Set-Valued Queueing Approximations Given Partial Information

Yan Chen
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, yc3107@columbia.edu

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027, ww2040@columbia.edu

In order to understand queueing performance given only partial information about the model, we propose determining intervals of likely performance measures given that limited information. We illustrate this approach for the steady-state waiting time distribution in the $GI/GI/K$ queue given the first two moments of the interarrival-time and service-time distributions plus additional information about these underlying distributions, including support bounds, higher moments and Laplace transform values. As a theoretical basis, we apply the theory of Tchebycheff systems to determine extremal models (yielding tight upper and lower bounds) on the asymptotic decay rate of the steady-state waiting-time tail probability, as in the Kingman-Lundberg bound and large deviations asymptotics. We then can use these extremal models to indicate likely intervals of other performance measures. We illustrate by constructing such intervals of likely mean waiting times. Without extra information, the extremal models involve two-point distributions, which yield a wide range for the mean. Adding constraints on the third moment and a transform value produces three-point extremal distributions, which significantly reduce the range, yielding practical levels of accuracy.

Key words: performance approximations, queues, multi-server queues, bounds, mean waiting time, extremal queues, October 25, 2019; revised: November 29, 2019

History:

1. Introduction

Despite many significant research contributions in queueing theory over the years, what J. F. C. Kingman (1970) wrote fifty years ago largely remains true today:

> It is a fair criticism of the theory of queues as it has been developed over the years that, even in the simple cases for which explicit analytical solutions can be found, these solutions are too complicated to be of practical use. It has been argued elsewhere (Kingman (1966)) that the criticism is to be met to some degree by the analysis of situations where robust approximations exist, such as that of "heavy traffic." It is, however, important to know how accurately such approximations represent the true solution, and the significance of inequalities for the various quantities of interest thus become apparent.

Just as Kingman (1970) did, we consider this problem for the $GI/GI/K$ queue, which is a K-server queue with unlimited waiting room and service in order of arrival by the first available server, where the interarrival times and service times come from independent sequences of independent and identically distributed (i.i.d.) random variables distributed as U and V with general cumulative distribution functions (cdf's) F and G. We are especially interested in exposing the performance impact of the variability of these underlying cdf's F and G. To describe the extent of the variability independent of the mean, we let c_a^2 and c_s^2 be the squared coefficient of variation (scv, variance divided by the square of the mean) of U and V. We start by considering the special case $K=1$, but it is significant that our approach extends directly to $K>1$.

The complication is well illustrated by the formula for the mean of steady-state waiting time W (before starting service) for $K=1$,

$$E[W]=\sum_{k=1}^{\infty}\frac{E[S_k^+]}{k}<\infty, \tag{1}$$

where $[x]^+\equiv\max\{x,0\}$, S_k is the k^{th} partial sum of k i.i.d. random variables distributed as $X\equiv V-U$ and $\equiv$ means equality by definition. And there is no analog of (1) for $K>1$. Formula (1) is reviewed in Abate et al. (1993), which is devoted to algorithms to compute $E[W]$ and $P(W>t)$ when $K=1$ for general F and G based on alternative integral representations. In general, simulation remains an attractive method, although it applies to only one specified model, does not yield the insight of formulas, and is a relatively time consuming numerical procedure.

A candidate simple and insightful approximation formula for $E[W]$ is provided by the heavy-traffic approximation (HTA). Choose measuring units by setting $\mathbb{E}[U]=1$, so that $E[V]=\rho K$, where ρ is the traffic intensity. Then the second moments are $E[U^2]=c_a^2+1$ and $E[V^2]=\rho^2K^2(c_s^2+1)$. In this context, the HTA for the mean with $K\geq 1$ is

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

For $K=1$, the HTA in (2) is obtained by combining the $M/GI/1$ Pollaczek-Khintchine exact formula for the special case of a Poisson arrival process, where $c_a^2=1$, with the heavy traffic limit in Kingman (1961). The extension to $K>1$ was provided by Borovkov (1965), Iglehart and Whitt (1970a,b), Kollerstrom (1974). (We do not consider the many-server heavy-traffic scaling in Halfin and Whitt (1981) or Gamarnik and Goldberg (2013).)

The limit shows that the approximation is asymptotically correct in the sense that

$$E[W]=HTA+o(1-\rho)\quad\text{as}\quad\rho\uparrow 1, \tag{3}$$

where $o(x)$ is a quantity $h(x)$ such that $h(x)/x\to 0$ as $x\to 0$.

In this context, the problem posed by Kingman (1970) can be expressed as: How accurate is formula (2)? In part, that question is answered for the case $K = 1$ by the large literature on bounds for $E[W]$, given the partial specification by the parameter 4-tuple

$$(E[U], E[U^2], E[V], E[V^2]) \equiv (1, c_a^2, \rho, c_s^2), \tag{4}$$

starting from Kingman (1962, 1970) and continuing with Daley (1977), Daley et al. (1992), Wolff and Wang (2003), Chen and Whitt (2018, 2019a) and the many references therein.

Unfortunately, however, this program has not yet been very successful. As shown by Table 1 in Chen and Whitt (2018), the range of possible values of the mean $E[W]$ in the $GI/GI/1$ model given the first two moments of U and V is quite wide, and so is of limited value. Consequently, we would want to add a little more information. However, relatively little is known about the impact of additional information, beyond the early results for the $GI/M/1$ model in Whitt (1984a,b), Klincewicz and Whitt (1984) and queues with phase-type distributions in Johnson and Taaffe (1991, 1993). Almost nothing is known about the case $K > 1$, but it is known that the range given the first two moments of U and V is even wider; see Daley (1997) and Gupta et al. (2010).

1.1. Applying T Systems to the Asymptotic Decay Rate

In order to make progress, we propose a new approach based on the asymptotic decay rate. To do so, we restrict attention to the light-tailed case, where the service-time cdf G has finite moments of all orders. We then typically have

$$P(W > t) \sim \alpha e^{-\theta_W t} \quad \text{as} \quad t \to \infty, \tag{5}$$

where $f(t) \sim g(t)$ as $t \to \infty$ means that $f(t)/g(t) \to 1$, e.g., see Abate and Whitt (1995). Then we call θ_W the (asymptotic) decay rate. Then we have the rough approximations $E[W] \approx \alpha/\theta \approx 1/\theta$. (We remark that the assumption supporting (5) does not reduce the range for the mean waiting time, but it is essential to even have a well-defined decay rate.) The key observation is that, in great generality, but under regularity conditions, the asymptotic decay rate θ_W in (5) is attained as the unique positive real root of an equation involving the Laplace transforms of U and V, e.g, $\hat{f}(s) \equiv \int_0^\infty e^{-st}\, dF(t)$. In particular, the equation for the decay rate is

$$\hat{f}(s)\hat{g}(-s) = 1. \tag{6}$$

In this light-tailed setting, we show that the theory of Tchebycheff (T) systems from Karlin and Studden (1966), as used in Rolski (1972), Holtzman (1973), Eckberg (1977), Whitt (1984a,b), Johnson and Taaffe (1991, 1993), Gupta and Osogami (2011), can be applied to determine extremal models (yielding tight upper and lower bounds) on the asymptotic decay rate θ_W above.

We start in §2 by giving background on T systems. Compared to previous papers that apply T systems to queueing systems, we contribute in §2.3 by exposing tractable sufficient conditions for a system of functions to be a T system in terms of Wronskians. In Lemma 2 we show that the systems of functions that we consider satisfy these conditions. After that, it suffices to apply the Markov-Krein theorem for T systems.

In §3 we obtain the extremal distributions for the decay rate. In §3.1, we provide technical background on the decay rate; in §3.1.2 we show that this approach also applies to the $GI/GI/K$ model for $K > 1$. In §3.2 we obtain two-point extremal distributions given only the first two moments of U and V and bounded intervals of support. Then in §3.3 we obtain more useful three-point extremal distributions when we are also given the third moment and values of the Laplace transform of U and V. Finally, in §3.4 we obtain extremal distributions in the common case when F and G have unbounded support.

1.2. Applications to Reveal Likely Intervals for the Mean Waiting Time

In §4 we apply these extremal distributions for the decay rate to determine associated intervals of likely values for the mean steady-state waiting time given the limited information. We do not establish rigorous bounds for the mean waiting time in this way, but nevertheless we think that this approach yields useful insight. Moreover, this general approach can be applied to other performance measures besides the mean and other stochastic models.

1.2.1. Starting from one concrete model. Our proposed method is based on Theorem 6 in §3.3. We start with a concrete model determined by the pair of cdf's (F, G). We first calculate: (i) the decay rate θ_W for that model by solving for the unique positive root of the single equation (6) involving the Laplace transforms $\hat{f}$ and $\hat{g}$ of F and G and (ii) four parameters from each of the underlying cdf's F and G: the first three moments and one argument of each Laplace transform. We have two alternatives for each of the arguments μ_s of $\hat{g}(-s)$ and μ_a of $\hat{f}(s)$: either $\leq \theta_W$ or $\geq \theta_W$ Our experiments indicate that we can set

$$\mu \equiv \theta_W/R \quad \text{if} \quad \mu \leq \theta_W \quad \text{and} \quad \mu \equiv R\theta_W \quad \text{if} \quad \mu \geq \theta_W \tag{7}$$

for suitable R, e.g., $R \in \{1, 5, 10, 20\}$. We find that it is better to have $\mu_s \leq \theta_W$; see Theorem 6 and §4.3.

For the concrete model, we also directly calculate the mean steady state waiting time $E[W]$, but the goal is to determine a set of likely values of that mean given *any* model with the two-moment parameters in (4) and the small set of additional parameters. For that purpose, we add the additional parameters to reduce the range to a smaller interval of likely values.

1.2.2. Fast application starting from the parameters in (4). To simplify applications of this method starting from the basic two-moment parameters in (4), we suggest working with standard distributions having the specified parameters in (4), as in §3 of Whitt (1982). For $c^2 \geq 1$, we use the H_2 (hyperexponential, mixture of two exponential) distribution with balanced means; For $c^2 = 1/k$, we use the E_k (Erlang) distribution. These both reduce to the exponential distribution when $c^2 = 1$. These distributions are fully specified by their first two moments. All examples here are for the cases $c^2 \in \{0.5, 1.0, 4.0\}$.

These H_2 and E_k distributions are fully specified by their first two moments. As in §5 of Whitt (1983), for parameter pair $(1, c^2)$, the third moments are $m_3 = 3c^2(1+c^2)$ if $c^2 > 1$ and $m_3 = (2c^2+1)(c^2+1)$ if $c^2 < 1$.

For these models, numerical values of $E[W]$ and θ_W can be found in the tables of Seelen et al. (1985) and are easily computed by available algorithms. Alternatively we can use heavy-traffic approximations. If we use HT approximations, then we can give a closed-form expression for all parameters needed to determine the extremal distributions by the application of the T-system methodology in §2.

For the mean $E[W]$, we can use the HT approximation in (2). For the decay rate, we can use the associated HT approximation

$$\theta_W \approx \frac{2(1-\rho)}{\rho(c_a^2 + c_s^2)}, \tag{8}$$

which is obtained by combining the $M/M/1$ exact formula $\theta_W = (1-\rho)/\rho$ with the heavy-traffic asymptotic expansion established in Abate and Whitt (1994); i.e.,

$$\theta_W(\rho) = \frac{2(1-\rho)}{c_a^2 + c_s^2} + C(1-\rho)^2 + O(1-\rho)^3 \quad \text{as} \quad \rho \uparrow 1, \tag{9}$$

where C is an (explicit) function of the first three moments of the mean-1 random variables U and V/ρ. Related asymptotics and approximations for the $GI/GI/s$ and $BMAP/GI/1$ models are established in Abate et al. (1995), Choudhury and Whitt (1994) and Corollary 3 of Glynn and Whitt (1994).

1.2.3. Determining the Extremal Models and the Interval of Mean Values. We also use bounded intervals of support for F and G. We propose using support bounds M_a and M_s that should have negligible impact on $E[W]$ in typical cases of interest; see §4.1. Given the required parameters, as indicated in §4.2, we can either apply a nonlinear equation solver or linear programming to solve the system of equations specified by the T-system theory to determine the two extremal models. We then simulate these extremal models to calculate the mean $E[W]$ and any other desired steady-state performance measures, for which we exploit the Minh and Sorli (1983) algorithm as we did in Chen and Whitt (2019a). In §4.3 we indicate how we select the arguments for the Laplace transforms. In §4.4 we report the results of simulation experiments to test the procedure. In §5 we draw conclusions. We present additional simulation results and other supporting material in the appendix, Chen and Whitt (2019b).

2. Tcheycheff System Foundations

To put the T system results in perspective, we start in §2.1 by reviewing the classical moment problem. Then in §2.2 we specify the additional conditions needed to get a T system and state the Markov-Krein theorem. In §2.3 and develop lemmas under smoothness conditions that are convenient for establishing the T system property, including for the systems we consider.

2.1. The Classical Moment Problem

We first review the classical moment problem, as in Lasserre (2010), Smith (1995) and references therein. Let u_i, $0 \le i \le n$, be $n+1$ continuous real-valued functions on the closed interval $[a,b]$. The expectations of these functions are assumed to be known, and are called the moments m_i, $0 \le i \le n$. The canonical example is $u_i(t) \equiv t^i$, $0 \le i \le n$, in which case these functions yield the usual moments. In this setting, we want to draw conclusions about the unspecified underlying probability measure P on $[a,b]$ such that:

$$m_i \equiv E_P[u_i] \equiv \int_a^b u_i \, dP, \quad 0 \le i \le n. \tag{10}$$

We assume that $u_0(t) \equiv 1$, $a \le t \le b$, and $m_0 \equiv 1$, so that the measure is necessarily a probability measure.

For the general moment problem, let $\mathcal{P}_n$ be the set of all probability measures P on $[a,b]$ with $n+1$ moments as defined above. We assume that $\mathcal{P}_n$ is nonempty. let $\mathcal{P}_{n,k}$ be the subset of probability measures in $\mathcal{P}_n$ that have support on at most k points in $[a,b]$. The following is a generalization of a standard result in linear programming (LP), stating that the supremum (or infimum) is attained at a basic feasible solution or an extreme point. (The notion of extreme point extends to more general spaces; e.g., see §III.6 of Karlin and Studden (1966).)

Theorem 1. (*a version of the classic moment problem, §2.1 of Smith (1995)*) *In addition to the $n+1$ functions u_i introduced above, let $\phi : [a,b] \to \mathbb{R}$ be another continuous real-valued function. Assume that $\mathcal{P}_n$ is not empty. Then there exists $P^* \in \mathcal{P}_{n,n+1}$ such that*

$$\sup\{\int_a^b \phi \, dP : P \in \mathcal{P}_n\} = \sup\{\int_a^b \phi \, dP : P \in \mathcal{P}_{n,n+1}\} = \sum_{k=1}^{n+1} \phi(t_k) P^*(\{t_k\}), \tag{11}$$

where $\{t_k : 1 \le k \le n+1\}$ is the support of P^*. The same result holds for the infimum.

.

let $\sigma(P)$ denote the cardinality of the support of P. Let P_U^* and P_L^* denote upper and lower extremal distributions, yielding the supremum and infimum in (11). Theorem 1 implies that there exist extremal distributions with $\sigma(P_U^*) \leq n+1$ and $\sigma(P_L^*) \leq n+1$.

In Chen and Whitt (2018) we applied Theorem 1 plus other arguments to show that the extremal distributions of both the transient and steady-state mean in the $GI/GI/1$ model, given the first two moments of U and V, are attained on three-point distributions. The T systems that we consider next provide a way to improve that result to uniquely-determined distributions, often with smaller support.

2.2. Tchebycheff Systems and the Markov-Krein Theorem

If we make additional assumptions about the functions u_i, then we can identify concrete extremal distributions P_U^* and P_L^* for (11). This refinement of Theorem 1 can be achieved by applying the theory of Tchebycheff systems, commonly called T systems, as in the seminal book Karlin and Studden (1966) and the review papers Johnson and Taaffe (1993) and Zalik (1996).

2.2.1. Upper and Lower Principle Representations. We first discuss what is possible and what is achieved under the T system assumption. To do so, we impose a regularity condition involving the moment space

$$\mathcal{M}_n \equiv \{(m_1,\ldots,m_n) \in \mathbb{R}^n : \text{there exists}\, P \in \mathcal{P}_n\, \text{such that} \int_a^b u_i\, dP = m_i, \quad \text{for all} \quad i\}. \tag{12}$$

If $(m_1,\ldots,m_n)$ is contained in the boundary of $\mathcal{M}_n$, then the probability measure is uniquely determined. We rule out that case by assuming that $(m_1,\ldots,m_n)$ is contained in the interior of $\mathcal{M}_n$. That assumption tends to be without loss of generality if we can adjust the support interval $[a,b]$.

To see what is possible, note that if $\sigma(P)=k$, then P is specified by $2k$ parameters: the k atoms x_i in $[a,b]$ and the k probabilities p_i. Given the $n+1$ constraints in (10), a solution P to (11) must have $2k \geq n+1$. When n is odd, we must have $\sigma(P) \geq (n+1)/2$. When n is even, we must have $\sigma(P) \geq n/2$. The final story under the T-system assumption is different in these two cases. It is summarized in (5) of Eckberg (1977) and on p. 342 of Gupta and Osogami (2011).

The story (the conclusions, not the proof) is relatively simple when n is even. Then, under the regularity conditions the extremal distributions have the minimum possible number, $k=n/2$, of points in the support. But that leaves one extra parameter. Then there is a one-parameter family of distributions satisfying all the constraints. Then upper (lower) extremal distributions P_U^* and P_L^* (called upper and lower principal representations in Karlin and Studden (1966)), are the ones that attach mass to the upper (lower) endpoints a (b) of the interval $[a,b]$. Given that additional specification, the remaining number of unknowns matches the number of constraints, so that the extremal distributions are uniquely determined.

The story is more complicated when n is odd. Now there is a unique distribution on (a,b) with the least number of points in the support $k=(n+1)/2$. That distribution turns out to be the lower extremal distribution P_L^*. The upper extremal distribution P_U^* has mass on both endpoints a and b. That leaves $2k-2$ unknowns, so we have $2k \geq n+3$ or $k \geq (n+3)/2$, so that $\sigma(P_U^*) = \sigma(P_L^*)+1$. The remaining $(n-1)/2$ points inside the open interval (a,b) are then uniquely determined.

2.2.2. T Systems and the Markov-Krein Theorem.

In this setting, the fundamental result supporting the conclusions above is the Markov-Krein theorem. It says that the description above holds under the condition that certain collections of functions constitute a T system. In Karlin and Studden (1966), T system theory and the Markov-Krein theory are first developed for continuous functions on a compact interval in Chapters I-III that we are considering here. The results are then extended to unbounded intervals and discrete subsets in later chapters, but a totally ordered set is needed. In this paper we primarily consider the basic case $[a,b]$, but we use it to obtain results for unbounded intervals of support in §3.4.

Definition 1. (T System) Consider the same set of $n+1$ continuous real-valued functions $\{u_i(t): 0 \leq i \leq n\}$ defined on $[a,b]$ introduced in §2.1. Assume that the moment vector lies in the interior of the moment space. This set of functions constitutes a T system if the $(n+1)^{\text{st}}$-order determinant of the $(n+1)\times(n+1)$ matrix formed by $u_i(t_j)$, $0 \leq i \leq n$ and $0 \leq j \leq n$, is strictly positive for all $a \leq t_0 < t_1 < \cdots < t_n \leq b$.

Equivalently, except for an appropriate choice of sign, we could instead require that every nontrivial real linear combination $\sum_{i=0}^n a_i u_i(t)$ of the $n+1$ functions (called a u-polynomial; see §I.4 of Karlin and Studden (1966)) possesses at most n distinct zeros in $[a,b]$. (Nontrivial means that $\sum_{i=0}^n a_i^2 > 0$.)

The main extremal result under this stronger condition is the Markov-Krein theorem; see Theorem 1.1 in §III.1 of Karlin and Studden (1966) and Theorem 1 of Gupta and Osogami (2011).

Theorem 2. (*Markov-Krein, §III.1 of Karlin and Studden (1966)*) *In the setting of Theorem 1 extended by requiring that the moment vector is in the interior of the moment space, if* $\{u_0,\ldots,u_n\}$ *and* $\{u_0,\ldots,u_n,\phi\}$ *are T systems on the interval* $[a,b]$, *then the upper and lower extremal distributions* P_U^* *and* P_L^* *described above uniquely attain the supremum and infimum of the optimization problem in* (11).

2.3. Convenient Sufficient Conditions for Smooth Functions: Wronskians

The major challenge for applications is showing that the two sets of functions in Theorem 2 are indeed T systems. However, it turns out that there is a very tractable sufficient condition for suitably smooth functions (having continuous derivatives of all relevant orders). This sufficient condition is expressed using the Wronskian.

Definition 2. (Wronskian) Let $u_i^{(j)}(t)$ be the j^{th} derivative of u_i at the argument t. The Wronskian of the $n+1$ functions $\{u_i(t): 0 \leq i \leq n\}$ is the determinant of the $(n+1) \times (n+1)$ matrix $\{u_i^{(j)}(t): 0 \leq i, j \leq n\}$ of the functions and their derivatives

$$W_n(u_i : 0 \leq i \leq n) \equiv det(u_i^{(j)}(t) : 0 \leq i, j \leq n). \tag{13}$$

An example makes it clear. For $s > 0$, let $w_3 \equiv w(1, t, t^2, -e^{-st})$ be the Wronskian of the $3+1=4$ indicated functions of t, i.e., the determinant of the matrix (as a function of t)

$$\begin{bmatrix} 1 & t & t^2 & -e^{-st} \\ 0 & 1 & 2t & se^{-st} \\ 0 & 0 & 2 & -s^2e^{-st} \\ 0 & 0 & 0 & s^3e^{-st} \end{bmatrix}$$

which clearly is $2s^3e^{-st} > 0$.

In order to verify the required T system properties, instead of looking at $n+1$ functions at $n+1$ arguments, we look at the same functions and their first n derivatives at a single argument. The Wronskian is intimately related to extended complete T systems or ECT systems, which is a special case of a T system.

Definition 3. (complete T system, p. 1 of Karlin and Studden (1966)) If each (ordered) subset $\{u_i(t): 0 \leq i \leq m\}$ for $1 \leq m \leq n$ of the T system of $n+1$ functions is itself a T system, then the T system is called a complete T system or CT system or a Markov system.

The classical CT system is the set of functions $u_i(t) \equiv t^i$, $0 \leq i \leq n$. Then the determinant is the Vandermonde determinant

$$det(u_i(t_j) : 0 \leq i, j \leq m) = \prod_{0 \leq i < j \leq m} (t_j - t_i) \quad \text{for all} \quad 1 \leq m \leq n, \tag{14}$$

which clearly is strictly positive for all $a \leq t_0 < t_1 < \cdots < t_m \leq b$, $1 \leq m \leq n$.

The direct definition of an extended T system in §I.2 of Karlin and Studden (1966) is somewhat complicated. Thus, we give an equivalent definition

Definition 4. (extended T system, §I.2 of Karlin and Studden (1966) and Theorem 1 of Zalik (2011)) An extended T system or ET system is characterized, except for the sign, by the property that every nontrivial real linear combination $\sum_{i=0}^{n} a_i u_i(t)$ of the $n+1$ functions (called a u-polynomial; see §I.4 of Karlin and Studden (1966)) possesses at most n distinct zeros in $[a, b]$, counting multiplicities.

The main point is that the definition of an ET system is more restrictive than the definition of a T system; i.e., every ET system is necessarily a T system. Completeness is defined the same for ET systems as for T systems. Hence every ECT system is necessarily an CT system, which in turn is necessarily a T system.

It turns out that an ECT system can be characterized completely by the Wronskian; see Definition I.2.4 on p. 6 and Theorem XI.1.1 on p. 376 of Karlin and Studden (1966), Theorem 5 and Corollary 1 of Zalik (1996), and Theorem 29 of Johnson and Taaffe (1993).

Theorem 3. (*Wronskians and ECT systems, p. 376 of Karlin and Studden (1966)*) *Under the smoothness condition, the system of $n+1$ functions $\{u_i : 0 \le i \le n\}$ is an ECT system on $[a,b]$, and thus necessarily a CT system, if and only if the Wronskians w_k of the first $k+1$ functions and their first k derivatives are strictly positive at all of its arguments in the interval $[a,b]$ for all k, $0 \le k \le n$.*

For smooth functions, Theorem 3 tends to be easy to apply, as illustrated by the example above. For one function in addition to the standard moments, the following lemma applies.

Lemma 1. *If $u_i(t) \equiv t^i$, $0 \le i \le n$, and ϕ has $n+1$ continuous derivatives, then $\{u_0(t), u_1(t), \ldots, u_n(t), \phi(t)\}$ is an ECT system if and only if the $(n+1)^{\mathrm{st}}$ derivative of ϕ, $\phi^{(n+1)}(t)$, is strictly positive on $[a,b]$.*

Proof. The triangular structure of the matrix of functions and their derivatives implies that the k^{th} Wronskian's take the constant value $w_k(t) = 1! \times \cdots \times k!$, $0 \le k \le n$, while the last Wronskian takes the value $w_n(t)\phi^{(n+1)}(t)$. ■

In this paper we will consider only a limited class of ECT systems. All the cases we consider will be covered by the following lemma about ECT systems.

Lemma 2. (*sufficient conditions for this paper*) *Consider three ordered sets of continuous real-valued functions on the interval $[0,M]$: $\mathcal{A}_1(m) \equiv \{t^k : 0 \le k \le m\}$, $\mathcal{A}_2 \equiv \{(-1)^{m+1}e^{-s_i t} : s_i > s_{i+1} > 0 \quad for\ all \quad i\}$ and $\mathcal{A}_3 \equiv \{e^{z_i t} : 0 < z_i < z_{i+1} \quad for\ all \quad i\}$. Let $\mathcal{F}$ be a finite ordered subset of $\mathcal{A}_2 \bigcup \mathcal{A}_3$ (with the elements of $\mathcal{A}_2$ appearing first and the order within each set). For any m and M, $0 \le m < \infty$ and $0 < M < \infty$, the ordered set $\mathcal{A}_1(m) \bigcup \mathcal{F}$ constitutes an ECT system over $[0,M]$ and thus a CT system over $[0,M]$.*

Before giving the proof, we give an example of an ordered subset of functions in $\mathcal{A}_1(m) \bigcup \mathcal{F}$. For $m = 2$ and two elements from each of $\mathcal{A}_2$ and $\mathcal{A}_3$, the ordered subset is $(1, t, t^2, -e^{-s_1 t}, -e^{-s_2 t}, e^{z_1 t}, e^{z_2 t})$ where $s_1 > s_2 > 0$ and $0 < z_1 < z_2$, so that $-s_1 < -s_2 < z_1 < z_2$. Here $m = 2$, so $(-1)^{m+1} = -1$. Overall, the exponential arguments are increasing as in (3.1) on p. 9 of Karlin and Studden (1966) or Example 6 of Zalik (1996).

Proof. These special functions have derivatives of all orders. Moreover, it is easy to evaluate the Wronskian. The first k derivatives of t^j are 0 when $k \geq j$. Thus the first m Wronskians are positive constants. The order $(m+1)$ determinant is a positive constant times $(-1)^{m+1} e^{-s_1 t} > 0$. Then, by induction, all higher-order determinants among the initial functions reduce to positive constant multiple of the determinant of a matrix of exponential functions. Finally, the the determinant of the $n \times n$ matrix containing elements $e^{x_i y_j}$, $1 \leq i, j \leq n$, is strictly positive for all $-\infty < x_1 < x_2 < \cdots < x_n < +\infty$ and $-\infty < y_1 < y_2 < \cdots < y_n < +\infty$; see (3.1) in §I.3 on p. 9 of Karlin and Studden (1966) and Example 6 of Zalik (1996). ■

3. Bounds for the Asymptotic Decay Rate

In this section we identify tight upper and lower bounds for the decay rate and the extremal distributions that attain those bounds. We start in §3.1 by introducing additional definitions and assumptions. In §3.1.1 we elaborate on the key equation (6) determining the decay rate. In §3.1.2 we show that the results also apply to the $GI/GI/K$ model. In §3.2 we establish the two-point extremal models given two moments and finite support for U and V. In §3.3 we establish the three-point extremal models given three moments, a Laplace transform value and finite support for U and V.

3.1. Theory for the Asymptotic Decay Rate

To increase the level of generality, instead of (5), we can let θ_W be defined by the critical exponent in the Kingman-Lundberg bound for the $GI/GI/1$ queue, as in Kingman (1964) and §XIII.5 of Asmussen (2003), defined by

$$\theta_W \equiv \inf\{x \geq 0 : P(W > t) \leq e^{-xt}, \quad t \geq 0\}, \tag{15}$$

so that large waiting times correspond to small values of θ_W. Under regularity conditions, θ_W in (15) coincides with the asymptotic decay rate studied in large-deviations theory, defined by

$$\theta_W \equiv \lim_{x \to \infty} \frac{-\log P(W > x)}{x}. \tag{16}$$

We assume that a strictly positive infimum exists in (15) and a strictly positive limit exists in (16), which requires that the service-time V must have a finite moment generating function $E[e^{sV}]$ for some $s > 0$. (We obtain $\theta_W = \infty$ if $P(V - U \leq 0) = 1$ and thus $P(W = 0) = 1$.) Thus, we are considering the light-tail case as in the discussion of exponential change of measure in Chapter XIII in Asmussen (2003), large deviation limits in Corollary 1 in §1.2 of Glynn and Whitt (1994) and approximations in Abate et al. (1995). More about the asymptotic decay rate can be found in discussions of the caudal characteristic curve of queues in Neuts (1986) and effective bandwidths in Choudhury et al. (1996), Kelly (1996), Whitt (1993) and references therein.

Part of the appeal of this approach is that it extends directly to $K > 1$, as we show in §3.1.2. Moreover, it has been observed that the approximation $P(W > t | W > 0) \approx e^{-\theta_W t}$ is quite good for $K \geq 1$; see Seelen et al. (1985). Indeed, for that reason, θ_W is displayed in the tables there (with different scaling, i.e., with $E[V] = 1$).

3.1.1. The Key Equation Determining the Decay Rate. For our queueing application, the key observation is that, under regularity conditions, the asymptotic decay rate θ_W in (5), (15) or (16) is attained as the unique positive real root of equation (6) involving the Laplace transforms of U and V, e.g, $\hat{f}(s) \equiv \int_0^\infty e^{-st}\, dF(t)$. Equivalently, as in §XIII.1 of Asmussen (2003), $\kappa_F(s) + \kappa_G(-s) = 0$, where $\kappa_F(s) \equiv \log(\hat{f}(s))$ is the cumulant generating function. (The function $\hat{g}(-s) \equiv E[e^{sV}]$ for $s > 0$ is the moment generating function (mgf).)

Indeed, it is well known that the distribution of W depends on $V - U$, which has Laplace transform $\hat{f}(-s)\hat{g}(s)$. Moreover, Chapter II.5 of Cohen (1982) shows that the distribution of W can be characterized by all complex roots of equations related to (6).

Given the simple structure in (6), the extremal result and alternative ones follow from the theory of T systems, as in §2 above. To state the result, we impose some technical conditions. In contrast to the mean $E[W]$, which is finite for all models given the partial moment information in (4), as can be seen from §X.2 of Asmussen (2003), the decay rate is not well defined for all these models. Hence, in order to establish extremal results for the decay rate in (15) given the partial moment information in (4), we make the following assumption.

Assumption 1. (*finite moment generating function*) *Assume that there exists s^*, $0 < s^* \leq \infty$, such that the service-time cdf G has a finite moment generating function $\hat{g}(-s) = \int_0^\infty e^{sx}\, dG(x)$ for all s, $0 < s < s^*$.*

In general, we need to impose additional regularity conditions to have the limit for the decay rate in (16) be well defined, as can be seen from Corollary 1 and Proposition 2 in Glynn and Whitt (1994) and Theorems 2.1, 5.5 and 5.3 in Chapter XIII in Asmussen (2003). Instead of adding additional assumptions, we allow the decay rate to be defined by (15). It coincides with (16) when the limit exists.

We still need extra conditions for (6) to have a solution; see Example 5 in §3 and Theorem 5 in §7 of Abate et al. (1995). However, no extra condition is needed when G has support in $[0, M_s]$, because then $E[e^{tV}] \leq e^{tM_s}$ for all $t > 0$, so that $s^* = \infty$ in Assumption 1.

3.1.2. Extension to the $GI/GI/K$ Model. As indicated in Abate et al. (1995), the asymptotic decay rate also is well defined for the more general $GI/GI/K$ model. We have fixed $E[U] = 1$, If instead we had fixed $E[V] = 1$, then $\theta_W(K) = K\theta_W(1)$, as in (5) of Abate et al. (1995), where

$U(K) = U(1)/K$ to keep ρ fixed. Since we fix $E[U] = 1$, we get $\theta_W(K) = \theta_W(1)$. (As a sanity check, this can easily be verified for the $P(W > t|W > 0) = e^{-\theta_W t}$ in the $M/M/K$ model; see Theorem 9.1 in §III.9 on p. 108 of Asmussen (2003).) However, we must adjust the service-time V to maintain $\rho = E[V]/KE[U]$. Thus, we leave U independent of K, but we let $V(K) = KV(1)$. Thus the finite support of $V(K)$ becomes $[0, \rho K M_s]$, the p^{th} moment of $E[V(K)^p] = K^p E[V(1)^p]$ and the laplace transforms are related by $\hat{g}_{V(K)}(s) = \hat{g}_{V(1)}(Ks)$. This implies that we can apply the extremal distributions for $K = 1$ to directly obtain the corresponding extremal distributions for $K > 1$: If $V^*(K)$ is the extremal random variable as a function of K, then $V^*(K) = KV^*(1)$.

In Abate et al. (1995), it was observed that the extension to $K > 1$ in (5) there was proved for the $GI/PH/K$ by Neuts and Takahashi (1981), but a continuity result implies that result applies to the general $GI/GI/K$ model.

Theorem 4. (*extension of decay rate to $GI/GI/K$*) *If the decay rate $\theta_W(1)$ is well defined for the $GI/GI/1$ model with (U, V) having cdf's (F, G) where $E[U] = 1$, then it is well defined in the associated $GI/GI/K$ model with (U, KV) with the same cdf F and*

$$\theta_W(K) = \theta_W(1) \quad \textit{for} \quad K > 1. \tag{17}$$

Proof. Fix the interarrival-time cdf F and consider a sequence of phase-type service-time $\{G_n : n \geq 1\}$ such that G_n is phase-type for each n and $G_n \Rightarrow G$, where G is the given cdf, which is possible because phase-type distributions are dense in the family of all distributions. By Neuts and Takahashi (1981), (17) holds for each n, as explained above. The convergence in distribution implies the associated convergence $\hat{g}_n(s) \to \hat{g}(s)$ for each s. Since the Laplace transform $\hat{g}(s)$ is continuous and strictly decreasing in the real variables s, (17) must hold in the limit as well. We could also work from the result for the $Ph/Ph/c$ model in Takahashi (1981) by taking such a limit for both F and G. ■

Remark 1. The models $GI/Ph/K$ and $Ph/Ph/K$ are special because $P(V - U > 0) > 0$, so that θ_W is always finite, but that is not the case for the $GI/GI/K$ model. However, if we consider such a general model with infinite decay rate, then we will get an infinite limit as we let the phase-type distribution approach the given distribution.

3.2. Two-Point Extremal Distributions Given Two Moments and Finite Support

We first consider the classical case in which we specify two moments. Let $\mathcal{P}_2(m, m^2(c^2+1), M)$ be the set of all cdf's with mean m, support mM and second moment $m^2(c^2+1)$, where c^2 is the scv with $c^2 + 1 < M < \infty$. (The last property ensures that the set $\mathcal{P}_2(m, m^2(c^2+1), M)$ is non-empty.) The extremal distributions for the decay rate will be the extremal distributions P_U^* and P_L^* for T systems in §2.2.

In this classical setting, the extremal distributions P_U^* and P_L^* are special two-point distributions.The set of two-point distributions is a one-dimensional parametric family. In particular, any two-point distribution with mean m, scv c^2 and support mM has probability mass $c^2/(c^2+(b-1)^2)$ at mb, and mass $(b-1)^2/(c^2+(b-1)^2)$ on $m(1-c^2/(b-1))$ for $1+c^2 \le b \le M$.

Let subscripts a and s denote sets for the interarrival and service times, respectively. Let F_0 and F_u (G_0 and G_u) be the two-point extremal interarrival-time (service-time) cdf's corresponding to P_L^* and P_U^*, respectively, in the space $\mathcal{P}_{a,2}(1, c_a^2+1, M_a)$ ($\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1), M_s)$) from §2.2.1. (Recall our convention that $E[U]=1$ and $E[V]=\rho$. Hence, the support of V is $[0, \rho M_s]$.)

Consequently, F_0 has probability mass $c_a^2/(1+c_a^2)$ at 0 and probability mass $1/(c_a^2+1)$ at $m(c_a^2+1)$, while F_u has mass $c_a^2/(c_a^2+(M_a-1)^2)$ at the upper bound of the support, M_a, and mass $(M_a-1)^2/(c_a^2+(M_a-1)^2)$ on $m(1-c_a^2/(M_a-1))$.

We are especially interested in the map

$$\theta_W : \mathcal{P}_{a,2}(1, 1+c_a^2, M_a) \times \mathcal{P}_{s,2}(\rho, \rho^2(1+c_s^2), M_s) \to \mathbb{R}, \tag{18}$$

where $0<\rho<1$ and $\theta_W(F,G)$ is the asymptotic decay rate of the steady-state waiting time $W(F,G)$ with interarrival-time cdf $F \in \mathcal{P}_{a,2}(1, 1+c_a^2, M_a)$ and service-time cdf $G \in \mathcal{P}_{s,2}(\rho, \rho^2(1+c_s^2), M_s)$. We also consider case in which one cdf is specified, in which case it need not have bounded support.

Theorem 5. (*two-point extremal distributions for the decay rate*) *Let F_0, F_u, G_0 and G_u be the two-point extremal cdf's for the $GI/GI/1$ queue defined above.*

(*a*) *For any specified $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1))$ satisfying Assumption 1 such that there is a root $\bar{s}$ to equation (6) for the $F_u/G/1$ model (with service cdf G) such that $0<\bar{s}<s^*$, where s^* is defined in Assumption 1,*

$$\theta_W(F_0, G) \le \theta_W(F,G) \le \theta_W(F_u, G) \tag{19}$$

for all $F \in \mathcal{P}_{a,2}(1, c_a^2+1, M_a)$.

(*b*) *For any specified $F \in \mathcal{P}_{a,2}(1, (c_a^2+1))$,*

$$\theta_W(F, G_u) \le \theta_W(F,G) \le \theta_W(F, G_0) \tag{20}$$

for all $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1), M_s)$

(*c*) *for all $F \in \mathcal{P}_{a,2}(1, c_a^2+1, M_a)$ and $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1, M_s)$,*

$$\theta_W(F_0, G_u) \le \theta_W(F,G) \le \theta_W(F_u, G_0). \tag{21}$$

Proof. We make extra conditions in part (a) to ensure that equation (6) has a solution $\bar{s}$ strictly less than the upper limit s^*, but no extra conditions are needed in parts (b) and (c) because then G has bounded support, implying that $s^* = +\infty$.

We apply (6) to see that order for the Laplace transforms translates into order for θ_W, recalling that (i) (6) is equivalent to $\hat{f}(s) = 1/\hat{g}(-s)$, (ii) Laplace transforms are continuous strictly decreasing functions of a real variable argument and (ii) large waiting times are associated with smaller θ_W. For part (a), we see that

$$\hat{f}_u(s) \le \hat{f}(s) \le \hat{f}_0(s) \quad \text{for} \quad s > 0. \tag{22}$$

From (6) and (22), we see that, for any $\hat{g}$, θ_W is maximized by $\hat{f}_u$ in (22). Hence, (6) holds for all F in $\mathcal{P}_{a,2}(1, c_a^2 + 1, M_a)$ if it holds for F_u.

To establish (b), we use

$$1/\hat{g}_u(-s) \le 1/\hat{g}(-s) \le 1/\hat{g}_0(-s) \quad \text{for} \quad s > 0. \tag{23}$$

From (6) and (23), we see that, for any $\hat{f}$, θ_W is maximized by $1/\hat{g}_0(-s)$ in (23).

To justify all the inequalities, we can apply the T-system theory working with bounded support sets, as in §2.2 and §2 of Eckberg (1977). To treat F, we apply Lemma 2 to show that $\{1, t, t^2\}$ and $\{1, t, t^2, -e^{-st}\}$ are T systems on $[0, M_a]$ for any $s > 0$ and $M_a > 0$; to treat G, we apply Lemma 2 again to show that and $\{1, t, t^2\}$ and $\{1, t, t^2, e^{st}\}$ is a T-system on $[0, \rho M_s]$ for any $s > 0$ and $M_s > 0$. We obtain the extremal distributions from §2.2.1 the case $n = 2$ in §2.2.1 or in (5) of Eckberg (1977). ■

Based on Theorem 5, the overall extremal $GI/GI/1$ models are thus (F_0, G_u) and (F_u, G_0). Our assumption that the distributions have bounded support plays an important role. That is evident from the following elementary proposition.

Proposition 1. (*limits as the support increases*) *Under the assumptions of Theorem 5, for all* $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1, M_a)$ *and* $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), M_s)$,

$$\theta_W(F, G_u) \to 0 \quad \text{as} \quad M_s \to \infty, \tag{24}$$

while

$$\theta_W(F_u, G) \to \theta_W(F_1, G) \quad \text{as} \quad M_a \to \infty, \tag{25}$$

where F_1 *is the cdf of the unit point mass on* 1, *associated with the* $D/GI/1$ *model.*

Remark 2. (the decay rates of other steady-state distributions.) Analogs of Theorem 5 (and the later Theorem 6) hold for the steady-state continuous-time queue length and workload, because there are simple relations among all these decay rates. That follows from Theorem 6, Proposition 9 and Proposition 2 of Glynn and Whitt (1994). For the workload, the decay rate is the same; for the queue length, $\theta_Q = \hat{g}(-\theta_W)$.

Remark 3. (comparison to the mean.) The extremal model (F_0, G_u) in Theorem 5 yielding the smallest decay rate coincides with the conjectured upper bound model for the mean $E[W]$, but the extremal model (F_u, G_0) in Theorem 5 yielding the largest decay rate does not coincide with the lower bound for the mean; see §§5.2, 7 and EC.6 of Chen and Whitt (2018).

Remark 4. (more on the application of Theorem 5) More on the application of Theorem 5 appears in §§2-5 of the appendix Chen and Whitt (2019b).

3.3. Laplace Transform Constraints to Reduce the Range

We now add additional constraints on the cdf's F and G. In particular, we add a third moment and a value of the Laplace transform. With (6) in mind, we now impose constraints on the Laplace transform $\hat{f}(s)$ at $s = \mu_a > 0$ and on the reciprocal of the mgf, $1/\hat{g}(-s)$, at $s = \mu_s$, $0 < \mu_s < s^*$, for s^* in Assumption 1.

For the new extremal distributions, let $\mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a)$ be the subset of F in $\mathcal{P}_{a,2}(1, c_a^2 + 1, M_a)$ having specified third moment $m_{a,3}$ and Laplace transform value $\hat{f}(\mu_a)$. Since we are working with the mgf $\hat{g}(-s)$ for $s > 0$, let $\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$ be the subset of G in $\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), M_s)$ having specified third moment $m_{s,3}$ and mgf value $\hat{g}(-\mu_s)$ at μ_s for $0 < \mu_s < s^*$. (Recall that $s^* = +\infty$ if G has bounded support.)

Let F_L and F_U (G_L and G_U) be the three-point extremal interarrival-time (service-time) cdf's corresponding to P_L^* and P_U^*, respectively, in the space $\mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a)$ ($\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$) based on §2.2.1. (Recall our convention that $E[U] = 1$ and $E[V] = \rho$.) In particular, F_L (F_U) is the unique element of $\mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a)$ with support on the set $\{0, x_1, x_2\}$ (on the set $\{x_1, x_2, M_a\}$) for $0 < x_1 < x_2 < M_a$, while G_L (G_U) is the unique element of $\mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$ with support on the set $\{0, \bar{x}_1, \bar{x}_2\}$ (on the set $\{\bar{x}_1, \bar{x}_2, \rho M_s\}$) for $0 < \bar{x}_1 < \bar{x}_2 < \rho M_s$.

Theorem 6. (*three-point extremal distributions for the decay rate*) *Let F_L, F_U, G_L and G_U be the three-point extremal cdf's for the $GI/GI/1$ queue defined above.*

(*a*) *For any $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1, m_{a,3}, \mu_a, M_a)$ with $\mu_a > 0$ and $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1))$ satisfying Assumption 1 such that equation (6) holds for the $F_L/G/1$ and $F_U/G/1$ models (with service cdf G), the unique positive solution of (6), $\theta_W(F, G)$, is well defined. Moreover, if $\mu_a \geq \theta_W$, then*

$$\theta_W(F_U, G) \leq \theta_W(F, G) \leq \theta_W(F_L, G); \tag{26}$$

if $\mu_a \leq \theta_W$, then

$$\theta_W(F_L, G) \leq \theta_W(F, G) \leq \theta_W(F_U, G). \tag{27}$$

(*b*) *For any $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1)$ and $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$, the unique positive solution of (6), $\theta_W(F, G)$, is well defined. Moreover, if $\mu_s \leq \theta_W$, then*

$$\theta_W(F, G_U) \leq \theta_W(F, G) \leq \theta_W(F, G_L); \tag{28}$$

if $\theta_W < \mu_s < s^*$, then

$$\theta_W(F, G_L) \le \theta_W(F, G) \le \theta_W(F, G_U). \tag{29}$$

(*c*) *As a consequence, for all $F \in \mathcal{P}_{a,2}(1, c_a^2+1, m_{a,3}, \mu_a, M_a)$ with $\mu_a > 0$ and $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), m_{s,3}, \mu_s, M_s)$ with $\mu_s > 0$, the unique positive solution of* (6)*, $\theta_W(F, G)$, is well defined. Moreover, for all (F, G) in these sets, the following four pairs of lower and upper bounds for $\theta_W(F, G)$ are valid:*

$$\begin{array}{rlll}
(i) & \theta_W(F_L, G_U) \le \theta_W(F, G) \le \theta_W(F_U, G_L) & \textit{if} & \mu_s, \mu_s \le \theta_W \\
(ii) & \theta_W(F_U, G_U) \le \theta_W(F, G) \le \theta_W(F_L, G_L) & \textit{if} & \mu_s \le \theta_W \le \mu_a \\
(iii) & \theta_W(F_U, G_L) \le \theta_W(F, G) \le \theta_W(F_L, G_U) & \textit{if} & \theta_W \le \mu_s, \mu_a,\ \mu_s < s^* \\
(iv) & \theta_W(F_L, G_L) \le \theta_W(F, G) \le \theta_W(F_U, G_U) & \textit{if} & \mu_a \le \theta_W \le \mu_s < s^*.
\end{array} \tag{30}$$

(*d*) *The bounds on θ_W get tighter as μ_a and μ_s move closer to $\theta_W(F, G)$. The bounds coincide with θ_W when $\mu_a = \theta_W$ in* (*a*) *and $\mu_s = \theta_W$ in* (*b*).

Proof. The proof is essentially the same as for Theorem 5, but now we have $n = 4$ for (a) and (b) instead of $n = 2$ in §2.2.1 and (5) of Eckberg (1977). As before, we apply the T-system theory from §2, but care is needed with the sign of the exponential arguments when we apply Lemma 2. To treat F, we apply Lemma 2 to show, first, that $\{1, t, t^2, t^3, e^{-\mu_a t}\}$ is a T system on $[0, M_a]$ for all $\mu_a > 0$. (Recall that $m = 3$ now, so that $(-1)^{m+1} = 1$.) But we also need to consider the set $\{1, t, t^2, t^3, e^{-\mu_a t}, e^{-st}\}$. For this second collection of functions, we require that $-\mu_a < -s$ or $\mu_a > s > 0$. If instead $s > \mu_a > 0$, then the set of functions becomes a T system if we change the order of the last two functions. But changing the order of two adjacent columns of a square matrix causes the sign of the determinant to change. That means that the supremum and infimum get switched.

For part (a), we see that all possible cases for F are covered by the two cases $\mu_a > s > 0$ and $s > \mu_a > 0$. Hence, if the decay rate θ_W is well defined for the two models $F_L/G/1$ and $F_U/G/1$ models, it is well defined for all F with the given constraints. The we get (26) and (27) in the two cases.

To treat G in part (b), the root θ_W is always well defined because G has bounded support. We apply Lemma 2 to show, first, that $\{1, t, t^2, t^3, e^{\mu_s t}\}$ is a T system on $[0, \rho M_s]$ for all $\mu_s > 0$, but then we also need to consider the set $\{1, t, t^2, t^3, e^{\mu_s t}, e^{st}\}$. For this second collection of functions, we require that $\mu_s < s < s^*$. If instead $0 < s < \mu_s$, then the set of functions becomes a T system if we change the order of the last two functions. But changing the order of two adjacent columns of a square matrix causes the sign of the determinant to change. That means that the supremum and infimum get switched. For G, the order also gets switched when we consider $1/\hat{g}(-s)$ instead of $\hat{g}(-s)$. The stated inequalities hold by combining the conclusions above.

Finally, part (c) is obtained by combining parts (a) and (b). ■

Remark 5. (choice of the Laplace transform values) Part (d) of Theorem 6 has important practical implications. It shows that, for any given model with a specified decay rate, the range of possible decay rate values consistent with the partial information becomes smaller as the arguments of the Laplace transforms become closer to the final decay rate.

3.4. Extending the Extremal Models to Unbounded Support

The T-system theory and the Markov-Krein theorem extend to unbounded support intervals as shown by Karlin and Studden (1966) and as indicated in Eckberg (1977) and Gupta and Osogami (2011). The extension is easy if the extremal distribution places no mass on the upper endpoint. Then the same extremal distribution holds for all larger support bounds, including the unbounded interval $[0,\infty)$.

First, in the setting of the two-point extremal distributions in Theorem 5, the extremal cdf's F_0 and G_0 have support on $\{0,x\}$ for appropriate x and so remain valid if we increase M_a and M_s. (The x depends on the cdf.)

Similarly, in the setting of the three-point extremal distributions in Theorem 5, the extremal cdf's F_L and G_L have support on $\{0,x_1,x_2\}$ for appropriate x_1 and x_2 and so remain valid if we increase M_a and M_s. (Again, the points x_1 and x_2 depend on the cdf.)

Consequently, we need to make no adjustments for truncation provided we use the following special case of (30):

$$\begin{aligned} \theta_W(F_L,G_L) &\leq \theta_W(F,G) \quad \text{for} \quad \mu_a \leq \theta_W \leq \mu_s < s^* \\ \theta_W(F_L,G_L) &\geq \theta_W(F,G) \quad \text{for} \quad \mu_s \leq \theta_W \leq \mu_a. \end{aligned} \tag{31}$$

This recipe also eliminates the need to consider multiple cases.

We state the result formally in the following corollary. To simplify, we make the following stronger assumption.

Assumption 2. (*uniformly good cdf G*) *In addition to Assumption 1, assume that, for the service-time cdf G, equation (6) has a finite solution for all $F \in \mathcal{P}_{a,2}(1,c_a^2+1)$.*

Note that Assumption 2 is satisfied by the M, H_k and E_k distributions considered here and many others, but we need to avoid pathological examples like Example 5 of Abate et al. (1995).

Corollary 1. (*extension to unbounded support*) *Consider the setting of Theorem 6 extended by allowing unbounded support, i.e., $M_a = M_s = \infty$.*

(*a*) *For any* $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1))$ *satisfying Assumption* 2, *the unique positive solution of* (6), $\theta_W(F,G)$, *is well defined. Moreover, if* $\mu_a \le \theta_W$, *then*

$$\theta_W(F_L, G) \le \theta_W(F, G) \tag{32}$$

for all $F \in \mathcal{P}_{a,2}(1, c_a^2+1, m_{a,3}, \mu_a)$.

(*b*) *For any* $F \in \mathcal{P}_{a,2}(1, c_a^2+1)$ *and* $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2+1), m_{s,3}, \mu_s)$ *satisfying Assumption* 2, *the unique positive solution of* (6), $\theta_W(F,G)$, *is well defined. Moreover, if* $\theta_W \le \mu_s < s^*$, *then*

$$\theta_W(F, G_L) \ge \theta_W(F, G). \tag{33}$$

for all (F,G).

(*c*) *For all* (F,G) *such that Assumption* 2 *holds, the unique positive solution of* (6), $\theta_W(F,G)$, *is well defined and* (31) *holds.*

4. Application to Produce a Practical Range for the Mean

We now apply the theoretical results for the decay rate established in §3 to develop a practical way to identify intervals of likely values for the mean steady-state waiting time given the basic moment parameters in (4) and the additional parameters introduced in Theorems 5 and 6. This analysis is heuristic, because we have no explicit relation between the decay rate and the mean, but the general idea is that the mean should be decreasing in the decay rate.

We have already outlined our approach in §1.2. We elaborate here. We start in §4.1 by discussing the support bounds used in Theorems 5 and 6. Then in §4.2 we indicate how we can obtain the extremal models with three-point distributions derived in Theorem 6. In §4.4 we report numerical results of our application of this method to the $GI/GI/1$ queue. Finally, in §4.5 we report numerical results for the $GI/GI/2$ queue.

4.1. Choosing the Support Bounds M_a and M_s

Before considering the support bounds, we emphasize that the range of possible values for the mean $E[W]$ in the $GI/GI/1$ model given only the first two moments of U and V tends to be remarkably wide. That is shown in Tables 1-2 and EC4-EC5 in Chen and Whitt (2018). The relative errors tend to increase in c_a^2 but decrease in ρ and c_s^2; see §2 of Chen and Whitt (2019b).

As indicated in §1.2.3, we use the support bounds M_a and M_s to give a good indication of the likely set of possible values given only the moments. Hence, starting from a specific model or data with unbounded U and V, we suggest choosing the support bounds M_a and M_s so that

$$P(U > M_a E[U]) = P(U > M_a) = P(V > M_s E[V]) = P(V > \rho M_s) = \epsilon \tag{34}$$

for a suitably small ϵ such as 0.001; see §3 of Chen and Whitt (2019b) for more discussion.

For our numerical experiments, as indicated in §1.2.2, we use the M, E_k and H_2 distributions, which are determined by their first two moments., For M, $c^2 = 1$; for E_k, $c^2 = 1/k < 1$; for H_2 distributions, $c^2 \geq 1$. We suggest using a simple exponential approximation based on the asymptotic decay rates of these distributions, which are well defined. Thus, we choose M_s so that

$$\epsilon = P(V/E[V] > M_s) \approx e^{-\theta_V M_s}, \tag{35}$$

where θ_V is the asymptotic decay rate of V.

For M, the decay rate of V is $\theta_V = 1/\rho$; for E_k, the scv is $1/k$, while the decay rate of V is $\theta_V = k/\rho$, so we let $\theta_V(\rho, c_s^2) = 1/\rho c_s^2$ for $c_s^2 \geq 0.01$, and $\theta(\rho, c_s^2) = 100/\rho$ for $c_s^2 \leq 0.01$ to avoid the deterministic case with $c_s^2 = 0$. Our examples use $c_s^2 = 0.5$, for which $\theta_V(\rho) = 2/\rho$. In the case of H_2 with balanced means, by (37) in Whitt (1982), the asymptotic decay rate of $V/E[V]$ is

$$\theta_V(1, c_s^2) = 1 - \sqrt{(c_s^2 - 1)/(c_s^2 + 1)}. \tag{36}$$

Our examples use $c_s^2 = 4.0$, for which we use $\theta_V(\rho, c_s^2) = (1 - \sqrt{3/5})/\rho = 0.2254/\rho$.

We now see how the extremal UB model $F_0/G_u/1$ and LB model $F_u/G_0/1$ for the decay rate from Theorem 5 apply to the mean $E[W]$ with $K = 1$ when we introduce the support bounds M_a and M_s following the prescription above. Table 1 show results for five cases: $(c_a, c_s^2) = (1.0, 1.0)$, $(4.0, 4.0)$, $(0.5, 0.5)$, $(4.0, 0.5)$, $(0.5, 4.0)$. (We show more results for other traffic intensities in §4 of Chen and Whitt (2019b).) We show two candidate support bounds for each case, based on $\epsilon = 0.01$ and 0.001 in (34). For comparison, Table 1 shows the heavy-traffic approximation (HTA) and the tight UB and LB given only the moments as well as the values of the mean with the support bounds.

Table 1 shows that the range decreases as the traffic intensity increases and as the support bounds decrease. For $\rho = 0.7$, the tight UB is not too far above the HTA approximation, but the tight LB tends to be far below. The mean for the $F_u/G_0/1$ model with M_a is significantly larger than the tight LB, but still the final range is very large, except for the one case $(c_a^2, c_s^2) = (0.5, 4.0)$. Note that the relative error is only about 5% for $\rho = 0.7$ in that good case; see Chen and Whitt (2018, 2019a) for additional details.

To obtain these estimates of $E[W]$ and later ones, we use simulation. We implement standard Monte-Carlo simulation to estimate the sample mean of the steady-state waiting time with a run length (number of arrivals) $N = 5 \times 10^8$ and 20 independent replications for the model $F_u/G_0/1$, but it helps to use an efficiency-improvement algorithm for the $F_0/G_u/1$ model with the atom at the upper support bound, as discussed in Chen and Whitt (2019a). We implement the Minh and Sorli (1983) simulation algorithm with total simulation length $T = 1 \times 10^7$ and 20 independent replications for the model $F_0/G_u/1$. We can construct 95% confidence interval by using statistical $t-$test.

Table 1 **Comparing bounds for $E[W]$ using $F_u/G_0/1$ (UB) and $F_0/G_u/1$ (LB) with (M_a, M_s) from §4.1**

	ρ	Tight LB	$M_a = 9$	$M_a = 7$	HTA (2)	$M_s = 7$	$M_s = 9$	Tight UB
$c_a^2 = c_s^2 = 1$	0.50	0.000	0.122	0.162	0.500	0.810	0.821	0.846
	0.70	0.467	0.970	1.130	1.633	2.025	2.036	2.071
	0.90	3.600	7.265	7.596	8.100	8.564	8.579	8.620
			$M_a = 39.9$	$M_a = 31.1$		$M_s = 31.1$	$M_s = 39.9$	
$c_a^2 = c_s^2 = 4$	0.50	0.750	1.013	1.097	2.000	3.419	3.430	3.470
	0.70	2.917	4.303	4.748	6.533	8.384	8.394	8.441
	0.90	15.750	28.924	30.239	32.400	34.658	34.671	34.721
			$M_a = 4.5$	$M_a = 3.5$		$M_s = 31.1$	$M_s = 39.9$	
$c_a^2 = 0.5, c_s^2 = 4$	0.50	0.750	0.957	0.988	1.125	1.263	1.270	1.289
	0.70	2.917	3.464	3.494	3.675	3.841	3.851	3.875
	0.90	15.750	17.973	17.993	18.225	18.408	18.427	18.470
			$M_a = 39.9$	$M_a = 31.1$		$M_s = 3.5$	$M_s = 4.5$	
$c_a^2 = 4, c_s^2 = 0.5$	0.50	0.000	0.000	0.000	1.125	2.556	2.559	2.595
	0.70	0.058	0.342	0.450	3.675	5.524	5.533	5.583
	0.90	1.575	9.075	11.988	18.225	20.469	20.486	20.546
			$M_a = 4.5$	$M_a = 3.5$		$M_s = 3.5$	$M_s = 4.5$	
$c_a^2 = 0.5, c_s^2 = 0.5$	0.50	0.000	0.000	0.000	0.250	0.377	0.388	0.414
	0.70	0.058	0.410	0.530	0.817	0.966	0.982	1.017
	0.90	1.575	3.613	3.771	4.050	4.207	4.229	4.295

The worst-case confidence interval length for Monte-Carlo simulation achieves 10^{-3} level which happens at the highest ρ, while the worst-case confidence interval length for the Minh and Sorli (1983) simulation is around 10^{-4} level. (See Chen and Whitt (2019a) for more discussion.)

In §4.2 of Chen and Whitt (2019b) we show that we could also start from the HT approximations in (2) and (8) instead of the exact models based on E_2 and H_2 distributions, Table 6 there compares the exact values of θ_W and $E[W]$ to these heavy-traffic approximations.

4.2. Determining the Extremal Models from Theorem 6

We now investigate how we can apply Theorem 6 to obtain a better indication of typical values of the mean $E[W]$. For specified parameters, it suffices to solve the equations characterizing the extremal models and calculate $E[W]$ for those extremal models.

First, we can solve the system of equations provided by the T-system theory by using a non-linear equation solver (we used MATLAB). Second, A convenient way to calculate the extremal distributions approximately (to any desired accuracy) is to assume finite support and apply linear programming to minimize (or maximize) the Laplace transform given the constraints. We can let the support be $\{kM_a/n : 0 \leq k \leq n\}$, so that the only variables are the probabilities p_k assigned to the points $x_k \equiv kM_a/n$. As in Theorem 1, there will necessarily be five-point extremal distributions given the four constraints using this approach. The solution converges to the three-point solution for the original support set $[0, M_a]$ as $n \to \infty$. Moreover, we can see that the optimal solution does not depend on the argument of the Laplace transform provided that the sign of $\mu - \theta_W$ does not change. See §6.1 of Chen and Whitt (2019b) for numerical examples.

4.3. Choosing the Laplace Transform Arguments

From Theorem 6 and Corollary 1, we see that we have five candidate ways to set the positive arguments of the Laplace transform $\hat{f}(s)$ and the mgf $\hat{g}(-s)$: the four cases with bounded support in (30) and the single composite version with unbounded support in (31). Both of these have advantages and disadvantages. First, the finite support bounds in (30) require truncation, so we either must calculate new parameters for the truncated model with bounded support or use the parameters of the original distributions with unbounded support without altering them. In addition, we must choose among the four alternatives in (30).

The alternative with unbounded support in (31) is appealing because it requires no truncation and we need not choose among four cases. On the other hand, it uses different parameter specifications for the minimum and maximum, which can distort the results, leading to anomalies such as the lower bound for the mean exceeding the upper bound.

We performed extensive experiments to test these alternatives and deduced that it is better to use the finite support bounds in (30) even though they require truncation, provided that the support bounds are chosen to have negligible impact, as in §4.1. In particular, we found that the parameters were not significantly altered by the truncation. For example, for the $E_2/H_2/1$ model with $\rho = 0.7$, the second and third moments of V with truncation were $s_2 = 2.44, s_3 = 20.19$ and $s_2 = 2.45, 20.58$ without truncation.) Hence, our procedure for the mean $E[W]$ uses the parameters taken directly from the base model with unbounded support or the heavy-traffic approximations, but then applies the results in (30) with the constructed support bounds.

It still remains to select one of the four alternatives in (30). From our experiments, we conclude that a good robust approximation is obtained by doing all four cases, and using the minimum of the four lower bounds for $E[W]$ for the final lower bound, and the maximum of the four upper bounds for $E[W]$ as the final upper bound. However, that requires more computational effort. Hence, we also propose a way to select one of the four alternatives.

We first observe that F_L (F_U) in (30) of Theorem 6 is the natural analog of F_0 (F_u) from Theorem 5, having 0 (M_a) as one of the mass points. Thus, case (i) in (30) is the natural choice. Nevertheless, we examine all four cases for the models we consider. To start, Table 2 below shows results for all four cases associated with the $M/M/1$ model with $\rho = 0.7$ and three possible values of R in (7). In the implementation, we do not allow $\mu_s > s^*$. Thus, if we are considering one of the cases with $\mu_s \geq \theta_W$, then we first check to see if $R\theta_W > s^*$ for our largest value of R, which we take to be $R = 20$. If it is, then we create alternative values of μ_s in the interval (θ_W, s^*). In particular, we use

$$\mu_s \equiv \theta_W + \left(\frac{R}{25}\right)(s^* - \theta_W), \quad 1 \leq k \leq 4, \tag{37}$$

so that the values of R remain in $\{5, 10, 15, 20\}$, but all values are within the interval (θ_W, s^*).

From the analytical formulas $\theta_W = (1-\rho)/\rho = 0.4286$ and $E[W] = \rho^2/(1-\rho) = 1.63$, we see that θ_W ($E[W]$) is strictly decreasing (increasing) in ρ. Of course, we are considering a large collection of models with $c_a^2 = c_s^2 = 1$, not simply $M/M/1$, but it is our reference case from which we extract parameters.

Table 2 **Bounds for θ_W (exact) and $E[W]$ (approximate) for $\rho = 0.7$ and $c_a^2 = c_s^2 = 1$ based on $M/M/1$ (For reference, exact values for $M/M/1$ are $\theta_W = (1-\rho)/\rho = 0.4286$ and $E[W] = \rho^2/(1-\rho) = 1.63$.)**

case	θ_W			$E[W]$			case	θ_W			$E[W]$		
(30)	$R=5$	10	20	$R=5$	10	20	(30)	$R=5$	10	20	$R=5$	10	20
(i)	0.426	0.425	0.425	1.67	1.67	1.68	(ii)	0.421	0.418	0.415	1.59	1.62	1.68
	0.432	0.432	0.439	1.65	1.65	1.56		0.434	0.437	0.446	1.53	1.56	1.61
(iii)	0.422	0.417	0.409	1.71	1.72	1.71	(iv)	0.426	0.424	0.418	1.61	1.60	1.57
	0.434	0.436	0.436	1.65	1.63	1.62		0.431	0.432	0.429	1.60	1.61	1.63

Consistent, with Theorem 6, Table 2 shows that the decay rate associated with the UB (LB) for $E[W]$ is decreasing (increasing) in R in each case, while the reverse order tends to hold for $E[W]$ too. There are minor exceptions in cases (iii) and (iv), because we get the decay rates from the original $M/M/1$ model.

From Table 2, we obtain the composite bounds for $E[W]$ based on all four cases. With $R = 20$, the composite bounds are

$$\min_{1 \le i \le 4} \{E[W_{LB,i}(R=20)]\} = 1.56 < E[W] = 1.63 < 1.71 = \max_{1 \le i \le 4} \{E[W_{UB,i}(R=20)]\}. \tag{38}$$

Notice that the interval $[1.56, 1.71]$ in (38) is not too different from the intervals $[1.56, 1.68]$ in case (i) with $\mu_s, \mu_a < \theta_W$ and $[1.61, 1.68]$ in case (ii) with $\mu_s < \theta_W < \mu_a$. On the other hand, the LB 1.62 for $E[W]$ in case (iii) is too large, while the UB 1.57 for $E[W]$ in case (iv) is too small. Thus, we tentatively conclude that it is better to have $\mu_s \le \theta_W$. For this case, the choice of μ_s seems to be more important than μ_a. We tentatively conclude that the cases (i) and (ii) in (30) are both consistently effective for the $M/M/1$ base model, while the other alternatives are not.

4.4. Numerical Experiments for the Non-Exponential Base Models

We now extend the study to the four models with $c_a^2, c_s^2 \in \{0.5, 4.0\}$ based on the H_2 and E_2 distributions. Table 3 shows the approximate upper bounds (top) and lower bounds (bottom) for $E[W]$ with $\rho = 0.7$ and $c_a^2, c_s^2 \in \{0.5, 4.0\}$ based on the E_2 and H_2 models in each of the four cases in (30) of Theorem 6 for three values in R in (7). The cases are labeled at the left by the base model. (The exact values of $E[W]$ for $H_2/H_2/1$, $H_2/E_2/1$, $E_2/H_2/1$ and $E_2/E_2/1$ are 6.61, 3.37, 3.56 and 0.725, respectively.)

Table 3 reinforces the conclusions about Table 2 for the case $c_a^2 = c_s^2 = 1$ based on the $M/M/1$ model. Table 3 shows that the UB exceeds the LB for all models and all values of R in case (i) with

Table 3 **Approximate upper and lower bounds for $E[W]$ for $\rho = 0.7$ and $c_a^2, c_s^2 \in \{0.5, 4.0\}$ based on the E_2 and H_2 models in each of the four cases in (30) of Theorem 6 for three values in R in (7) (The exact values of $E[W]$ for $H_2/H_2/1$, $H_2/E_2/1$, $E_2/H_2/1$ and $E_2/E_2/1$ are 6.61, 3.37, 3.56 and 0.725.)**

		(i)			(ii)			(iii)			(iv)	
model	$R=5$	10	20	$R=5$	10	20	$R=5$	10	20	$R=5$	10	20
H_2/H_2	6.93	6.94	6.73	6.28	6.19	7.20	6.93	7.08	7.20	6.72	6.72	6.66
	6.53	6.52	6.12	6.49	6.44	6.41	6.70	6.56	6.47	6.26	6.25	6.21
H_2/E_2	3.57	3.61	3.63	3.92	4.19	4.33	3.57	3.60	3.63	3.57	3.60	3.63
	3.06	3.08	3.06	2.95	2.82	2.69	3.06	3.08	3.06	3.06	3.08	3.06
E_2/H_2	3.62	3.68	3.68	3.53	3.54	3.56	3.51	3.51	3.52	3.52	3.52	3.49
	3.52	3.55	3.51	2.95	2.82	2.69	3.59	3.59	3.57	3.53	3.53	3.53
E_2/E_2	0.738	0.738	0.729	0.721	0.719	0.734	0.766	0.767	0.762	0.701	0.689	0.673
	0.737	0.733	0.704	0.642	0.625	0.642	0.730	0.730	0.721	0.736	0.738	0.753

$\mu_a, \mu_a \le \theta_W$, while this good property holds for case (ii) except for the case $c_a^2 = c_s^2 = 4.0$ based on the $H_2/H_2/1$ model, but it holds there as well for $R = 20$. In contrast, cases (iii) and (iv) perform significantly worse. In case (iii) the LB exceeds the UB for the case $c_a^2 = 0.5, c_s^2 = 4.0$ based on the $E_2/H_2/1$ model. In case (iv) the LB exceeds the UB for the case $c_a^2 = 0.5, c_s^2 = 4.0$ based on the $E_2/H_2/1$ model.

Table 17 in Chen and Whitt (2019b) displays the corresponding rates obtained in deriving the extremal distributions used for the mean $E[W]$ in Table 3. That table confirms Theorem 6, just like Table 2. (Again there are minor discrepancies because we get the decay rates from the original models.)

We offer two possible explanations for the better performance of cases (i) and (ii) in (30) of Theorem 6. First, since large waiting times tend to be caused by large service times and short interarrival times (leading to clumps of arrivals), we should pin down $E[W]$ most effectively from parameters with case (ii) with $\mu_s < \theta_W < \mu_a$ as in (7). A second consideration is the nature of the distribution itself. Given an E_k distribution that has a pdf h with $h(0) = 0$, large values of μ are not likely to help much. In contrast, a more variable H_2 distribution could be helped by additional specification wherever it appears. Thus, cases (iii) and (iv) with $c_a^2 = 0.5$ involving an E_2 arrival process are likely to not perform well, as we have seen.

4.5. Examples for Multi-Server Queues

Table 4 confirms that the procedure extends directly to $GI/GI/K$ queues with $K > 1$.

Table 4 **The improved UB and LB for $E[W]$ in $GI/GI/2$ for $(c_a^2, c_s^2) \in \{(1,1),(4.0,4.0),(4.0,0.5),(0.5,4,0),(0.5,0.5)\}$, $\rho \in \{0.5,0.7,0.9\}$ and $R \in \{5,10,20\}$**

$\rho = 0.5$	$c_a^2 = c_s^2 = 1$			$\rho = 0.7$	$c_a^2 = c_s^2 = 1$			$\rho = 0.9$	$c_a^2 = c_s^2 = 1$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	0.353	0.405	0.427	UB	1.34	1.39	1.41	UB	7.69	7.69	7.71
LB	0.290	0.262	0.251	LB	1.30	1.31	1.33	LB	7.67	7.62	7.61
$\rho = 0.5$	$c_a^2 = c_s^2 = 0.5$			$\rho = 0.7$	$c_a^2 = c_s^2 = 0.5$			$\rho = 0.9$	$c_a^2 = c_s^2 = 0.5$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	0.129	0.152	0.162	UB	0.590	0.606	0.608	UB	3.68	3.70	3.66
LB	0.092	0.087	0.086	LB	0.531	0.522	0.534	LB	3.64	3.66	3.64
$\rho = 0.5$	$c_a^2 = c_s^2 = 4$			$\rho = 0.7$	$c_a^2 = c_s^2 = 4$			$\rho = 0.9$	$c_a^2 = c_s^2 = 4$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	1.34	1.44	1.68	UB	5.29	5.37	5.76	UB	30.6	30.4	31.6
LB	1.30	1.27	1.21	LB	5.58	5.54	5.49	LB	30.9	30.7	30.8
$\rho = 0.5$	$c_a^2 = 4, c_s^2 = 0.5$			$\rho = 0.7$	$c_a^2 = 4, c_s^2 = 0.5$			$\rho = 0.9$	$c_a^2 = 4, c_s^2 = 0.5$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	1.33	1.49	1.59	UB	3.64	3.78	4.02	UB	17.9	17.9	18.1
LB	0.356	0.286	0.230	LB	2.65	2.56	2.43	LB	17.5	17.5	17.6
$\rho = 0.5$	$c_a^2 = 0.5, c_s^2 = 4$			$\rho = 0.7$	$c_a^2 = 0.5, c_s^2 = 4$			$\rho = 0.9$	$c_a^2 = 0.5, c_s^2 = 4$		
R	5	10	20	R	5	10	20	R	5	10	20
UB	0.540	0.548	0.556	UB	2.56	2.56	2.58	UB	16.6	16.6	17.0
LB	0.588	0.591	0.593	LB	2.73	2.74	2.72	LB	16.7	16.7	16.4

Indeed, we can apply the result for $K = 1$ to derive the decay rate. To apply the results for $K = 1$ to $K > 1$, we use the same extremal interarrival-time distribution, but multiply the extremal service-time random variable by K. We then can apply simulation to estimate $E[W]$ just as before. Table 4 shows the approximate upper and lower bounds for $E[W]$ obtained by this method for $\rho \in \{0.5, 0.7, 0.9\}$ and the five pairs of variability parameters $(c_a^2, c_s^2) \in \{(1,1),(4.0,4.0),(4.0,0.5),(0.5,4,0),(0.5,0.5)\}$ in case (ii) of (30) in Theorem 6 for $R \in \{5,10,20\}$. To illustrate for larger K, Table 5 shows set-valued approximations for $E[W]$ in the $M/M/10$ and $E_2/E_2/10$ models for $\rho \in \{0.7, 0.9\}$.

Table 5 **The set-valued approximations of $E[W]$ in $M/M/10$ (upper) and $E_2/E_2/10$ (lower) using case (ii) of (30) for $\rho = 0.7$ (left) and $\rho = 0.9$ (right)**

$\rho = 0.7$		θ_W			$E[W]$		$\rho = 0.9$		θ_W			$E[W]$	
	$R = 5$	10	20	$R = 5$	10	20		$R = 5$	10	20	$R = 5$	10	20
	0.421	0.418	0.415	0.520	0.523	0.539		0.111	0.111	0.110	5.97	6.05	6.07
	0.434	0.437	0.446	0.524	0.520	0.469		0.111	0.111	0.111	6.01	5.94	5.94
$\rho = 0.7$		θ_W			$E[W]$		$\rho = 0.9$		θ_W			$E[W]$	
	$R = 5$	10	20	$R = 5$	10	20		$R = 5$	10	20	$R = 5$	10	20
	0.842	0.833	0.825	0.176	0.177	0.179		0.222	0.221	0.221	2.76	2.71	2.74
	0.880	0.889	0.893	0.162	0.162	0.161		0.222	0.223	0.223	2.73	2.74	2.73

From readily available algorithms for $M/M/10$, we see that the exact values of $E[W]$ for $\rho = 0.7$ and 0.9 are 0.519 and 6.03, respectively, which fall right in the middle of the interval $[LB, UB]$ in each case. In contrast, the HTA in (2) are 1.633 and 8.10, which seriously overestimates the mean for $K = 10$. However, it is well known that the HTA, which tends to be good for $K = 1$, typically overestimates the mean for $K > 1$; e.g., see Whitt (2004) and references therein.

5. Conclusions

In this paper we investigated how additional support bounds M_a and M_s and other constraints on an interarrival time U with cdf F and a service times V with cdf G in the $GI/GI/1$ queue can help understand the quality of simple approximations for steady-state performance measures given partial information provided by the first two moments of U and V as specified by the parameter four-tuple $(1, c_a^2, \rho, c_s^2)$ in (4). The idea is to obtain an interval of likely values for performance measures given the partial information.

As a theoretical basis, we applied the theory of Tchebycheff systems to determine extremal models (yielding tight upper and lower bounds) for the asymptotic decay rate of the steady-state waiting-time tail probability, as in (5), (15) or (16). We reviewed the T system theory in §2 and exposed a relatively simple way to show that a system of functions is a T system in terms of Wronskians in §2.3. Lemma 2 verifies that the systems of functions we consider is a T system. Theorems 5 and 6 establish new tight upper and lower bounds for the decay rate in the $GI/GI/1$ queue and identify the extremal distributions. §3.1.2 shows that these results extend to the $GI/GI/K$ queue. Moreover, the extremal distributions for $K > 1$ are simple modifications of the extremal distributions for $K = 1$.

In §4 we showed that we can apply the theoretical results for the decay rate established in §3 to develop a practical way to identify intervals of likely values for the mean steady-state waiting time $E[W]$ given the basic moment parameters in (4) and the additional parameters introduced in Theorems 5 and 6, namely, support bounds, the third moments and values of the Laplace transform. We conducted extensive numerical experiments to study our proposed approach. We found that the proposed method based on cases (i) and (ii) in (30) of Theorem 6 is consistently effective for a range of base $GI/GI/K$ models. This performance is illustrated in §§4.3-4.5. For example, with these bounds, Table 3 shows that the maximum error of the midpoint of each interval in case (i) is less than 10% for all four models.

We emphasize that this good performance in our estimates of $E[W]$ depends critically on the extra parameters introduced in Theorem 6. With only the parameters in (4), the range is usually very wide, as shown in §2 of Chen and Whitt (2019b). However, the good performance can be expected if the actual model is near one of the base models using E_2 and H_2 cdf's that were used to generate the new parameters. Moreover, that remains true using the HTA for the decay rate in (8) and the third moments constructed from the base model in §1.2.2.

Overall, we contributed to a better understanding of simple queueing approximations such as (2) in typical $GI/GI/K$ cases. (At the end of §3.1.2 we noted that (2) tends to be quite good for $K = 1$, but seriously overestimates the true value for $K = 10$.) More generally, we presented a case for general set-valued performance approximations, given partial information about the model. We showed that with appropriate partial information it is possible to give a better idea of the range of likely values. A highlight is the unified application to $K > 1$ as well as $K = 1$.

There are many directions for future research. First, it remains to expose the precise relation between $E[W]$ and θ_W. Second, it remains to explore the approximation for other performance measures such as the tail probability $P(W > t)$. We expect even better results for large t, but then worse results for $t = 0$; see Abate et al. (1995). Third, there is opportunity for improved rare-event simulation for the extremal queues with $K > 1$ paralleling Minh and Sorli (1983) used for $K = 1$ in Chen and Whitt (2019a); see Minh (1989) for some. Finally, we think that there is great potential for applying this approach to other stochastic models.

Acknowledgments

We thank Richard Zalik of Auburn University for assistance with T systems.

References

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

Abate J, Choudhury GL, Whitt W (1995) Exponential approximations for tail probabilities in queues, i: Waiting times. *Operations Research* 43(5):885–901.

Abate J, Whitt W (1994) A heavy-traffic expansion for the asymptotic decay rates of tail probabilities in multi-channel queues. *Operations Research Letters* 15:223–230.

Abate J, Whitt W (1995) Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing* 7:36–43.

Asmussen S (2003) *Applied Probability and Queues* (New York: Springer), second edition.

Borovkov AA (1965) Some limit theorems in the theory of mass service, II. *Theor. Probability Appl.* 10:375–400.

Chen Y, Whitt W (2018) Extremal $GI/GI/1$ queues given two moments, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Chen Y, Whitt W (2019a) Algorithms for the upper bound mean waiting time in the $GI/GI/1$ queue, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Chen Y, Whitt W (2019b) Appendix to set-valued performance of queues given partial information, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Choudhury GL, Lucantoni D, Whitt W (1996) Squeezing the most out of ATM. *IEEE Transactions on Communications* 44(2):203–217.

Choudhury GL, Whitt W (1994) Heavy-traffic asymptotic expansions for the asymptotic decay rates in the $BMAP/G/1$ queue. *Stochastic Models* 10(2):453–498.

Cohen JW (1982) *The Single Server Queue* (Amsterdam: North-Holland), second edition.

Daley DJ (1977) Inequalities for moments of tails of random variables, with queueing applications. *Zeitschrift fur Wahrscheinlichkeitsetheorie Verw. Gebiete* 41:139–143.

Daley DJ (1997) Some results for the mean waiting-time and workloads in the $GI/GI/k$ queue. Dshalalow JH, ed., *Froniers in Queueing: Models and Applicatiions in Science and Engineering*, 35–59 (CRC Press, Boca Raton, FL).

Daley DJ, Kreinin AY, Trengove C (1992) Inequalities concerning the waiting-time in single-server queues: a survey. Bhat UN, Basawa IV, eds., *Queueing and Related Models*, 177–223 (Clarendon Press).

Eckberg AE (1977) Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems. *Mathematics of Operations Research* 2(2):135–142.

Gamarnik D, Goldberg DA (2013) Steady-state $GI/GI/n$ queue in the Halfin-Whitt regime. *Ann. Appl. Probability* 23(6):2382–2419.

Glynn PW, Whitt W (1994) Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *J. Appl. Prob.* 31:131–156.

Gupta V, Dai J, Harchol-Balter M, Zwart B (2010) On the inapproximability of $M/G/K$: why two moments of job size distribution are not enough. *Queueing Systems* 64:5–48.

Gupta V, Osogami T (2011) On Markov-Krein characterization of the mean waiting time in $M/G/K$ and other queueing systems. *Queueing Systems* 68:339–352.

Halfin S, Whitt W (1981) Heavy-traffic limits for queues with many exponential servers. *Operations Research* 29(3):567–588.

Holtzman JM (1973) The accuracy of the equivalent random method with renewal inouts. *Bell System Techn ical Journal* 52(9):1673–1679.

Iglehart DL, Whitt W (1970a) Multiple channel queues in heavy traffic, I. *Advances in Applied Probability* 2(1):150–177.

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

Johnson MA, Taaffe MR (1991) An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queieing Systems* 8(1-2):129–148.

Johnson MA, Taaffe MR (1993) Tchebycheff systems for probability analysis. *American Journal of Mathematical and Management Sciences* 13(1-2):83–111.

Karlin S, Studden WJ (1966) *Tchebycheff Systems; With Applications in Analysis and Statistics*, volume 137 (New York: Wiley).

Kelly FP (1996) Notes on effective bandwidths. F P Kelly SZ, Ziedins I, eds., *Stochastic Networks: Theory and Applications*, 141–168 (Clarendon Press, Oxford).

Kingman JFC (1961) The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.* 77:902–904.

Kingman JFC (1962) Inequalities for the queue $GI/G/1$. *Biometrika* 49(3/4):315–324.

Kingman JFC (1964) A martingale inequality in the theory of queues. *Proc. Camb. Phil. Soc.* 59:359–361.

Kingman JFC (1966) The heavy traffic approximation in the theory of queues. Smith WL, Wilkinson WE, eds., *Proceedings of the Symposium on Congestion Theory*, 137–159 (The University of North Carolina Press, Chael Hill, NC).

Kingman JFC (1970) Inequalities in the theory of queues. *J. Roy. Statist. Soc., Series B* 32(1):102–110.

Klincewicz J, Whitt W (1984) On approximations for queues, ii: Shape constraints. *AT&T Bell Laboratories Technical Journal* 63(1):115–138.

Kollerstrom J (1974) Heavy traffic theory for queues with several servers. *J. Appl. Prob.* 11(3):544–552.

Lasserre JB (2010) *Moments, Positive Polynomials and Their Applications* (Imperial College Press).

Minh DL (1989) Simulating $GI/G/k$ queues in heavy traffic. *Management Science* 33(9):1192–1199.

Minh DL, Sorli RM (1983) Simulating the $GI/G/1$ queue in heavy traffic. *Operations Research* 31(5):966–971.

Neuts MF (1986) The caudal characteristic curve of queues. *Adv. Appl. Prob.* 18:221–254.

Neuts MF, Takahashi Y (1981) Asymptotic behavior of stationary distributions in thee $GI/PH/C$ queue with hterogeneous servers. *Z. Wahrscheinlichkeiteth.* 57:441–452.

Rolski T (1972) Some inequalities for $GI/M/n$ queues. *Zast. Mat.* 13(1):43–47.

Seelen LP, Tijms HC, van Hoorn MH (1985) *Tables for Multi-Server Queues* (Amsterdam: North-Holland).

Smith J (1995) Generalized Chebychev inequalities: Theory and application in decision analysis. *Operations Research* 43:807–825.

Takahashi Y (1981) Asymptotic exponentiality of the tail of the waiting time distribution in a $Ph/Ph/c$ queue. *Adv. Appl. Prob.* 13(3):619–630.

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

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

Whitt W (1984a) On approximations for queues, I. *AT&T Bell Laboratories Technical Journal* 63(1):115–137.

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

Whitt W (1993) Tail probabilities with statistical multiplexing and effective bandwidths in multiclass queues. *Telecommunication Systems* 2:71–107.

Whitt W (2004) A diffusion approximation for the $G/GI/n/m$ queue. *Operations Research* 52(6):922–941.

Wolff RW, Wang C (2003) Idle period approximations and bounds for the $GI/G/1$ queue. *Advances in Applied Probability* 35(3):773–792.

Zalik RA (1996) Chebychev and weak Chebychev systems. Gasca M, Miccelli AA, eds., *Total Positivity and its Applications*, 301–332, MAIA Volume 359 (Kluwer Academic Publishers, Dordrecht).

Zalik RA (2011) Some properties of Chebycheff systems. *J. Comput. Anal. Appl.* 13(1):20–26.