The Advantage of Indices of Dispersion in Queueing Approximations

Ward Whitt and Wei You

Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027-6699, USA

Abstract

A recent robust queueing approximation for open queueing networks exploits partial characterizations of each arrival process by its rate and index of dispersion for counts (IDC), which is a scaled version of the variance-time curve. Even though only means and variances are involved, we show that the IDC provides a basis for more accurate approximations than traditional two-moment partial characterizations. For the $GI/GI/1$ queue, this approach applied to the arrival and service processes fully characterizes the model.

Keywords: queues, open queueing networks, index of dispersion for counts, robust queueing, generalized Jackson networks, queueing approximations

1. Introduction

The purpose of this paper is to describe the advantages of a new non-parametric method for approximating steady-state performance measures in queueing models. The main idea is to approximately characterize each arrival counting process by its rate and *index of dispersion for counts* (IDC) instead of a few parameters, such as the first two moments. The IDC is a scaled version of the variance-time function. In particular if $A(t)$ is an arrival counting process, assumed to be stationary with rate λ, then as in §4.5 of [1], the IDC is

$$I_c(t) \equiv \frac{Var(A(t))}{E[A(t)]} = \frac{Var(A(t))}{\lambda t}, \quad t \geq 0, \tag{1}$$

where $\equiv$ denotes equality by definition. The IDC is independent of the rate, showing how the variability evolves over time.

It is important that the IDCs be available for each model counting process. First, they can be estimated from simulation output or from the large data sets that are becoming common for applications. For many stochastic models, closed-form expressions are available, as we illustrate in (20) in §5 here. They often can be accurately and rapidly computed. For a renewal arrival process specified by an interarrival-time cdf F having a probability density function f with Laplace transform (LT)

$$\hat{f}(s) \equiv \mathcal{L}(f)(s) \equiv \int_0^\infty e^{-st} f(t)dt = \int_0^\infty e^{-st} dF(t), \tag{2}$$

its IDC is readily computed by numerical inversion of the LT of the associated variance function, as we indicate here in §2.

The biggest challenge is using this complex partial specification of a queueing model to generate effective approximations for performance measures. While the main idea is quite general, much depends on the context. We have extended the robust queueing (RQ) in [2] to include IDCs in [3] and applied it to approximate the mean steady-state workload in a $G/GI/1$ queue, with general arrival process specified by its rate and IDC. We subsequently have shown that this approach can be extended to yield useful steady-state performance approximations for generalized Jackson queueing networks (GJQNs).

A GJQN can be viewed as open network generalizations of the $GI/GI/1$ queue. A GJQN has Markovian routing of a single class of customers through single-server queues with unlimited waiting space and the first-come first-served service discipline. There are mutually independent renewal external arrival processes and sequences of independent and identically distributed (i.i.d.) service times, where each interarrival time and service time has a general cumulative distribution function (cdf) with finite first two moments.

Previous approximations for the steady-state performance measures (assuming stability) primarily depend on each interarrival-time and service-time cdf only through its first two moments or, equivalently, its mean and squared coefficient of variation (scv, variance divided by the square of the mean). The two-parameter approach is applied in parametric-decomposition approximations such as the Queueing Network Analyzer (QNA) [4] and the first Robust Queueing Network Analyzer (RQNA) [2] as well as in the QNET [5] and sequential bottleneck decomposition (SBD) [6] approximations based on multidimensional reflected Brownian motion (RBM) stemming from the

heavy-traffic limit established in [7]. In [3, 8, 9, 10] we use IDCs with RQ to obtain a new RQNA and show that it is effective in simulation experiments.

In this paper we expose the advantage of using the rate and the index of dispersion for counts (IDC) of each arrival counting process instead of the first two moments. Unlike the papers [3, 8, 9, 10] which develop the new RQNA based on IDC's and conduct simulation experiments to demonstrate its effectiveness, *our purpose here is to show that the new partial characterization (i) contains much more information about the model than traditional two-parameter characterizations and (ii) that the extra information can have a significant impact upon performance.* Our new RQNA shows that this additional model information can be used, but that does not directly (without experimental evidence) imply that RQNA is necessarily effective. When RQNA falls short, we now see that the IDCs provide room for improvement.

To meet our goal, we consider different models with the same two-moment parameters, so that the previous approximations necessarily give the same answer, but the IDC captures important differences. While the IDC is indeed just a normalized variance, it is for the stationary version of the arrival process and it is a function of time, and therefore provides much more information about the arrival process. Here we provide strong theoretical support for this claim for a single $GI/GI/1$ queue and we provide simulation evidence for several queues in series.

Here is how this paper is organized: In §2 we show (i) how to calculate the IDC of a stationary renewal process and (ii) that a stationary renewal arrival process is fully characterized by its rate and IDC. Thus *the RQNA algorithm in [10] which uses the rate and IDC of both the arrival and service processes has full information about the* $GI/GI/1$ *queue.* (This result follows quite directly from existing theory.) In §3 we review the theory of extremal queues in [11, 12], which exposes the wide range of possible values given the usual two-moment partial characterization. In §4 we briefly review the RQ algorithm for the $G/GI/1$ queue from [3] and then its extension to GJQNs. We give numerical results for the $H_2/M/1$ queue in §5 and simulation results for queues in series in §6.

2. The IDC of a Renewal Process

We show how to calculate the IDC of a renewal process and observe that the rate and the IDC provide a full characterization of a renewal process and the $GI/GI/1$ queue. First, it is important to note that the IDC is defined

in terms of the equilibrium renewal process instead of the ordinary renewal process; e.g, see [13] or §3.4 and §3.5 of [14]. Thus the IDC is independent of the rate.

We start with a rate-λ renewal process $N \equiv \{N(t) : t \geq 0\}$. Let F be the cdf of the interval U between points having mean $E[U] = \lambda^{-1}$ and finite second moment. As a regularity condition, we also assume that F has a *probability density function* (pdf) f.

The stationary or equilibrium renewal process differs from the ordinary renewal process only by the distribution of the first interarrival time. Let F_e be the cdf of the equilibrium distribution, which has pdf $f_e(t) = \lambda(1 - F(t))$. Note that we can construct F_e given F, but we cannot construct F given F_e, but the pair (λ, F_e) fully characterizes F: $F(t) = 1 - \lambda^{-1} f_e(t)$, $t \geq 0$.

Let $E^e[\cdot]$ and and let $E^0[\cdot]$ denote the expectations with first interval distributed according to F_e and F, respectively. Conditioning on the first arrival, the renewal equations for the mean and second moment of $N(t)$, the number of points in an interval $[0, t]$, are:

$$
\begin{aligned}
m(t) &\equiv E^0[N(t)] = F(t) + \int_0^t m(t-s)dF(s), \\
m_e(t) &\equiv E^e[N(t)] = F_e(t) + \int_0^t m(t-s)dF_e(s), \\
\sigma(t) &\equiv E^0[N^2(t)] = F(t) + 2\int_0^t m(t-s)dF(s) + \int_0^t \sigma(t-s)dF(x), \\
\sigma_e(t) &\equiv E^e[N^2(t)] = F_e(t) + 2\int_0^t m(t-s)dF_e(s) + \int_0^t \sigma(t-s)dF_e(x).
\end{aligned}
$$

Note that the IDC is simply $I_c(t) \equiv \sigma_e(t)/\lambda$ and $m(t)$ is the familiar renewal function. To simplifying the relations among these expressions involving convolution integrals, we use the Laplace Transform (LT) in (2). Let a subscript e denote a quantity associated with the equilibrium renewal process. Then the LT of f_e is

$$
\hat{f}_e(s) = \frac{\lambda(1 - \hat{f}(s))}{s} \text{ and } \hat{F}_e(s) = \frac{\hat{f}_e(s)}{s},
$$

where $\lambda^{-1} \equiv \int_0^\infty t f(t)\, dt$ is the mean.

Applying the LT to the renewal equations, we obtain

$$\hat{m}(s) = \frac{\hat{f}(s)}{s(1-\hat{f}(s))}, \quad \hat{m}_e(s)\frac{\hat{f}_e(s)}{s(1-\hat{f}(s))} = \frac{\lambda}{s^2}, \tag{3}$$

$$\hat{\sigma}(s) = \frac{\hat{f}(s) + 2s\hat{m}(s)\hat{f}(s)}{s(1-\hat{f}(s))} = \frac{\hat{f}(s)(1+\hat{f}(s))}{s(1-\hat{f}(s))^2}, \tag{4}$$

$$\hat{\sigma}_e(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) = \frac{\lambda(1+\hat{f}(s))}{s^2(1-\hat{f}(s))}. \tag{5}$$

From (3), we see that $E^e[N(t)] = \lambda$, $t \geq 0$. From (5), we see that, given λ, we can compute $I_c(t) = \sigma_e(t)/\lambda$ by numerically inverting the LT $\hat{\sigma}_e(s)$, e.g., as in §13 of [15]. We are now ready to state the basic characterization theorem, which is not new, e.g., [13] and [1], but deserves to be better known.

Theorem 2.1. (*renewal process characterization theorem*) *A renewal process with an inter-renewal distribution having pdf f and cdf F having finite first two moments with positive mean λ^{-1} is fully characterized by any one of the following:*

1. *the pdf $f(t)$ of the time between renewals;*
2. *the cdf $F(t)$ of the time between renewals;*
3. *the LT $\hat{f}(s)$;*
4. *the renewal function $m(t)$;*
5. *the LT $\hat{m}(s)$;*
6. *the rate λ and the variance function of the equilibrium renewal process $\sigma_e(t)$;*
7. *the rate λ and the LT $\hat{\sigma}_e(s)$;*
8. *the rate λ and the IDC $I_e(t) \equiv \sigma_e(t)/\lambda t$ of the equilibrium renewal process.*

Proof. The equivalence of the time functions and their transforms follows from the basic theory of Laplace transforms. Hence, we obtain the equivalence by explicit expressions in terms of the Laplace transforms, i.e.,

$$\hat{f}(s) \;=\; \frac{s\hat{m}(s)}{1+s\hat{m}(s)} \quad \text{and} \quad \hat{f}(s) = \frac{s^2\hat{\sigma}_e(s) - \lambda}{s^2\hat{\sigma}_e(s) + \lambda}. \tag{6}$$

Then, from the definition of the IDC, we obtain $\sigma_e(t) = \lambda(t)I_e(t)$, $t \geq 0$. ∎

Corollary 2.1. (*full characterization of a $GI/GI/1$ queue*) *The $GI/GI/1$ queue with interarrival-time cdf F and service-time cdf G having finite second moments is fully characterized by the four-tuple $(\lambda, I_a(t), \tau, I_s(t))$, where τ is the mean service time and $I_a(t)$ ($I_s(t)$) is the IDC of the equilibrium renewal process associated with the interarrival (service) times.*

Corollary 2.1 is exploited strongly in the approximations for departure processes in (16). The final formula is an approximation, exploiting heavy-traffic limits, but Corollary 2.1 implies that the true formula must be a function of the four-tuple $(\lambda, I_a(t), \tau, I_s(t))$.

3. The Limitations of Two Moments for the $GI/GI/1$ Queue

To show that the extra information beyond the first two moments of F and G in the $GI/GI/1$ queue can be important, we review the theory of extremal queues in [11, 12]; also see the references there. It focuses on the mean steady-state waiting time, which is related to the mean steady-state workload via (10).

For the $GI/GI/1$ queue partially specified by the vector $(\lambda, c_a^2, \tau, c_s^2)$, a commonly used approximation for the mean steady-state waiting time is

$$E[W] \approx \frac{\tau\rho(c_a^2 + c_s^2)}{2(1-\rho)}, \tag{7}$$

because it is exact (being the classical Pollaczek-Khintchine formula) for the $M/GI/1$ special case, when the interarrival time has an exponential distribution, in which case $c_a^2 = 1$, and is asymptotically correct in heavy-traffic, i.e., $2(1-\rho)E[W(\rho)]/\tau\rho \to c_a^2 + c_s^2$ as $\rho \to 1$.

An insightful way to examine the quality of approximations for $E[W]$ given the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$ is to examine the range of possible values, which is the interval $[LB^*, UB^*]$, where LB^* is the tight lower bound (LB) and UB^* is the tight upper bound (UB). These tight bounds are studied in [11]. The most familiar UB given $(\lambda, c_a^2, \tau, c_s^2)$ is the Kingman[16] UB,

$$E[W] \le \frac{\tau\rho([c_a^2/\rho^2] + c_s^2)}{2(1-\rho)}, \tag{8}$$

which is asymptotically correct in heavy traffic, just like (7). A better UB depending is the Daley [17] UB, which replaces the term c_a^2/ρ^2 by $(2-\rho)c_a^2/\rho$.

But even the Daley bound is not tight; a better (not tight) UB is given in Theorem 1 of [11], while a numerical algorithm to compute the tight UB is given in [12]. The explicit formula for the tight LB, which has long been known, is

$$E[W(LB)] = \frac{\tau\rho((1+c_s^2)\rho - 1)^+}{2(1-\rho)}, \tag{9}$$

where $x^+ \equiv \max\{x, 0\}$.

Tables 1 and 2 plus Tables EC.1 and EC.2 in [11] give a numerical overview of the upper and lower bounds for $E[W]$, given the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$, in the four cases for which c_a^2 and c_s^2 assume all combinations of the two values 0.5 (less variable than exponential) and 4.0 (more variable than exponential). We illustrate by reproducing a portion of Table 1 of [11] here in Table 1. Paralleling (14), to focus on the impact of the variability independent of the traffic intensity ρ, so in Table 1 we display values for the normalized or *scaled mean waiting time* $c_W^2(\rho) \equiv 2(1-\rho)E[W(\rho)]/\rho\tau$, which shows the impact of the total variability in the arrival and service processes as a function of ρ, and assumes the constant value $c_a^2 + c_s^2$ for the approximation in (7).

Table 1: A comparison of the scaled bounds and approximations for $E[W]$, i.e., for $c_W^2(\rho) \equiv 2(1-\rho)E[W(\rho)]/\rho\tau$, given the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$ as a function of ρ for the case $c_a^2 = c_s^2 = 4.0$.

ρ	Tight LB	(7)	Tight UB	Daley	Kingman (8)
0.10	0.0	8.0	76.0	80.0	404.0
0.30	0.0	8.0	23.4	26.6	48.4
0.50	1.0	8.0	13.8	16.0	20.0
0.70	3.0	8.0	10.4	11.4	12.2
0.90	3.8	8.0	8.6	8.8	9.0
0.99	4.0	8.0	8.0	8.0	8.0

One conclusion from Table 1 is that the basic approximation in (7) is consistent with the possible values for all ρ. A second conclusion is that the quality of approximations depends on the traffic intensity. Consistent with [18], it is not possible to obtain reliable approximations for $E[W]$ in light traffic based only on $(\lambda, c_a^2, \tau, c_s^2)$. In contrast, the heavy-traffic approximation in (7) and all the upper bounds are asymptotically equivalent in the heavy-traffic limit for the scaled mean waiting time $2(1-\rho)E[W(\rho)]/\rho\tau$ in heavy-traffic. Even in heavy traffic, this is an iterated limit; i.e., we first fix the interarrival-time distribution with given c_a^2 and then we let $\rho \uparrow 1$.

However, an important conclusion from Table 1 is that the lower bound is surprisingly low, even in heavy traffic, so that the range of possible values consistent with the parameters is surprisingly wide. That occurs because the LB is attained asymptotically at the associated $D/GI/1$ queue with $c_a^2 = 0$. That D distribution is approached by a distribution that has a very small mass at a very large value, and the rest of the mass just less than the mean. That allows very small values of $E[W]$.

The heavy-traffic limit of the lower bound is the heavy-traffic limit for the associated $D/GI/1$ model, which in Table 1 would be 4.0. The values in Table 1 are obtained by, first, fixing ρ and then letting the interarrival-time distribution approach D. Table 1 implies that, for any given ρ, the least possible value is attained in the corresponding $D/GI/1$ model. This is why the distance between the LB and the UB remains large for all ρ.

While we should regard the LB as something of an anomaly, these numerical results clearly indicate that the mean waiting time is not so well approximated by the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$. Moreover, the difficulty is primarily caused by the arrival process. For example, for $M/GI/1$, the mean $E[W]$ is fully determined by (7), but for $GI/M/1$ there is a wide range.

In contrast, Corollary 2.1 implies that there would be no error at all if we found the exact value $E[W]$ determined by the rate and the IDC of the arrival and service processes. Unfortunately, that is not achieved by RQNA, but there is potential to do better with the information being used. We illustrate by examples in the next two sections.

4. Brief Review of the Robust Queueing Approximation

To show how the IDC can be used, we first focus on the $GI/GI/1$ queue with interarrival times U_n and service times V_n distributed as U and V, partially characterized by the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$, where $\lambda^{-1} \equiv E[U]$, $c_a^2 \equiv c_U^2 \equiv Var(U)/E[U]^2$ and $\tau \equiv \mu^{-1} \equiv E[V]$, $c_s^2 \equiv c_V^2 \equiv Var(V)/E[V]^2$, where $\rho \equiv \lambda/\mu < 1$ to ensure stability.

We will focus on the expected steady-state waiting time (for each arrival until starting service) $E[W]$ and workload (remaining work in the system at each time) $E[Z]$ at each queue. These are related by the conservation law $H = \lambda G$ or Brumelle's formula, [19], which for the $G/GI/1$ model is

$$E[Z] = \lambda \left(E[WV] + \frac{E[V^2]}{2} \right) = \rho E[W] + \rho\tau \frac{(c_s^2 + 1)}{2}. \tag{10}$$

These in turn are related to the mean number in queue and in system by Little's law.

The main RQ approximation in [3] is for the expected steady-state workload at each queue. It uses the *index of dispersion for work* (IDW) associated with the cumulative input process Y, defined by

$$I_w(t) \equiv \frac{Var(Y(t))}{E[V_1]E[Y(t)]} \quad \text{and} \quad Y(t) \equiv \sum_{k=1}^{A(t)} V_k, \quad t \geq 0. \tag{11}$$

as in [20]. For the $G/GI/1$ model, where the arrival process is general but independent of an i.i.d. sequence of service times, the IDW is related to the IDC by

$$I_w(t) = I_c(t) + c_s^2, \quad t \geq 0; \tag{12}$$

see §4.3.1 of [3]. In both (11) and (12), the arrival process is assumed to be stationary, just as in (1).

Given the IDW, the RQ approximation for the mean workload as a function of the traffic intensity ρ when the mean service time is fixed at $\tau = 1$ appears in (28) in §4.1 of [3], being simply

$$E[Z] \equiv E[Z_\rho] \approx Z_\rho^* \equiv \sup_{x \geq 0} \left\{ -(1-\rho)x/\rho + b_f \sqrt{x I_w(x)} \right\}, \tag{13}$$

where b_f is a parameter to be specified, which we take to be $\sqrt{2}$, which we explain below. (See [21] for additional background on the RQ approximations.)

Strong positive results for the RQ approximation in (13) with $b_f \equiv \sqrt{2}$ for the $G/GI/1$ queue appear in Theorems 2-5 of [3]. Theorem 2 states it is exact for the $M/GI/1$ queue, while Theorem 5 states that it is asymptotically correct in both light and heavy traffic. To state it, we define the normalized or scaled (steady-state) workload by comparing to what it would be in the associated $M/D/1$ model; i.e.,

$$c_Z^2(\rho) \equiv \frac{E[Z_\rho]}{E[Z_\rho; M/D/1]} = \frac{2(1-\rho)E[Z_\rho]}{\tau\rho} \quad \text{and} \quad c_{Z^*}^2(\rho) \equiv \frac{2(1-\rho)E[Z^*]}{\tau\rho}. \tag{14}$$

The normalizations in (14) expose the impact of variability separately from the traffic intensity. Theorem 5 of [3] states, that

$$\lim_{\rho \uparrow 1} c_{Z^*}^2(\rho) = I_w(\infty) = \lim_{\rho \uparrow 1} c_Z^2(\rho) \quad \text{and} \quad \lim_{\rho \downarrow 0} c_{Z^*}^2(\rho) = I_w(0) = \lim_{\rho \downarrow 0} c_Z^2(\rho), \tag{15}$$

where (12) holds with $I_c(\infty)$ being the scaled version of the asymptotic variance of the arrival process (the normalization constant in the central limit theorem). The heavy-traffic limit in (15) implies that RQ is asymptotically exact in a $G/GI/1$ with general stationary (including non-renewal) arrival process, provided that we use the exact IDC.

The GJQNs are substantially more complicated. Drawing on (74) and (75) of [8], each departure IDC is approximated by a convex combination of the arrival and service IDC's, i.e.,

$$I_{d,\rho}(t) \approx w_\rho(t) I_a(t) + (1 - w_\rho(t) I_s(t), \quad t \geq 0, \tag{16}$$

where the ρ-dependent weight is

$$w_\rho(t) \equiv w^*((1-\rho)^2 \lambda t)/\rho c_s^2) \quad \text{for} \quad w^*(t) \equiv 1 - (1 - c^*(t))/2t, \tag{17}$$

with $c^*(t)$ being the correlation function of the stationary version of canonical (drift - 1, variance 1) RBM; see (24)-(27) of [8]. The new RQNA for GJQNs in [10] is supported by heavy-traffic limits and simulation experiments. However, if all queues go into heavy traffic together, then SBD in [6] reduces to QNET in [5] and only it is asymptotically correct in heavy-traffic. On the other hand, if there is a single bottleneck (only one queue gues into heavy traffic), then SBD and RQNA are asymptotically correct in heavy-traffic. The main potential advantage of RQNA is away from the heavy-traffic regime.

5. The Case of an $H_2/M/1$ Queue

For the $GI/M/1$ queue with interarrival-time pdf f, the steady-state performance depends on a single root of a transform equation. In particular,

$$E[W] = \frac{\tau\sigma}{1-\sigma} \quad \text{and} \quad E[Z] = \rho\tau \left(\frac{\sigma}{1-\sigma} + \frac{c_s^2 + 1}{2} \right), \tag{18}$$

where σ is the unique root in $(0,1)$ of the equation $\hat{f}(\mu(1-\sigma)) = \sigma$.

Consider an $H_2/M/1$ queue, which is a $GI/GI/1$ queue with an exponential service distribution and a hyperexponential (H_2) interarrrival-time distribution, i.e., a mixture of two exponential distributions, which has pdf

$$f(t) \equiv p\lambda_1 e^{-\lambda_1 t} + (1-p)\lambda_2 e^{-\lambda_2 t}, \quad t \geq 0, \tag{19}$$

and thus the parameter triple $(p, \lambda_1, \lambda_2)$. Equivalently, it has as parameters its first three moments or the mean λ^{-1}, scv c_a^2 and the ratio between the two components of the mean $r \equiv p_1/\lambda_1/(p_1/\lambda_1+p_2/\lambda_2)$ where $\lambda_1 > \lambda_2$. The third parameter is often specified by stipulating balanced means,i.e., $r = 0.5$, as in (37) on p. 137 on [22]. The behavior as a function of the third parameter has been studied in [23]. As far as the congestion in the queue is concerned, the H_2 arrival process can be as smooth as a Poisson process with the same rate or as bursty as a batch Poisson process with the same scv; see (9) in §IV of [23]. The consequence is illustrated for $c_a^2 = 2$ and $c_a^2 = 12$ in Tables I and II on p. 170 of [23].

From p. 50 of [13], the IDC is

$$I_a(t) = c_a^2 - \frac{2\beta}{\gamma t}(1 - e^{-\gamma t}), \quad t > 0, \tag{20}$$

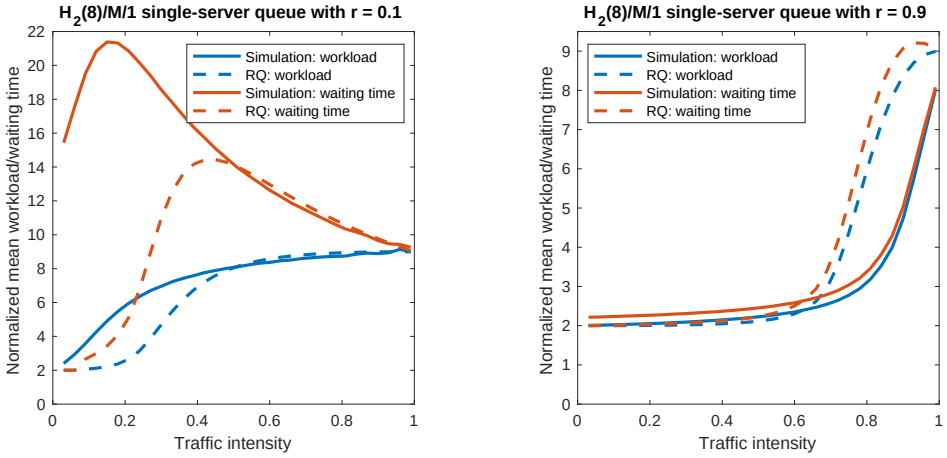
where $\beta \equiv (c_a^2 - 1)/2$ and $\gamma \equiv 2\beta((m_3/3) - (c_s^2 + 1)^2/2)$ with m_3 being the third moment, which is increasing in r. Note that $I_a(t)$ in (20) in strictly increasing in t from 1 at $t = 0$ to c_s^2 at $t = \infty$.

By virtue of Theorem 2.1, the set of H_2 arrival processes for given rate and scv c_a^2 are in fact fully characterized by their rate and IDC. The range of possible behavior of these IDC's can be seen from the IDC's of the two extremal H_2 arrival processes. The IDC of the lower bound is identically 1, while the IDC of the upper bound is identically c_s^2. All other IDC's increase from 1 at 0 to c_s^2 at infinity.

Figure 1 compares the RQ approximation for the workload via (13) and then the waiting time via (10) with simulation estimates for the $H_2/M/1$ model with $c_a^2 = 8$ and $r = 0.1$ and $r = 0.9$. The IDW and normalized mean workload increase from 2 to 9.

Figure 1 shows that the RQ approximation is not nearly exact even though the RQNA algorithm has full information about the $GI/GI/1$ model; there is room for improvement. For $\rho \geq 0.5$ (the region of primary interest), the RQ results for $r = 0.1$ are remarkably good. Figure 1 is consistent with the heavy-traffic and light-traffic limits for the mean workload in [3]. The poor performance for the mean waiting time in light traffic occurs because the light-traffic limit does not hold for it. Focusing on the third parameter r, we see that the performance and the approximations remain closer to the $M/M/1$ lower bound with $r = 1$ for $r = 0.9$ than $r = 0.1$. Finally, note that the standard two-moment approximations yield constant values independent of ρ in each case. Clearly no such approximation can be effective for all ρ.

Figure 1: A comparison of the robust queueing approximation for the mean workload and waiting time in (13) and (10) for the $H_2/M/1$ model with $c_a^2 = 8$ and $r = 0.1$ (left) and $r = 0.9$ (right) to simulation estimates.



6. Simulation Comparisons for a Series of Queues

We now consider simulation experiments for a series of single-server queues. We consider the heavy-traffic bottleneck examples from [24], which are for a non-Poisson arrival process feeding a series of 9 single-server queues, each with an exponential service cdf, where $\rho_i = 0.6$ for $1 \leq i \leq 8$, and $\rho_9 = 0.9$, so that the last queue is a bottleneck queue. In particular, we compare the new RQNA algorithm in [8] and RQ, the algorithm in (13) from [3] with the exact estimated IDC, to the performance of QNA from [4], QNET from [5] and SBD from [6]. The QNET method uses the multi-dimensional reflected Brownian motion resulting from the heavy-traffic limit in [7].

Table 2 compares five approximation methods to simulation for 9 exponential (M) queues in series fed by a highly-variable rate-1 H_2 renewal arrival process with $c_a^2 = 8$ and the usual balanced means. Table 2 compares the various approximations of the mean steady-state waiting time at each station, as well as the total waiting time in the system, to simulation estimates. The simulation estimates for $E[W_i]$ and the IDC are based on a C^{++} program and run of length 5×10^7, discarding the initial 10^5 customers to approach steady state. The half width of the confidence interval at the final bottleneck queue was about 0.2% in every case. The QNA approximation appears in [24]; the heavy-traffic approximations QNET from [5] and SBD appear in [6].

Table 2 shows poor performance of QNA [4] at the last bottleneck queue

Table 2: A comparison of five approximation methods to simulation for the expected waiting time $E[W]$ at each of 9 exponential (M) queues in series with $\rho_i = 0.6$, $1 \leq i \leq 8$, and $\rho_9 = 0.9$ fed by a highly-variable rate-1 H_2 renewal arrival process with $c_a^2 = 8$ and $r = 0.5$, i.e., the usual balanced means.

node	Sim	QNA	QNET	SBD	RQNA	RQ
1	3.36	4.05	4.05	4.05	3.95	3.95
2	2.32	2.92	1.81	1.82	1.58	2.61
3	1.96	2.19	1.47	1.49	0.98	2.04
4	1.77	1.73	1.16	1.19	0.92	1.72
5	1.64	1.43	1.07	1.10	0.90	1.53
6	1.56	1.24	1.03	1.06	0.90	1.41
7	1.49	1.12	1.00	1.03	0.90	1.32
8	1.44	1.04	0.98	1.01	0.90	1.27
9	29.2	8.9	6.0	36.4	29.1	37.1
sum	45.3	24.6	18.6	49.8	40.1	52.9

originally exposed in [24]. It also shows the significant improvement provided by the sequential bottleneck approximation (SBD) reported in [6], which is largely matched by RQNA and RQ.

To illustrate the impact of additional information about the arrival process, Table 3 is the analog of Table 2 in which we use three alternative rate-1 H_2 arrival processes, all with $c_a^2 = 8.0$ but different r, in particular for $r = 1.0$, 0.9 and 0.1. The case $r = 1$ is the lower-bound H_2 renewal arrival process with the same mean and $c_s^2 = 8$, which is the Poisson process, for which both RQNA and RQ are exact. For $r = 0.1$, the arrival process is close to a batch Poisson process. For these cases, the QNA, QNET and SBD approximations are the same as in Table 2.

From Tables 2 and 3, we see that the mean waiting time increases as r decreases. We also see that both RQNA and RQ are very accurate at the first $H_2/M/1$ queue, where the arrival process is a renewal process, but are far less reliable at later queues, which have non-renewal arrival processes. For queues 3-8, RQNA seriously underestimates $E[W_i]$; since RQ does not, we include the difficulty lies in the IDC approximation in (16) under lighter loads. Consistent with the heavy-traffic limit in [8], RQNA performs well at the final bottleneck queue, although RQ does not, which is partly explained by the relevant times are those where the IDC experineces most of its increase. RQNA also does reasonably well predicting the sum of the waiting times.

Table 3: A comparison of RQNA and RQ to simulation for the expected waiting time at each queue for the same model in Table 2 except except the third parameter of the H_2 renewal arrival process with $c_a^2 = 8$ is changed from $r = 0.5$ to $r = 1.0$, 0.9 and 0.1.

r	1.0	0.9			0.1		
node	exact	Sim	RQNA	RQ	Sim	RQNA	RQ
1	0.90	1.16	1.13	1.13	5.69	5.84	5.83
2	0.90	1.16	0.95	1.12	2.46	2.71	2.40
3	0.90	1.15	0.91	1.11	1.98	1.28	1.83
4	0.90	1.14	0.90	1.10	1.76	0.97	1.56
5	0.90	1.14	0.90	1.10	1.63	0.91	1.41
6	0.90	1.13	0.90	1.09	1.54	0.90	1.31
7	0.90	1.13	0.90	1.08	1.48	0.90	1.24
8	0.90	1.12	0.90	1.08	1.42	0.90	1.20
9	8.10	19.6	27.2	36.5	29.6	29.3	36.3
sum	15.3	28.8	33.8	45.3	47.5	43.7	53.1

The RQNA from [2] seems to perform far worse, as shown in Tables 1 and 2 of [21], but it provides tuning parameters that can yield significant improvement given additional information. In [21] we show that the specific version of RQNA from §7.2 of [2] corresponds to the asymptotic method from [22] for all the arrival processes and the Kingman upper bound from [16] at each queue.

Finally, we observe that the cases $r = 0.9$ and $r = 0.1$ in Table 3 provide a rough estimate of the range of reasonable approximation values for unspecified third parameter.

7. Conclusions

In this paper we have shown that the partial characterization of an arrival process by its rate and IDC provides significantly more information than the familiar two-moment partial characterizations, and so can serve as a basis for improved performance approximations. The new robust queueing approximations show how the IDC can be used, but the examples in §5 and §6 show that there remains room for improvement.

Acknowledgement. The authors thank NSF for research support (grant CMMI 1634133).

References

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

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

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

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

[5] J. M. Harrison, V. Nguyen, The QNET method for two-moment analysis of open queueing networks, Queueing Systems 6 (1) (1990) 1–32.

[6] J. Dai, V. Nguyen, M. I. Reiman, Sequential bottleneck decomposition: an approximation method for generalized Jackson networks, Operations Research 42 (1) (1994) 119–136.

[7] M. I. Reiman, Open queueing networks in heavy traffic, Math. Oper. Res. 9 (3) (1984) 441–458.

[8] W. Whitt, W. You, Heavy-traffic limit of the $GI/GI/1$ stationary departure process and its variance function, Stochastic Systems 8 (2) (2018) 143–165.

[9] W. Whitt, W. You, Heavy traffic limits for the stationary flows in generalized Jackson networks, submitted to Stochastic Systems. Available at: http://www.columbia.edu/∼ww2040/allpapers.html (2019).

[10] W. Whitt, W. You, A robust queueing network analyzer based on indices of dispersion, submitted to INFORMS Journal on Computing. Available at: http://www.columbia.edu/∼ww2040/allpapers.html (2019).

[11] Y. Chen, W. Whitt, Extremal $GI/GI/1$ queues given two moments, submitted to Operations Research. Available at http://www.columbia.edu/∼ww2040/allpapers.html (2018).

[12] Y. Chen, W. Whitt, Algorithms for the upper bound mean waiting time in the $GI/GI/1$ queue, submitted to INFORMS Journal on Computing. Available at http://www.columbia.edu/∼ww2040/allpapers.html (2018).

[13] D. R. Cox, Renewal Theory, Methuen, London, 1962.

[14] S. M. Ross, Stochastic Processes, 2nd Edition, Wiley, New York, 1996.

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

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

[17] D. J. Daley, Inequalities for moments of tails of random variables, with queueing applications, Zeitschrift fur Wahrscheinlichkeitsetheorie Verw. Gebiete 41 (1977) 139–143.

[18] D. Daley, T. Rolski, A light-traffic approximation for a single-server queue, Mathematics of Operations Research 9 (4) (1984) 624–628.

[19] S. Brumelle, On the relation between customer averages and time averages in queues, J. Appl. Prob. 8 (3) (1971) 508–520.

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

[21] W. Whitt, W. You, Supplement on robust queueing approximations for the $GI/GI/1$ queue and series of these queues, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html (2018).

[22] W. Whitt, Approximating a point process by a renewal process: two basic methods, Oper. Res. 30 (1982) 125–147.

[23] W. Whitt, On approximations for queues, III: Mixtures of exponential distribtuoons, AT&T Bell Laboratories Technical Journal 63 (1) (1984) 163–175.

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