S_n^x states that

$$(S_n^x - nE[X_1])/\sqrt{n\sigma_X^2} \Rightarrow N(0,1) \quad \text{as} \quad n \rightarrow \infty, \tag{26}$$

where $N(0,1)$ is a standard (mean-0, variance-1) normal random variable and σ_X^2 is the asymptotic variance constant in (18).

At the beginning of this section we observed that the RQ framework in §4 still applies in this setting and the optimal solution is unchanged except for interpretation of the parameter b in (7) when we relate it to the variances. However, the CLT (as well as the FCLT) can be written in a different way that supports the promising new RQ problem in (1) plus (16). Instead of (26), we can also write

$$[S_n^x - E[S_n^x]]/\sqrt{Var(S_n^x)} \Rightarrow N(0,1) \quad \text{as} \quad n \to \infty. \tag{27}$$

The numerators in (26) and (27) are identical because $E[S_n^x] = nE[X_1]$. The full statements in (26) and (27) are asymptotically equivalent as $n \to \infty$ by the CMT, because

$$\frac{S_n^x - nEU_1}{\sqrt{Var(S_n^x)}} = \frac{S_n^x - nE[U_1]}{\sqrt{n\sigma_X}} \times \frac{\sqrt{n\sigma_X}}{\sqrt{Var(S_n^x)}} \Rightarrow N(0,1) \times 1 = N(0,1).$$

Thus, formulation (27) leads to the RQ formulation in (1) plus (16), where we need not have $\sqrt{Var(S_k^x)} = \sqrt{kVar(X_k)}$. The same is true for the the RQ formulation in (13) plus (17).

6.3. The Associated Heavy-Traffic FCLT

Theorem 2 and Corollary 4 also can be used as a basis for establishing HT FCLT's for the waiting-time and workload processes. To state the HT FCLT, we let $\rho_n \to 1$ as $n \to \infty$ at the usual rate; see (29) below. Let $\hat{\mathbf{W}}^n$ and $\hat{\mathbf{Z}}^n$ be the random elements associated with the waiting time and workload processes, defined by

$$\left(\hat{\mathbf{W}}^n(t), \hat{\mathbf{Z}}^n(t)\right) = \left(n^{-1/2} W_{n,\lfloor nt \rfloor}, n^{-1/2} Z_n(nt)\right), \quad t \ge 0. \tag{28}$$

Let $\psi : D \to D$ be the one-dimensional reflection map with impenetrable barrier at the origin, assuming $x(0) = 0$, i.e., $\psi(x)(t) \equiv x(t) - \inf_{0 \leqslant s \leqslant t} x(s)$; see §13.5 of Whitt (2002). Here is the HT FCLT; it is is a variant of Theorem 2 of Iglehart and Whitt (1970); see §5.7 and 9.6 in Whitt (2002). Given Corollary 4, it suffices to apply the CMT with the reflection map ψ.

Theorem 3. (*heavy-traffic FCLT*) *Consider the sequence of* $G/G/1$ *models as specified in* §3. *If, in addition to the conditions of Theorem* 2,

$$n^{1/2}(1-\rho_n)\rightarrow\eta,\quad 0<\eta<\infty, \tag{29}$$

then

$$\left(\hat{\mathbf{W}}_n,\hat{\mathbf{Z}}_n\right)\Rightarrow\left(\psi(\hat{\mathbf{S}}^x-\eta\mathbf{e}),\psi(\hat{\mathbf{S}}^x-\eta\mathbf{e})\right)\quad in\quad D^2\quad as\quad n\rightarrow\infty, \tag{30}$$

jointly with the limits in (20)*, where* ψ *is the reflection map and* $\hat{\mathbf{S}}^x-\eta\mathbf{e}\overset{\mathrm{d}}{=}\sigma_Y\mathbf{B}-\eta\mathbf{e}$ *is BM with variance constant* σ_Y^2 *in* (24) *and drift* $-\eta<0$*, so that* $\psi(\hat{\mathbf{S}}^x-\eta\mathbf{e})$ *is reflected BM* (*RBM*)*.*

The HT approximation for the mean steady-state wait and workload stemming from Theorem 3 is

$$E[W(\rho)]\approx E[Z_\rho]\approx\frac{\sigma_Y^2}{2\eta}\approx\frac{\sigma_Y^2}{2(1-\rho)} \tag{31}$$

for σ_Y^2 in (24), which is independent of ρ, using the mean of the exponential limiting distribution of the RBM $\psi(\sigma_x\mathbf{B}-\eta\mathbf{e})(t)$ as $t\rightarrow\infty$, as in (6).

Remark 4. (the limit-interchange problem) the standard HT limits for the processes do not directly imply limits for the steady-state distributions. Strong results have been obtained with i.i.d. assumptions, e.g., see Gamarnik and Zeevi (2006) and Budhiraja and Lee (2009), but the case with dependence is more difficult. Nevertheless, supporting results for the $G/G/1$ queue when dependence is allowed appear in Szczotka (1990, 1999).

These theorems have important implications for RQ with the uncertainty sets in (7) and (14).

Corollary 5. (*RQ with dependence and the original uncertainty sets*) *If RQ is applied with the* $G/G/1$ *model satisfying the conditions of Theorem* 2 *using the uncertainty sets in* (7) *and* (14)*, where the constants are chosen to satisfy,* $b_x=\beta_x\sigma_X$ *and* $b_z=\beta_z\sigma_Y$*, for* $\sigma_X^2=\sigma_X^2(\rho)$ *in* (22) *and* σ_Y^2 *in* (24)*, then the solutions of these RQ optimizations for the mean steady-state wait and workload are different, but they agree asymptotically in HT as* $\rho\rightarrow1$*. They both are asymptotically correct in HT as* $\rho\rightarrow1$ *if* $\beta_x=\beta_z=\sqrt{2}$ *as in Corollary* 2*.*

Remark 5. (the asymptotic method) The RQ approach in Corollary 5 corresponds to approximating the arrival and service processes in the $G/G/1$ queue by the asymptotic method in Whitt (1982), which develops approximations for the arrival and service processes using all the correlations. That is in contrast to the stationary-interval method discussed just before §6.1, which uses none of the correlations. Below we use RQ to develop intermediate methods in between those two extremes.

7. Robust Queueing with Dependence

The following is our main generalization of Corollaries 1 and 3. As regularity conditions for $Y(t)$, we assume that $V(t) \equiv Var(Y(t))$ is differentiable with derivative $\dot{V}(t)$ having finite positive limits as $t \to \infty$ and $t \to 0$, i.e.,

$$\dot{V}(t) \to \sigma_Y^2 \quad \text{as} \quad t \to \infty \quad \text{and} \quad \dot{V}(t) \to \dot{V}(0) > 0 \quad \text{as} \quad t \to 0, \tag{32}$$

for σ_Y^2 in (24). These assumptions are known to be reasonable; see §7.5.

Theorem 4. (*RQ exposing the impact of the dependence*) *Consider the general stationary $G/G/1$ queue with $\rho < 1$ with the assumptions in* §6.1. (*a*) *The solution of the RQ optimization* (1) *with the single uncertainty set in* (16) *is*

$$\begin{aligned} W_n^* &\equiv \max\{W_n : \tilde{X}_n \in \mathcal{U}^{x'}\} = \max\{S_k^x : \tilde{X}_n \in \mathcal{U}^{x'}, 1 \le k \le n\} \\ &= \max\{-(1-\rho)k/\rho + b_x' \sqrt{Var(S_k^x)} : 1 \le k \le n\} < \infty. \end{aligned} \tag{33}$$

For each ρ, $0 < \rho < 1$, there is an $n^* \equiv n^*(\rho) < \infty$ such that a finite maximum is attained at n^* for all $n \ge n^*$. This index $n^*(\rho)$ is unique if the differences $Var(S_k^x) - Var(S_{k-1}^x)$ are either strictly increasing or strictly decreasing for $k \ge 1$.

(*b*) *The solution of the RQ optimization* (13) *with the single uncertainty set in* (17) *is*

$$\begin{aligned} Z^*(t) &\equiv \sup\{Z(t) : \tilde{N}(t) \in \mathcal{U}^{z'}\} = \sup\{N(s) : \tilde{N}(t) \in \mathcal{U}^{z'}, 0 \le s \le t\} \\ &= \sup\{-(1-\rho)s + b_z' \sqrt{V(s)} : 0 \le s \le t\} < \infty, \end{aligned} \tag{34}$$

For each ρ, $0 < \rho < 1$, there is an $x^* \equiv x^*(\rho)$, such that a finite maximum is attained at x^* for all $t \geq x^*$. In addition, $0 < x^* < \infty$ and x^* satisfies the equation

$$(1-\rho) = \dot{h}(x) \quad \textit{where} \quad h(x) \equiv b_z' \sqrt{V(x)}. \tag{35}$$

This time $x^*(\rho)$ is unique for all ρ, $0 < \rho < 1$, if $h(x)$ is strictly concave or strictly convex, i.e., if $\dot{h}(x)$ is strictly increasing or strictly decreasing.

Proof. The inequalities can be satisfied as equalities just as before. There are finite values k_0 and s_0 such that $\sqrt{Var(S_k^x)} \leq \sqrt{2\sigma_X^2 k}$ for all $k \geq k_0$ and $\sqrt{V(s)} \leq \sqrt{2\sigma_Y^2 s}$ for all $s \geq s_0$ by virtue of the limits in (18) and (25). That shows that the optimization can be regarded as being over closed bounded intervals. The assumed differentiability of V implies that it is continuous, which implies that the supremeum is attained over the compact interval. Because $\dot{V}(x) \to \dot{V}(0) > 0$, we see that there exists a small s' such that

$$-(1-\rho)s + b_z' \sqrt{V(s)} \geq -(1-\rho)s + b_z' \sqrt{s\dot{V}(0)/2} > 0 \quad \text{for all} \quad s \leq s'.$$

As a consequence, the maximum in (34) must be strictly positive and must be attained at a strictly positive time. ■

Henceforth, we primarily focus on the continuous-time workload.

7.1. Positive and Negative Dependence

A common case in models for applications is to have positive dependence in the input process Y, which holds if

$$Cov(Y(t_2) - Y(t_1), Y(t_4) - Y(t_3)) \geq 0 \quad \text{for all} \quad 0 \leq t_1 < t_2 \leq t_3 < t_4. \tag{36}$$

Negative dependence holds if the inequality is reversed. These are strict if the inequality is a strict inequality. From (17) and (18) of §4.5 in Cox and Lewis (1966), which is restated in (48) and (49) of Fendick and Whitt (1989), with positive (negative) dependence, under appropriate regularity conditions, $\dot{V}(t) \geq 0$ and $\ddot{V}(t) \geq (\leq) 0$.

The following is a consequence of Theorem 4 and (36).

Corollary 6. (*positive and negative dependence*) *The variance function* $V(x)$ *is convex* (*concave*), *so that the function* $h(x) \equiv \sqrt{V(x)}$ *is concave if there is positive* (*negative*) *dependence, as in* (36) (*with sign reversed*). *Moreover, a strict inequality is inherited. Thus, there exists a unique solution to the RQ if there is strict positive dependence or strict negative dependence. Moreover, the optimal time* $x^*(\rho)$ *is strictly increasing in* ρ, *approaching* 1 *as* $\rho \uparrow 1$, *so that* $Z^*_\rho \to \dot{V}(\infty) = I_w(\infty) = \sigma_Y^2$ *as* $\rho \uparrow 1$.

Proof. The results for $\sqrt{V(x)}$ with positive dependence follow from convexity properties of compositions. First, with positive dependence, $-\sqrt{V(x)}$ is a convex function of an increasing convex function, and thus convex so that $\sqrt{V(x)}$ is concave. Second, with negative dependence, we have $V \ge 0$, $\dot{V}(t) \ge 0$ and $\ddot{V}(t) \le (\le) 0$. Thus, by direct differentiation

$$\ddot{h}(x) = \frac{1}{\sqrt{V(x)}} \left(\frac{\ddot{V}(x)}{2} - \frac{\dot{V}(x)}{4V(x)} \right) \le 0,$$

with strictness implying a strict inequality. ■

7.2. Approximation from the Asymptotic Expansion

As we explain at the end of §7.5, for a large class of stochastic models the variance $V(t)$ has the asymptotic representation

$$V(t) = \sigma_Y^2 t + \zeta + O(e^{-\gamma t}) \quad \text{as} \quad t \to \infty, \tag{37}$$

with $\gamma > 0$ and $\zeta \le (\ge) 0$ in the case of positive (negative) dependence, which supports that approximation $V(t) \approx \sigma_Y^2 t + \zeta$ for t suitably large. Thus, it is natural to use this approximation in (35) for all ρ not too small. If we do so, then we get the approximation

$$t^*(\rho) \approx \frac{b_z'^2 \sigma_Y^2}{4(1-\rho)^2} - \frac{\zeta}{\sigma_Y^2}. \tag{38}$$

We can then insert (38) into (34) to obtain, after some algebra,

$$Z^*_\rho \approx -(1-\rho)t^*(\rho) + b'_z \sqrt{\sigma_Y^2 t^*(\rho) + \zeta} = \frac{b_z'^2 \sigma_Y^2}{4(1-\rho)} - \frac{(1-\rho)\zeta}{\sigma_Y^2}. \tag{39}$$

The first term in (39) is the HT approximation, while the second term is the refinement.

7.3. Discrete Time: Indices of Dispersion for Intervals

We now recast the RQ solutions in Theorem 4 in terms of indices of dispersion, starting with the discrete-time RQ solution in (33). We create scaled versions of the discrete-time variance-time functions $Var(S^x_k)$, $Var(S^a_k)$ and $Var(S^s_k)$ as functions of k with *indices of dispersion for intervals* (IDI), as in Chapter 4 of Cox and Lewis (1966), defined by

$$I_a(k) \equiv \frac{kVar(S^a_k)}{(E[S^a_k])^2}, \quad I_s(k) \equiv \frac{kVar(S^s_k)}{(E[S^s_k])^2} \quad \text{and} \quad I_{a,s}(k) \equiv \frac{kCov(S^a_k, S^s_k)}{E[S^a_k]E[S^s_k]}. \tag{40}$$

With (40),

$$\sqrt{Var(S^x_k)} = E[U_1]\sqrt{kI_x(k)}, \quad k \ge 1, \quad \text{and} \quad \sigma^2_X \equiv \lim_{k \to \infty} \{k^{-1}Var(S^x_k)\} = E[U_1]^2 I_x(\infty) \tag{41}$$

where

$$I_x(k) \equiv I_a(k) + \rho^2 I_s(k) - 2\rho I_{a,s}(k) \quad \text{for} \quad \rho \equiv E[V_1]/E[U_1 < 1. \tag{42}$$

These three IDI's $I_a(k)$, $I_s(k)$ and $I_{a,s}(k)$ were used to develop queueing approximations in Fendick et al. (1989).

As a consequence, (16) can be rewritten as

$$\mathcal{U}^{x'} \equiv \{\tilde{X}_n : S^x_k \le kE[X_k] + b^{''}_x \sqrt{kI_x(k)}, \, k_L \le k \le n\}. \tag{43}$$

where $b^{''}_x \equiv b'_x E[U_1]$ for b'_x in (16). To apply (43) with $k_L = 0$, we let $I_x(0) = 0$.

7.4. Continuous Time: The Indices of Dispersion for Counts and Work

The workload process is not only convenient because it leads to the continuous RQ optimization problem in (34), but also because the workload process scales with ρ in a more elementary way than the waiting times, as indicated after (12). In particular, with our scaling of the interarrival times, we obtain a simple representation of the arrival processes as a function of the traffic intensity via $A_\rho(t) \equiv A(\rho t)$ and $Y_\rho(t) \equiv Y(\rho t)$, $t \ge 0$, where A and Y are defined in terms of the mean-1 variables. In contrast, the scaling of the waiting times in previous sections is more complicated because the interarrival times are scaled with ρ but the service times are not, so that $X_k(\rho) \equiv$

$V_{n-k}(1)-\rho^{-1}U_{n-k}(1)$. That leads to the relatively complicated way that ρ appears in the IDI $I_k(x)$ in (42).

Paralleling §7.3, in the stationary framework it is useful to relate the variances of the arrival counting process $A(s)$ and the cumulative work input process $Y(s)$ to associated continuous-time indices of dispersion, studied in Fendick and Whitt (1989) and Fendick et al. (1991). With that convention, we define the *index of dispersion for counts* (IDC) associated with the rate-1 arrival process A as in §4.5 of Cox and Lewis (1966) by

$$I_c(t) \equiv \frac{Var(A(t))}{E[A(t)]} = \frac{Var(A(t))}{t}, \quad t \ge 0.$$

and the *index of dispersion for work* (IDW) associated with the rate-1 cumulative input process Y by

$$I_w(t) \equiv \frac{Var(Y(t))}{E[V_1]E[Y(t)]} = \frac{V(t)}{t}, \quad t \ge 0.$$

Fendick and Whitt (1989) showed that the IDW I_w is intimately related to a scaled workload $c_Z^2(\rho)$, which can be defined by comparing to what it would be in the associated $M/D/1$ model; i.e.,

$$c_Z^2(\rho) \equiv \frac{E[Z_\rho]}{E[Z_\rho; M/D/1]} = \frac{2(1-\rho)E[Z_\rho]}{E[V_1]\rho} = \frac{2(1-\rho)E[Z_\rho]}{\rho}, \tag{44}$$

Indeed, under regularity conditions, the following finite positive limits exist and are equal:

$$\begin{aligned} \lim_{t\to\infty}\{I_w(t)\} &\equiv I_w(\infty) = I_a(\infty) + I_s(\infty) - 2I_{a,s}(\infty) = \sigma_Y^2 = c_Z^2(1) \equiv \lim_{\rho\to 1}\{c_Z^2(\rho)\} \\ \lim_{t\to 0}\{I_w(t)\} &\equiv I_w(0) = 1 + c_s^2 = c_Z^2(0) \equiv \lim_{\rho\to 0}\{c_Z^2(\rho)\} \end{aligned} \tag{45}$$

for c_Y^2 in (24) and (42) and $c_s^2 \equiv Var(V_1)/E[V_1]^2$. The limits for I_w above and the differentiability of I_w follow from the assumed differentiability for $V(t)$ and limits in (32). For $t \to 0$ and $\rho \to 0$, see §IV.A of Fendick and Whitt (1989). The IDW limits are related to the IDC limits, with the large-time limit related to the corresponding limit for the IDI I_a in §7.3:

$$\lim_{t\to\infty}\{I_c(t)\} \equiv I_c(\infty) = \sigma_A^2 = I_a(\infty) \equiv \lim_{k\to\infty}\{I_a(k)\} \quad \text{and} \quad \lim_{t\to 0}\{I_c(t)\} \equiv I_c(0) = 1 \tag{46}$$

for σ_A^2 in (18).

The challenge is to relate $c_Z^2(\rho)$ to the IDW I_w for $0 < \rho < 1$. For that, RQ can help. Paralleling (43), we can express the uncertainty set $\mathcal{U}^z(\rho)$ in (17) as

$$\begin{aligned}
\mathcal{U}^z(\rho) &= \{\tilde{N}_\rho(t) : N_\rho(s) \leq -(1-\rho)s + b'_z\sqrt{V(\rho s)},\, 0 \leq s \leq t\}, \\
&= \{\tilde{N}_\rho(t) : N_\rho(s) \leq \frac{-(1-\rho)x}{\rho} + b'_z\sqrt{V(x)},\, 0 \leq x \leq t/\rho\} \\
&= \{\tilde{N}_\rho(t) : N_\rho(s) \leq -\frac{(1-\rho)x}{\rho} + b'_z\sqrt{xI_w(x)},\, 0 \leq x \leq t/\rho\}, \quad 0 < \rho < 1,
\end{aligned} \tag{47}$$

where we have introduced ρ as a time scaling of $V(s)$ in the first line and made the change of variables $x \equiv \rho s$ in the second line. Unlike the IDI I_x in (43), the variance $V(x) \equiv Var(Y(x))$ and the IDW $I_w(x)$ in (47) are independent of ρ. Note that (47) differs from (17) by the presence of $I_w(x)$. These are essentially equivalent if $I_w(x)$ is approximately constant. However, as shown in Fendick and Whitt (1989), Fendick et al. (1989, 1991), the IDW's are often far from constant.

From (34) and (47), we see that the extreme points occur where the slope of $h(x) \equiv \sqrt{2V(x)} = \sqrt{2xI_w(x)}$ equals $(1-\rho)/\rho$. From (45), we anticipate that the slope of $h(x)$ is likely to be strictly increasing from 0 to ∞ over $(0,\infty)$. The optimal value $x^*(\rho)$ as a function of ρ thus should be relatively easy to see from plots of the function h.

The RQ approach allows us to establish versions of the variability fixed-point equation suggested in (9), (15) and (127) of Fendick and Whitt (1989). For the steady-state workload Z_ρ, we let $t \to \infty$ in the RQ optimization (47).

Theorem 5. (*candidate RQ solutions*) *Any optimal solution of the RQ in* (13) *with uncertainty set* (47)*, where* $t \to \infty$*, is attained at* $s^*(\rho) \equiv x^*/\rho$*, where* $x^* \equiv x^*(\rho)$ *satisfies the equation*

$$x^* = \frac{{b'_z}^2\rho^2 I_w(x^*)}{4(1-\rho)^2}\left(1 + \frac{x^*\dot{I}_w(x^*)}{I_w(x^*)}\right)^2 \tag{48}$$

for b'_z *in* (47)*. The associated RQ optimal workload is*

$$Z_\rho^* = \frac{b_z'^2\rho I_w(x^*)}{4(1-\rho)}\left(1 - \left(\frac{x^*\dot{I}_w(x^*)}{I_w(x^*)}\right)^2\right), \tag{49}$$

which is a valid nonnegative solution provided that $x^*\dot{I}_w(x^*) \le I_w(x^*)$*. If* $b_z' = \sqrt{2}$*, then* Z_ρ^* *in* (49) *approaches the heavy-traffic limit of the mean workload as* $\rho \to 1$*. The associated scaled workload satisfies*

$$c_{Z^*}^2(\rho) \equiv \frac{Z_\rho^*}{Z_\rho^*(M/D/1)} = I_w(x^*)\left(1 - \left(\frac{x^*\dot{I}_w(x^*)}{I_w(x^*)}\right)^2\right), \tag{50}$$

Proof. Note that $xI_w(x) = V(x)$. Because we have assumed that $V(x)$ is differentiable, so is I_w. We obtain (48) by differentiating with respect to x in (47) and setting the derivative equal to 0. After substituting (48) into (47), algebra yields (49). The limits in (32) imply that $x^*\dot{I}_w(x^*) \to 0$ and $I_w(x^*) \to I_w(\infty)$ as $\rho \to 1$. ■

Significantly, the scaled workload $c_{Z^*}^2(\rho)$ in (50) is independent of the constant b_z' and depends on ρ only through the solution $x^*(\rho)$ of equation (48). Given that $x\dot{I}_w(x) \to 0$ as $x \to \infty$, it is natural to consider the approximation

$$x^*(\rho) \approx \left(\frac{b_z'\rho}{2(1-\rho)}\right)^2 I_w(x^*(\rho)) \quad \text{so that} \quad Z_\rho^* \approx \frac{b_z'^2 \rho I_w(x^*(\rho))}{4(1-\rho)} \quad \text{and} \quad c_{Z^*}^2(\rho) = I_w(x^*(\rho)). \tag{51}$$

The first equation in (51) is a variability fixed-point equation of the form in suggested in (15) of Fendick and Whitt (1989).

7.5. Estimation and Calculation

For applications, it is significant that the IDW I_w used in §7.4 can readily be estimated from data from system measurements or simulation and calculated in a wide class of stochastic models. The time-dependent variance functions can be estimated from the time-dependent first and second moment functions, as discussed in §III.B of Fendick et al. (1991). Calculation depends on the specific model structure.

7.5.1. The $G/GI/1$ Model. If the service times are i.i.d. with a general distribution having mean τ and scv c_s^2 and are independent of a general stationary arrival process, then as indicated in (58) and (59) in §III.E of Fendick and Whitt (1989),

$$I_w(t) = c_s^2 + I_c(t), \quad t \ge 0, \tag{52}$$

where c_s^2 is the scv of a service time and I_c is the IDC of the general arrival process.

7.5.2. The Multi-Class $\sum_i (G_i/G_i)/1$ Model. As indicated in (56) and (57) in §III.E of Fendick and Whitt (1989), if the input comes from independent sources, each with their own arrival process and service times, then the overall IDC and IDW are revealing functions of the component ones. Let λ_i be the arrival rate, τ_i the mean service time of class i, and $\rho_i \equiv \lambda_i \tau_i$ be the traffic intensity for class i with $\lambda \equiv \sum_i \lambda_i$, $\tau \equiv \sum_i (\lambda_i/\lambda)\tau_i = 1$ so that $\rho = \lambda$. With our scaling conventions,

$$I_c(\lambda t) \equiv \frac{Var(A(t))}{E[A(t)]} = \frac{\sum_i Var(A_i(t))}{\lambda t} = \sum_i \left(\frac{\lambda_i}{\lambda}\right) I_{c,i}(\lambda_i t) \tag{53}$$

and

$$I_w(\lambda t) \equiv \frac{Var(X(t))}{\tau E[X(t)]} = \frac{\sum_i V_i(t)}{\rho t} = \sum_i \left(\frac{\rho_i \tau_i}{\rho \tau}\right) I_{w,i}(\lambda_i t) \quad \text{for all} \quad t \ge 0. \tag{54}$$

From (53) and (54), we see that I_c and I_w are convex combinations of the component $I_{c,i}$ and $I_{w,i}$ modified by additional time scaling. The interaction with the time scaling in (53) and (54) with the time scaling by $n = (1-\rho_n)^{-2}$ in (28) for the HT limits in Theorem 3 can have an important implications for performance, as we illustrate in §7.5.4.

7.5.3. The IDC's for Common Arrival Processes. The two previous subsections show that for a large class of models the main complicating feature is the IDC of the arrival process from a single source. The only really simple case is a Poisson arrival process with rate λ. Then $I_c(t) = 1$ for all $t \ge 0$. A compound (batch) Poisson process is also elementary because the process Y has independent increments; then the arrival process itself is equivalent to M/GI source. However, for a large class of models, the variance $Var(A(t))$ and thus the IDC $I_c(t)$ can either be calculated directly or can be characterized via their Laplace transforms and thus calculated by inverting those transforms and approximated by performing asymptotic analysis. For all models, we assume that the processes A and Y have stationary increments.

An important case for A is the renewal process; to have stationary increments, we assume that it is the equilibrium renewal process, as in §3.5 of Ross (1996). Then $Var(A(t))$ can be expressed in terms of the renewal function, which in turn can be related to the interarrival-time distribution and its transform. The explicit formulas for renewal processes appear in (14), (16) and (18) in §4.5

of Cox (1962). The required Numerical transform inversion for the renewal function is discussed in §13 of Abate and Whitt (1992). The hyperexponential (H_2) and Erlang (E_2) special cases are described in §III.G of Fendick and Whitt (1989).

It is also possible to carry out similar analyses for much more complicated arrival processes. Neuts (1989) applies matrix-analytic methods to give explicit representations of the variance $Var(A(t))$ for the versatile Markovian point process or Neuts process; see §5.4, especially Theorem 5.4.1. Explicit formulas for the Markov modulated Poisson process (MMPP) are given on pp. 287-289.

All of these explicit formulas above have the asymptotic form

$$Var(A(t)) = \sigma_A^2 + \zeta + O(e^{-\gamma t}) \quad \text{as} \quad t \to \infty.$$

Combining this with (52) yields the asymptotic expansion for $V(t)$ in (37).

7.5.4. The Superposition of Many Component Sources. To better understand the complex multi-class examples, consider the $\sum_i GI_i/GI/1$ model where the arrival process is the superposition of n i.i.d. renewal processes, each with rate ρ/n, so that the overall arrival rate is ρ. From (53) and (54),

$$I_{c,n}(\rho t) = I_{c,1}(\rho t/n) \quad \text{and} \quad I_{w,n}(\rho t) = I_{w,1}(\rho t/n), \quad t \geq 0, \tag{55}$$

so that the superposition IDI and IDW differ from those of a single component process only by the time scaling. In support of the IDC and IDW as useful partial characterizations, we see that the expressions in (53)-(52) are consistent with the known complex behavior of queues with superposition arrival processes, as discussed in §9.8 of Whitt (2002). As $n \to \infty$, we see evidence of the convergence to a Poisson process; As $t \to \infty$ we see the same limit as for a single component renewal process, i.e., $I_{c,n}(\infty) = I_{c,1}(\infty)$. We see that the RQ approach can capture the complex interaction between n and ρ.

8. Simulation Comparisons

We illustrate how the new RQ approach can be used with system data from queueing networks by applying simulation to analyze two common but challenging network structures: (i) a queue with a superposition arrival process and (ii) several queues in series. The specific examples are chosen to capture a known source of difficulty: The relevant variability parameter of the arrival process at each queue can depend strongly on the traffic intensity of that queue, as discussed in Whitt (1995).

8.1. A Queue with a Superposition Arrival Process

We start by looking at an example of a $\sum_i G_i/GI/1$ single-server queue with a superposition arrival process, where (55) can be applied. Let the rate-1 arrival process A be the superposition of $n = 10$ i.i.d. renewal processes, each with rate $1/n$, where the times between renewals have a lognormal distribution with mean n and scv $c_a^2 = 10$. Let the service-times distribution be hyperexponential (H_2), a mixture of two exponential distributions) with mean 1, $c_s^2 = 2$ and balanced means as on p. 137 of Whitt (1982). Then (55) and (45) imply that the IDW has limits $I_w(0) = 1 + c_s^2 = 3$ and $I_w(\infty) = c_a^2 + c_s^2 = 12$, so that the IDW is not nearly constant.

Figure 1 (left) shows a comparison between the simulation estimate of the normalized workload $c_Z^2(\rho)$ in (44) and the approximation $c_{Z^*}^2(\rho)$ in (50) for this example. Two important observations are: (i) the normalized mean workload $c_Z^2(\rho)$ in (44) as a function of ρ is not nearly constant, and (ii) there is a close agreement between the RQ approximation $c_{Z^*}^2(\rho)$ in (50) from the RQ in (47) and the direct simulation estimate; the close agreement for all traffic intensities is striking.

For this example, we see that $c_Z^2(\rho) \approx 3$ for $\rho \leq 0.5$, which is consistent with the Poisson approximation for the arrival process and the associated $M/G/1$ queue, where $c_Z^2(\rho) = 3$ for all ρ, but the normalized workload increases steadily to 12 after $\rho = 0.5$, as explained in §9.8 of Whitt (2002).

The estimates for Figure 1 were obtained for ρ over a grid of 99 values, evenly spaced between 0.01 and 0.99. Similarly, the RQ optimization was performed using (47) with a discrete-time estimate of the IDW. By doing multiple runs, we ensured that the statistical variation was not an issue. For the main simulation of the arrival process and the queue we used 5×10^6 replications, discarding a large initial portion of the workload process to ensure that the system is approximately in steady state. (The component renewal arrival processes thus can be regarded as equilibrium renewal processes, as in §3.5 of §Ross (1996).) We let the run length and amount discarded be increasing in ρ, as

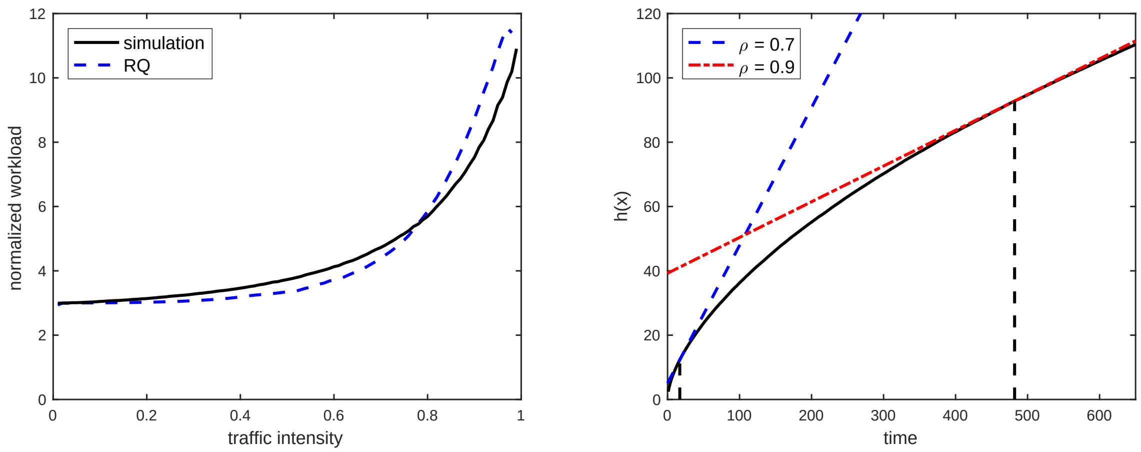

Figure 1 A comparison between simulation estimates of of the normalized mean workload $c_Z^2(\rho)$ in (44) and its approximation $c_{Z^*}^2(\rho)$ in (50) from the RQ in (47) as a function of ρ for the $\sum_i^n GI_i/H_2/1$ model with $c_s^2 = 2$ and a superposition of n i.i.d. lognormal renewal arrival processes for $n = 10$ and $c_a^2 = 10$ (left). On the right is the graphical RQ solution showing $h(x) \equiv \sqrt{2xI_w(x)}$ and the tangent line with slope $(1-\rho/\rho$ at $x^* \approx 482$ for $\rho = 0.9$ and at $x^* \approx 17$ for 0.7, as dictated by (35).

dictated by Whitt (1989). We provide additional details about our simulation methodology in the appendix.

8.2. Ten Queues in Series

This second example is a variant of examples in Suresh and Whitt (1990), exposing the complex impact of variability on performance in a series of queues if the external arrival process and service times at a previous queue have very different levels of variability. This example has 10 single-server queues in series. The external arrival process is a rate-1 renewal process with H_2 interarrival times having $c_a^2 = 10$. (We use the same distribution as for the service time in §8.1.) The first 9 queues all have deterministic service times. The first 8 queues have mean service time and thus traffic intensity 0.6, while the 9$^{\text{th}}$ queue has mean service time and thus traffic intensity 0.95. The last (10$^{\text{th}}$) queue has an exponential service-time distribution. with mean and traffic intensity ρ; we explore the impact of ρ on the performance of that last queue.

The deterministic queues act to smooth the arrival process at the last queue. Thus, for sufficiently low traffic intensities ρ at the last queue, the last queue should behave essentially the same as a

$D/M/1$ queue, which has $c_a^2 = 0$, but as ρ increases, the arrival process at the last queue should inherit the variability of the external arrival process, and behave like an $H_2/M/1$ queue with scv $c_a^2 = 10$. This behavior is substantiated by Figure 2, which compares simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (44) at the last queue of ten queues in series as a function of the mean service time and traffic intensity ρ there with the corresponding values in the $D/M/1$ queue (left) and with the RQ approximation $c_{Z^*}^2(\rho)$ in (50) from (47) (right). Figure 2 (left) shows that the last queue behaves like a $D/M/1$ queue for all traffic intensities ≤ 0.8, but then starts behaving more like an $H_2/M/1$ queue as the traffic intensity approaches the value 0.95 at the 9th queue. Figure 2 (right) shows that RQ successfully captures this phenomenon and provides an accurate approximation for all ρ.

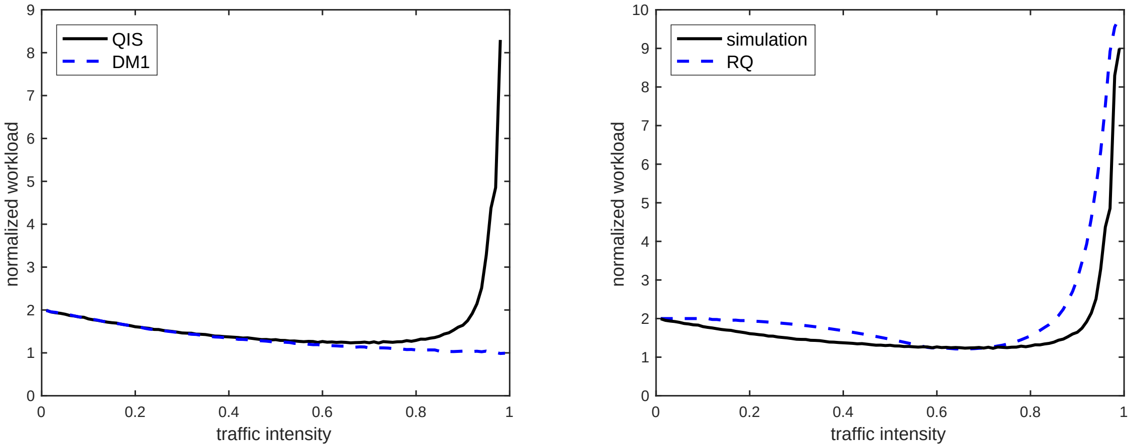

Figure 2 A comparison between simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (44) at the last queue of the ten queues in series with highly variable external arrival process, but low-variability service times, as a function of the mean service time and traffic intensity ρ there with the corresponding value in the $D/M/1$ queue (left) and with the RQ approximation $c_{Z^*}^2(\rho)$ in (50) from (47) (right).

To elaborate on this series-queue example, we show the IDW for the last queue in Figure 3. The plot on the left shows the IDW over the long interval $[0, 10^5]$, while the plots in the middle and right give a closer view of the IDW over the initial segments $[0, 20]$ and $[0, 400]$. On the right, we plot the IDW assuming continuous-time stationarity (which we use) together with the plot using the discrete-time Palm stationarity (see Sigman (1995)), which acts as if there is an arrival at time 0, so that the plot is 0 over the initial interval of length 0.95 (the deterministic service time at the previous queue). The good performance in Figure 2 for small values of ρ depends on using the proper (continuous-time) version.

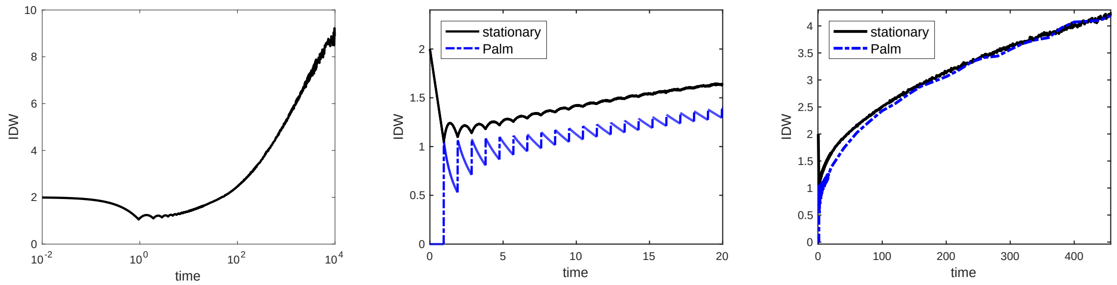

Figure 3 The IDW at the last queue over the interval $[0,10000]$ in log scale (left), $[0,20]$ (middle) and $[0,400]$ (right). The continuous-time stationary version used for RQ with the workload is contrasted with the discrete-time Palm version over the initial segment on the middle and right.

We conclude this example by illustrating the discrete-time approach for approximating the expected steady-state waiting time $E[W]$ using the RQ optimization in (1) with the uncertainty set in (43). Figure 4 is the discrete analog of Figure 2. Figure 4 compares simulation estimates of the normalized mean waiting time $c_W^2(\rho)$, defined just as in (44), at the last queue of ten queues in series as a function of the mean service time and traffic intensity ρ there with the corresponding values in the $D/M/1$ queue (left) and with the RQ approximation $c_{W*}^2(\rho)$, defined just as in (50). Figure 4 and 2 look similar, except that there is a significant difference for small velues of ρ. In general, we do not expect RQ to be effective for extremely low ρ, because (i) the CLT is not appropriate for only a few summands and (ii) the mean waiting time is known to depend on other properties when ρ is small. The mean waiting time and mean workload actually are quite different in light traffic; see §IV.A of Fendick and Whitt (1989). As explained there, the mean workload tends to be more robust to model detail.

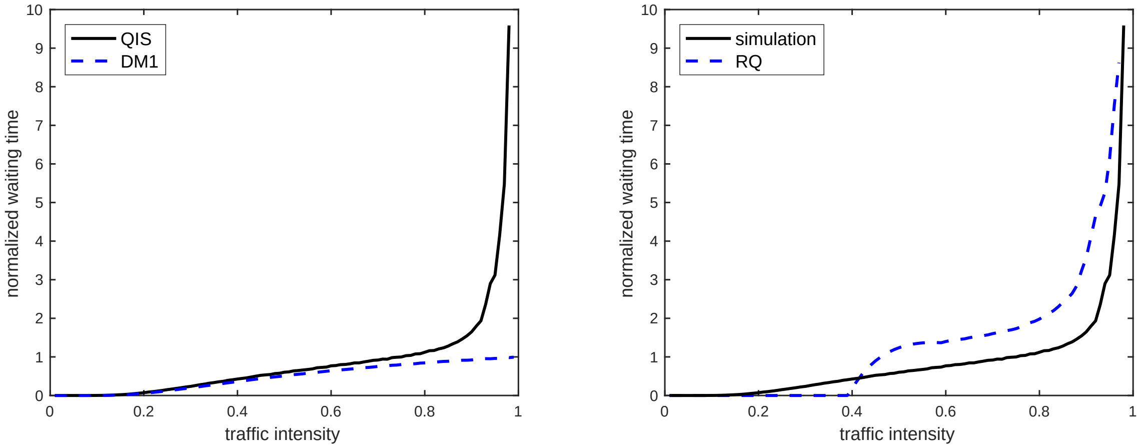

Figure 4 Contrasting the discrete-time and continuous-time views: the analog of Figure 2 for the waiting time. Simulation estimates of the normalized mean waiting time $c^2_W(\rho)$, defined as in (44), at the last queue of the ten queues in series with highly variable external arrival process, but low-variability service times, as a function of the mean service time and traffic intensity ρ there with the corresponding value in the $D/M/1$ queue (left) and with the RQ approximation $c^2_{W*}(\rho)$, defined as in (47) (right).

9. Conclusions

9.1. Summary

We have shown that the robust queueing (RQ) approach in Bandi et al. (2015) can be usefully extended to expose the impact of dependence among interarrival times and service times upon the expected waiting time and expected workload in a general $G/G/1$ single-server queue as a function of the traffic intensity ρ. First, we showed that it can be useful to replace the original pair of uncertainty sets in (2) by the version with one uncertainty set in (7); e.g., Corollary 2 shows that RQ is asymptotically correct for the $GI/GI/1$ queue with the single uncertainty set, but not for the two uncertainty sets. We have also shown that it also can be advantageous to focus on the continuous-time workload using (14), primarily because the total workload $Y(t)$ in (11) scales with the traffic intensity ρ in a more elementary way, as can be seen from the asymptotic variance parameters σ^2_X in (22) and σ^2_Y in (24). (In particular, σ^2_X in (22) depends on ρ in a more complicated way.) We showed that the impact of the dependence can be captured by including versions of the variance-time functions in these uncertainty sets, as in (16) and (17). It can then be helpful to express the versions in (16) and (17) in terms of indices of dispersion, as in (43) and (47).

In §3 and §6 we exposed the intimate connection between RQ and heavy-traffic theory. Corollaries 2 and 5 show that the main RQ methods for the waiting time and the workload here are both asymptotically correct as the traffic intensity ρ increases to its critical level 1. We have also shown that the RQ can usefully supplement previous approximations for the performance of complex $G/G/1$ queues with dependence among interarrival times and service times in Fendick and Whitt (1989). Theorem 5 shows that the solution of the continuous-time RQ optimization for the workload identifies a time $x^*(\rho)$ as a function of the traffic intensity ρ such that the RQ workload Z^*_ρ depends on the IDW I_w primarily through the single value $I_w(x^*(\rho))$. Particularly attractive are the formulas for the scaled RQ workload $c^2_{Z^*}(\rho)$ in (50) and (51), which can generate useful approximations for the scaled workload $c^2_Z(\rho)$ defined in (44). In this way, we obtain new insight into the way dependence affects the performance of the queue as a function of the traffic intensity in the queue.

We conducted simulation experiments in §8 that show that the RQ approximations can be effective. These experiments also dramatically demonstrate the inadequacy of methods that either (i) ignore the dependence within the flows or (ii) act as if a single-variability parameter can characterize an arrival process, independent of the traffic intensity at the queue.

9.2. How Can the Results Here Be Applied?

This paper helps develop useful diagnostic tools to study complex queueing systems. This paper adds additional support to Fendick and Whitt (1989) by showing how to measure flows (arrival processes, possibly together with service times) in complex queueing systems and the value for doing so in understanding congestion at a queue, as characterized by the mean workload and the mean waiting time. In particular, we see how the variance time curves and indices of dispersion can provide useful descriptions of the flows, enabling us with the aid of RQ to predict congestion as a function of the traffic intensity quite accurately. These measurements can fruitfully be applied with either system measurements or simulations. As we indicated in §7.5, the indices of dispersion can also be calculated for quite complex models.

As in Bandi et al. (2015), the new RQ can help develop improved performance analysis tools for complex queueing networks. In particular, the methods here provide a basis for improving parametric-decomposition approximations such as QNA in Whitt (1983b) by exploiting variability functions instead of variability parameters, as proposed in Whitt (1995).

One concrete way the RQ here can be applied is to analyze the consequence of changing the service mechanism and/or the arrival process associated with a single-server queue in a complex queueing network. For example, assuming that (i) the same arrival process would come to a new service mechanism and (ii) the new service mechanism produces i.i.d. service times with a distribution that can be predicted, then we could first measure the IDC of the arrival process and combine that with (52) to obtain an estimate of the full IDW. Then we could apply RQ to estimate the mean workload at the queue. If we are contemplating several alternative service mechanisms, we can apply the same techniques to compare their performance impact.

As a second example, suppose that the arrival rate will increase. If that will occur in a way that corresponds approximately to deterministic scaling of the arrival counting process, then we can directly apply RQ to predict the performance consequence. On the other hand, if the arrival rate increases by superposing more streams, as in Sriram and Whitt (1986), then we can apply RQ with (53)-(55) to predict the performance consequence.

9.3. Directions for Future Research

There are many important directions for future work. It remains to use RQ with dependence to estimate the mean waiting time and mean workload in multi-server queues. It remains to use RQ to usefully bound and approximate the full distribution of the workload instead of just the mean. It also remains to use RQ to obtain bounds and approximations for the range of possible values of the mean waiting time and workload, given various constraints, in the spirit of Klincewicz and Whitt (1984) and Johnson and Taaffe (1990, 1993), where optimization was used to expose the range of possible values for the mean steady-state performance measures given constraints on the moments and the shape of the interarrival-time and service-time distribution. Most important, it remains to apply the new RQ approach to develop improved approximations of the performance in complex queueing networks with a variety of service disciplines. It remains to apply the present paper to enhance the variability function approach in Whitt (1995).

Acknowledgments

Support was received from NSF grant CMMI 1265070.

References

Abate, J., G. L. Choudhury, W. Whitt. 1993. Calculation of the GI/G/1 steady-state waiting-time distribution and its cumulants from Pollaczek's formula. *Archiv fr Elektronik und bertragungstechnik* **47**(5/6) 311–321.

Abate, J., W. Whitt. 1992. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems* **10** 5–88.

Asmussen, S. 2003. *Applied Probability and Queues*. 2nd ed. Springer, New York.

Badidi, E., L. Esmahi, M. A. Serhani. 2005. A queueing model of multi-class QOS-aware web services. *Proceedings of the Third European Confereence on Web Services (ECOWS'05)*. IEEE Coputer Society, 1–9.

Bandi, C., D. Bertsimas, N. Youssef. 2015. Robust queueing theory. *Operations Research* **63**(3) 676–700.

Ben-Tal, A., L. El-Ghaoui, A. Nemirovski. 2009. *Robust Optimization*. Princeton University Press, Princeton, NJ.

Bertsimas, D., D. B. Brown, C. Caramanis. 2011. Theory and applications of robust optimization. *SIAM Review* **53**(3) 464–501.

Beyer, H. G., B. Sendhoff. 2007. Robust optimzation - a comprehensive survey. *Computer Methods in Applied Mechanics and Engineering* **196**(33-34) 3190–3218.

Billingsley, P. 1999. *Convergence of Probability Measures*. Wiley, New York.

Budhiraja, A., C. Lee. 2009. Stationary distribution convergence for generalized jackson networks in heavy traffic. *Mathematics of Operations Research* **34**(1) 45–56.

Cochran, J. K., K. T. Roche. 2009. A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers and Operations Research* **36** 1497–1512.

Cohen, J. W. 1982. *The Single Server Queue*. 2nd ed. North-Holland, Amsterdam.

Cox, D. R. 1962. *Renewal Theory*. Methuen, London.

Cox, D. R., P. A. W. Lewis. 1966. *The Statistical Analysis of Series of Events*. Methuen, London.

Fendick, K. W., V. Saksena, W. Whitt. 1989. Dependence in packet queues. *IEEE Trans Commun.* **37** 1173–1183.

Fendick, K. W., V. Saksena, W. Whitt. 1991. Investigating dependence in packet queues with the index of dispersion for work. *IEEE Trans Commun.* **39**(8) 1231–1244.

Fendick, K. W., W. Whitt. 1989. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE* **71**(1) 171–194.

Gamarnik, D., A. Zeevi. 2006. Validity of heavy traffic steady-state approximations in generalized jackson networks. *Advances in Applied Probability* **16**(1) 56–90.

Gayon, J.-P., F. de Vericourt, F. Karaesmen. 2009. Stock rationing in an $M/E_r/1$ mutli-class make-to-stock queue with backorders. *IIE Transactions* **41**(12) 1096–1109.

Hall, R. W. 2006. *Patient Flow: Reducing Delay in Healthcare Delivery*. Springer, New York.

Hall, R. W., ed. 2012. *Handbook of Healthcare System Scheduling*. Springer, New York.

Harrison, J. M. 1985. *Brownian Motion and Stochastic Flow Systems*. Wiley, New York.

Iglehart, D. L., W. Whitt. 1970. Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Advances in Applied Probability* **2**(2) 355–369.

Johnson, M. A., M. R. Taaffe. 1990. Matching moments to phase distributions: Density function shapes. *Stochastic Models* **6**(2) 283–306.

Johnson, M. A., M. R. Taaffe. 1993. Tchebycheff systems for probability analysis. *American Journal of Mathematical and Manageement Sciences* **13**(1-2) 83–111.

Kingman, J. F. C. 1962. Inequalities for the queue $GI/G/1$. *Biometrika* **49**(3/4) 315–324.

Klincewicz, J. G., W. Whitt. 1984. On approximations for queues, II: Shape constraints. *AT&T Bell Laboratories Technical Journal* **63**(1) 139–161.

Loynes, R.M. 1962. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society* **58**(3) 497–520.

Neuts, M. F. 1989. *Structured Stochastic Matrices of $M/G/1$ Type and their Application.* Marcel Dekker, New York.

Ross, S. M. 1996. *Stochastic Processes.* 2nd ed. Wiley, New York.

Sigman, K. 1995. *Stationary Marked Point Processes: An Intuitive Approach.* Chapman and Hall/CRC, New York.

Sriram, K., W. Whitt. 1986. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications* **SAC-4**(6) 833–846.

Suresh, S., W. Whitt. 1990. The heavy-traffic bottleneck phenomenon in open queueing networks. *Operations Research Letters* **9**(6) 355–362.

Szczotka, W. 1990. Exponential approximation of waiting time and queue size for queues in heavy traffic. *Advances in Applied Probability* **22**(1) 230–240.

Szczotka, W. 1999. Tightness of the stationary waiting time in heavy traffic. *Advances in Applied Probability* **31**(3) 788–794.

Whitt, W. 1982. Approximating a point process by a renewal process: two basic methods. *Oper. Res.* **30** 125–147.

Whitt, W. 1983a. Queue tests for renewal processes. *Oper. Res. Letters* **2**(1) 7–12.

Whitt, W. 1983b. The queueing network analyzer. *Bell Laboratories Technical Journal* **62**(9) 2779–2815.

Whitt, W. 1989. Planning queueing simulations. *Management Science* **35**(11) 1341–1366.

Whitt, W. 1995. Variability functions for parametric-decomposition approximations of queueing networks. *Management Science* **41**(10) 1704–1715.

Whitt, W. 2002. *Stochastic-Process Limits.* Springer, New York.

Additional Examples

In this e-companion we present some additional examples illustrating more complex behavior that can be seen in the IDW $I_W(t)$ and in the normalized mean workload $c_Z^2(\rho)$. All examples are for single-server queues in series, as in §8.2. For background on this example, we refer to §4.5 of Whitt (1983b), Suresh and Whitt (1990) and §§5 and 6 of Whitt (1995).

Recall that Figure 2 illustrated the performance impact in an $H_2/D/1 \rightarrow \cdot/D/1 \ldots \rightarrow \cdot/D/1 \rightarrow \cdot/M/1$ model with a rate-1 H_2 renewal external arrival process, where the interarrival times has scv $c_a^2 = 10$, followed by nine single-server queues with deterministic D service times and then a final 10th queue with an exponential service time distribution. The first 8 queues all have mean service times and thus traffic intensities of $\rho_k = 0.6$, while the 9th queue has mean service time and thus traffic intensity $\rho_9 = 0.95$. We look at the performance at the last queue as a function of the traffic intensity $\rho \equiv \rho_{10}$ there. Figure 2 shows that the normalized workload at the last queue as a function of ρ. From (45), we know that the left and right limits of the normalized mean workload are $c_Z^2(0) = 1 + c_s^2 = 2.0$ and $c_Z^2(1) = c_a^2 + c_s^2 = 11.0$. Figure 2 shows that the performance is consistent with these limits, even though we cannot see the right hand limit, because the simulation considered traffic intensities bounded above by a quantity less than 1. Nevertheless, we see that the performance varies as a function of ρ approximately as predicted by these two limits.

Figure 2 also shows a dip in the middle consistent with the smoothing provided by the the low variability at the first 9 queues, but the performance does not oscillate too much. Now we illustrate more complex performance functions that can be obtained with more complex models.

In general, experience indicates that for 10 queues in series the normalized mean workload can be bounded above and below, approximately, by

$$\min\{1, c_a^2, c_{s,k}^2, 1 \le k \le 9\} + c_{s,10}^2 \le c_Z^2(\rho) \le \max\{c_a^2, c_{s,k}^2, 1 \le k \le 9\} + c_{s,10}^2. \tag{EC.1}$$

(The "1" appears in the minimum because the left limit at 0 is $1+c_s^2$.) For example, this approximate bound is consistent with the approximatioon for the variability parmeter c_d^2 of the departure process froma $GI/GI/1$ queue in formula (38) in Whitt (1983a), i.e.,

$$c_d^2 \approx (1-\rho^2)c_a^2 + \rho^2 c_s^2. \tag{EC.2}$$

The bound can be obtained by iterating that approximation forward to get an approximation for $c_{d,9}^2$ and then allowing the previous traffic intensities to vary.

For this example, the bound in (EC.1) is not too informative, concluding that $1 \le c_Z^2(\rho) \le 11$, which corresponds to the left and right limits. Our goal is to say more about $c_Z^2(\rho)$ for $0 < \rho < 1$ by using the IDW and RQ.

However, so far, the examples do not show that too much is going on in the middle except for moving from one limit to the other. That motivates us to look at the next examples.

EC.1. The $EHEHE \rightarrow M$ Example with Four Internal Modes

We now consider an example of 5 single-server queues in series where the variability increases and then decreases 5 times, with the traffic intensities at successive queues decreasing. That makes the external arrival process and the earlier queues relevant only as the traffic intensity increases. Specifically, the example can be donoted by

$$E_{10}/H_2/1 \rightarrow \cdot/E_{10}/1 \rightarrow \cdot/H_2/1 \rightarrow \cdot/E_{10}/1 \rightarrow\rightarrow \cdot/M/1. \tag{EC.3}$$

In particular, the external arrival process is a rate-1 renewal process with E_{10} interarrival times, thus $c_a^2 = 0.1$. The 1st queue has H_2 service times with mean 0.99 and $c_s^2 = 10$ (and also balanced means, as before), thus the traffic intensity at this queue is 0.99. The 2nd queue has E_{10} service time with mean and thus traffic intensity 0.98. The 3rd queue has H_2 service times with mean 0.70 and $c_s^2 = 10$. The 4th queue has E_{10} service times with mean and thus traffic intensity 0.5. The last (5th) queue has an exponential service-time distribution. with mean and traffic intensity ρ. As before, we explore the impact of ρ on the performance of that last queue.

Looking backwards starting from the 4^{th} queue, i.e., the queue just before the last queue, the Erlang service act to smooth the arrival process at the last queue. Thus, for sufficiently low traffic intensities ρ at the last queue, the last queue should behave essentially the same as a $E_{10}/M/1$ queue, which has $c_a^2 = 0.1$, but as ρ increases, the arrival process at the last queue should inherit the variability of the previous service times and the external arrival process, and altering between $H_2/M/1$ and $E_{10}/M/1$ as the traffic intensity at the last queue increases. This implies that the normalized workload $c_Z^2(\rho)$ in (44) as a function of ρ should have four internal modes. (If we also count the left and right ends, there will be six modes.

This behavior is substantiated by Figure EC.1 (left), which compares simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (44) at the last queue with the RQ approximation $c_{Z^*}^2(\rho)$ in (50) from (47). It shows that the the normalized workload at the last queue fluctuates and each mode corresponds to a previous service process or the external arrival process. Figure EC.1 (left) also shows that RQ successfully captures all modes and provides a reasonably accurate approximation for all ρ. Note that a new scale in the horizontal x axis is used in Figure EC.1 (left), namely $-\ln(1-\rho)$. Since 4 out of 6 modes lies in $\rho > 0.8$, the new scale act to stretch out the crowded plot under heavy traffic.

To conclude on this series-queue example, we show the IDW for the last queue in Figure EC.1 (right). The x axis of the figure is in log scale for easier display. We see a more irregular plot at the right because it is hard to directl estimate the IDW $IW(t)$ for very large t. Clearly, the IDW has the same qualitative property as the normalized workload as well as the RQ approximation, as we expect from equation (51).

EC.2. A Similar Example with Highly Variable Input

In this section, we consider a similar example where the normalized workload as a function of ρ also has several modes, but the external arrival here has high variability.

In this example we use groups of queues in series with the same distribution and traffic intensity in order to better bring about an adjustment in the level of variability. This device is motivated by

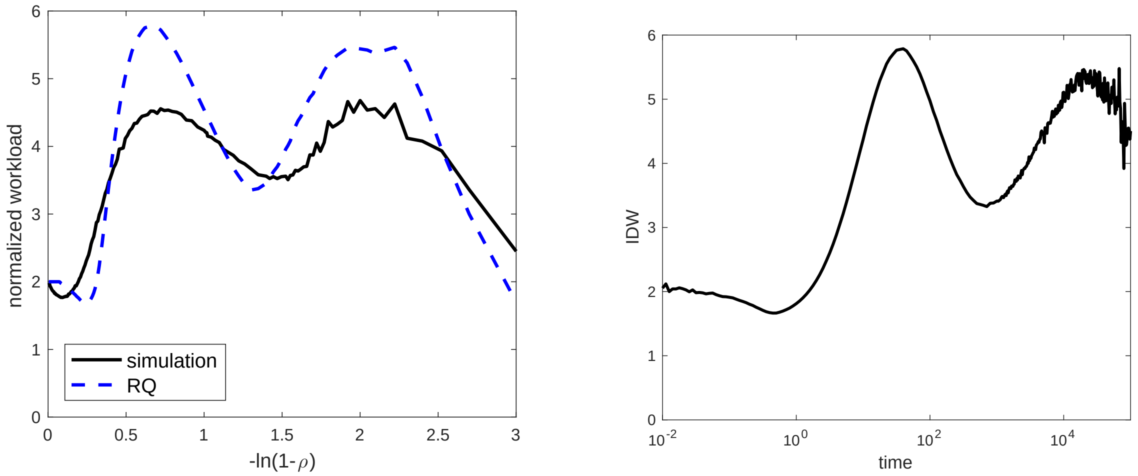

Figure EC.1 A comparison between simulation estimation of the normalized workload $c_Z^2(\rho)$ at the last queue as a function of traffic intensity ρ with the RQ approximation $c_{Z^*}^2(\rho)$ in (50) from (47) (left), and the IDW at the last queue over the interval $[0, 10000]$ in log scale (right).

the convex-copmbination approximation in (EC.2). Specifically, this example has 13 single-server queues in series. The external arrival process is a rate-1 renewal process with H_2 interarrival times with $c_a^2 = 10$. A group of three queues having E_{10} service times with mean 0.99 is then added to smooth the highly variable external arrivals. The next group of three queues has H_2 service times with mean 0.92 and squared coefficient of variation 5. These queues will bring up the variability of the departure process. Then, another group of three queues with mean 0.9 has E_{10} service times to smooth the departure process again. The variability is then raised by yet another group of three queues having H_2 service times with mean 0.3 and $c_S^2 = 10$. Finally, the last (13^{th}) queue has exponential service times with mean and traffic intensity ρ. As before, we explore the impact of ρ on the performance of that last queue.

As explained in last example, for sufficiently low traffic intensities ρ at the last queue, the last queue should behave approximately the same as an $H_2/M/1$ queue, which has $c_a^2 = 10$, but as ρ increases, the arrival process at the last queue should inherit the variability of the previous service times and the external arrival process, and altering between $E_{10}/M/1$ and $H_2/M/1$ as the traffic intensity at the last queue increases. This implies that the normalized workload $c_Z^2(\rho)$ in (44) as a function of ρ should have several modes, corresponding to the variability of the external arrival process and the service processes at the first 4 groups of queues.

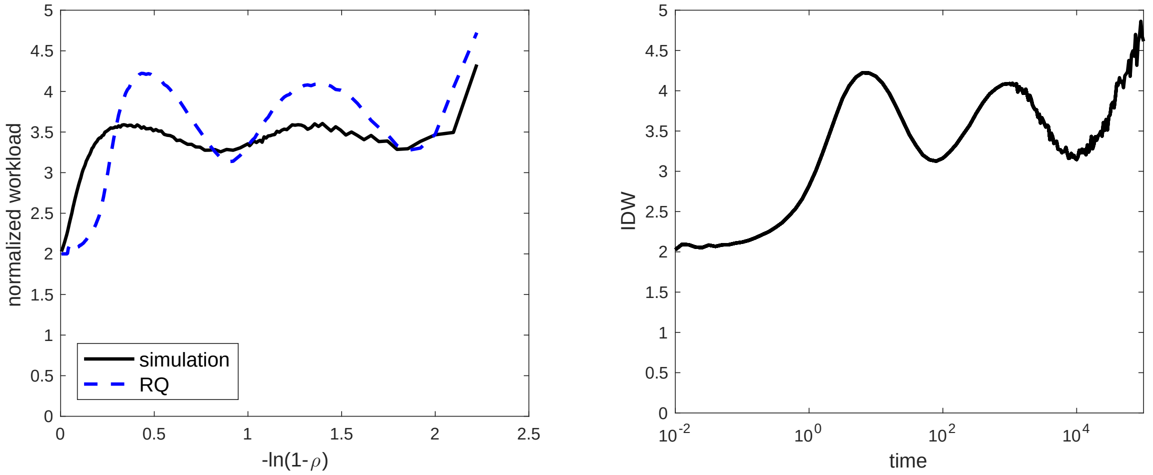

Figure EC.2 A comparison between simulation estimation of the normalized workload $c_Z^2(\rho)$ at the last queue as a function of traffic intensity ρ with the RQ approximation $c_{Z^*}^2(\rho)$ in (50) from (47) (left), and the IDW at the last queue over the interval $[0, 10000]$ in log scale (right).

We then have the similar plots in Figure EC.2, which compares simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (44) at the last queue with the RQ approximation $c_{Z^*}^2(\rho)$ in (50) from (47) (left) and shows the IDW for this example (right). Again, we are using the same scale as in Figure EC.1 (left), i.e., $-\ln(1-\rho)$, to stretch out the plot under heavy traffic.

Figure EC.2 (left) shows that the the normalized workload at the last queue again has four internal modes and that RQ successfully captures all modes and provides a reasonably accurate approximation for all ρ. Figure EC.2 (right) shows that the IDW has the same qualitative property as the RQ approximation, which is explained in (51). However, the fluctuations in the simulation values for $0 < \rho < 1$ in Figure EC.2 are much less than in Figure EC.1.

We conclude that (i) the IDW and RQ do capture the qualititative behavior and (ii) the RQ approximation based on the IDW is reasonably accurate in these difficult examples.