Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Time-Varying Robust Queueing

Ward Whitt
Industrial Engineering and Operations Research, Columbia University, ww2040@columbia.edu
Wei You
Industrial Engineering and Operations Research, Columbia University, wy2225@columbia.edu

We develop a time-varying robust-queueing (TVRQ) algorithm for the continuous-time workload in a single-server queue with a time-varying arrival-rate function. We apply this TVRQ to develop approximations for the periodic steady-state expected workload in models with a periodic arrival-rate function. We apply simulation and asymptotic methods to examine the performance of periodic TVRQ (PRQ). We find that PRQ predicts the mean of the periodic distribution and even the full distribution (specified by the quantiles) remarkably well. We show that the PRQ converges to a proper limit in appropriate long-cycle and heavy-traffic regimes, and coincides with long-cycle fluid limits and heavy-traffic diffusion limits for long cycles.

Key words: robust queueing theory, time-varying arrival rates, nonstationary queues, periodic queues, heavy traffic

History: Submitted, August 27, 2016; Revision: December 8, 2018

1. Introduction

Queueing has long played a prominent role in operations research applications. For example, early OR studies include traffic delays at tool booths by Edie (1954), letter delays at post offices by Oliver and Samuel (1962), airplane landing delays at airports by Koopman (1972) and dispatching delays for police patrol cars by Kolesar et al. (1975). As in many other OR applications, the arrival processes in these applications all have time-varying (TV) arrival rates. Thus, the natural queueing models require simulation or nonstandard analysis techniques beyond elementary stochastic textbooks.

Those four OR studies also illustrate two of the most important analytical techniques for analyzing TV queueing models. First, the papers by Edie (1954) and Oliver and Samuel (1962) illustrate that a relatively simple deterministic analysis can be employed when the TV arrival rate tends to dominate the randomness. The other papers by Koopman (1972) and Kolesar et al. (1975) illustrate how numerical methods for systems of TV ordinary differential equations (ODE's) can be applied to calculate TV performance measures for the TV Markovian $M_t/M_t/s_t$ queueing model, which has a nonhomogeneous Poisson process (NHPP, the M_t) as its arrival process, and possibly a TV service rate and number of servers as well, because the number of customers in the system evolves as a TV birth-and-death process, so that its TV transition probability density function evolves according to a system of ODE's, often called the Kolmogorov forward equations.

The ODE approach to the TV $M_t/M_t/s_t$ queueing model has become the accepted analytical approach. The ODE approach is complicated by the fact that there are infinitely many ODE's in the system of equations, but that difficulty can be circumvented by truncating to a finite system, as was done by Koopman (1972) and Kolesar et al. (1975). Improved computer power has made this approach easier to apply.

Further progress with the ODE approach has also been made by introducing other approximations. Much more efficient ODE algorithms for the TV mean and variance were subsequently obtained by Rothkopf and Oren (1979) by employing closure approximations to dramatically reduce the number of equations; also see Taaffe and Ong (1987), Ong and Taaffe (1989) and others.

Despite the successes of the ODE approach to TV queues, there are two deficiencies. First, the ODE approach only applies to TV Markov processes. Second, just like computer simulation and some other numerical approaches, such as the numerical-transform-inversion algorithm of Choudhury et al. (1997a), the ODE approach yields the numerical values of performance measures, but it does not otherwise provide any structural insight.

This second deficiency has recently been addressed by Massey and Pender (2013) and Pender and Massey (2017) by developing closure approximations for the $M_t/M_t/s$ model and more general

TV Markovian systems in the context of many-server heavy-traffic (MSHT) limits as in Mandelbaum et al. (1998), which yield deterministic fluid and stochastic diffusion approximations. They use the closure approximation to greatly improve the numerical accuracy of the MSHT diffusion approximations.

However, no such link has yet been provided between numerical algorithms and the very different conventional heavy-traffic (HT) limits for single-server models. In fact, the HT limits for TV single-server queues tend to be quite intractable themselves, as can be seen from Mandelbaum and Massey (1995) and Whitt (2014, 2016), so that we need new tractable approximation methods.

1.1. Main Contributions

1. In this paper, we introduce a time-varying robust queueing (TVRQ) approach to single-server queueing systems that addresses the two deficiencies mentioned above. In particular, we develop a TVRQ algorithm to approximate the TV workload in the non-Markov $G_t/G_t/1$ single-server queue. Like Rothkopf and Oren (1979), we focus on the special case of the dynamic steady-state behavior of a system with a periodic arrival rate. In doing so, we establish new periodic TVRQ (PRQ). This paper evidently is the first application of robust optimization to study the performance of a queueing model with time-varying arrival rates.

2. Even for the stationary model, we contribute by extending Whitt and You (2018b) to approximate all quantiles as well as the mean. The PRQ provides remarkably tractable approximations; e.g., see (20), (22) and (28). Extensive simulation experiments confirm that the quantile connection is remarkably effective.

3. As in Whitt and You (2018b), we develop a non-parametric approximation by exploiting the index of dispersion for work (IDW) to represent the variability of the total input of work over time, independent of its mean. We use the IDW to develop TVRQ and PRQ for models with stochastic dependence as well as a time-varying arrival-rate function. The IDW is convenient for separately characterizing these two important causes of congestion. The non-parametric approach also provides a vehicle to connect the modeling to large datasets.

4. We establish new HT limits for PRQ in the $G_t/G/1$ model. These new HT limits exploit the HT scaling introduced in Whitt (2014, 2016), and so go beyond the earlier HT literature. In particular, time scaling is used within the deterministic arrival-rate function, so that the length of the periodic cycle grows with the traffic intensity ρ. We show that the HT limits for PRQ and the original model do not coincide in general, but they do in associated long-cycle and heavy-traffic double limits; see §6.

1.2. Related Literature

There is a substantial literature on TV single-server queues, which can be divided into three main categories: (i) structural results (e.g., definition and existence of processes), illustrated by Harrison and Lemoine (1977), Heyman and Whitt (1984), Lemoine (1981, 1989) and Rolski (1989), (ii) numerical algorithms, as discussed above, and (iii) asymptotic methods and approximations by Newell (1968a,b,c), Keller (1982), Massey (1985), Mandelbaum and Massey (1995) and Whitt (2014, 2016). The present paper falls in the last two categories.

Robust optimization is a relatively new approach to difficult stochastic models. As in Bertsimas et al. (2011a), Ben-Tal et al. (2009), Beyer and Sendhoff (2007); the main idea is to replace a difficult stochastic model by a tractable optimization problem. We replace an "average-case" expected value by a "worst-case" optimization, where stochastic process sample paths are constrained to belong to uncertainty sets. From a pure-optimization-centric view of the operations research landscape, robust optimization might be viewed as a way to replace stochastic modeling entirely. However, we think of robust optimization as a useful tool that supplements existing tools in our stochastic toolkit. Accordingly, much of this paper is devoted to establishing connections between PRQ and established queueing theory.

Our work on TVRQ builds on our previous paper, Whitt and You (2018b), which developed robust queueing (RQ) algorithms to approximate the expected steady-state waiting-time and workload in stationary single-server queues, aiming especially to capture the impact of dependence among interarrival times and service times. In turn that paper builds on the RQ formulation of

Bandi et al. (2015), which has precedents in earlier work such as Bertsimas and Thiele (2006), Bertsimas et al. (2011b) and references cited there. The principal difference here is that we focus on the TV performance of a TV model instead of the steady-state performance of a stationary model.

Bandi et al. (2018) have also developed an RQ formulation for the transient behavior of stationary models, which tends to be a quite different (but still challenging) problem (and for which there is a large literature, which we do not review here). We remark that the performance of a queueing model with time-varying arrival-rate function can be approximated by the iterative transient analysis of the associated model with a piecewise-constant arrival-rate function, but that approach introduces another level of approximation and is not easy to implement. Indeed, the iterative transient approach to TV queues has evidently has been attempted only once, by Choudhury et al. (1997a).

1.3. Organization

In §2 we formulate TVRQ. In §3 we narrow our scope to focus on PRQ, introduce our framework to approximate the quantiles of the steady-state workload and describe the simulation experiments. In §4 and §5 we study PRQ for underloaded models and overloaded models, respectively. In §6 we establish heavy-traffic limits for PRQ. Supplementary material appears in the e-companion (EC), including proofs and additional simulation examples.

2. TVRQ for the Steady-State Workload in the $G_t/G_t/1$ Queue

In §2.1 we introduce the general time-varying $G_t/G_t/1$ model and define the steady-state workload at each time in that model. In §2.2 we develop the time-varying robust queueing (TVRQ) approximation and in §2.3 we express it in terms of the *index of dispersion for work* (IDW).

2.1. The Steady-State Workload in the $G_t/G_t/1$ Queue

We consider a time-varying version of the standard single-server queue with unlimited waiting space and the first-come first-served service discipline, which we call the $G_t/G_t/1$ queue. As in

Whitt and You (2018b), we will exploit a reverse-time construction of the workload process, but here we will directly construct the steady-state workload at time t. For that purpose, let $A_t(s)$ be the number of arrivals in interval $[t-s,t]$. As in Whitt (2015), let the service requirements be specified separately from the rate at which service is provided. Let service be provided at a time-varying rate $\mu(u)$ at time u, where μ is a right-continuous deterministic nonnegative function with left limits, so that the cumulative service rate available in the interval $[t-s,t]$ is

$$M_t(s) \equiv \int_{t-s}^{t} \mu(u)\, du, \quad s \geq 0, \tag{1}$$

Let the service requirement of customer k be V_k, indexed going backwards from time t. Let the (potential) net-input of work in the interval $[t-s,t]$, $s \geq 0$, be

$$X_t(s) \equiv \sum_{k=1}^{A_t(s)} V_k - M_t(s), \quad t \geq 0. \tag{2}$$

Then the steady-state workload at time t is

$$W_t \equiv \sup_{s \geq 0} \{X_t(s)\}, \tag{3}$$

which we assume is almost surely finite.

For our supporting mathematical results and simulation examples, we will impose more structure. We impose one-dimensional partial characterizations of the variability of the arrival and service processes by assuming that the arrival process A takes the form of

$$A_t(s) = N(\Lambda_t(s)), \quad s \geq 0, \tag{4}$$

where the base process N is a unit-rate stationary point process satisfying the FCLT

$$\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - nt] \Rightarrow c_a B_a \quad \text{in} \quad \mathcal{D}, \tag{5}$$

with $\Rightarrow$ denoting convergence in distribution, B_a being standard (drift 1, diffusion 1) Brownian motion (BM) and $\mathcal{D}$ the function space of (right-continuous with left limits) sample paths as in Whitt (2002b), while the cumulative arrival rate function is

$$\Lambda_t(s) \equiv \int_{t-s}^{t} \lambda(u)\, du, \quad t \geq 0, \tag{6}$$

with the arrival-rate function λ being a deterministic nonnegative function in $\mathcal{D}$ (e.g., ensuring that the integral is well defined). If N is a Poisson process, then A is a nonhomogeneous Poisson process, but we allow other possibilities. Similarly, we assume that $\{V_k\}$ is a stationary sequence, independent of the process N, with $E[V_k] = 1$ satisfying the FCLT

$$\hat{S}_n(t) \equiv n^{-1/2}\left[\sum_{k=1}^{\lfloor nt \rfloor} V_k - nt\right] \Rightarrow c_s B_s \quad \text{in} \quad \mathcal{D}, \tag{7}$$

where B_s is a BM independent of B_a. The actual service times are relatively complicated; see §3.1 of Whitt (2015). However, we will primarily focus on the standard special case $\mu(s) \equiv 1$, where the service times coincide with the service requirements. If $\mu(t) \equiv 1$, then W_t is the usual virtual waiting time. More generally, the virtual waiting time can be expressed in terms of the workload as a first passage time, as in Lemma 4.1 of Ma and Whitt (2018a).

From all past work, e.g., Theorem 1 of Massey (1985), it is known that the performance at time t depends strongly on the loading, which depends on the history of the rates before time t, as characterized by the *time-varying traffic intensity*

$$\rho^*(t) \equiv \sup_{s \geq 0} \{\Lambda_t(s)/M_t(s)\} \tag{8}$$

for Λ_t in (6) and M_t in (1), which is to be distinguished from the *instantaneous traffic intensity*

$$\rho(t) \equiv \lambda(t)/\mu(t). \tag{9}$$

The model is called overloaded (OL), underloaded (UL) and critically loaded (CL) at time t if $\rho^*(t) > 1$, < 1 and $= 1$, respectively.

Remark 1. (an alternative representation) Combining (2), (3) and (6), we have the following equivalent representation of the steady-state workload

$$W_t = \sup_{s \geq 0}\left\{\sum_{k=1}^{N\left(\Lambda_t\left(M_t^{-1}(s)\right)\right)} V_k - s\right\},$$

which can be viewed as an equivalent system with alternative arrival-rate function $\Lambda_t\left(M_t^{-1}(s)\right)$.

2.2. Time-Varying Robust Queueing (TVRQ)

From (3), we see that the steady-state workload at time t can be formulated directly a supremum. For our TVRQ, we apply robust optimization in the setting of §2.1 by replacing the stochastic model of the reverse-time net input process $X_t(s)$ in (2) and (3) by an appropriate deterministic uncertainty set $\mathcal{U}_t$ and then analyzing the worst case scenario. In particular, we let the TVRQ approximation of the steady-state workload at time t be

$$W_t^* \equiv \sup_{X_t \in \mathcal{U}_t} \sup_{s \geq 0} \{X_t(s)\}, \tag{10}$$

where $\mathcal{U}_t$ is the deterministic uncertainty set

$$\mathcal{U}_t \equiv \{X_t(s) \in \mathbb{R} : X_t(s) \leq E[X_t(s)] + bSD(X_t(s))), \quad s \geq 0\}, \tag{11}$$

with SD being the standard deviation and b being a parameter to be specified.

The uncertainty set in (11) is a natural time-varying generalization of the uncertainty sets in Whitt and You (2018b), which are similar to the ones used in Bandi et al. (2015). The main idea is that (11) can be based on a Gaussian approximation for $X_t(s)$, assuming that the supremum is attained for s not too small, which in turn is supported by a FCLT for $X_t(s)$ in (2), which follows from the assumed FCLTs in (5) and (7); see the EC of Whitt and You (2018b).

For applications, the practical meaning of the Gaussian approximation for the net input process $X_t(s)$ supporting (11) is that our TVRQ approximation is intended for high-volume systems. High-volume means high arrival rates and service rates, which we achieve by scaling time. We are also primarily aiming to treat large-scale systems. Large scale is achieved by having the system operate under heavy-traffic conditions, i.e., by having high instantaneous traffic intensities over extended periods of time. For large-scale high-volume systems, the supporting FCLTs are appropriate, being intimately related to the heavy-traffic limits for the queueing model. We will establish new heavy-traffic limits that will further justify the connection.

As in Lemma EC.1 of Whitt and You (2018b), we can interchange the order of the suprema in (10) and write

$$W_t^* \equiv \sup_{s \geq 0} \{E[X_t(s)] + bSD(X_t(s))\}, \tag{12}$$

where again $X_t(s)$ is defined in (2).

2.3. TVRQ Formulation Using the Index of Dispersion for Work (IDW)

As in Whitt and You (2018b), let the *index of dispersion for work* in the underlying (time-homogenous) process be

$$I_w(t) \equiv \frac{Var\left(\sum_{k=1}^{N(t)} V_k\right)}{E[\sum_{k=1}^{N(t)} V_k]} = t^{-1} Var\left(\sum_{k=1}^{N(t)} V_k\right), \quad t \geq 0, \tag{13}$$

with the last relation holding because $E[N(t)] = t$ and $E[V_k] = 1$. Clearly, the IDW is just a scaled version of the variance function of the total input process, but it is conveniently scaled to be independent of the rate. When the service requirements are independent and identically distributed (i.i.d.) with squared coefficient of variation (scv, variance divided by the square of the mean) c_s^2,

$$I_w(t) = I_a(t) + c_s^2, \quad t \geq 0, \tag{14}$$

where $I_a(t)$ is the *index of dispersion for counts* (IDC) of the base arrival process N, defined by

$$I_a(t) \equiv \frac{Var(N(t))}{E[N(t)]} = t^{-1} Var(N(t)), \quad t \geq 0, \tag{15}$$

as in §4.5 of Cox and Lewis (1966). When N is Poisson, $I_a(t) = 1$, $t \geq 0$.

For the net input process $X_t(s)$ in (2),

$$E[X_t(s)] = \Lambda_t(s) - M_t(s) \quad \text{and} \quad Var(X_t(s)) = Var\left(\sum_{k=1}^{N(\Lambda_t(s))} V_k\right) = \Lambda_t(s) I_w(\Lambda_t(s)), \tag{16}$$

so that we can express the TVRQ representation for the steady-state workload at time t in terms of the IDW as

$$W_t^* \equiv \sup_{s \geq 0} \{\Lambda_t(s) - M_t(s) + b\sqrt{\Lambda_t(s) I_w(\Lambda_t(s))}\}, \tag{17}$$

where Λ_t and M_t are defined in (6) and (1), while I_w is the IDW defined in (13).

Example 1. (A Markov Model) An important special case is the associated Markov model, where N is a rate-1 Poisson process while $\{V_k\}$ is an i.i.d. sequence of mean-1 random variables with scv c_s^2, so that the total input of work over $[0, t]$ is a nonhomogeneous compound Poisson process. In this case, by (14), $I_w(t) = 1 + c_s^2$ for all t, so that the IDW plays a relatively trivial role. In this case,

$$W_t^* = \sup_{s \geq 0} \{\Lambda_t(s) - s + b\sqrt{(1 + c_s^2)\Lambda_t(s)}\}, \tag{18}$$

for Λ_t in (6). ∎

3. Periodic Robust Queueing (PRQ)

Henceforth in this paper we will narrow the scope and focus on the special case of periodic TVRQ (PRQ), but much of what follows should be applicable more generally. In particular, we will assume that $\mu(s) \equiv 1$, $s \geq 0$, and λ is a periodic nonnegative function with period c and average rate

$$\rho \equiv c^{-1} \int_0^c \lambda(s)\, ds < 1, \tag{19}$$

which makes the steady-state workload W_t in (3) and the TVRQ W_t^* in (17) periodic with period c as well. We then let

$$W_y^* \equiv \sup_{s \geq 0} \{\Lambda_{yc}(s) - s + b\sqrt{\Lambda_{yc}(s) I_w(\Lambda_{yc}(s))}\}, \quad 0 \leq y \leq 1 \tag{20}$$

be the TVRQ at time yc, which we refer to as "position y in the cycle." As before, Λ_t comes from (6) and I_w comes from (13)-(15). We understand that W_y^* is an approximation for W_{yc}.

In §3.1 we introduce a new framework for exploiting the PRQ parameter b to approximate the full distribution of W_y. In §3.2 we describe our simulation experiments that we use to study PRQ.

3.1. Approximating the Full Distribution of W_y

In this section, we show how PRQ W_y^* in (20) with the PRQ parameter b can be used to approximate the full distribution of the stochastic steady-state workload W_{yc} in (3) as a function of y, $0 \leq y \leq 1$, which we do via quantiles. Hence, we refer to as the PRQ(b) algorithm.

In Whitt and You (2018b), we established the connection between RQ and stochastic queues in the case of a stationary model. In particular, we found that the steady-state mean is often well approximated by letting $b = \sqrt{2}$; that choice makes RQ correct for the Kingman bound for $GI/GI/1$ (Corollary 1), the Pollaczek-Khintchine formula for $M/GI/1$ (Corollary 3), heavy-traffic and light-traffic limits for $G/G/1$ (Theorem 5) and can be explained by an exact analysis of Levy processes (§EC.3.2).

From the form of PRQ(b), it is evident that as b increases, the approximation should apply more to the tail of the distribution. We find that a useful connection can be made between the parameter

b and the quantiles of the distribution of the steady-state workload W_{yc} at position y within a cycle. For a nonnegative random variable Z and $0<p<1$, let the p^{th} *quantile* of (the distribution of) Z be

$$Z(p) \equiv \inf\{z \geq 0 : P(Z \leq z) = p\}, \quad 0<p<1, \tag{21}$$

i.e., the inverse of the cumulative distribution function (cdf). We propose the approximation

$$W_{yc}(\Pi(b)) \approx W_y^*(b), \tag{22}$$

where $W_y^*(b)$ denotes PRQ in (20), while $\Pi : (-\infty, \infty) \to (0,1)$ is a one-to-one continuous function chosen to map the PRQ parameter b into the quantile level p of W_{yc}.

As indicated in §2.1, we find that the form of the mapping $\Pi(b)$ should depend on the loading. To proceed, we focus on the maximum TV traffic intensity, defined by

$$\rho^{\uparrow} \equiv \sup\{\rho^*(t) : 0 \leq t \leq c\}, \tag{23}$$

for ρ^* in (8). The periodic model is called overloaded (OL), underloaded (UL) and critically loaded (CL) if $\rho^{\uparrow} > 1$, $\rho^{\uparrow} < 1$ and $\rho^{\uparrow} = 1$, respectively. In §4 and §5 we examine PRQ in the UL and OL cases. We discuss PRQ in the CL case in §EC.6.

3.2. Simulation Experiments

For simulation comparisons, we will focus on the sinusoidal special case

$$\lambda(t) \equiv \rho + \beta \sin(2\pi\gamma t), \quad t \geq 0, \quad \text{and} \quad c \equiv c(\gamma) \equiv 1/\gamma. \tag{24}$$

with parameter vector (ρ, β, γ). We assume that $\beta \leq \rho < 1$ to ensure that the arrival rate is always nonnegative and periodic steady state is well defined. In §6 when we consider heavy-traffic limits, we will let the parameter pair (β, γ) depend on ρ.

For these simulations, we consider the $GI_t/GI/1$ model with arrival rate function in (24) and i.i.d. service times $\{V_k\}$ with $E[V_k] = 1$ and scv c_s^2 that are independent of a base rate-1 stationary renewal process N used to generate the arrival process via (4). Let c_a^2 be the scv of an interarrival

time in the ordinary renewal process associated with N. Our examples use exponential (M), Erlang (E_k), hyperexponential (H_2, mixture of two exponentials with balanced means, p. 137 of Whitt (1982)) and lognormal distributions, with the scv specified in parentheses for each experiment. By varying the level of variability in the arrival and processes, we can expose and separate the impact of the stochastic variability from the impact of the deterministic time-variability provided by the time-varying arrival rate in (24). We describe the simulation methodology in §EC.2.

4. Underloaded Models

In this section we investigate PRQ for UL models, which of course includes the stationary model as a special case. At first glance, the proposed scheme in (22) deviates from our previous approximation that focused on the steady-state mean in the stationary model in Whitt and You (2018b), but in §4.1 we show that RQ can be generalized to a RQ(b) algorithm that approximates the quantiles in addition to the mean. In §4.2, we show that PRQ(b) is quite effective in approximating the quantiles of the steady-state workload for UL models.

4.1. RQ(b) for Stationary Queueing Models

For stationary queues, the standard heavy-traffic approximation implies that the steady-state workload W should be approximately exponentially distributed; see §5.7 and §9.3 in Whitt (2002b). In particular, for mean-1 service and traffic intensity ρ,

$$P(W > x) \approx e^{-x/m}, \quad x \geq 0, \quad \text{for} \quad m \equiv \frac{\rho c_x^2}{2(1-\rho)}. \tag{25}$$

Thus, for quantile p of W, denoted by $W(p)$, we have $P(W \leq W(p)) \approx 1 - e^{-W(p)/m} = p$, so that

$$W(p) \approx -\ln(1-p)m, \tag{26}$$

for m in (25).

On the other hand, if we apply Theorem 2 of Whitt and You (2018b) to the $M/GI/1$ queue or the RBM approximation, then we get

$$W^*(b) = \frac{b^2 m}{2}, \tag{27}$$

To match the actual mean in $M/GI/1$ for all ρ and to match the mean in heavy-traffic and light-traffic limits, Corollary 3 and Theorem 5 of Whitt and You (2018b) imply that we should chose $b^2 = \sqrt{2}$ in Whitt and You (2018b). Hence, further connection can be made by equating (26) and (27) to obtain an approximation for the desired function Π in (22), getting

$$p \approx \Pi(b) \equiv 1 - e^{-b^2/2}. \tag{28}$$

For the stationary model, we propose the RQ(b) algorithm as in (20) with (22) and (28), where we restrict (20) to stationary arrival rate functions.

By (26), for an exponential random variable, the mean coincides with the $p = 1 - e^{-1} \approx 0.632$ quantile. By (28), this quantile corresponds to $b = \sqrt{2}$. Hence, the RQ(b) algrithm reduces to the RQ algorithm for the steady-state mean workload in Whitt and You (2018b).

4.2. PRQ(b) for Underloaded Models

We now return to the periodic model. To start, we note that an alternative approximation for UL models is the *pointwise stationary approximation* (PSA) as in Green and Kolesar (1991), Massey and Whitt (1998), Whitt (1991b). The idea in PSA is to approximate the time-varying performance at time t in the UL $G_t/G_t/1$ model by using the steady-state performance of the stationary $G/G/1$ model having the parameters that prevail at time t. In our setting, the PSA is appropriate if the cycle length is sufficiently long that the arrival rate does not change too quickly (relative to the service times). The periodic queue then performs at each time approximately the same as the PSA stationary queue, which is discussed in §4.1. As a result, we propose the same mapping $\Pi(b)$ in (28) for UL periodic queues.

Figure 1 demonstrates the performance of PRQ(b) in the UL case. The first three plots in Figure 1 show the simulation estimates of the quantiles at the level of $p = 0.95, 0.8, 0.632, 0.4$ or 0.2 for three models. In each plot, we overlay the PRQ approximations of the quantiles in broken curves, calculated from (22) and (28). Figure 1 shows that (i) our PRQ framework for approximating the full distribution of W_y is very effective; (ii) the estimated mean is close to the 0.6321 quantile and

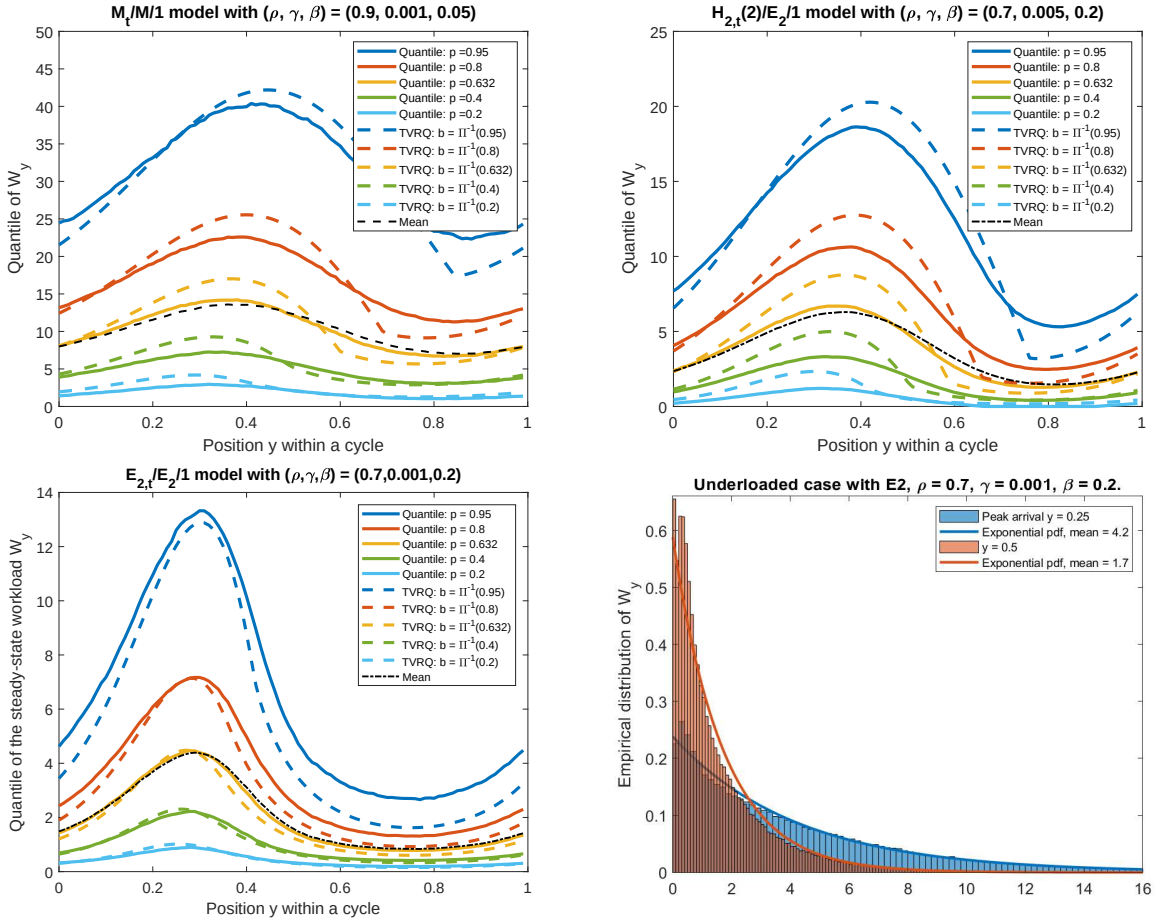

Figure 1 Comparisson of the PRQ quantile approximation in (22) and (28) to simulation estimates of the quantiles in the $M_t/M/1$ model (upper left), $H_{2,t}(2)/E_2/1$ model (upper right) and $E_{2,t}(2)/E_2/1$ model (lower left). The arrival-rate function is (24) with parameters specified in the title of the plot. For the quantile level, we consider $p = 0.95, 0.8, 0.632, 0.4$ and 0.2. The lower right shows the empirical distribution of W_y for the $E_{2,t}(2)/E_2/1$ model at two locations of the cycle: $y = 0.25$ and $y = 0.50$.

PRQ(b) with $b = \sqrt{2}$ serves as a good approximation for the mean in the UL case, as discussed in §4.1; and (iii) even though the exponential approximation draws on the HT limits, we see that our approximation works well under moderate traffic intensity, as demonstrated by the upper right and lower left plots. For the lower right plot, we show the empirical distribution of W_y at two location of the cycle $y = 0.25$ and $y = 0.5$. Both of them are well fitted by exponential distributions, showing that the exponential approximation is appropriate in our settings here.

In Figure 1 the cycle lengths are $c = 1000$ with $\gamma = 0.001$, which is quite long, representing high-volume systems. In contrast, Figure 2 shows the performance of the $M_t/M/1$ model for shorter cycles. Figure 2 shows plots for all combinations of $\rho = 0.7$ and 0.9 and $\gamma = 0.01$ and 0.1. Figure 2

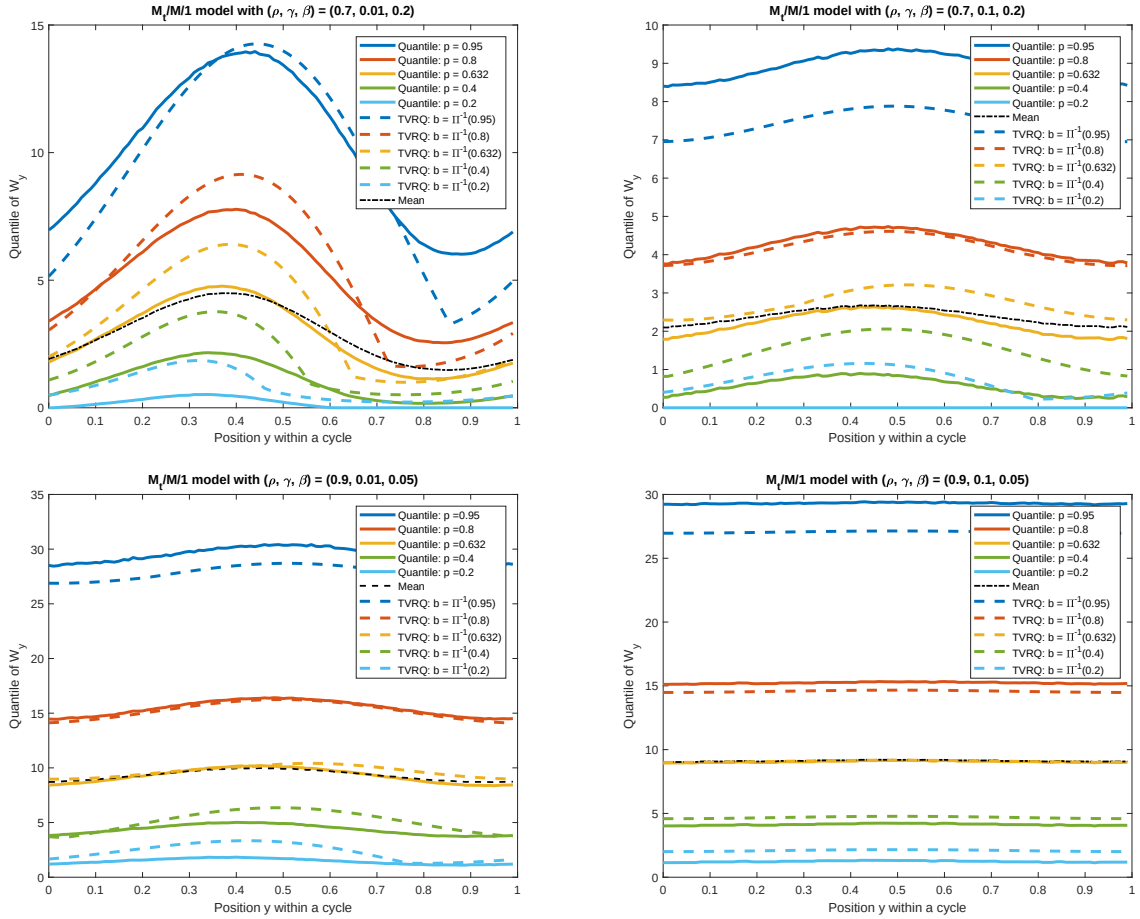

Figure 2 Coomparing the PRQ quantile approximation in (22) and (28) to simulation estimates of the quantiles in the $M_t/M/1$ model for $(\rho, \gamma) = (0.7, 0.01)$ (upper left), $(0.7, 0.1)$ (upper right), $(0.9, 0.01)$ (lower left) and $(0.9, 0.01)$ (lower right),

again shows that PRQ can be effective to approximate both the quantiles and the mean. Figure 2 also shows that PRQ accurately captures the asymptotically stationary performance that prevails in heavy traffic without the extra scaling of the arrival-rate function introduced in Whitt (2014). It also motivates our use of the scaling from Whitt (2014) in our heavy-traffic limits in §6.

To conclude this section, we return to consider PSA, which motivated our use of (28) for periodic UL models as well as stationary models. Unlike the righthand plots in Figure 2, PSA predicts relatively rapid oscillations for short cycles, much like the PSA plot in Figure 1 of Jennings et al. (1996) for many-server models. Figure 3 shows that PSA makes sense for long cycles, but that PRQ provides an improvement. In the present context, we can combine RQ with PSA to create a to obtain PSA-RQ. It suffices to change (17) (with $M(t) \equiv t$) to

$$X^*_{PSA,t} \equiv \sup_{s\geq 0}\left\{\Lambda(t)s - s + b\sqrt{\Lambda(t)sI_w(\Lambda(t)s)}\right\} = \sup_{s\geq 0}\left\{-(1-\rho(t))s + \sqrt{\rho(t)sI_w(\rho(t)s)}\right\}, \qquad (29)$$

which corresponds to a the RQ formula (27) in Whitt and You (2018b) with $\rho \equiv \rho(t) = \lambda(t) < 1$.

Figure 3 compares PRQ and PSA-RQ to simulation estimates for three different models with (24) for $\rho = 0.7$, $\beta = 0.2$ and $\gamma = 0.001$ (left) and $\gamma = 0.01$ (right). As in (25) of Whitt and You (2018b), shows the normalized mean workload $2(1-\rho)E[W_y]/\rho$ (which would be 1 in the $M/D/1$ model) as a function of the position y within the cycle.

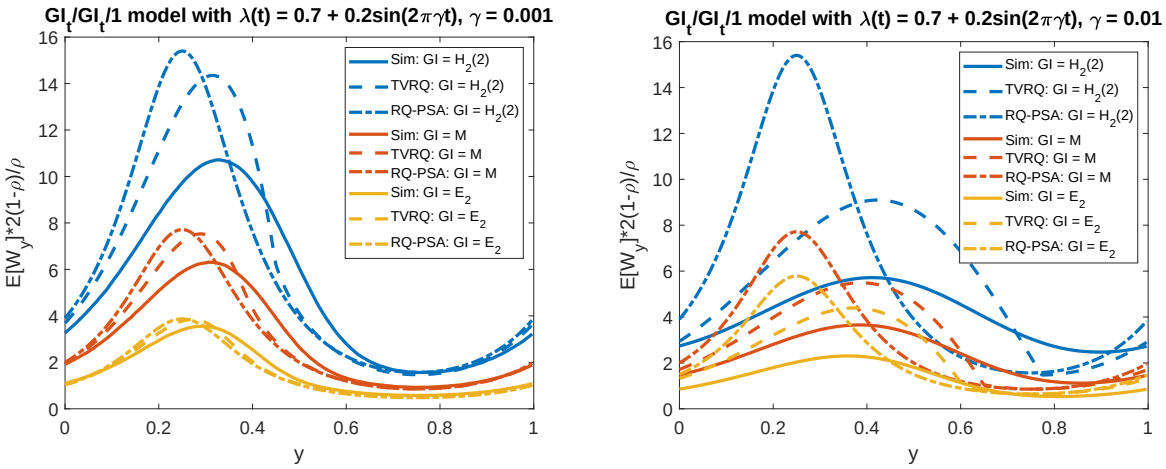

Figure 3 A comparison of PRQ in (20) and PSA-RQ in (29) to simulation estimates of the normalized steady-state mean workload $2(1-\rho)E[W_y]/\rho$ in the UL $GI_t/GI/1$ model with sinusoidal arrival rate in (24) having $(\rho, \beta) = (0.7, 0.2)$ for $\gamma = 0.001$ (left) and $\gamma = 0.01$ (right), as a function of the position y in a cycle. Three cases for the underlying distributions are displayed $(H_2(2), M, E_2)$, being identical for arrival and service.

Figure 3 shows that PRQ provides only a slight improvement over PSA-RQ for $\gamma = 0.001$ (left), but a significant improvement for $\gamma = 0.01$ (right). As before, Figure 3 shows that the quality of

the approximation is excellent for the exponential distribution (M) and lower levels of variability, but degrades for higher variability, serving as an upper bound at the peak (but not uniformly in y). Unlike PSA-RQ, PRQ provides remarkably good estimates of the location of the peak congestion. See §EC.8.2 for more simulation comparisons.

5. Overloaded Models

The behavior of OL models is quite different, especially at the peak. Since $\rho^{\uparrow} > 1$ for $\rho^{\uparrow}$ in (23), PSA does not apply at the peak.

5.1. Deterministic Approximations.

For OL models, it makes sense to consider relatively simple deterministic approximations, which we obtain by assuming that there is no stochastic variability. One way to do so is to assume that $X(t) \equiv \Lambda(t) - M(t) = \Lambda(t) - t$ for all t. As a consequence,

$$W^*_{det,t} = W_{det,t} = \sup_{s\geq 0}\{X_t(s)\} = \sup_{s\geq 0}\{\Lambda_t(s) - s\}. \tag{30}$$

Since the model is deterministic, TVRQ cannot provide an improved performance approximation, but we see that in this case TVRQ is giving the exact time-varying workload. We discuss this model further in §EC.4, but we make two important observations. First, Proposition EC.1 shows that in the periodic case it suffices to do the supremum over one cycle. Second, the deterministic model is very helpful to identify the position $y^{\uparrow}$ where W_y attains its peak; e.g., for the OL sinusoidal model in (24) with $\rho^{\uparrow} > 1$ in (23), measuring time in units of a cycle length, Corollary EC.2 implies that

$$W_{det,y^{\uparrow}} \equiv \sup_{0\leq y\leq 1}\{W_{det,y}\} \quad \text{for} \quad y^{\uparrow} = 0.5 - \arcsin\left(1-\rho\right)/\beta/2\pi. \tag{31}$$

Since the arrival rate has its peak at $y = 0.25$, the time lag in the peak of $W_{det,y}$ is $0.25 - \arcsin\left(1-\rho\right)/\beta/2\pi$, both measured in units of a cycle length.

5.2. Long-Cycle Fluid Limits.

The deterministic model in (30) also arises by taking a long-cycle limit, for which we consider a family of periodic $G_t/GI/1$ stochastic models with growing cycle length indexed by the parameter γ. We assume that model γ has arrival-rate function

$$\lambda_\gamma(t) \equiv \lambda(\gamma t), \quad t \geq 0, \tag{32}$$

for a base periodic arrival-rate function λ. Thus, the arrival rate in model γ is periodic with cycle length $c_\gamma \equiv c/\gamma$. We will let $\gamma \downarrow 0$, so that $c_\gamma \to \infty$.

As regularity conditions for N, we assume that

$$t^{-1}N(t) \to 1 \quad \text{as} \quad t \to \infty \quad \text{w.p.1} \tag{33}$$

and, for all $\epsilon > 0$, there exists $t_0 \equiv t_0(\epsilon)$ such that

$$|t^{-1}N(t) - 1| < \epsilon \quad \text{for all} \quad t \geq t_0 \quad \text{w.p.1.} \tag{34}$$

Both conditions hold when N is a Poisson process and can be anticipated more generally. We prove the following result and provide additional discussion in §EC.4.3.

THEOREM 1. (*long-cycle fluid limit*) *For the periodic $G_t/GI/1$ model under conditions* (33) *and* (34)*, including the scaling in* (20) *as a function of γ,*

$$(\gamma W_{\gamma,y}, \gamma W^*_{\gamma,y}(b)) \to (W_{det,y}, W_{det,y}) \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1} \tag{35}$$

for any b, where $W_{det,y}$ is the deterministic workload in (30) *at time yc within a cycle of length c, while $W^*_{\gamma,y}(b)$ is the PRQ(b) approximation in (20).*

5.3. A Gaussian Approximation for the Quantiles.

The connection to quantiles changes for OL models. Heavy-traffic theory indicates that W_{yc} in the OL period in the cycle should be approximately Gaussian, approximately equal in law to $X_{yc}(s)$

for an appropriate s (where the OL begins in the cycle); e.g., see Newell (1968b), regions B and E in Figure 4.1 of Mandelbaum and Massey (1995), Theorems 5.3.3 (b) and 13.4.2 of Whitt (2002b).

To illustrate, Figure 4 (left) compares PRQ in (20) using $b = 0.50$ developed below and the deterministic approximation in (30) to simulation estimates of the normalized steady-state mean workload $E[W_y]\gamma$, which is consistent with Theorem 1, in the $E_{2,t}/E_2/1$ model with sinusoidal arrival rate in (24), $\rho = 0.7$, $\beta = 0.5$ and three values of γ, as functions of the position y in a cycle. The deterministic approximation is not sensitive to changing cycle length as well as stochastic variability, but it is asymptotically exact as the cycle length grows to infinity. Moreover, both PRQ and the deterministic approximation predicts the location of the peak congestion very well, showing that it lags substantially after the peak of $\lambda(t)$, which is 0.25, again measuring time in cycle lengths. In particular, formula (31) predicts the peak congestion occurs at $y^{\uparrow} = 0.3975$, which is a significant time lag of 0.1475. Figure 4 (left) shows that both the deterministic approximation and PRQ predict this time lag very accurately. We have found that to be consistently true for both OL and UL models.

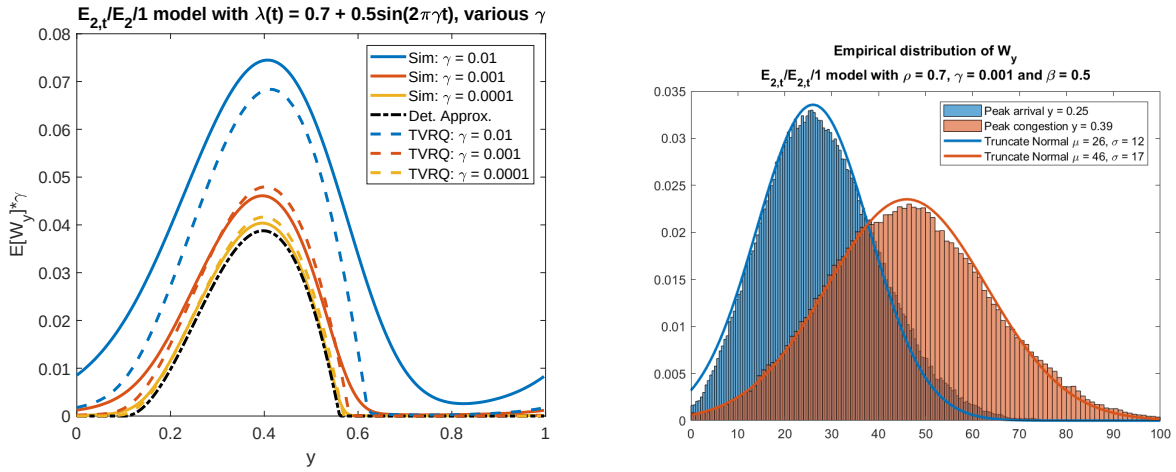

Figure 4 Left: PRQ in (20) and the deterministic approximation in (30) compared to simulation estimates of the normalized steady-state mean workload $\gamma E[W_y]$ in the OL $E_{2,t}/E_2/1$ model with sinusoidal arrival rate in (24), $\rho = 0.7$, $\beta = 0.5$ and three values of γ, as functions of the position y in a cycle. Right: Estimates of the distribution of W_y at the location of the peak of the arrival rate and of W_y.

At this point, we proceeded experimentally. We looked at multiple $G_t/G_t/1$ models to estimate the function $\Pi(b)$ in (22) that relates the TVRQ parameter b to the sample quantiles. To illustrate, Figure 5 compares the quantiles for p ranging from 0.9 to 0.1 estimated by simulation to the PRQ(b) values associated with the parameter b to make PRQ(b) agree as closely as possible. In particular, we focus on $E_{4,t}/E_{4,t}/1, H_{2,t}(8)/H_{2,t}(8)/1$ and $E_{4,t}/H_{2,t}(8)/1$ models and a arrival rate function of $\lambda(t) = 0.9 + 0.8\sin(0.001 * 2\pi t)$.

First, Figure 5 shows that the match is remarkably good for all y. Second, Figure 5 (lower right) shows these numerical results fit to normal cdf's, for which there is remarkable consensus. As a simple overall approximation, we choose

$$\Pi(b) \approx \Phi(b; 0.5, 1.0), \tag{36}$$

where $\Phi(x; m, \sigma^2) \equiv P(N(m, \sigma^2) \le x) = P(N(0, 1) \le (x - m)/\sigma)$ for mean m and variance σ^2. If we want to approximate the mean, then we use $b = 0.5$ because $\Pi(0.5) = 0.5$, the median.

We then tested PRQ(b) with Π in (36) for a range of OL models. Figure 6 illustrates by showing the results for the $M_t/M/1$ model for the parameter vectors $(\rho, \beta, \gamma) = (0.9, 0.5, 0.001)$ and $(0.7, 0.5, 0.01)$. See §EC.8.3 for more simulation comparisons.

6. Heavy-Traffic Limits for Periodic Queues

We now apply heavy-traffic limits to further study periodic robust queueing (PRQ). In §6.1 we first review a heavy-traffic limit for periodic queues from Whitt (2014) and Ma and Whitt (2018b,a). In addition to the conventional heavy-traffic scaling of time in space, as in Ch. 9 of Whitt (2002b), these heavy-traffic limits involve an additional scaling of the arrival rate function. In §6.2 we show how it can be used to generate a diffusion-based parametric PRQ. We then compare our proposed functional PRQ, the diffusion-based parametric PRQ and the direct heavy-traffic diffusion approximation to simulation estimates of the time-varying mean workload. In §6.3 we develop new heavy-traffic limits for PRQ approximation. In §6.4 we establish new heavy-traffic limits combined with long-cycle limits. These involve the three cases: underloaded (UL), overloaded (OL) and critically loaded (CL).

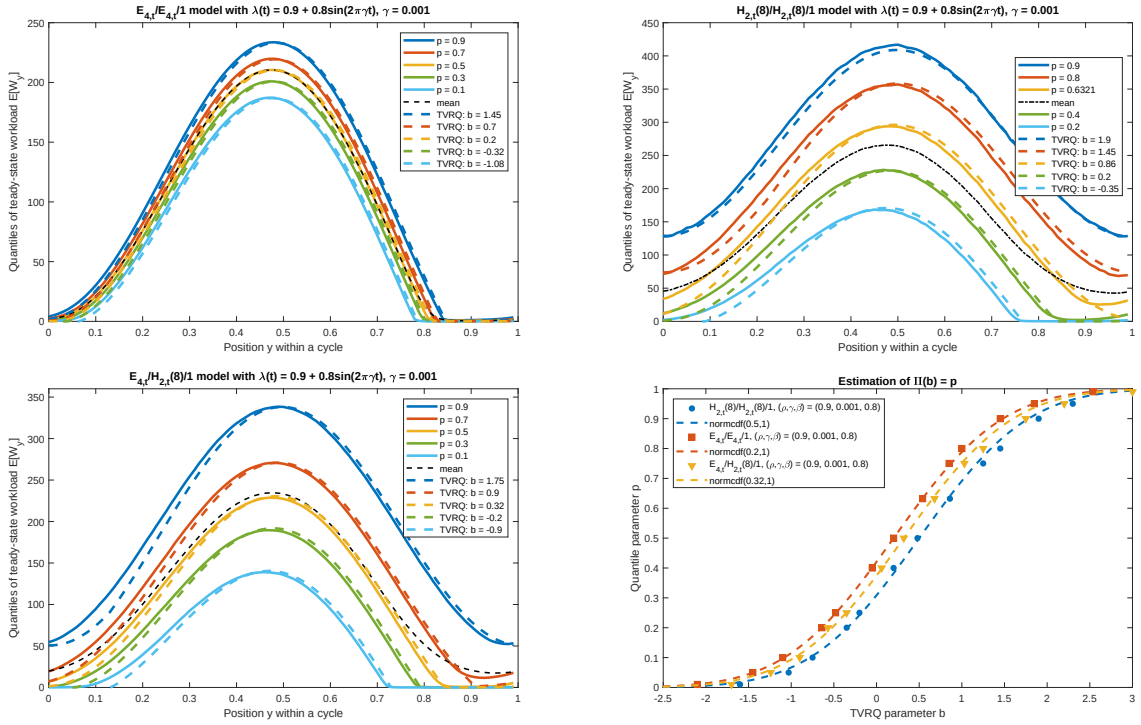

Figure 5 A comparison of quantiles p ranging from 0.9 to 0.1 estimated by simulation to the PRQ(b) values associated with the parameter b to make PRQ(b) agree as closely as possible. for the $E_{4,t}/E_{4,t}/1, H_{2,t}(8)/H_{2,t}(8)/1$ and $E_{4,t}/H_{2,t}(8)/1$ models and the sinusoidal arrival rate function in (24) with $(\rho, \beta, \gamma) = (0.9, 0.8, 0.001)$. These are fit to Gaussian cdf's in the lower right.

6.1. Heavy-Traffic Limit for the Workload Process in the Stochastic Model

We consider a family of models indexed by the long-run average traffic intensity ρ. To avoid notational confusion, we add a subscript d to the diffusion quantities. We let the cumulative arrival-rate function in model ρ be

$$\Lambda_{\gamma,\rho}(t) \equiv \rho t + (1-\rho)^{-1}\Lambda_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0, \tag{37}$$

so that the associated arrival-rate function is

$$\lambda_{\gamma,\rho}(t) \equiv \rho + (1-\rho)\lambda_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0, \tag{38}$$

where

$$\Lambda_{d,\gamma}(t) \equiv \int_0^t \lambda_{d,\gamma}(s)\, ds, \quad \lambda_{d,\gamma}(t) \equiv h(\gamma t), \quad \text{and} \quad \int_0^1 h(t)\, dt = 0 \tag{39}$$

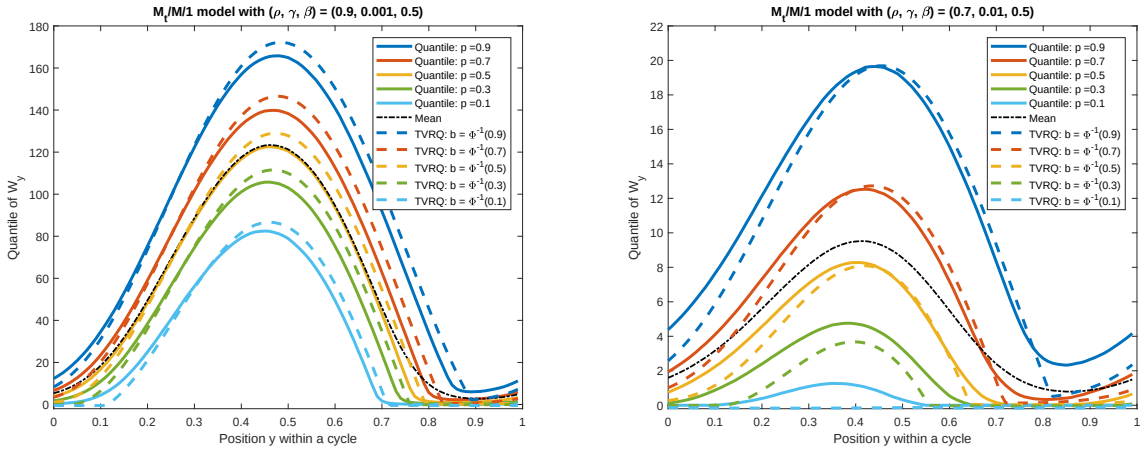

Figure 6 A comparison of quantiles p ranging from 0.9 to 0.1 estimated by simulation to the PRQ(b) based on Π in (36) for the $M_t/M/1$ model and the sinusoidal arrival rate function in (24) with $(\rho, \beta, \gamma) = (0.9, 0.5, 0.001)$ (left) and $(0.7, 0.5, 0.01)$ (right).

with $h(t)$ being a periodic function with period 1. As a consequence, $\lambda_{d,\gamma}(t)$ is a periodic function with period $c_\gamma = 1/\gamma$ and $\lambda_{\gamma,\rho}(t)$ is a periodic function with period $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$. To ensure that $\lambda_{\gamma,\rho}$ is nonnegative, we assume that

$$h(t) \geq -\rho/(1-\rho), \quad 0 \leq t < 1, \tag{40}$$

which will be satisfied for all ρ sufficiently close to the critical value 1 provided that h is bounded below. In fact, we directly assume that

$$-\infty < h^\downarrow \equiv \inf_{0\leq t\leq 1}\{h(t)\} < \sup_{0\leq t\leq 1}\{h(t)\} \equiv h^\uparrow < \infty. \tag{41}$$

There are two primary cases of interest $h^\uparrow < 1$ and $h^\uparrow > 1$. When $h^\uparrow < 1$, the instantaneous traffic intensity, which is $\lambda_{\gamma,\rho}(t)$, satisfies $\lambda_{\gamma,\rho}(t) < 1$ for all t and ρ. On the other hand, when $h^\uparrow > 1$, $\lambda_{\gamma,\rho}(t) > 1$ for some t. When $\lambda_{\gamma,\rho}(t) > 1$ for some t, the workload can reach very high values when time is scaled, because the cycles are very long. That takes us into the setting of Choudhury et al. (1997b).

Theorem 3.2 of Whitt (2014) and Theorem 2 of Ma and Whitt (2018b) provide a heavy-traffic limit as $\rho \uparrow 1$ when $h^\uparrow < 1$. for the workload at time t starting empty at time 0, which we denote by

$W_{\gamma,\rho}(t)$, in the periodic $G_t/GI/1$ model. This heavy-traffic limit is for the time-varying behavior starting empty, but it also applies to the periodic steady-state distribution except for the usual problem of interchanging the order of the limits as $\rho \uparrow 1$ and as $t \uparrow \infty$. We use the periodic steady-state of the limit to approximate the periodic steady-state of the periodic $G_t/GI/1$ queue.

To express the heavy-traffic limits, we use (37) and let

$$A_{\gamma,\rho}(t) \equiv N\left(\Lambda_{\gamma,\rho}(t)\right), \quad Y_{\gamma,\rho}(t) \equiv \sum_{k=1}^{A_{\gamma,\rho}(t)} V_k, \quad \text{and} \quad X_{\gamma,\rho}(t) \equiv Y_{\gamma,\rho}(t) - t, \quad t \geq 0. \tag{42}$$

Then $X_{\gamma,\rho}(t)$ is the net-input process and $W_{\gamma,\rho}(t)$ is the workload process, which is the image of $X_{\gamma,\rho}$ under the reflection map Ψ, i.e.,

$$W_{\gamma,\rho}(t) = \Psi(X_{\gamma,\rho})(t) = \sup_{0 \leq s \leq t} \{X_{\gamma,\rho}(t) - X_{\gamma,\rho}(t-s)\}. \tag{43}$$

For the heavy-traffic functional central limit theorem (FCLT), we introduce the scaled processes

$$\hat{N}_n(t) \equiv n^{-1/2}[N\left(nt\right) - nt], \quad \hat{A}_{\gamma,\rho}(t) \equiv (1-\rho)[A_{\gamma,\rho}\left((1-\rho)^{-2}t\right) - (1-\rho)^2 t],$$
$$\hat{X}_{\gamma,\rho}(t) \equiv (1-\rho)X_{\gamma,\rho}\left((1-\rho)^{-2}t\right) \quad \text{and} \quad \hat{W}_{\gamma,\rho}(t) \equiv (1-\rho)W_{\gamma,\rho}\left((1-\rho)^{-2}t\right), \quad t \geq 0. \tag{44}$$

Let $\mathcal{D}^k$ be the k-fold product space of the function space $\mathcal{D}$. Again let e be the identity map in $\mathcal{D}$, i.e., $e(t) \equiv t$, $t \geq 0$. Recall that $g(x) = o(x)$ as $x \to 0$ if $g(x)/x \to 0$ as $x \to 0$.

THEOREM 2. *(heavy-traffic FCLT, Theorem 3.2 of Whitt (2014) and Theorem 2 of Ma and Whitt (2018b)) For the family of $G_t/GI/1$ models indexed by (γ, ρ) with cumulative arrival-rate functions in* (37)*, if $\hat{N}_n \Rightarrow c_a B_a$ as $n \to \infty$, where B_a is a standard Brownian motion, then*

$$(\hat{A}_{\gamma,\rho}, \hat{X}_{\gamma,\rho}, \hat{W}_{\gamma,\rho}) \Rightarrow (\hat{A}_\gamma, \hat{X}_\gamma, \hat{W}_\gamma) \quad \textit{in} \quad \mathcal{D} \quad \textit{as} \quad \rho \uparrow 1, \tag{45}$$

where

$$(\hat{A}_\gamma, \hat{X}_\gamma, \hat{W}_\gamma) \equiv (c_a B_a + \Lambda_{d,\gamma} - e, \hat{A}_\gamma + c_s B_s, \Psi(\hat{X}_\gamma)), \tag{46}$$

Ψ is the reflection map in (43)*, $\Lambda_{d,\gamma}$ is defined in* (39)*, and B_a and B_s are two independent standard (mean 0 variance 1) Brownian motions; i.e., $\hat{W}_\gamma$ is reflected periodic Brownian motion (RPBM) with*

$$\hat{W}_\gamma = \Psi(c_a B_a + c_s B_s + \Lambda_{d,\gamma} - e) \stackrel{\mathrm{d}}{=} \Psi(c_x B + \Lambda_{d,\gamma} - e), \tag{47}$$

where $c_x^2 = c_a^2 + c_s^2$. The result remains valid if a term of order $o(1-\rho)$ is added to $\Lambda_{\gamma,\rho}$ in (37)*.*

6.2. Three Periodic Approximations from Theorem 2

We directly can obtain three approximations for the mean workload in the periodic $G_t/G_t/1$ model from Theorem 2. In particular, the workload at fixed place y within a cycle for a system which started empty and has run for t time units is

$$W_{\gamma,\rho,y}(t) \stackrel{\mathrm{d}}{=} \sup_{0\le s\le t}\left\{\sum_{k=1}^{A_{\gamma,\rho,y}(s)} V_k - s\right\}, \tag{48}$$

where $A_{\gamma,\rho,y}(s) \equiv A_{\gamma,\rho}(y) - A_{\gamma,\rho}(y-s)$, $A_{\gamma,\rho}(t)$ is defined in (42) and V_k is a generic service time.

As a consequence, first there is the direct diffusion approximation based on (47)

$$\tilde{W}_{\gamma,\rho,y} \equiv \sup_{s\ge 0}\{\Lambda_{\gamma,\rho,y}(s) - s + c_x B(s)\}. \tag{49}$$

Second, there is the parametric PRQ (for the diffusion approximation) obtained from (49) using the mean and variance of BM in (49), namely,

$$\tilde{W}^{**}_{\gamma,\rho,y}(b) \equiv \sup_{s\ge 0}\{\Lambda_{\gamma,\rho,y}(s) - s + bc_x\sqrt{s}\}, \tag{50}$$

where we use $b=\sqrt{2}$ if we are interested in the mean, because this model is UL.

Finally, there is our proposed functional PRQ,

$$\tilde{W}^{*}_{\gamma,\rho,y}(b) \equiv \sup_{s\ge 0}\{\Lambda_{\gamma,\rho,y}(s) - s + b\sqrt{\Lambda_{\gamma,\rho,y}(s) I_w(\Lambda_{\gamma,\rho,y}(s))}\}, \tag{51}$$

where we again use $b=\sqrt{2}$ if we are interested in the mean. Note that (51) does not exploit the diffusion approximation, and so should have advantages away from heavy traffic.

For all simulation examples in the section, we use the base sinusoidal arrival function in (24) with the scaling in (37)-(39), so that

$$\lambda_{\gamma,\rho} = \rho + (1-\rho)h^{\uparrow}\sin(2\pi(1-\rho)^2\gamma t). \tag{52}$$

Figure 7 compares these three approximations for the mean in three cases. First, we consider a case for which the heavy-traffic approximation should perform well. In particular, we first consider the $H_{2,t}(4)/H_{2,t}(4)/1$ model with $(\rho,\gamma)=(0.8,0.01)$ (left). Figure 7 (left) shows that the diffusion performs best, as expected.

Then we consider two cases that should favor PRQ more. For the $LN_t(16)/H_{2,t}(4)/1$ model with $(\rho,\gamma) = (0.55, 0.0001)$ (middle), which has lighter traffic and longer cycles, we see that all three approximations perform about the same, although functional PRQ does better away from the peak. Finally, for the for the $\sum_{i=1}^{10} LN_{i,t}(16)/H_{2,t}(4)/1$ model with the more complex arrival process from the superposition of 10 i.i.d. stationary $LN_t(16)$ renewal processes having $(\rho,\gamma) = (0.6, 0.01)$, we see that functional PRQ performs far better than the others, evidently because the IDC is able to capture the complex dependence in the superposition arrival process.

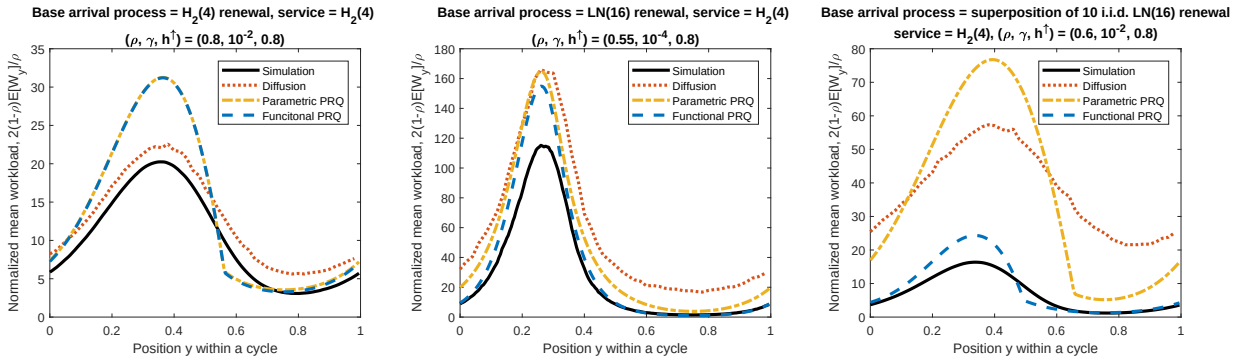

Figure 7 A comparison of the diffusion approximation in (49), the parametric PRQ in (50) and the functional PRQ in (51) for the normalized mean workload $2(1-\rho)E[W_y]/\rho$ as a function of the position y within a cycle to simulation estimates in three cases: standard model (left), lighter traffic and longer cycles (middle) and complex superposition arrival process (right).

6.3. The Heavy-Traffic Limit for PRQ

We now establish a heavy-traffic limit for PRQ as given in (51) above. The proofs for the following results appear in §EC.5.

LEMMA 1. *For a fixed place y within a cycle in the periodic $G_t/G_t/1$ model indexed by (ρ,γ),*

$$\Lambda_{\gamma,\rho,y}(s) = \rho s + \frac{1}{\gamma(1-\rho)} H_{\gamma,\rho,y}(s) \tag{53}$$

where

$$H_{\gamma,\rho,y}(s) \equiv \int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt. \tag{54}$$

and $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$ is the cycle length.

To express the heavy-traffic limit, we define two functions. The first function

$$f(t) \equiv -t + 2\sqrt{t} \tag{55}$$

is a variant of the function to be optimized with the stationary $M/GI/1$ model, as can be seen from Theorem 1 of Whitt and You (2018b). The second function

$$g_{\gamma,\rho,y}(t) \equiv \frac{4}{b^2 c_x^2 \gamma \rho^2} H_{\gamma,\rho,y}\left(\frac{b^2 c_x^2 \rho}{4(1-\rho)^2} t\right) = \frac{4}{b^2 c_x^2 \gamma \rho^2} \int_{y-\frac{b^2 c_x^2 \gamma \rho}{4} t}^{y} h(s) ds \tag{56}$$

is a periodic function that captures the time-varying part of the arrival rate function. The period of $g_{\gamma,\rho,y}(t)$ is $4/b^2 c_x^2 \gamma \rho$. When the arrival-rate function is constant, $g_{\gamma,\rho,y}(t) = 0$ because $h(t) = 0$.

We remark that the constant $\rho c_x^2/2(1-\rho)$ is the exact steady-state mean waiting time in a $M/GI/1$ model, $f(t)$ attains maximum value of 1 at $t = 1$, $g_{\gamma,\rho,y}$ is a periodic function fluctuating around 0 with limits in Lemma EC.3 in §EC.5. Now, we present the heavy traffic limit for PRQ.

THEOREM 3. (*heavy traffic limit for PRQ*) *For the $G_t/G/1$ model with $W^*_{\gamma,\rho,y}(b)$ in* (51)*, f in* (55) *and g in* (56)*,*

$$\lim_{\rho \uparrow 1} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y}(b) = \sup_{t \geq 0} \{f(t) + g_{\gamma,1,y}(t)\}. \tag{57}$$

We immediately obtain an upper bound for the PRQ in the special case of a sinusoidal arrival rate, which reveals the essential shape of the solution, as we shall see in later examples.

COROLLARY 1. *Suppose $h(x) = \beta \sin(2\pi x)$, then*

$$\lim_{\rho \uparrow 1} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} W^*_{\gamma,\rho,y} \leq \lim_{\rho \uparrow 1} f(t) + \lim_{\rho \uparrow 1} g_{\gamma,\rho,y}(t) \leq 1 + \frac{2\beta}{\pi b^2 c_x^2 \gamma}(1 - \cos(2\pi y)), \quad 0 \leq y < 1. \tag{58}$$

REMARK 2. (The heavy traffic limits do not coincide in this case.) Our numerical experiments show that PRQ in Theorem 3 does not coincide with the mean in Theorem 2 in general, but we will get agreement in double limits in the next section.

6.4. Long-Cycle Limits for PRQ in Heavy Traffic

For useful approximations of periodic queues, it is helpful to combine the heavy-traffic perspective with the long-cycle perspective considered in §5.2 and §EC.4.3. When we let the cycles get long in heavy-traffic, we see that there are three very different cases, depending on h in (38) or, equivalently upon the loading $\rho^{\uparrow}$ defined in (23). In the heavy-traffic setting of §6.1-6.3, the three cases are the *underloaded* case in which $h^{\uparrow} < 1$, the *overloaded* case in which $h^{\uparrow} > 1$ and the *critically loaded* case in which $h^{\uparrow} = 1$. We consider the critically loaded case in §EC.6.

6.4.1. Underloaded Queues. In the underloaded case, there will be no times at which the net input rate is positive. We will show that if we let the cycles get long for PRQ in an underloaded model, PRQ is asymptotically consistent with the heavy-traffic limit and PSA.

THEOREM 4. (*long-cycle heavy-traffic limit for PRQ in an underloaded queue*) *Assume that h in* (38) *is continuously differentiable with $h^{\uparrow} < 1$, then the PRQ workload in* (51) *for the $G_t/G/1$ model admits the double limit*

$$\lim_{\substack{\gamma \downarrow 0 \\ \rho \uparrow 1}} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y}(b) = \frac{b^2}{2} \frac{1}{1-h(y)}, \tag{59}$$

so that PRQ is asymptotically consistent with PSA, i.e., the instantaneous traffic intensity is $\rho(y) = \rho + (1-\rho)h(y)$, so that

$$W^*_y(b) = \frac{b^2}{2} \cdot \frac{\rho(y) c_x^2}{2(1-\rho(y))} + o(1-\rho) + o(\gamma). \tag{60}$$

By (28)*, we have*

$$\lim_{\substack{\gamma \downarrow 0 \\ \rho \uparrow 1}} \frac{2(1-\rho)}{\rho c_x^2} \cdot W_y(p) = -\ln(1-p) \frac{1}{1-h(y)} = \lim_{\substack{\gamma \downarrow 0 \\ \rho \uparrow 1}} \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_y\left(\Pi^{-1}(b)\right), \tag{61}$$

where $W_y(p)$ is the p^{th} quantile of W_y and $\Pi(b)$ is defined in (28)*, so that PRQ captures the exact steady-state distribution of the workload W_y in long-cycle heavy-traffic limit.*

REMARK 3. (the iterated limit) We remark that the double limit in Theorem 4 is stronger than a natural iterated limit, which has been established for the $M_t/M/1$ queue and should hold more

generally. In particular, PSA has been proved to be asymptotically correct as $\gamma \downarrow 0$ for the $M_t/M/1$ model in Whitt (1991b). Then RQ has been shown to be asymptotically correct for the stationary model as $\rho \uparrow 1$ in Whitt and You (2018b).

Figure 8 (left) compares the PRQ approximation in (20) and the PSA approximation with the simulated steady-state mean workload. Under moderate traffic intensity $\rho = 0.5$ and moderate cycle length $\gamma = 0.01$, the PRQ provides substantial improvement over PSA. Figure 8 (right) demonstrate the performance of the PRQ approximation for a higher traffic intensity of $\rho = 0.7$ and a longer cycle length with $\gamma = 0.005$, validating Theorem 4.

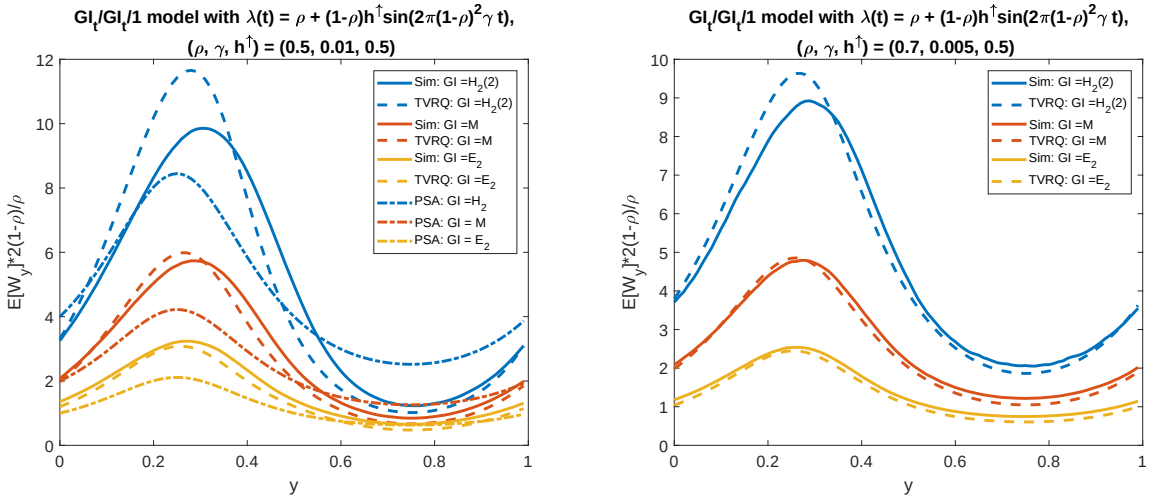

Figure 8 A comparison of PRQ in (20) as a function of the position y within a cycle to simulation estimations of the normalized mean workload $2(1-\rho)E\left[W_{\gamma,\rho,y}\right]/\rho$ for $W_{\gamma,\rho,y}$ in (48) and the limit in Theorem 4 in the underloaded $GI_t/GI_t/1$ model with arrival-rate function in (24) and (37) for the arrival rate function in (52) with $(\gamma, \rho, h^{\uparrow}) \in \{(0.5, 0.01), (0.7, 0.005)\}$. Several interarrival time and service time distribution is considered to demonstrate the robustness of the PRQ algorithm. Left plot also displays the corresponding PSA approximation.

6.4.2. Overloaded Queues. The overloaded case is very different. With long cycles, there will be long stretches of time over which the workload will build up. This will lead to limits with new scaling, as in Choudhury et al. (1997b). Finally, there is the more complicated critically loaded case, which we consider in §EC.6.

THEOREM 5. (*long-cycle limit for PRQ in an overloaded queue*) *For the $G_t/G/1$ model with the heavy-traffic scaling in* (37) *and $h^{\uparrow} > 1$, PRQ in* (51) *admits the long-cycle limit*

$$(1-\rho)\lim_{\gamma\downarrow 0}\gamma\cdot W^*_{\gamma,\rho,y}(b) = \sup_{t\geq 0}\left\{-t+\int_{y-t}^{y} h(s)ds\right\}, \quad 0\leq\rho<1. \tag{62}$$

Note that the long-cycle limit is independent of the parameter b, suggesting a deterministic workload. This is consistent with the long-cycle fluid limit in Theorem 1. Theorem 5 here goes beyond the long-cycle fluid limit by revealing the linear dependence on $(1-\rho)$. This is confirmed in Figure 9, where we observe that the same scaling constant in the simulated mean workload.

REMARK 4. (the space scaling) When the queue is not overloaded, Theorem 5 yields the trivial limit 0, as does Theorem EC.2. That implies that the scaling constant γ in (62) then becomes too much to generate an interesting limit. For underloaded queues, we saw in §6.4.1 that the long-cycle scaling constant γ is not needed. For critically loaded queues, the long-cycle scaling is much more interesting; we discuss that case in §EC.6.

To illustrate, Figure 9 compares PRQ in (20) with parameter $b = 0.5$ as a function of the position y within a cycle to simulation estimates of the normalized mean workload $\gamma(1-\rho)E[W_{\gamma,\rho,y}]$ for $W_{\gamma,\rho,y}$ in (48) and the limit in Theorem 5 in the overloaded $G_t/LN(1)/1$ model with arrival-rate function in (24) and (38) for three values of γ (left) and three values of ρ (right). Figure 9 (left) shows that both simulated values and PRQ approximations converge to the theoretical limit calculated from Proposition EC.1, confirming Theorem 1 and Corollary EC.2, while Figure 9 (right) demonstrates that the scaling constant $(1-\rho)$ also appears in the simulated mean workload. Overall, Figure 9 shows that PRQ serves as a reasonable approximation for the overloaded queues even in moderate cycle length and traffic intensities.

7. Conclusions

In this paper, we have developed a time-varying robust queueing (TVRQ) algorithm to approximate the time-varying workload in a general $G_t/G_t/1$ single-server queue with time-varying arrival-rate and service-rate functions. Exploiting a reverse-time construction of the steady-state workload at

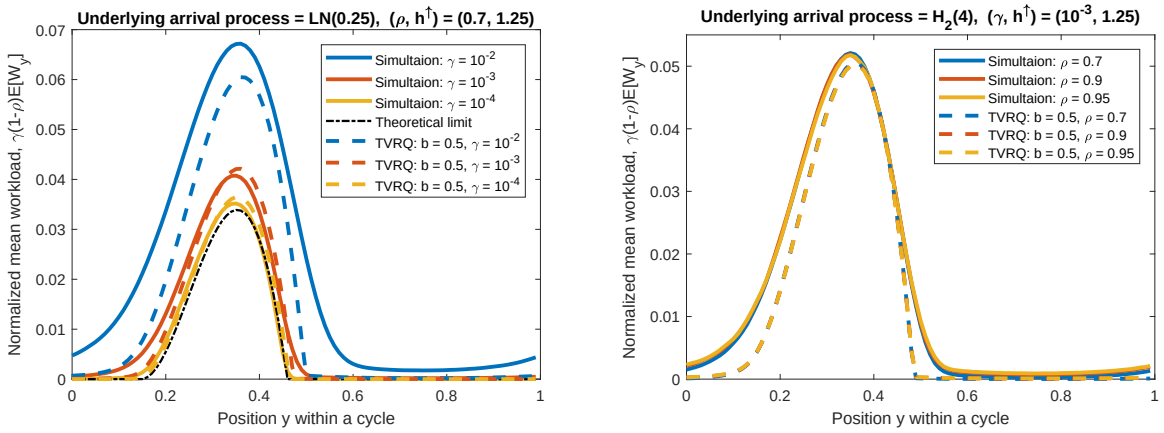

Figure 9 A comparison of PRQ(b) in (20), (22) and (36) as a function of b and the position y within a cycle to simulation estimates of the normalized mean workload $\gamma(1-\rho)E\left[W_{\gamma,\rho,y}\right]$ for $W_{\gamma,\rho,y}$ in (48) and the limit in Theorem 5 in the overloaded $G_t/LN(1)/1$ model with arrival-rate function in (38) and (24) for three values of γ (left) and three values of ρ (right). The arrival rate function is (52) with the parameters specified in each plot.

time t in §2.1, in §2.2 we developed a general TVRQ representation of the steady-state workload at time t as the supremum over an uncertainty set. In (17) in §2.3 we expressed it in terms of the index of dispersion for work (IDW).

The rest of the paper focused on the special case of periodic RQ (PRQ) with unit service rate. In that case we consider the periodic steady-state workload at place yc, $0 \le y \le 1$, within a periodic cycle of length c, focusing especially on high-volume systems (reflected by long cycles) with heavy loading (associated with high traffic intensities). The general representation of the PRQ workload as a function of y appears in (20). We found that the control parameter b can be used to approximate different quantiles of the workload distribution, as indicated in (22). We also found that the function Π in (22) and the performance of the queue depends on the loading $\rho^{\uparrow}$ as defined in (23).

In §4 we found that Π in (28) is effective for underloaded (UL) models with $\rho^{\uparrow} < 1$, and is consistent with RQ for the stationary model in Whitt and You (2018b). In contrast, in §5 for overloaded (OL) models with $\rho^{\uparrow} > 1$, we found that the Gaussian approximation for Π in (36) performs remarkably well. Both PRQ and the more elementary deterministic approximation approximate

the location of the peak remarkably well, as illustrated in Figure 4. Overall, the figures in §4, §5 and the EC provide strong support for PRQ.

In §6 we established heavy-traffic limits as the long-run average traffic intensity ρ increases toward 1 for both the actual periodic workload and the PRQ, using the scaling in Whitt (2014), but in general these limits do not agree. In §6.4 we established double limits as the traffic intensity increases and the cycle length increases. These limits expose three important cases: First, for underloaded models in which the maximum instantaneous traffic intensity remains less than 1, the limit for PRQ is the same as the pointwise stationary approximation (PSA) version of the heavy-traffic limit for the stationary model, which has been shown to be asymptotically correct in Whitt and You (2018b). Second, for the overloaded case, we obtain limits with very different scaling that captures the long periods of overloading, just as in Choudhury et al. (1997b). Third, for critically loaded cases, we obtained the limit for PRQ in Theorem EC.3, consistent with Whitt (2016). In each case, we reported results of simulation experiments that confirm the limit theorems and show that PRQ is remarkably effective. Overall, we conclude that TVRQ can provide helpful insight into complex time-varying queueing models.

We regard this paper is an exploration, opening a promising new line of research. There are many directions for further research. For example, it remains to develop theoretical explanations for the function Π in (36) yielding $b = 0.5$ for OL models and the choice $b = 1$ for CL models in §EC.6. There are opportunities for new insightful asymptotics. It also remains to explore various applications and consider extensions to networks of queues, paralleling Whitt and You (2018a), and queues with multiple servers.

Acknowledgments

Support was received from NSF grants CMMI 1265070 and 1634133.

References

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

Bandi, C., D. Bertsimas, N. Youssef. 2018. Robust transient analysis of multi-server queueing systems and feed-forward networks. *Queueing Systems* **89**(3-4) 351–413.

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

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

Bertsimas, D., D. Gamarnik, A. A. Rikun. 2011b. Performance analysis of queueing networks via robust optimization. *Operations Research* **59**(2) 455–466.

Bertsimas, D., A. Thiele. 2006. A robust optimization approach to inventory theory. *Operations Research* **54**(1) 150–168.

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

Choudhury, G. L., D. L. Lucantoni, W. Whitt. 1997a. Numerical solution of piecewise-stationary $M_t/G_t/1$ queues. *Operatioons Research* **45**(3) 451–463.

Choudhury, G. L., A. Mandelbaum, M. I. Reiman, W. Whitt. 1997b. Fluid and diffusion limits for queues in slowly changing random environments. *Stochastic Models* **13**(1) 121–146.

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

Davis, J. L., W. A. Massey, W. Whitt. 1995. Sensitivity to the service-time distribution in the nonstationary Erlang loss model. *Management Sci.* **41**(6) 1107–1116.

Edie, L. C. 1954. Traffic delays at toll booths. *Operations Research* **2**(2) 107–138.

Eick, S. G., W. A. Massey, W. Whitt. 1993. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** 731–742.

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.

Green, L. V., P. J. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* **37** 84–97.

Harrison, J. M., A. J. Lemoine. 1977. Limit theoorems for periodic queues. *Journal of Applied Probability* **14** 566–576.

Heyman, D. P., W. Whitt. 1984. The asymptoic behavior of queues with time-varying arrivals. *Journal of Applied Probability* **21**(1) 143–156.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42** 1383–1394.

Keller, J. 1982. Time-dependent queues. *SIAM Review* **24** 401–412.

Kolesar, P. J., P. J. Rider, T. B. Craybill, W. E. Walker. 1975. A queueing-linear-programming approach to scheduling police patrol cars. *Operations Research* **23** 1045–1062.

Koopman, B. O. 1972. Air-terminal queues under time-dependent conditions. *Operations Research* **20** 1089–1114.

Lemoine, A. J. 1981. On queues with periodic Poisson input. *Journal of Applied Probability* **18** 889–900.

Lemoine, A. J. 1989. Waiting time and workload in queues with periodic Poisson input. *Journal of Applied Probability* **26**(2) 390–397.

Ma, N., W. Whitt. 2015. Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. *Statistics and Probability Letters* **102** 202–207.

Ma, N., W. Whitt. 2018a. Minimizing the maximum expected waiting time in a periodic single-server queue with a service-rate control. *Stochastic Systems* **8**(4).

Ma, N., W. Whitt. 2018b. A rare-event simulation algorithm for periodic single-server queues. *INFORMS Journal on Computing* **30**(1) 71–89.

Mandelbaum, A., W. A. Massey. 1995. Strong approximations for time-dependent queues. *Mathematics of Operations Research* **20**(1) 33–64.

Mandelbaum, A., W. A. Massey, M. I. Reiman. 1998. Strong approximations for Markovian service networks. *Queueing Systems* **30** 149–201.

Massey, W. A. 1985. Asymptotic analysis of the time-varying $M/M/1$ queue. *Mathematics of Operations Research* **10**(2) 305–327.

Massey, W. A., J. Pender. 2013. Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems* **75**(2-4) 243–277.

Massey, W. A., W. Whitt. 1998. Uniform acceleration expansions for Markov chains with time-varying rates. *Annals of Applied Probability* **9**(4) 1130–1155.

Newell, G. F. 1968a. Queues with time dependent arrival rates, I. *Journal of Applied Probability* **5** 436–451.

Newell, G. F. 1968b. Queues with time dependent arrival rates, II. *Journal of Applied Probability* **5** 579–590.

Newell, G. F. 1968c. Queues with time dependent arrival rates, III. *Journal of Applied Probability* **5** 591–606.

Oliver, R. M., A. H. Samuel. 1962. Reducing letter delays in post offices. *Operations Research* **10** 839–892.

Ong, K. L., M. R. Taaffe. 1989. Nonstationary queues with interrupted poisson arrivals and unreliable/repairable servers. *Queueing Systems* **4** 27–46.

Pender, J., W. A. Massey. 2017. Approximating and stabilizing dynamic rate Jackson networks with abandonment. *Probability in the Engineering and Information Sciences* **31** 1–42.

Rolski, T. 1989. Queues with nonstationary inputs. *Queueing Systems* **5** 113–130.

Rothkopf, M. H., S. S. Oren. 1979. A closure approximation for the nonstationary $M/M/s$ queue. *Management Science* **25**(6) 522–534.

Taaffe, M. R., K. L. Ong. 1987. Approximating $Ph(t)/M(t)/S/C$ queueing systems. *Annals of Operations Research* **8** 103–116.

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

Whitt, W. 1991a. The efficiency of one long run versus independent replications in steady-state simulation. *Management Science* **37**(6) 645–666.

Whitt, W. 1991b. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science* **37**(3) 307–314.

Whitt, W. 2002a. Internet supplement to the book, Stochastic-Process Limits. Available online at: http://www.columbia.edu/∼ww2040.

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

Whitt, W. 2014. Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters* **42** 458–461.

Whitt, W. 2015. Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems* **81** 341–378.

Whitt, W. 2016. Heavy-traffic limits for a single-server queue leading up to a critical point. *Operations Research Letters* **44** 796–800.

Whitt, W., W. You. 2018a. A robust queueing network analyzer based on indices of dispersion. Under review, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Whitt, W., W. You. 2018b. Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research* **66**(1) 184–199.

e-companion

EC.1. Overview

This is an online e-companion to the main paper. It has five more sections, roughly in order of their appearance in the main paper. First, in §EC.2 we describe the simulation methodology, which is applied throughout the main paper, starting in §4. Second, in §EC.3 we present an alternative framework for the function Π in (22) and (28) based on the $M/M/1$ queue instead of heavy-traffic. Next, in §EC.4 we elaborate on the deterministic model in §5.1 and the long-cycle limit in §5.2. In §EC.5 we provide the proofs for §6. We establish a long-cycle heavy-traffic limit for the critically loaded case in §EC.6. We provide heavy-traffic theory for the $G_t/G_t/1$ model in §EC.7. Finally, in §EC.8 we present additional simulation examples that provide further insight into PRQ.

EC.2. Simulation Methodology

The simulations were conducted with C++ on a personal computer. Each simulation run was for 10^8 time units, but the first 10^7 time units were discarded to allow the system to approach steady-state. Since we have unit-rate service times, this amounts to $10^8\rho$ customers, where ρ is the average arrival rate (and traffic intensity). We then divide the cycle into 100 segments. For each segment, we collect the time average of the workload in that segment for each cycle, so that we have a sample size of $10^6\gamma$ for each segment. The mean of the workload at the start of each segment is then estimated by the sample average while the quantiles are estimated by the sample quantiles. That is, we used one long run instead of independent replications; see Whitt (1991a). The run lengths are long enough to make tight confidence intervals, see Figure EC.2 in §EC.8.

We apply the efficient simulation algorithm proposed in Ma and Whitt (2015). In particular, the inversion table for both the cumulative arrival-rate function and the cumulative service-rate function are generated using Algorithm 1 there. For the time-varying external arrival process, a base renewal arrival process is generated and then converted to its time-varying version by using the inversion table, see Algorithm 2 there. For time-varying service process, we first generate the

(stationary) base service time and record the time that the customer enters service. The base service time is then converted to the service time under the time-varying service rate by using the service-rate inversion table starting from the time that the customer enters service. In the special case of periodic service-rate function, only one inversion table is needed, regardless of the starting time of the service.

A rough estimate of the required run time is 6 minutes to conduct the simulation estimates for one case, while the PRQ calculations are relatively negligible. For example, since the upper left plot in Figure 1 displays 5 percentiles, there are 5 cases, so that it would take about 30 minutes to create that plot. For display, the output is exported to MATLAB.

EC.3. An Alternative $M/M/1$ View of the Function Π

In order to consider possible refinements for the function Π in (28), we now consider a concrete queueing model instead of the HT limit. For the UL stationary $GI/GI/1$ queue, there is an atom at the origin with probability $1-\rho$. In particular, for the $M/M/1$ queue,

$$P(W \le x) = 1 - \rho e^{-\rho x/m}, \quad x > 0, \quad \text{for} \quad m \equiv \rho/(1-\rho). \tag{EC.1}$$

Hence, the p quantile is

$$W(p) = -(m/\rho)\ln((1-p)/\rho). \tag{EC.2}$$

If we apply Theorem 2 and Corollary 3 of Whitt and You (2018b) and apply (27) above, then we can equate (27) and (EC.2) to get the more complex formula for Π:

$$\Pi(b) \equiv \Pi(b,\rho) = 1 - \rho e^{-\rho b^2/2}. \tag{EC.3}$$

Note that Π here is not a surjective mapping. For any $p < 1-\rho$, there is no pre-image b. However, the atom at the origin of the workload W has a probability of $1-\rho$, so that $W(p) = 0$ for any $p < 1-\rho$. In this case, we can set $b = 0$ as the pre-image of any $p < 1-\rho$, so that RQ algorithm gives an approximation of 0 for the quantile, which is exact. Consistent with intuition, formula $\Pi(b,\rho)$ in (EC.3) coincides with (28) when $\rho = 1$. The derivative of $\Pi(b,\rho)$ with respect to ρ is $[(\rho b^2/2) - 1]e^{-\rho b^2/2}$.

EC.4. Supporting Theory for the Periodic Deterministic Model in 5.1

We now elaborate on the periodic deterministic model introduced in §5.1. In particular, we assume that $X(t) = \Lambda(t) - t$, where the arrival rate function λ is periodic and the service rate is constant, so that the TVRQ coincides with the exact workload

$$W_t^* = W_t \equiv W_{det,t} \equiv \sup_{s \ge 0} \{\Lambda_t(s) - s\}, \tag{EC.4}$$

as shown in (30).

EC.4.1. Supremum Over Only One cycle

We start with the arrival-rate function $\lambda(t)$ with period c. In order for the model to be interesting (i.e., for there to be positive workload at some time), we also assume that

$$\lambda^\uparrow \equiv \sup_{0 \le s < c} \{\lambda(s)\} > 1. \tag{EC.5}$$

Now the main quantity we focus on is

$$\Lambda_t(s) \equiv \Lambda(t) - \Lambda(t-s), \quad s \ge 0, \quad 0 \le t < c. \tag{EC.6}$$

We now observe that the workload at time t is determined by the input over the cycle ending at time t.

PROPOSITION EC.1. *For the deterministic model, the workload at time t within the cycle $[0, c)$ defined in* (EC.4) *reduces to the supremum over one cycle, i.e.,*

$$W_t = \sup_{0 \le u \le c} \{\Lambda_t(u) - u\}, \quad 0 \le t < c. \tag{EC.7}$$

Proof. Let $s = kc + t$, $0 \le t < c$ and $k \ge 0$. Then

$$\begin{aligned}
W_t &= \sup_{0 \le s \le \infty} \{\Lambda_t(s) - s\}, \quad 0 \le t < c, \\
&= \sup_{0 \le u \le c,\, k \ge 0} \{\Lambda_t(kc+u) - (kc+u)\}, \quad 0 \le t < c, \\
&= \sup_{0 \le u \le c,\, k \ge 0} \{(\Lambda_t(kc+u) - \Lambda_t(u) - kc) + (\Lambda_y(u) - u)\}, \quad 0 \le t < c, \\
&= \sup_{0 \le u \le c,\, k \ge 0} \{-(1-\rho)kc + (\Lambda_t(u) - u)\}, \quad 0 \le t < c, \\
&= \sup_{0 \le u \le c} \{\Lambda_t(u) - u\}, \quad 0 \le t < c,
\end{aligned} \tag{EC.8}$$

because the function inside the supremum is strictly decreasing in k. ∎

EC.4.2. Common Special Cases

We now consider a common structural property that holds in many special cases. In many cases, if we start the periodic cycle at an appropriate point, then we can express the arrival-rate function so that the net input rate is positive on an initial subinterval and then negative thereafter. That is, we assume that there exists δ, $0 < \delta < c$, such that

$$\lambda(t) - 1 \geq 0, \quad 0 \leq t < \delta, \quad \text{and} \quad \lambda(t) - 1 \leq 0, \quad \delta \leq t < c. \tag{EC.9}$$

Often we may require a time shift to satisfy condition (EC.9). In this setting it is easy to determine the periodic fluid W_t, $0 \leq t \leq c$.

Proposition EC.2. *If conditions* (EC.5) *and* (EC.9) *hold, then there exists one and only one* δ^* *with* $0 < \delta < \delta^* < c$ *such that* $\Lambda(\delta^*) = \delta^*$. *Moreover,* $\Lambda(t) - t$ *is nondecreasing over* $[0, \delta]$ *and nonincreasing over* $[\delta, c]$, *so that*

$$W_t = \Lambda(t) - t, \quad 0 \leq y \leq \delta^*, \quad \textit{and} \quad W_t = 0, \quad \delta^* \leq t \leq c, \tag{EC.10}$$

and

$$W^{\uparrow} \equiv \sup_{0 \leq t \leq c} \{W_t\} = W_\delta = \Lambda(\delta) - \delta > 0. \tag{EC.11}$$

We now apply Proposition EC.2 to three special cases. The easiest case appears to be the piecewise-constant case with two pieces.

Corollary EC.1. (*piecewise-constant case*) *If, in addition to the conditions of Proposition EC.2,* $\lambda(t) = a1_{[0,\delta)}(t) + b1_{[\delta,c)}(t)$, *where* $a > 1 > b > 0$, *then*

$$W_t = (a-1)yt, \quad 0 \leq t \leq \delta, \quad W^{\uparrow} = W_\delta = (a-1)\delta, \tag{EC.12}$$

and

$$W_t = (a-1)\delta - (1-b)(t-\delta), \quad \delta \leq t \leq \delta^* \equiv (a-b)\delta/(1-b) \quad \textit{and} \quad W_t = 0, \quad \delta^* \leq t \leq c. \tag{EC.13}$$

The following corollary shows that, for a sinusoidal arrival rate function, the maximum workload is attained shortly before the middle of the arrival-rate cycle.

COROLLARY EC.2. (*sinusoidal case*) *If, in addition to the conditions of Proposition EC.2,* $\lambda(t) = \rho + \beta \sin(2\pi\gamma t)$*, so that a cycle has period* $c(\gamma) \equiv 1/\gamma$ *and the peak is at* $c(\gamma)/4$ *, and* $t_0 = \arcsin((1-\rho)/\beta)/2\pi\gamma = c(\gamma)\arcsin((1-\rho)/\beta)/2\pi$*, then* $\lambda(t_0 + t))$ *satisfies condition* (EC.9) *and* $\delta = c/2 - 2t_0$*, so that in terms of the original* Λ

$$W^{\uparrow} = W_{c/2-t_0} = \Lambda(c/2 - t_0) - \Lambda(t_0) - (c/2) + 2t_0, \tag{EC.14}$$

so that the time lag in the peak is $(c/4) - t_0 = c(0.25 - \arcsin((1-\rho)/\beta)/2\pi)$*. As* $\rho \uparrow 1$*,* $t_0 \equiv t_0(\rho) \downarrow 0$*,* $\delta(\rho) \uparrow 0.5$ *and* $W^{\uparrow} \to \Lambda(c/2) - 0.5$*.*

Finally, to treat general non-sinusoidal examples it may be useful to consider Taylor series expansions of the arrival rate function in order to obtain simple approximation formulas, as in Remark 10 and §3 of Eick et al. (1993), which focuses on the infinite-server model. If we consider arrival-rate functions with a single peak, then that leads to a quadratic approximation in the neighborhood of the peak.

Thus, we next consider a quadratic function, defined so that the condition in (EC.9) holds. Thus, let

$$\lambda(t) \equiv [a - b(t-p)^2]^+, \quad t \geq 0 \quad \text{for} \quad a > 1 \quad \text{and} \quad b > 0, \tag{EC.15}$$

where $[x]^+ \equiv \max\{x, 0\}$, $a > 1$ because OL and $b > 0$ because the peak is at

$$p \equiv \sqrt{(a-1)/b}. \tag{EC.16}$$

We let the arrival-rate function be periodic with cycle length $c > 2p$, chosen so that the average arrival rate is strictly less than 1.

We have chosen p in (EC.15) and (EC.16) so that the net input rate is initially $\lambda(0) - 1 = 0$, but $\lambda(t) - 1 > 0$ for suitably small t with $t > 0$. This value of p is obtained by solving the equation

$$\lambda(t) - 1 = 0 \quad \text{or} \quad a - 1 = b(t-p)^2, \tag{EC.17}$$

which has solution $t = p$ in (EC.16). Clearly, λ is positive over the interval $(0, 2p)$, symmetric about p with $\lambda(0) = \lambda(2p) = 0$. Thus, we can apply Proposition EC.2 to this quadratic example in (EC.15) for $\delta = 2p$.

COROLLARY EC.3. (*quadratic case*) *If $\lambda(t)$ is a quadratic function as defined in* (EC.15) *with cycle length $c > 2p$ and average arrival rate strictly less than* 1*, then the condition in* (EC.9) *holds for $\delta = 2p$, so that the time lag in the peak is p in* (EC.16) *and*

$$\begin{aligned} W^{\uparrow} &\equiv \sup_{0\le t\le c}\{W_t\} = W_\delta = W_{2p} = \Lambda(2p) - 2p \\ &= 2(\Lambda(p) - p) = 2p\left(\frac{2a+1}{3}\right). \end{aligned} \tag{EC.18}$$

To illustrate how the Taylor series approximatin would work, we consider the sinusoidal example in (24). In the setting of (24), we can move the peak to the origin by replacing sine by cosine. Then using the asymptotic expansion $\cos x = 1 - x^2/2 + O(x^4)$ as $x \downarrow 0$, we get that $a = \rho + \beta$ and $b = 2\beta\pi^2\gamma^2$ in (EC.15). Thus the approximate time lag in the peak of W_0^* in (30) and (EC.4) is

$$\text{time lag} \approx \sqrt{(\rho+\beta-1)/2\beta\pi^2\gamma^2} = \left(\frac{1}{\gamma}\right)\left(\frac{1}{2\pi}\right)\sqrt{2(\rho+\beta-1)/\beta}. \tag{EC.19}$$

The final expression separates out the cycle length γ^{-1} and expresses the time lag relative to 2π, so that $\sqrt{2(\rho+\beta-1)/\beta} = 2\pi/4 \approx 1.56$ means that the time lag would be one quarter of a sine cycle; i.e., $y^* - 0.25 = \sqrt{2(\rho+\beta-1)/2\pi\beta}$.

For example, in the setting of Figure 4, where $\rho = 0.7$ and $\beta = 0.5$, the approximate time lag from the quadratic approximation (EC.19) above is $y^* - 0.25 = 0.4/\pi \approx 0.127$, which indicates the peak congestion should be at about 0.377. This is somewhat smaller than the exact time lag of 0.1475 we obtain from applying Corollary EC.2.

EC.4.3. The Long-Cycle Fluid Limit in §5.2

For periodic queues, it is helpful to consider the case of long cycles relative to a fixed service-time distribution. (This case is equivalent to letting the service times become short relative to a fixed arrival rate function.) We now consider a family of periodic $G_t/GI/1$ stochastic models with growing cycle length indexed by the parameter γ. We assume that model γ has arrival-rate function

$$\lambda_\gamma(t) \equiv \lambda(\gamma t), \quad t \ge 0, \tag{EC.20}$$

for the base arrival-rate function λ satisfying (EC.5). Thus, the arrival rate in model γ is periodic with cycle length $c_\gamma \equiv c/\gamma$. We will let $\gamma \downarrow 0$, so that $c_\gamma \to \infty$.

In the stochastic model we can also let the cumulative arrival-rate function be defined in terms of the base cumulative arrival-rate function Λ. In particular, we let

$$\Lambda_\gamma(t) \equiv \gamma^{-1}\Lambda(\gamma t) \quad \text{and} \quad \Lambda_{\gamma,y}(t) \equiv \Lambda_\gamma(\gamma^{-1}y) - \Lambda_\gamma(\gamma^{-1}y - t), \quad 0 \le y < c, \tag{EC.21}$$

so that the associated arrival-rate function is as in (EC.20). The periodic structure with (EC.5) implies the following bound.

LEMMA EC.1. *In the setting above with* (EC.5),

$$\max\{\Lambda(t), \Lambda_y(t)\} \le \rho t + \lambda^\uparrow c \quad \textit{and} \quad \max\{\Lambda_\gamma(t), \Lambda_{\gamma,y}(t)\} \le \rho t + \lambda^\uparrow c/\gamma \quad \textit{for all} \quad t \ge 0. \tag{EC.22}$$

Let $A_\gamma(t)$ and $X_\gamma(t)$ be the associated arrival and net input processes in the $G_t/GI/1$ model, defined by

$$A_\gamma(t) \equiv N(\Lambda_\gamma(t)) \quad \text{and} \quad X_\gamma(t) \equiv \sum_{k=1}^{A_\gamma(t)} V_k - t, \quad t \ge 0, \tag{EC.23}$$

where N is a rate-1 stochastic process and $\{V_k\}$ is the i.i.d. sequence of service times with $E[V_k] = 1$ independent of N and thus of A_γ.

As regularity conditions for N, we assume that

$$t^{-1}N(t) \to 1 \quad \text{as} \quad t \to \infty \quad \text{w.p.1} \tag{EC.24}$$

and, for all $\epsilon > 0$, there exists $t_0 \equiv t_0(\epsilon)$ such that

$$|t^{-1}N(t) - 1| < \epsilon \quad \text{for all} \quad t \ge t_0 \quad \text{w.p.1.} \tag{EC.25}$$

Condition (EC.24) is a strong law of large numbers (SLLN), which is equivalent to the stronger functions SLLN (FSLLN), see §3.2 of Whitt (2002a), while condition (EC.25) is implied by refinements such as the law of the iterated logarithm. Condition (EC.25), together with Lemma EC.1, is needed for Theorem 1 to guarantee that a supremum over the entire real line is attained over a

bounded subinterval, which allows us to apply a continuous mapping argument. Both conditions hold when N is a Poisson process and can be anticipated more generally.

The basis for the fluid limit is a functional law of large numbers for A_γ and X_γ after introducing extra time and space scaling.

LEMMA EC.2. *For the periodic $G_t/GI/1$ model under condition* (EC.24),

$$\gamma A_\gamma(\gamma^{-1}(t)) \to \Lambda(t) \quad \textit{and} \quad \gamma X_\gamma(\gamma^{-1}(t)) \to \Lambda(t) - t \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1} \tag{EC.26}$$

Proof. Observe that

$$\begin{aligned} \gamma A_\gamma(\gamma^{-1}t) &= \gamma N(\Lambda_\gamma(\gamma^{-1}t)) = \gamma N(\gamma^{-1}\Lambda(\gamma(\gamma^{-1}t))) \\ &= \gamma N(\gamma^{-1}\Lambda(t)) \to \Lambda(t) \quad \text{as} \quad \gamma \downarrow 0 \quad \text{w.p.1} \end{aligned} \tag{EC.27}$$

because $\gamma N(\gamma^{-1}t) \to t$ uniformly over bounded intervals w.p.1 by the FSLLN in (EC.24). A further application of the composition mapping yields the corresponding limit for X_γ in (EC.23):

$$\gamma X_\gamma(\gamma^{-1}t) = \gamma \sum_{k=1}^{\gamma^{-1}(\gamma A_\gamma(\gamma^{-1}t))} V_k - t \to \Lambda_f(t) - t \quad \text{as} \quad \gamma \downarrow 0 \quad \text{w.p.1},$$

because

$$\gamma \sum_{k=1}^{\gamma^{-1}t} V_k \to t \quad \text{as} \quad \gamma \downarrow 0 \quad \text{w.p.1}$$

uniformly over bounded intervals w.p.1 by the FSLLN. ∎

Let $W_{\gamma,y}$ be the periodic steady-state workload at time y/γ for $0 \le y < c$ in $G_t/GI/1$ model γ with arrival rate function $\lambda_\gamma(t)$, i.e.,

$$W_{\gamma,y} = \sup_{s \ge 0} \{X_{\gamma,y}(s)\}, \tag{EC.28}$$

where

$$X_{\gamma,y}(t) \equiv X_\gamma(y\gamma^{-1}) - X_\gamma(y\gamma^{-1} - t), \quad t \ge 0, \quad 0 \le y < c, \tag{EC.29}$$

for X_γ in (EC.23). We get a fluid limit for $W_{\gamma,y}$, again after scaling.

THEOREM EC.1. (*long-cycle fluid limit*) *For the periodic $G_t/GI/1$ model under conditions* (EC.24) *and* (EC.25),

$$\gamma W_{\gamma,y} \to W_y \quad \textit{as} \quad \gamma \downarrow 0 \quad \textit{w.p.1}, \tag{EC.30}$$

where W_y is the deterministic workload at time y within a cycle of length c.

Proof. From (EC.28) and (EC.29),

$$\gamma W_{\gamma,y} = \sup_{s\geq 0}\{\gamma Y_{\gamma,y}(\gamma^{-1}s) - s\} \to \sup_{s\geq 0}\{\Lambda_y(s) - s\} = W_y \quad \text{as} \quad \gamma\downarrow 0 \quad w.p.1, \tag{EC.31}$$

where W_y is the periodic workload in the limiting periodic model by virtue of Lemma EC.2 and a further continuity argument. Lemma EC.2 and condition (EC.25) guarantee that it suffices to consider the supremum over a bounded interval, so that the supremum is continuous. ∎

Let $W^*_{\gamma,y}$ be the PRQ workload at time y/γ for $0\leq y<c$.

THEOREM EC.2. (*PRQ is asymptotically correct in the long-cycle fluid limit*) *For the periodic $G_t/GI/1$ model, PRQ with any b, $0<b<\infty$, is asymptotically exact as $\gamma\downarrow 0$, i.e.,*

$$\gamma W^*_{\gamma,y} \to W_y \quad as \quad \gamma\downarrow 0, \tag{EC.32}$$

where W_y is the deterministic workload at time y within a cycle of length c, so that

$$|\gamma W^*_{\gamma,y} - \gamma W_{\gamma,y}| \to 0 \quad as \quad \gamma\downarrow 0 \quad w.p.1. \tag{EC.33}$$

Proof of Theorem EC.2. Observe that

$$\begin{aligned}
\gamma W^*_{\gamma,y} &= \sup_{s\geq 0}\{\gamma\Lambda_{\gamma,y}(\gamma^{-1}s) - s + \gamma\sqrt{b^2\Lambda_{\gamma,y}(\gamma^{-1}s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s))}\} \\
&= \sup_{s\geq 0}\{\Lambda_y(s) - s + \sqrt{b^2\gamma\Lambda_y(s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s))}\} \\
&\to \sup_{s\geq 0}\{\Lambda_y(s) - s\} = W_y \quad \text{as} \quad \gamma\downarrow 0,
\end{aligned} \tag{EC.34}$$

where $\Lambda_{\gamma,y}(t)$ is defined in (EC.21) and again W_y is the workload in the periodic deterministic fluid model. To justify (EC.34), we apply Lemma EC.1 to see that, $b^2\gamma\Lambda_y(s)I_w(\Lambda_{\gamma,y}(\gamma^{-1}s)) \leq b^2\gamma I^{\uparrow}_w[\rho s + \lambda^{\uparrow}c] \leq \gamma(K_1 s + K_2)$ for constants $I^{\uparrow}_w = \sup_t I_w(t)$, K_1 and K_2 and , so that $\sqrt{2b^2\gamma\Lambda_y(s)} \leq \sqrt{\gamma(K_1 s + K_2)} \to 0$ uniformly over bounded interval as $\gamma\downarrow 0$. Hence, it suffices to consider the supremum in (EC.34) over a bounded interval, because the function is negative outside that interval for all sufficiently small γ. Since the limit W_y is the same as in Theorem 1, PRQ has been shown to be asymptotically correct as $\gamma\downarrow 0$. ∎

EC.5. Proofs of Heavy-Traffic Results from §6

Proof of Lemma 1. Observe that

$$\begin{aligned}\Lambda_{\gamma,\rho,y}(s) &\equiv \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho}) - \Lambda_{\gamma,\rho}((k+y)c_{\gamma,\rho} - s) = \Lambda_{\gamma,\rho}(yc_{\gamma,\rho}) - \Lambda_{\gamma,\rho}(yc_{\gamma,\rho} - s) \\ &= \rho s + (1-\rho)^{-1}\int_{y/\gamma-(1-\rho)^2 s}^{y/\gamma} h(\gamma t)dt = \rho s + \frac{1}{\gamma(1-\rho)}\int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt \\ &= \rho s + \frac{1}{\gamma(1-\rho)}H_{\gamma,\rho,y}(s), \end{aligned} \tag{EC.35}$$

where $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$ is the cycle length of $\Lambda_{\gamma,\rho,y}(s)$ and

$$H_{\gamma,\rho,y}(s) \equiv \int_{y-c_{\gamma,\rho}^{-1}s}^{y} h(t)dt. \quad \blacksquare \tag{EC.36}$$

The following lemma presents some basic limits for $g_{\gamma,\rho,y}(t)$.

LEMMA EC.3. *Let h be a differentiable 1-periodic function whose integral over one period is 0. Assume that h satisfies (41), then*

(a). $\lim_{(\gamma,\rho)\to(0,1)} g_{\gamma,\rho,y}(t) = h(y)t$ *uniformly for t in bounded intervals;*

(b). $\lim_{\gamma\to 0} g_{\gamma,\rho,y}(t) = h(y)t/\rho$ *uniformly for t in bounded intervals;*

(c). $\lim_{\gamma\to\infty} g_{\gamma,\rho,y}(t) = 0$ *uniformly for t over $[0,\infty)$;*

(d). $\lim_{\rho\to 1} g_{\gamma,\rho,y}(t) = g_{\gamma,1,y}(t)$ *uniformly for t in bounded intervals.*

Proof. (c) and (d) are trivial corollaries of the definition of $g_{\gamma,\rho,y}(\cdot)$. For (a) and (b), note that

$$\begin{aligned}|g_{\gamma,\rho,y}(t) - h(y)t/\rho| &\le \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |h(s)-h(y)|ds = \frac{4}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |h'(\xi)(s-y)|ds \\ &\le \frac{4M}{b^2c_x^2\gamma\rho^2}\int_{y-\frac{b^2c_x^2\gamma\rho}{4}t}^{y} |s-y|ds = \frac{4M}{b^2c_x^2\gamma\rho^2}\cdot\frac{1}{2}\left(\frac{b^2c_x^2\gamma\rho}{4}t\right)^2 = N\gamma t^2, \end{aligned} \tag{EC.37}$$

where $N \equiv Mb^2c_x^2/8$. Note that the second line requires $h(\cdot)$ to be differentiable. (b) follows directly from (EC.37). To prove (a), we note that $|g_{\gamma,\rho,y}(t) - h(y)t| \le |g_{\gamma,\rho,y}(t) - h(y)t/\rho| + |h(y)t|(1-\rho^{-1})$. $\blacksquare$

LEMMA EC.4. *With f and $g_{\gamma,\rho,y}$ defined in (55) and (56), we have*

$$W^*_{\gamma,\rho,y} = \frac{b^2}{2}\cdot\frac{\rho c_x^2}{2(1-\rho)}\cdot\sup_{t\ge 0}\left\{f(t) + \rho g_{\gamma,\rho,y}(t) + 2\left(\sqrt{(t+(1-\rho)g_{\gamma,\rho,y}(t))\,C_{\gamma,\rho,y}(t)} - \sqrt{t}\right)\right\}, \tag{EC.38}$$

where

$$C_{\gamma,\rho,y}(t) \equiv \frac{1}{c_x^2}\cdot I_w\left(\frac{b^2c_x^2\rho^2}{4(1-\rho)^2}\left(t+(1-\rho)g_{\gamma,\rho,y}(t)\right)\right).$$

Proof. We write

$$W^*_{\gamma,\rho,y} = \sup_{s\geq 0}\left\{(\rho s - s + bc_x\sqrt{\rho s}) + (\Lambda_{\gamma,y,\rho}(s) - \rho s) + bc_x\left(\sqrt{\Lambda_{\gamma,y,\rho}(s)\frac{I_w\left(\Lambda_{\gamma,\rho,y}(s)\right)}{c_x^2}} - \sqrt{\rho s}\right)\right\}.$$

Together with (55) and (56), the change of variable $s = b^2c_x^2\rho t/4(1-\rho)^2$ yields the desired expression. ∎

We remark that the constant $\rho c_x^2/2(1-\rho)$ is the exact steady-state mean waiting time in a $M/GI/1$ model, $f(t)$ attains maximum value of 1 at $t=1$, $g_{\gamma,\rho,y}$ is a periodic function fluctuating around 0 with limits in Lemma EC.3 and that the third component in (EC.38) is typically small, especially when $\rho \approx 1$. Furthermore, we have

$$\lim_{\rho\uparrow 1} C_{\gamma,\rho,y}(t) = \lim_{t\to\infty} I_w(t)/c_x^2 = 1$$

uniformly for t bounded away from 0, where the second equation holds under regularity conditions, see §IV.A of Fendick and Whitt (1989).

Proof of Theorem 3. First, for any small $\varepsilon > 0$, there exist $\delta > 0$ such that

$$\rho g_{\gamma,\rho,y}(t) + 2\left(\sqrt{(t + (1-\rho)g_{\gamma,\rho,y}(t))\, C_{\gamma,\rho,y}(t)} - \sqrt{t}\right) < \varepsilon$$

for all $t < \delta$ and $\rho > \delta$. Recall that $f(t)$ attains its maximum at $t=1$, it suffices to consider the maximization over interval $t \in [\delta, \infty)$ instead. Since $\lim_{\rho\uparrow 1} C_{\gamma,\rho,y}(t) = 1$ uniformly for all t bounded away from 0, $g_{\gamma,\rho,y}(t)$ and $C_{\gamma,\rho,y}(t)$ are bounded, we have

$$\lim_{\rho\uparrow 1}\sqrt{(t + (1-\rho)g_{\gamma,\rho,y}(t))\, C_{\gamma,\rho,y}(t)} - \sqrt{t} = 0$$

uniformly over $t \in [\delta, \infty)$.

Apply Lemma EC.4, and note that

$$\sup_x\{f(x)\} + \inf_x\{g(x)\} \leq \sup_x\{f(x) + g(x)\} \leq \sup_x\{f(x)\} + \sup_x\{g(x)\}$$

for any function $f(x)$ and $g(x)$, we have

$$\lim_{\rho\uparrow 1}\frac{2}{b^2}\cdot\frac{2(1-\rho)}{\rho c_x^2}W^*_{\gamma,\rho,y} = \lim_{\rho\uparrow 1}\sup_{t\geq 0}\left\{f(t) + \rho g_{\gamma,\rho,y}(t))\right\}.$$

Now, we need only consider a bounded interval of t, becasuse $g_{\gamma,\rho,y}(\cdot)$ is uniformly bounded by definition (56) and thus the objective function in the supremum will be negative outside a bounded interval. The result then follows from part (d) of Lemma EC.3. ∎

Proof of Theorem 4. From Lemma EC.4, we have

$$\frac{2}{b^2}\cdot\frac{2(1-\rho)}{\rho c_x^2}\cdot W^*_{\gamma,\rho,y}=\sup_{t\geq 0}\left\{f(t)+\rho g_{\gamma,\rho,y}(t)+2\left(\sqrt{(t+(1-\rho)g_{\gamma,\rho,y}(t))\,C_{\gamma,\rho,y}(t)}-\sqrt{t}\right)\right\}.$$

Now, let $F_{\gamma,\rho,y}(t)\equiv f(t)+\rho g_{\gamma,\rho,y}(t)+2\left(\sqrt{(t+(1-\rho)g_{\gamma,\rho,y}(t))\,C_{\gamma,\rho,y}(t)}-\sqrt{t}\right)$. For the same reason as discussed in the proof of Theorem 3, we can consider only the t's bounded away from 0. Furthermore, since $F_{\gamma,\rho,y}(\cdot)$ is negative outside a bounded interval and that $\sup_{t\geq 0}\{-(1-h(y))t+2\sqrt{t}\}=1/(1-h(y))$, it suffices to prove that $F_{\gamma,\rho,y}(t)$ converges uniformly to $-(1-h(y))t+2\sqrt{t}$ over all bounded interval of t as $(\gamma,\rho)\to(0,1)$. To this end, we write

$$\begin{aligned}\left|F_{\gamma,\rho,y}(t)-\left(-(1-h(y))t+2\sqrt{t}\right)\right|&=\left|\rho g_{\gamma,\rho,y}(t)-h(x)t+2\left(\sqrt{(t+(1-\rho)g_{\gamma,\rho,y}(t))\,C_{\gamma,\rho,y}(t)}-\sqrt{t}\right)\right|\\&\leq|g_{\gamma,\rho,y}(t)-h(x)t|+(1-\rho)|g_{\gamma,\rho,y}(t)|+2\sqrt{t|C_{\gamma,\rho,y}(t)-1|}\\&\quad+2\sqrt{(1-\rho)|g_{\gamma,\rho,y}(t)|C_{\gamma,\rho,y}(t)},\end{aligned}$$

where we used the concavity of the square root function. The result then follows from Lemma EC.3 and the fact that $\lim_{\rho\uparrow 1}C_{\gamma,\rho,y}(t)=1$ uniformly for $t\in[\delta,\infty]$ for any positive δ.

To see that this limit coincides with PSA, note that by (59), we have

$$W^*_y\approx\frac{b^2}{2}\cdot\frac{\rho c_x^2}{2(1-\rho)(1-h(y))}=\frac{b^2}{2}\cdot\frac{\rho c_x^2}{2(1-(\rho+(1-\rho)h(y)))}=\frac{b^2}{2}\cdot\frac{\rho c_x^2}{2(1-\rho(y))}$$

which is asymptotically correct up to $o(1-\rho)$ in the limit. ∎

Proof of Theorem 5. Note that

$$\begin{aligned}W^*_{\gamma,\rho,y}&=\sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s)-s+b\sqrt{\Lambda_{\gamma,\rho,y}(s)I_w(\Lambda_{\gamma,\rho,y}(s))}\right\}\\&=\sup_{s\geq 0}\left\{-(1-\rho)\,s+\frac{1}{\gamma(1-\rho)}\int_{y-c^{-1}_{\gamma,\rho}s}^{y}h(u)du+b\sqrt{\Lambda_{\gamma,\rho,y}(s)I_w(\Lambda_{\gamma,\rho,y}(s))}\right\}\\&=\frac{1}{\gamma(1-\rho)}\cdot\sup_{t\geq 0}\left\{-t+\int_{y-t}^{y}h(u)du+\gamma(1-\rho)bc_x\sqrt{\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t)I_w(\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t))}\right\},\end{aligned}\quad\text{(EC.39)}$$

where we applied a change of variable $c_{\gamma,\rho}t=s$ in the third line. The result follows from the fact that $I_w(t)$ is bounded and that $\Lambda_{\gamma,\rho,y}(c_{\gamma,\rho}t)$ is in the order of $\rho c_{\gamma,\rho}t=\rho t/(\gamma(1-\rho)^2)$ when $\gamma\to 0$. Then the third term in the curly brace will be $O\left(\gamma^{1/2}\right)$ and converges to 0 uniformly over bounded intervals of t. Note also that the function in the supremum is negative for all t sufficiently large, we need only consider a bounded interval for t. ∎

EC.6. Long-Cycle Heavy-Traffic Limits for Critically Loaded Queues

The critically loaded case is more complex in terms of space scaling. Though the space scaling does involve the cycle length parameter γ, it will depend on the detailed structure of the arrival rate function instead of a simple γ we see in Theorem 5. The following theorem reveals the relationship between the space scaling and γ.

THEOREM EC.3. (*long-cycle heavy-traffic limit for PRQ in a critically loaded queue*) *Assume that* $h(t)$ *satisfies*

$$h(t) = 1 - ct^p + o(t^p), \text{ as } t \to 0, \tag{EC.40}$$

for some positive real numbers c *and* p. *Then the long-cycle heavy-traffic limit of the PRQ solution for the* $G_t/G/1$ *model at the critical point* $y = 0$ *is in the order of* $O(\gamma^{-p/(2p+1)}(1-\rho)^{-1})$ *as* $(\rho, \gamma) \to (1, 0)$.

Proof. By (EC.40), we have

$$\begin{aligned} g_{\gamma,\rho,0}(t) = \frac{4}{b^2 c_x^2 \gamma \rho^2} \int_{-\frac{b^2 c_x^2 \gamma \rho}{4} t}^{0} h(s)ds &= \rho^{-1} \left(1 - \frac{c}{p+1} \left(\frac{b^2 c_x^2 \gamma \rho}{4}\right)^p t^{p+1} + o(\gamma^p t^{p+1})\right) \\ &= \rho^{-1} \left(t - M\gamma^p t^{p+1} + o(\gamma^p t^{p+1})\right) \end{aligned}$$

as $\gamma \downarrow 0$ for fixed t, where $M = c\left(b^2 c_x^2 \rho\right)^p / (4^p(p+1))$. Applying Theorem 3 yields

$$\frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,1,0} = \sup_{t \geq 0} \{f(t) + g_{\gamma,1,0}(t)\} = \sup_{t \geq 0} \left\{2\sqrt{t} - M\gamma^p t^{p+1} + o(\gamma^p)\right\}, \text{ as } \gamma \downarrow 0,$$

where the t^{p+1} is removed from the little-o expression by noting that it suffices to consider a bounded interval of t from the proof of Theorem 3. The supremum is then achieved at

$$t^* = \left(\frac{\gamma^{-p}}{(M + o(1))(p+1)}\right)^{2/(2p+1)},$$

with maximum value

$$(2 - 1/(p+1)) \left(\frac{1}{(M + o(1))(p+1)}\right)^{1/(2p+1)} \gamma^{-\frac{p}{2p+1}}$$

as $\gamma \downarrow 0$. ∎

We remark that the scaling in Theorem EC.3 coincides with the scaling in the heavy-traffic FCLT in Theorem 4.1 of Whitt (2016), where the space scaling needed at an isolated critical point was investigated. It was shown there that the space scaling of the heavy traffic limit depends on the detailed structure of the arrival-rate function.

We conducted simulation experiments to confirm Theorem EC.3. To illustrate, Figure EC.1 (left) shows that PRQ(b) with $b=1$ successfully captured the scaling with respect to γ and ρ, which in this sinusoidal case is $\gamma^{-p/(2p+1)}(1-\rho)^{-1}$ for $p=2$. Figure EC.1 (right) shows that both simulation estimation and the PRQ approximation after scaling is relatively insensitive to the traffic intensity ρ.

We end this section by remarking that, in this simulation example, we consider only the mean and applied the PRQ algorithm with robustness parameter $b=1$. This choice sits between our choice of $b=\sqrt{2}$ for the mean in the underloaded case in §4 and $b=0.5$ for the mean in the overloaded case in §5. Our choice of $b=1$ here was experimental. Finding a suitable function Π in (22) for the critically-loaded models remains to be an important direction for future research.

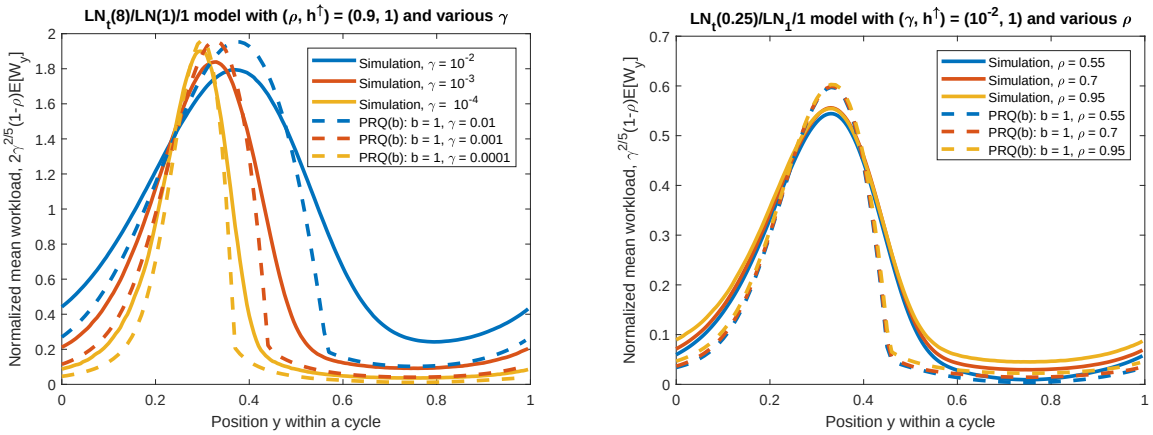

Figure EC.1 Comparing the simulation estimation of the steady-state mean workload to the PRQ(b) approximation in (20) with $b=1$ in two critically-loaded model. The arrival rate function is (52) with the parameters specified in each plot.

EC.7. Heavy-Traffic and Long-Cycle Limits in the $G_t/G_t/1$ model

In this section, we present heavy-traffic and long-cycle limits for the periodic $G_t/G_t/1$ model with sketches of the proofs. We follow the framework for variable service rate introduced in Remark 1, the heavy-traffic scaling in §6.1 and the periodic queueing setup in §6.3. In particular, we focus on the the steady-state workload at a fixed location y within a cycle

$$W_{\gamma,\rho,y} = \sup_{s\geq 0}\left\{\sum_{k=1}^{A_{\gamma,\rho,y}(s)} V_k - M_{\gamma,\rho,y}(s)\right\}$$

as in §6, where $A_{\gamma,\rho,y}(s) \equiv N\left(\Lambda_{\gamma,\rho,y}(s)\right)$. The corresponding PRQ is

$$W^*_{\gamma,\rho,y} = \sup_{s\geq 0}\left\{\Lambda_{\gamma,\rho,y}(s) - M_{\gamma,\rho,y}(s) + b\sqrt{\Lambda_{\gamma,\rho,y}(s) I_w\left(\Lambda_{\gamma,\rho,y}(s)\right)}\right\} \tag{EC.41}$$

as in (51). Here, we keep the same reverse-time cumulative arrival-rate function

$$\Lambda_{\gamma,\rho,y}(s) \equiv \Lambda_{\gamma,\rho}(yc_{\gamma,\rho}) - \Lambda_{\gamma,\rho}(yc_{\gamma,\rho} - s)$$

for $\Lambda_{\gamma,\rho}$ in (37) and $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$. Similarly, we define

$$M_{\gamma,\rho,y}(s) \equiv M_{\gamma,\rho}(yc_{\gamma,\rho}) - M_{\gamma,\rho}(yc_{\gamma,\rho} - s)$$

with

$$M_{\gamma,\rho}(t) \equiv t + (1-\rho)^{-1} M_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0 \tag{EC.42}$$

so that the associated service-rate function is

$$\mu_{\gamma,\rho}(t) \equiv 1 + (1-\rho)\mu_{d,\gamma}((1-\rho)^2 t), \quad t \geq 0,$$

where

$$M_{d,\gamma}(t) \equiv \int_0^t \mu_{d,\gamma}(s)\, ds, \quad \mu_{d,\gamma}(t) \equiv r(\gamma t), \quad \text{and} \quad \int_0^1 r(t)\, dt = 0 \tag{EC.43}$$

for a continuous function r with a cycle length of 1.

With the same heavy-traffic scalings as in (42), we generalize Theorem 2 as follows.

THEOREM EC.4. (*heavy-traffic FCLT for the $G_t/GI_t/1$ model*) *For the family of $G_t/GI_t/1$ models indexed by (γ,ρ) with cumulative arrival-rate functions in* (37) *and cumulative service-rate function in (EC.42), if $\hat{N}_n \Rightarrow c_a B_a$ as $n\to\infty$, where B_a is a standard Brownian motion, then*

$$(\hat{A}_{\gamma,\rho},\hat{X}_{\gamma,\rho},\hat{W}_{\gamma,\rho}) \Rightarrow (\hat{A}_{\gamma},\hat{X}_{\gamma},\hat{W}_{\gamma}) \quad in \quad \mathcal{D} \quad as \quad \rho\uparrow 1,$$

where

$$(\hat{A}_{\gamma},\hat{X}_{\gamma},\hat{W}_{\gamma}) \equiv (c_a B_a + \Lambda_{d,\gamma} - e, \hat{A}_{\gamma} + c_s B_s - M_{d,\gamma}, \Psi(\hat{X}_{\gamma})),$$

Ψ *is the reflection map in* (43), *and B_a and B_s are two independent standard* (*mean 0 variance 1*) *Brownian motions.*

Proof. By definition, we have

$$\begin{aligned}
\hat{X}_{\gamma,\rho}(t) &= (1-\rho)X_{\gamma,\rho}\left((1-\rho)^{-2}t\right) \\
&= (1-\rho)\sum_{k=1}^{A_{\gamma,\rho}\left((1-\rho)^{-2}t\right)} V_k - (1-\rho)M_{\gamma,\rho}\left((1-\rho)^{-2}t\right) \\
&= (1-\rho)\sum_{k=1}^{A_{\gamma,\rho}\left((1-\rho)^{-2}t\right)} V_k - (1-\rho)^{-1}t - M_{d,\gamma}\left(t\right) \\
&\equiv \Xi_{\gamma,\rho}(t) - M_{d,\gamma}\left(t\right).
\end{aligned}$$

where $\Xi_{\gamma,\rho}(t)$ denotes the quantity $\hat{X}_{\gamma,\rho}(t)$ exacly as it appears in Theorem 2, so the result follows. ∎

We remark that this generalized FCLT can be viewed as if we replace $\Lambda_{d,\gamma}$ by $\tilde{\Lambda}_{d,\gamma} \equiv \Lambda_{d,\gamma} - M_{d,\gamma}$ in a $G_t/GI/1$ model, or equivalently, replace h by $\tilde{h} \equiv h - r$ for h in (40) and r in (EC.43).

Next, we generalize the limit theorems for the PRQ problem in (EC.41). As preparation, we re-write $M_{\gamma,\rho,y}$ exactly the same as (53)

$$\begin{aligned}
M_{\gamma,\rho,y}(s) &\equiv M_{\gamma,\rho}((k+y)c_{\gamma,\rho}) - M_{\gamma,\rho}((k+y)c_{\gamma,\rho} - s) = M_{\gamma,\rho}(yc_{\gamma,\rho}) - M_{\gamma,\rho}(yc_{\gamma,\rho} - s) \\
&= s + (1-\rho)^{-1}\int_{y/\gamma-(1-\rho)^2 s}^{y/\gamma} r(\gamma t)dt = s + \frac{1}{\gamma(1-\rho)}\int_{y-c_{\gamma,\rho}^{-1}s}^{y} r(t)dt \\
&= s + \frac{1}{\gamma(1-\rho)}R_{\gamma,\rho,y}(s), \qquad \text{(EC.44)}
\end{aligned}$$

where $c_{\gamma,\rho} = 1/\gamma(1-\rho)^2$ is the cycle length of $M_{\gamma,\rho,y}$ and $R_{\gamma,\rho,y}(s) \equiv \int_{y-c_{\gamma,\rho}^{-1}s}^{y} r(t)dt$. Similar to (56), we define

$$\tilde{g}_{\gamma,\rho,y}(t) \equiv \frac{4}{b^2 c_x^2 \gamma \rho^2} \int_{y-\frac{b^2 c_x^2 \gamma \rho}{4} t}^{y} (h(s) - r(s))ds \tag{EC.45}$$

All generalizations are trivial in the way that we need only replace $g_{\gamma,\rho,y}$ in the original limits by $\tilde{g}_{\gamma,\rho,y}$ here in appropriate places. Equivalently, this can be done by replacing h by $\tilde{h} \equiv h - r$ appropriately as we observed in the generalized FCLT. We demonstrate this idea by proving a generalized version of Lemma EC.4.

LEMMA EC.5. *With f, $g_{\gamma,\rho,y}$ and $\tilde{g}_{\gamma,\rho,y}$ defined in* (55), (56) *and* (EC.45), *we have*

$$W^*_{\gamma,\rho,y} = \frac{b^2}{2} \cdot \frac{\rho c_x^2}{2(1-\rho)} \cdot \sup_{t\geq 0} \left\{ f(t) + \rho \tilde{g}_{\gamma,\rho,y}(t) + 2\left(\sqrt{(t + (1-\rho) g_{\gamma,\rho,y}(t))\, C_{\gamma,\rho,y}(t)} - \sqrt{t} \right) \right\}, \tag{EC.46}$$

where

$$C_{\gamma,\rho,y}(t) \equiv \frac{1}{c_x^2} \cdot I_w \left(\frac{b^2 c_x^2 \rho^2}{4(1-\rho)^2} \left(t + (1-\rho) g_{\gamma,\rho,y}(t) \right) \right).$$

Proof. From (EC.41), we write

$$\begin{aligned} W^*_{\gamma,\rho,y} = \sup_{s\geq 0} \{ &(\rho s - s + bc_x \sqrt{\rho s}) + ((\Lambda_{\gamma,y,\rho}(s) - M_{\gamma,y,\rho}(s) + s) - \rho s) \\ &+ bc_x \left(\sqrt{\Lambda_{\gamma,y,\rho}(s) I_w \left(\Lambda_{\gamma,\rho,y}(s) \right) / c_x^2} - \sqrt{\rho s} \right) \Big\}. \end{aligned}$$

Together with (55), (56) and (EC.45), the change of variable $s = b^2 c_x^2 \rho t / 4(1-\rho)^2$ yields the desired expression. ∎

Hence, we immediately obtain

THEOREM EC.5. (*heavy traffic limit for PRQ*) *The heavy traffic limit of the PRQ problem in* (EC.41) *for the $G_t/G_t/1$ model is*

$$\lim_{\rho \uparrow 1} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0} \{ f(t) + \tilde{g}_{\gamma,1,y}(t) \}. \tag{EC.47}$$

Before presenting the long-cycle heavy-traffic limits, we need to adjust the concept of underloaded, critically loaded and overloaded queues. In the case of a $G_t/G_t/1$ queue, the instantaneous traffic intensity becomes

$$\tilde{\rho}(y) = \frac{\rho + (1-\rho) h(y)}{1 + (1-\rho) r(y)} \tag{EC.48}$$

We now distinguish the three cases by the value of $\tilde{\rho}^{\uparrow} \equiv \sup_y\{\tilde{\rho}(y)\}$. So $\tilde{\rho}^{\uparrow} < 1$, $\tilde{\rho}^{\uparrow} = 1$ and $\tilde{\rho}^{\uparrow} > 1$ corresponds to the underloaded, critically loaded and overloaded case, separately. Equivalently, we can also use $\tilde{h}^{\uparrow}$ as the criteria, where $\tilde{h} = h - r$. Using $\tilde{h}^{\uparrow}$ is preferred because (i) it is more consistent with the notation in §5.2; (ii) it is consistent with our observation of replacing h by $\tilde{h}$ when generalizing to the case of $G_t/G_t/1$ models, as we discussed above.

The rest of the generalizations share the similar idea, and only minor adjustments are needed for the proofs. We list them below.

THEOREM EC.6. (*long-cycle heavy-traffic limit for PRQ in an underloaded queue*) *Assume that h is continuously differentiable with $\tilde{h}^{\uparrow} < 1$, then the PRQ problem in* (EC.41) *for the $G_t/G_t/1$ model admits the double limit*

$$\lim_{\substack{\gamma\downarrow 0\\ \rho\uparrow 1}} \frac{2}{b^2} \cdot \frac{2(1-\rho)}{\rho c_x^2} \cdot W^*_{\gamma,\rho,y} = \frac{1}{1-\tilde{h}(y)}, \tag{EC.49}$$

so that PRQ is asymptotically consistent with PSA, i.e.,

$$W^*_y = \frac{b^2}{2} \cdot \frac{\tilde{\rho}(y) c_x^2}{2(1-\tilde{\rho}(y))} + o(1-\rho). \tag{EC.50}$$

where $\tilde{\rho}(y)$ is the instantaneous traffic intensity in (EC.48)

THEOREM EC.7. (*long-cycle limit for PRQ in an overloaded queue*) *The PRQ problem in* (EC.41) *for the $G_t/G_t/1$ model with the heavy-traffic scaling in* (37) *and $\tilde{h}^{\uparrow} > 1$ admits the long-cycle limit*

$$(1-\rho) \lim_{\gamma\downarrow 0} \gamma \cdot W^*_{\gamma,\rho,y} = \sup_{t\geq 0} \left\{ -t + \int_{y-t}^{y} \tilde{h}(s) ds \right\}, \quad 0 \leq \rho < 1. \tag{EC.51}$$

THEOREM EC.8. (*long-cycle heavy-traffic limit for PRQ in a critically loaded queue*) *Assume that $\tilde{h}(t)$ satisfies*

$$\tilde{h}(t) = 1 - ct^p + o(t^p), \text{ as } t \to 0, \tag{EC.52}$$

for some positive real numbers c and p. Then the long-cycle heavy-traffic limit of the PRQ solution for the $G_t/G_t/1$ model at the critical point $y = 0$ is in the order of $O(\gamma^{-p/(2p+1)}(1-\rho)^{-1})$ as $(\rho, \gamma) \to (1, 0)$.

EC.8. Additional Examples

In this final section of the EC we make additonal simulation comparisons to provide further insight into the performance of PRQ.

EC.8.1. Statistical Precision

We start by showing the statistical precision of our estimated steady-state mean workload. Recall that the simulation methodology is described in §3.2. Figure EC.2 shows the estimation of the steady-state mean workload in two cases, together with the 95% confidence interval (CI). We conclude that the run time used here is sufficient to achieve high statistical precision.

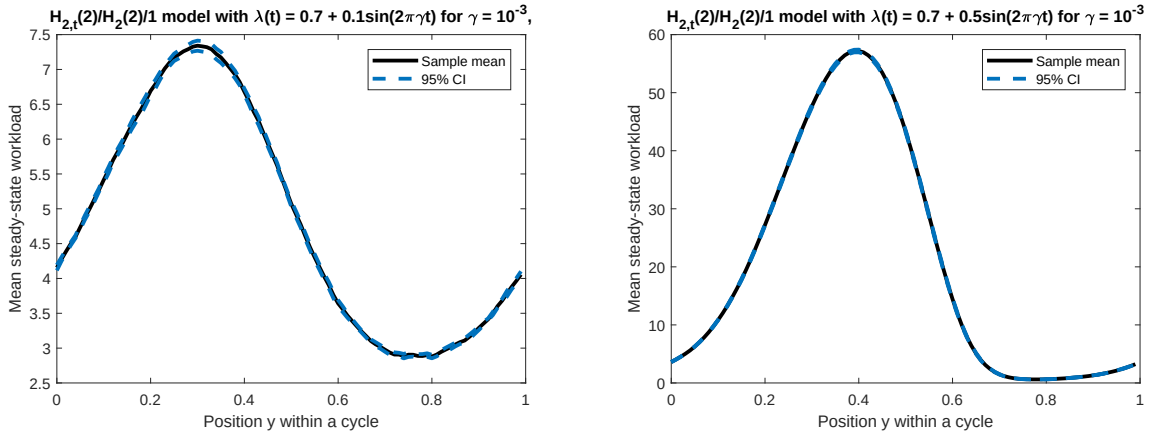

Figure EC.2 The estimated mean of the steady-state mean workload in the $H_{2,t}(2)/H_{2,t}/1$ model with arrival-rate function in (24). Both the UL and OL cases are shown here, with model parameters specified in the titles. The 95% confidence interval is displayed in dashed curves.

EC.8.2. Underloaded Models

For underloaded models, we will compare the simulation estimation of the mean or quantiles of the steady-state workload W_y to the PRQ(b) algorithm specified by (20), (22) and (28). For the mean, we use $b = \sqrt{2}$ as discussed in §4.2. For quantiles, we look at levels $p = 0.95, 0.8, 0.632, 0.4$ and 0.2.

In Figure EC.3, we show the robustness of the PRQ(b) algorithm by presenting four models that share the same arrival rate function but with different interarrival and service time distributions. In

particular, we consider three balanced $GI_t/GI/1$ models with GI being Erlang (E_2) distribution, exponential (M) distribution or hyperexponential ($H_2(2)$) distribution and also one unbalance $LN_t(4)/E_4/1$ model with $LN(4)$ being the Lognormal distirbution with $c_a^2 = 4$. Consistent with our previous observations, PRQ(b) performs very well across different choices of the underlying distribution.

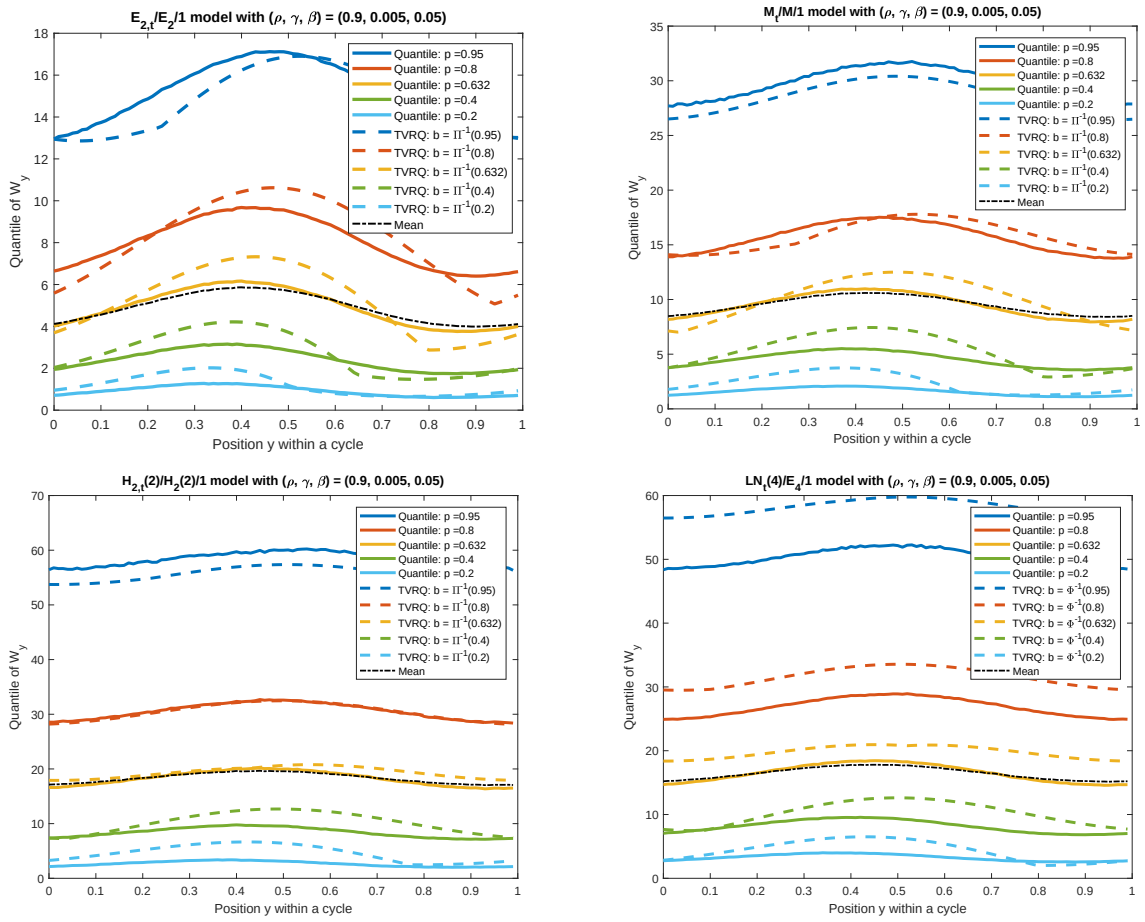

Figure EC.3 The estimated mean and quantiles of the steady-state mean workload in the $GI_t/GI/1$ model with arrival-rate function in (24) and model parameters specified in the titles. Models with four different distributions are displayed to demonstrate the robustness of the PRQ(b) algorithm.

Figure EC.4 supplements Figure 2 by presenting the corresponding long cycle models. In particular, we look at two $M_t/M/1$ models with traffic intensity $\rho = 0.7$ or 0.9. Both examples have a

cycle length of 1000, representing high volumn systems. Both plots compares the TVRQ approximation to the simulation estimation of the quantiles of the stead-state workload, as functions of the position y within a cycle. Again, PRQ(b) performs very well in approximating the full distribution of the steady-state workload.

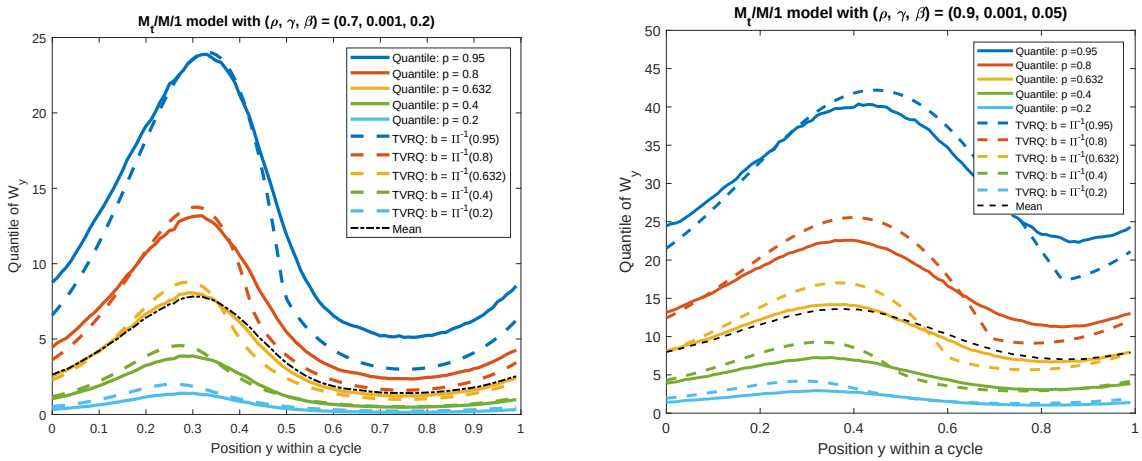

Figure EC.4 The estimated mean and quantiles of the steady-state mean workload in the $M_t/M/1$ underloaded model with arrival-rate function in (24) and model parameters specified in the titles. Left plot shows an example with moderate long-run traffic intensity of $\rho = 0.7$ and a highly variable arrival-rate function $\rho^{\uparrow} = 0.9$. Right plot shows a system with higher traffic intensity of $\rho = 0.9$.

EC.8.3. Overloaded Models

For overloaded models, we will compare the simulation estimation of the mean or quantiles of the steady-state workload W_y to the PRQ(b) algorithm specified by (20), (22) and (36). For the mean, we use $b = 0.5$ as discussed in §5.3. For quantiles, we look at levels $p = 0.9, 0.7, 0.5, 0.3$ and 0.1.

We now present more examples to demonstrate the performance of the PRQ(b) algorithm in the overloaded models. Figure EC.5 present four models with the same arrival-rate function parameters but different underlying distributions for the interarrival and service times. In particular, we consider a wide range of selections in terms of the variability parameter, including a low variability Erlang E_4 distribution and a highly variable hyperexponential $H_2(8)$ distribution. We see that

even the arrival-rate function is fixed, the variability in the underlying distribution can result in different forms of the quantile functions. On the other hand, the PRQ(b) algorithm successfully approximated the distribution of the steady-state workload W_y, as a function of the position y within a cycle.

Figure EC.5 demonstrate that the PRQ(b) algorithm adapts to the changing distribution quite well. However, we can still observe performance degradation as the variability increases, which is caused by our fixed choice of the $\Pi(b)$ function in (36). Further refinements are possible if we allow $\Pi(b)$ to be a function of the variability parameter. But we do not discuss such extensions in this paper.

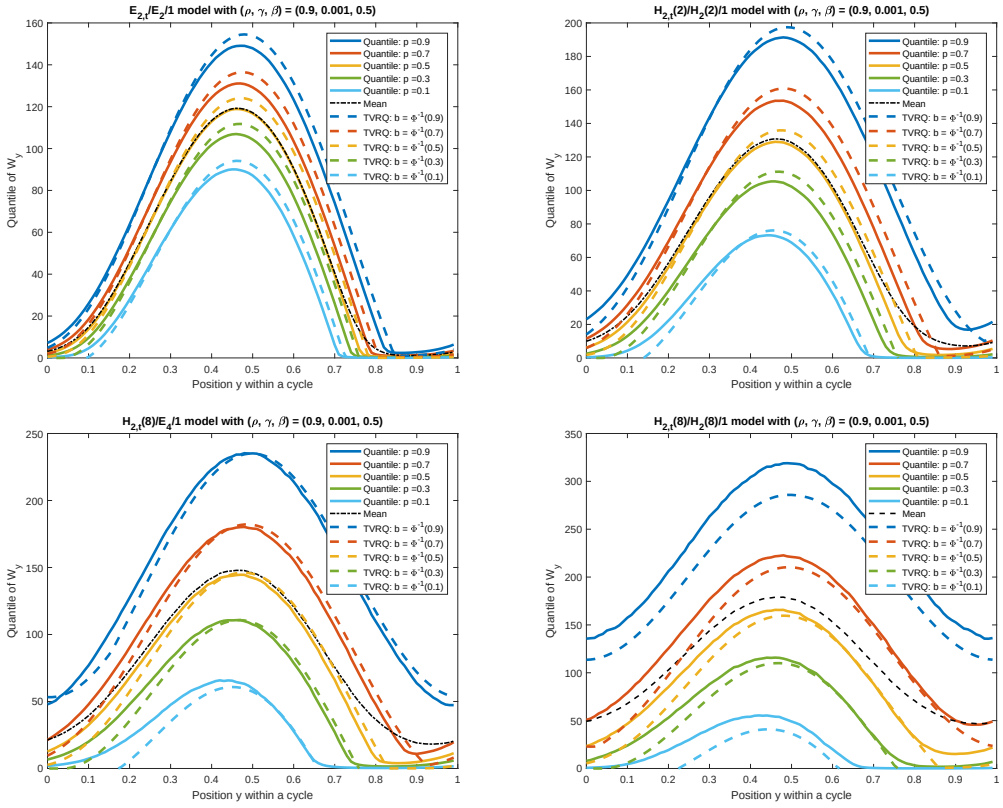

Figure EC.5 The estimated mean and quantiles of the steady-state mean workload in the $GI_t/GI/1$ model with arrival-rate function in (24) and model parameters specified in the titles. Models with four different distributions are displayed to demonstrate the robustness of the PRQ(b) algorithm.

EC.8.4. Long-Cycle Heavy-Traffic Limit

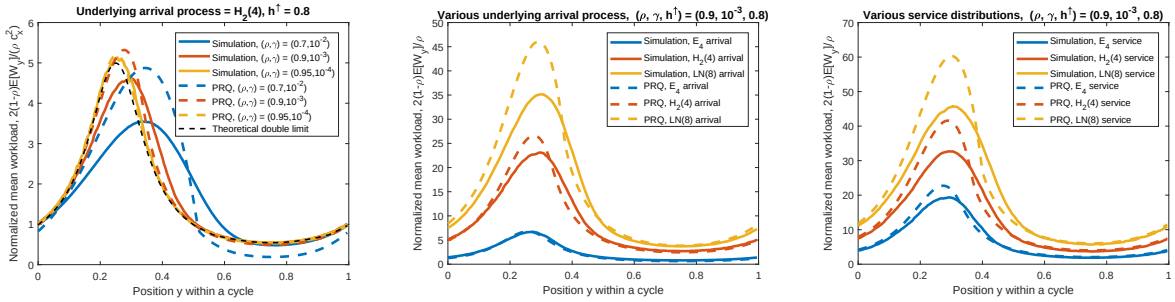

Figure EC.6 A comparison of PRQ in (17) as a function of the position y within a cycle to simulation estimations of the normalized mean workload $2(1-\rho)E\left[W_{\gamma,\rho,y}\right]/\rho$ for $W_{\gamma,\rho,y}$ in (48) and the limit in Theorem 4 in the underloaded $(H_2(4)_t/LN(1)/1$ model with arrival-rate function in (24) and (37) for $(\gamma,\rho) \in \{(0.7,10^{-2}),(0.9,10^{-3}),(0.95,10^{-4})\}$ (left), for three different arrival processes (middle) and for three different service-time distributions (right).

Figure EC.6 presents simulation comparisons illustrating Theorem 4. In each case, PRQ is compared to simulation estimates of the normalized mean workload $2(1-\rho)E\left[W_{\gamma,\rho,y}\right]/\rho$ for $W_{\gamma,\rho,y}$ in (48) in the underloaded $H_{2,t}(4)/LN(1)/1$ model with the sinusoidal model in (24) with the scaling in (37)-(39). In particular, the convergence as $\gamma \downarrow 0$ and $\rho \uparrow 1$ is illustrated by considering $(\gamma,\rho) \in \{(0.7,10^{-2}),(0.9,10^{-3}),(0.95,10^{-4})\}$ (left), while the improved performance of PRQ as the level of variability decreases in the arrival and service processes is illustrated in the middle and right.

Figure EC.6 (middle) and (right) show the impact of changing variability in the arrival process and the service-time distribution. Consistent with the stationary model, Figure EC.6 (middle) and (right) show that increased variability in either the arrival process or the service process tends to increase congestion. We remark that the story can be different; e.g., it is different from the impact of the service-time distribution on the blocking in the time-varying $M_t/GI/n/0$ loss model; see Davis et al. (1995).