Operations Research Letters 48 (2020) 770–776

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Extremal models for the $GI/GI/K$ waiting-time tail-probability decay rate

Yan Chen [1], Ward Whitt [*,1]

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

ARTICLE INFO

Article history:
Received 13 April 2020
Received in revised form 21 July 2020
Accepted 8 September 2020
Available online 11 September 2020

Keywords:
Queues
Tail probabilities
Tchebycheff systems
Bounds

ABSTRACT

We identify interarrival-time and service-time distributions that yield tight upper and lower bounds on the asymptotic decay rate of the steady-state waiting-time tail probability in the $GI/GI/K$ queue, given the first two moments of those underlying distributions plus alternative additional information. We apply Tchebycheff systems after providing a concise review.

1. Introduction

We consider the tail probability of the steady-state waiting time W, i.e., $P(W > t)$, in the $GI/GI/K$ queue, i.e., in the 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 (cdfs) F and G. We are especially interested in exposing the performance impact of the variability of these underlying cdfs 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 focus on the light-tailed case, where the service-time cdf G has finite moments of all orders. We then typically have

$$P(W > t) \sim \alpha e^{-\theta_W t} \quad \text{as} \quad t \to \infty, \tag{1}$$

where $f(t) \sim g(t)$ as $t \to \infty$ means that $f(t)/g(t) \to 1$, e.g., see [1]. Then we call θ_W the (asymptotic) decay rate. Under regularity conditions, the decay rate θ_W in (1) is attained as the unique positive real root of an equation involving the Laplace transforms of U and V, e.g., $\hat{f}(z) \equiv \int_0^\infty e^{-zt}\, dF(t)$. In particular, the equation for the decay rate is

$$\hat{f}(z)\hat{g}(-z) = 1. \tag{2}$$

In this light-tailed setting, we show that the theory of Tchebycheff (T) systems from [10], as used in [5,7–9,12,13], can be applied to determine extremal models (yielding tight upper and lower bounds) on the asymptotic decay rate θ_W above. We propose these extremal models as a way to provide heuristic set-valued approximations for a variety of performance measures in the challenging $GI/GI/K$ model given partial information. In [4] we provide evidence that this heuristic approach is effective for the steady-state mean $E[W]$. Here we start in Section 2 by giving background on T systems. In Section 3 we elaborate on (2) and obtain the extremal distributions for the decay rate.

2. Tchebycheff system foundations

To put the T system results in perspective, we start in Section 2.1 by reviewing the classical moment problem, as in [15]. Then in Section 2.2 we specify the additional conditions needed to get a T system and state the Markov–Krein theorem. In Section 2.3 we develop convenient lemmas under smoothness conditions.

2.1. The classical moment problem

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 $u_i(t) \equiv t^i$, $0 \le i \le n$, the usual moments. We want to draw conclusions about the unspecified underlying probability measure P on $[a, b]$ such that:

$$m_i \equiv E_P[u_i] \equiv \int_a^b u_i\, dP, \quad 0 \le i \le n. \tag{3}$$

* Corresponding author.
E-mail addresses: yc3107@columbia.edu (Y. Chen), ww2040@columbia.edu (W. Whitt).
[1] Department of Industrial Engineering and Operations Research, Columbia University, Mail Code 4704, S. W. Mudd Building, 500 West 120th Street, New York, NY 10027-6699, USA.

https://doi.org/10.1016/j.orl.2020.09.004
0167-6377/

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.

Let $\mathcal{P}_n$ be the set of all probability measures P on $[a, b]$ with $n+1$ moments, assumed to be nonempty. Let $\mathcal{P}_{n,k}$ be the subset of probability measures in $\mathcal{P}_n$ that have support on at most k points. 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 [10].)

Theorem 2.1 (*A Version of the Classic Moment Problem, §2.1 of [15]*)**.** *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*

$$\sup\{\int_a^b \phi\, dP : P \in \mathcal{P}_n\} = \sup\{\int_a^b \phi\, dP : P \in \mathcal{P}_{n,n+1}\} = \int_a^b \phi\, dP^*. \tag{4}$$

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 (4). Theorem 2.1 implies that there exist extremal distributions with $\sigma(P_U^*) \le n+1$ and $\sigma(P_L^*) \le n+1$.

2.2. Tchebycheff systems and the Markov–Krein theorem

If we make additional assumptions about the functions u_i, then we can apply T systems to identify concrete extremal distributions P_U^* and P_L^* in (4); see the seminal book [10] and the review papers [9,17].

2.2.1. Upper and lower principle representations.

We impose a regularity condition involving the moment space $\mathcal{M}_n$, i.e., on $\{(m_1, \ldots, m_n)\}$ in $\mathbb{R}^n$ such that there exists $P \in \mathcal{P}_n$ such that $\int_a^b u_i\, dP = m_i$ for all i. 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 $\mathcal{M}_n$.

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 (3), a solution P to (4) must have $2k \ge n+1$. When n is odd, we must have $\sigma(P) \ge (n+1)/2$. When n is even, we must have $\sigma(P) \ge 1+(n/2)$. The final story under the T-system assumption is different in these two cases. It is summarized in (5) of [5] and on p. 342 of [7].

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, $k = 1 + (n/2)$, 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 [10]), 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 $n-1$ unknowns. In fact, the remaining $(n-1)/2$ points inside the open interval (a, b) are then uniquely determined.

2.2.2. The Markov–Krein theorem

The Markov–Krein theorem says that the description above holds if certain collections of functions constitute a T system. In [10], T system theory is first developed for continuous functions on a compact interval in Chapters I–III and then extended to unbounded intervals and discrete subsets in later chapters, but a totally ordered set is needed. In this paper we consider the basic case $[a, b]$.

Definition 1 (*T System*)**.** Consider the same set of $n+1$ continuous real-valued functions $\{u_i(t) : 0 \le i \le n\}$ defined on $[a, b]$ introduced in Section 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 $u_i(t_j)$, $0 \le i \le n$ and $0 \le j \le n$, is strictly positive for all $a \le t_0 < t_1 < \cdots < t_n \le b$.

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 [10]) possesses at most n distinct zeros in $[a, b]$. (Nontrivial means that $\sum_{i=0}^n a_i^2 > 0$.)

Theorem 2.2 (*Markov–Krein, §III.1 of [10]*)**.** *In the setting of Theorem 2.1 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 $[a, b]$, the upper and lower extremal distributions P_U^* and P_L^* described above uniquely attain the supremum and infimum of the optimization problem in (4).*

2.3. Convenient sufficient conditions for smooth functions: Wronskians

The major challenge for applications is showing that the two sets of functions in Theorem 2.2 are indeed T systems. However, 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 jth derivative of u_i at the argument t. The Wronskian of the $n+1$ functions $\{u_i(t) : 0 \le i \le n\}$ is the determinant of the $(n+1) \times (n+1)$ matrix $\{u_i^{(j)}(t) : 0 \le i, j \le n\}$ of the functions and their derivatives

$$W_n(u_i : 0 \le i \le n) \equiv det(u_i^{(j)}(t) : 0 \le i, j \le n). \tag{5}$$

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

$$\begin{bmatrix} 1 & t & t^2 & -e^{-zt} \\ 0 & 1 & 2t & ze^{-zt} \\ 0 & 0 & 2 & -z^2e^{-zt} \\ 0 & 0 & 0 & z^3e^{-zt} \end{bmatrix}$$

which clearly is $2z^3e^{-zt} > 0$.

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 [10]*)**.** If each (ordered) subset $\{u_i(t) : 0 \le i \le m\}$ for $1 \le m \le n$ 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 $u_i(t) \equiv t^i$, $0 \le i \le n$. Then the determinant is the Vandermonde determinant

$$det(u_i(t_j) : 0 \le i, j \le m) = \prod_{0 \le i < j \le m} (t_j - t_i) \quad \text{for all} \quad 1 \le m \le n, \tag{6}$$

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

The direct definition of an extended T system in §I.2 of [10] is somewhat complicated. Thus, we give an equivalent definition

Definition 4 (*Extended T System, §I.2 of [10] and Theorem 1 of [18]*)**.** 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 [10]) possesses at most n 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 [10], Theorem 5 and Corollary 1 of [17], and Theorem 29 of [9].

Theorem 2.3 (*Wronskians and ECT Systems, p. 376 of [10]*)**.** *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 2.3 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 2.1. *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 $(n+1)^{\text{st}}$ 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 kth Wronskian takes 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 the limited class of ECT systems covered by the following lemma (where i, k and m are integers).

Lemma 2.2 (*Sufficient Conditions for this Paper*)**.** *Consider three ordered sets of continuous real-valued functions on the interval $[0, M]$: $\mathcal{A}_1(m) \equiv \{t^k : 0 \le k \le m\}$, $\mathcal{A}_2 \equiv \{(-1)^{m+1} e^{-z_i t} : z_i > z_{i+1} > 0 \text{ for all } i\}$ and $\mathcal{A}_3 \equiv \{e^{z_i t} : 0 < z_i < z_{i+1} \text{ for all } i\}$. 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 $(1, t, t^2, -e^{-z_1 t}, -e^{-z_2 t}, e^{z_3 t}, e^{z_4 t})$ where $z_1 > z_2 > 0$ and $0 < z_3 < z_4$, so that $-z_1 < -z_2 < z_3 < z_4$. Here $m = 2$, so $(-1)^{m+1} = -1$. Overall, the exponential arguments are increasing as in (3.1) on p. 9 of [10] or Example 6 of [17].

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 $k \ge j$. 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 determinant of the $n \times n$ matrix containing elements $e^{x_i y_j}$, $1 \le i, j \le n$, 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 [10] and Example 6 of [17]. ■

3. Extremal models for the asymptotic decay rate

In Section 3.1, we provide technical background on the decay rate for $K = 1$; in Section 3.2 we show that this approach also applies to the $GI/GI/K$ model for $K > 1$. In Section 3.3 we obtain two-point extremal distributions given only the first two moments of U and V and bounded intervals of support. Then in Section 3.4 we obtain more useful (as shown in [4]) three-point extremal distributions when we are also given the third moment and values of the Laplace transform of U and V. We give illustrative examples in Section 3.5. We discuss the extension to unbounded support in Section 3.6.

3.1. Theory for the asymptotic decay rate with $K = 1$

To increase the level of generality for $K = 1$, instead of (1), we can let θ_W be defined by the critical exponent in the Kingman–Lundberg bound for the $GI/GI/1$ queue, as in §XIII.5 of [2], defined by

$$\theta_W \equiv \inf\{x \ge 0 : P(W > t) \le e^{-xt}, \quad t \ge 0\}, \tag{7}$$

so that large waiting times correspond to small values of θ_W. Under regularity conditions, θ_W in (7) coincides with the asymptotic decay rate studied in large-deviations theory, defined by

$$\theta_W \equiv \lim_{x \to \infty} \frac{-\log P(W > x)}{x}. \tag{8}$$

We assume that a strictly positive infimum exists in (7) and a strictly positive limit exists in (8), which requires that the service-time V must have a finite moment generating function $E[e^{zV}]$ for some $z > 0$. (We obtain $\theta_W = \infty$ if $P(V - U \le 0) = 1$ 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 [2], large deviation limits in Corollary 1 in §1.2 of [6] and approximations in [1].

Part of the appeal of this approach is that it extends directly to $K > 1$, as we show in Section 3.2. Moreover, it has been observed that the approximation $P(W > t | W > 0) \approx e^{-\theta_W t}$ is quite good for $K \ge 1$; see [14]. Indeed, for that reason, θ_W is displayed in the tables there (with different scaling, i.e., with $E[V] = 1$).

Under regularity conditions, the asymptotic decay rate θ_W in (1), (7) or (8) is attained as the unique positive real root of equation (2) involving the Laplace transforms of U and V. Equivalently, as in §XIII.1 of [2], $\kappa_F(z) + \kappa_G(-z) = 0$, where $\kappa_F(z) \equiv \log(\hat{f}(z))$ is the cumulant generating function. (The function $\hat{g}(-z) \equiv E[e^{zV}]$ for $z > 0$ is the moment generating function (mgf).)

Given the simple structure in (2), the extremal result and alternative ones follow from the theory of T systems, as in Section 2. To state the result, we impose some technical conditions.

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

In general, we need to impose additional regularity conditions to have the limit for the decay rate in (8) be well defined, as can be seen from Corollary 1 and Proposition 2 in [6] and Theorems 2.1, 5.5 and 5.3 in Chapter XIII in [2]. Instead of adding additional assumptions, we allow the decay rate to be defined by (7). It coincides with (8) when the limit exists.

We still need extra conditions for (2) to have a solution; see Example 5 in §3 and Theorem 5 in §7 of [1]. However, no extra condition is needed when G has support in $[0, M_s]$, because then $E[e^{tV}] \le e^{tM_s}$ for all $t > 0$, so that $z^* = \infty$ in Assumption 1.

3.2. Extension to the GI/GI/K model

As indicated in [1], the asymptotic decay rate also is well defined for the more general $GI/GI/K$ model. We have fixed $E[U] = 1$, but instead in [1] there is fixed $E[V] = 1$. In (5) of [1], with $\theta_W(K)$ denoting the decay rate for the K-server model, $\theta_W(K) = K\theta_W(1)$, where $U(K) = U/K$ to keep ρ fixed. Since we fix $E[U] = 1$, we get $\theta_W(K) = \theta_W(1) \equiv \theta_W$. (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 [2].) 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 $V(K) = KV$. Thus the finite support of $V(K)$ becomes $[0, \rho KM_s]$, the pth moment of $E[V(K)^p] = K^p E[V^p]$ and the Laplace transforms are related by $\hat{g}_{V(K)}(z) = \hat{g}_V(Kz)$. 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 $V^*(K) = KV^*$.

In [1], it was observed that the extension to $K > 1$ in (5) there was proved for the $GI/PH/K$ by [11]. A continuity result implies that result applies to the general $GI/GI/K$ model.

Theorem 3.1 (*Extension of Decay Rate to GI/GI/K*)**.** *If the decay rate θ_W is well defined for the $GI/GI/1$ model with (U, V) having cdfs (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*

$$\theta_W(K) = \theta_W(1) \equiv \theta_W \quad \text{for} \quad K > 1. \tag{9}$$

Proof. Fix the interarrival-time cdf F and consider a sequence of phase-type service-time $\{G_n : n \ge 1\}$ 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 [11], (9) holds for each n, as explained above. The convergence in distribution implies the associated convergence $\hat{g}_n(z) \to \hat{g}(z)$ for each z. Since the Laplace transform $\hat{g}(z)$ is continuous and strictly decreasing in the real variables z, (9) must hold in the limit as well. ■

Remark 3.1. The $GI/Ph/K$ model is 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.

3.3. Two-Point extremal distributions given two moments

We now are able to present our main results. 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 cdfs 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 Section 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) cdfs 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 Section 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

$$\theta_W : \mathcal{P}_{a,2}(1, 1+c_a^2, M_a) \times \mathcal{P}_{s,2}(\rho, \rho^2(1+c_s^2), M_s) \to \mathbb{R}, \tag{10}$$

where $0 < \rho < 1$ and $\theta_W(F, G) \equiv \theta_W$ 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 3.2 (*Two-point Extremal Distributions for the Decay Rate*)**.** *Let F_0, F_u, G_0 and G_u be the two-point extremal cdfs 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 $\bar{z}$ to Eq. (2) for the $F_u/G/1$ model (with service cdf G) such that $0 < \bar{z} < z^*$, where z^* is defined in Assumption 1,

$$\theta_W(F_0, G) \le \theta_W(F, G) \le \theta_W(F_u, G) \tag{11}$$

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

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

$$\theta_W(F, G_u) \le \theta_W(F, G) \le \theta_W(F, G_0) \tag{12}$$

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

$$\theta_W(F_0, G_u) \le \theta_W(F, G) \le \theta_W(F_u, G_0). \tag{13}$$

Proof. We make extra conditions in part (a) to ensure that equation (2) has a solution $\bar{z}$ strictly less than the upper limit z^*, but no extra conditions are needed in parts (b) and (c) because then G has bounded support, implying that $z^* = +\infty$.

We apply (2) to see that order for the Laplace transforms translates into order for θ_W, recalling that (i) (2) is equivalent to $\hat{f}(z) = 1/\hat{g}(-z)$, (ii) Laplace transforms are continuous strictly decreasing functions of a real variable argument and (iii) large waiting times are associated with smaller θ_W. For part (a), we see that

$$\hat{f}_u(z) \le \hat{f}(z) \le \hat{f}_0(z) \quad \text{for} \quad z > 0. \tag{14}$$

From (2) and (14), we see that, for any $\hat{g}$, θ_W is maximized by $\hat{f}_u$ in (14). Hence, (2) 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

$$1/\hat{g}_u(-z) \le 1/\hat{g}(-z) \le 1/\hat{g}_0(-z) \quad \text{for} \quad z > 0. \tag{15}$$

From (2) and (15), we see that, for any $\hat{f}$, θ_W is maximized by $1/\hat{g}_0(-z)$ in (15).

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

Based on Theorem 3.2, 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 3.1 (*Limits as the Support Increases*)**.** *Under the assumptions of Theorem 3.2, for all* $F \in \mathcal{P}_{a,2}(1, c_a^2 + 1, M_a)$ *and* $G \in \mathcal{P}_{s,2}(\rho, \rho^2(c_s^2 + 1), M_s)$,

$$\theta_W(F, G_u) \to 0 \quad as \quad M_s \to \infty, \tag{16}$$

while

$$\theta_W(F_u, G) \to \theta_W(F_1, G) \quad as \quad M_a \to \infty, \tag{17}$$

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

Remark 3.2 (*The Decay Rates of Other Steady-State Distributions.*)**.** Analogs of Theorem 3.2 (and the later Theorem 3.3) 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 [6]. For the workload, the decay rate is the same; for the queue length, $\theta_Q = \hat{g}(-\theta_W)$.

Remark 3.3 (*Comparison to the Mean*)**.** In the $GI/GI/1$ queues, the extremal model (F_0, G_u) in Theorem 3.2 yielding the smallest decay rate coincides with the frequently conjectured upper bound model for the mean $E[W]$, but the extremal model (F_u, G_0) in Theorem 3.2 yielding the largest decay rate does not coincide with the lower bound for the mean; see [3].

3.4. Additional constraints

We now add additional constraints on the cdfs F and G. In particular, we add a third moment and a value of the Laplace transform. With (2) in mind, we now impose constraints on the Laplace transform $\hat{f}(z)$ at $z = \mu_a > 0$ and on the reciprocal of the mgf, $1/\hat{g}(-z)$, at $z = \mu_s$, $0 < \mu_s < z^*$, for z^* 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 $\hat{g}(-z)$ for $z > 0$, 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 $0 < \mu_s < z^*$. (Recall that $z^* = +\infty$ 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) cdfs 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 Section 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 3.3 (*Three-Point Extremal Distributions for the Decay Rate*)**.** *Let* F_L, F_U, G_L *and* G_U *be the three-point extremal cdfs 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, where Eq. (2) holds for the* $F_L/G/1$ *and* $F_U/G/1$ *models (with service cdf* G*), the unique positive solution of (2),* $\theta_W(F, G)$*, is well defined. Moreover, if* $\mu_a \geq \theta_W$*, then*

$$\theta_W(F_U, G) \leq \theta_W(F, G) \leq \theta_W(F_L, G); \tag{18}$$

if $\mu_a \leq \theta_W$*, then*

$$\theta_W(F_L, G) \leq \theta_W(F, G) \leq \theta_W(F_U, G). \tag{19}$$

(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 (2),* $\theta_W(F, G)$*, is well defined. Moreover, if* $\mu_s \leq \theta_W$*, then*

$$\theta_W(F, G_U) \leq \theta_W(F, G) \leq \theta_W(F, G_L); \tag{20}$$

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

$$\theta_W(F, G_L) \leq \theta_W(F, G) \leq \theta_W(F, G_U). \tag{21}$$

(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 (2),* $\theta_W(F, G)$*, 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:*

$$\begin{aligned}
&(i) \quad \theta_W(F_L, G_U) \leq \theta_W(F, G) \leq \theta_W(F_U, G_L) \quad if \quad \mu_s, \mu_s \leq \theta_W \\
&(ii) \quad \theta_W(F_U, G_U) \leq \theta_W(F, G) \leq \theta_W(F_L, G_L) \quad if \quad \mu_s \leq \theta_W \leq \mu_a \\
&(iii) \quad \theta_W(F_U, G_L) \leq \theta_W(F, G) \leq \theta_W(F_L, G_U) \quad if \quad \theta_W \leq \mu_s, \mu_a, \ \mu_s < z^* \\
&(iv) \quad \theta_W(F_L, G_L) \leq \theta_W(F, G) \leq \theta_W(F_U, G_U) \quad if \quad \mu_a \leq \theta_W \leq \mu_s < z^*.
\end{aligned} \tag{22}$$

(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 3.2, but now we have $n = 4$ for (a) and (b) instead of $n = 2$ in Section 2.2.1 and (5) of [5]. As before, we apply the T-system theory from Section 2, but care is needed with the sign of the exponential arguments when we apply Lemma 2.2. To treat F, we apply Lemma 2.2 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 $\{1, t, t^2, t^3, e^{-\mu_a t}, e^{-zt}\}$. For this second collection of functions, we require that $-\mu_a < -z$ or $\mu_a > z > 0$. If instead $z > \mu_a > 0$, 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 $\mu_a > z > 0$ and $z > \mu_a > 0$. 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 (18) and (19) 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 2.2 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 $\{1, t, t^2, t^3, e^{\mu_s t}, e^{zt}\}$. For this second collection of functions, we require that $\mu_s < z < z^*$. If instead $0 < z < \mu_s$, 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 $1/\hat{g}(-z)$ instead of $\hat{g}(-z)$. Then combine the conclusions above.

Finally, part (c) is obtained by combining (a) and (b), while (d) follows easily from (2). ■

3.5. Illustrative examples

We now illustrate Theorems 3.2 (c) and 3.3(c) (i) by showing how the extremal models perform for the steady-state mean EW. In doing so, we are providing a small sample from our companion study [4] to which we refer for more details and examples. We emphasize that we have not shown that the extremal models for the decay rate θ_W necessarily determine associated bounds for the mean EW. Nevertheless, our study indicates that we obtain useful estimates of the intervals of likely values for the mean EW.

Table 1 shows the mean steady-state waiting time EW along with the decay rate in four base models: $H_2/H_2/1$ with $c_a^2 = c_s^2 = 4.0$, $E_2/E_2/1$ with $c_a^2 = c_s^2 = 0.5$, $E_2/H_2/1$ with $c_a^2 = 0.5$, $c_s^2 = 4.0$, and $H_2/E_2/1$ with $c_a^2 = 4.0$, $c_s^2 = 0.5$, all with $\rho = 0.7$. The third parameter of the H_2 distribution is specified by stipulating balanced means as in (3.7) on p. 137 of [16]. For each model, the exact values are shown in the middle columns of Table 1.

Theorem 3.3(c) requires specification of μ_a and μ_s. Consistent with case (i), we let them be $\theta_W/20$. The lower-bound extremal model is then F_U/G_L, while the upper-bound extremal model is then F_L/G_U. The decay rates are shown below the mean in each case.

The outer columns of Table 1 refer to bounds based only on the first two moments of F and G; see §2 of [3]. First LB is the established tight lower bound, while UB is the conjectured tight upper bound, which is the limit of $EW(F_0, G_u)$ as the upper limit of support M_s approaches infinity. The mean values EW for the extremal models F_u/G_0 and F_0/G_u based on Theorem 3.2 require specification of the upper limits of support M_a for F and M_s for G. The specific values used were $M = 39.9$ for $c^2 = 4$ and 4.5 for $c^2 = 0.5$. These were chosen so that

$$P(V/EV > M_s) \approx e^{-\theta_V M_s} = \epsilon, \tag{23}$$

where θ_V is the decay rate of the distribution of V and $\epsilon = 0.001$.

In summary, Table 1 shows that the extremal models F_U/G_L and F_L/G_U obtained from Theorem 3.3(c) (i) with judiciously chosen parameters provide a reasonably short range for the mean EW, whereas the F_u/G_0 and F_0/G_u models from Theorem 3.2 do not.

Table 1
Comparing bounds and approximations for the steady-state mean $E[W]$ using Theorems 3.2(c) and 3.3(c) (i), starting with the base models $H_2/H_2/1$ with $c_a^2 = c_s^2 = 4.0$, $E_2/E_2/1$ with $c_a^2 = c_s^2 = 0.5$, $H_2/E_2/1$ with $c_a^2 = 4.0$, $c_s^2 = 0.5$ and $E_2/H_2/1$ with $c_a^2 = 0.5$, $c_s^2 = 4.0$, all with $\rho = 0.7$.

	LB	F_u/G_0	F_U/G_L	Exact	F_L/G_U	F_0/G_u	UB
$c_a^2 = 4.0, c_s^2 = 4.0$	2.92	4.30	6.12	6.61	6.73	8.39	8.44
decay rates		0.190	0.107	0.106	0.106	0.075	
$c_a^2 = 0.5, c_s^2 = 0.5$	0.058	0.470	0.704	0.725	0.729	0.982	1.017
decay rates		2.002	0.865	0.857	0.852	0.631	
$c_a^2 = 0.5, c_s^2 = 4.0$	2.92	2.92	3.51	3.56	3.68	3.85	3.88
decay rates		0.226	0.155	0.153	0.151	0.097	
$c_a^2 = 4.0, c_s^2 = 0.5$	0.058	0.342	3.06	3.37	3.63	5.53	5.58
decay rates		2.417	0.286	0.260	0.243	0.165	

3.6. Extending the extremal models to unbounded support

The T-system theory and the Markov–Krein theorem extend to unbounded support intervals as shown by [10] and as indicated in [5] and [7]. 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 3.2, the extremal cdfs 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 3.2, the extremal cdfs 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 (22):

$$\begin{aligned} &\theta_W(F_L, G_L) \le \theta_W(F, G) \quad \text{for} \quad \mu_a \le \theta_W \le \mu_s < z^* \\ &\theta_W(F_L, G_L) \ge \theta_W(F, G) \quad \text{for} \quad \mu_s \le \theta_W \le \mu_a. \end{aligned} \tag{24}$$

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, Eq. (2) 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 [1].

Corollary 3.1 (*Extension to Unbounded Support*)**.** *Consider the setting of Theorem 3.3 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 decay rate $\theta_W(F, G)$ is well defined as the unique positive solution of (2). Moreover, if $\mu_a \le \theta_W$, then

$$\theta_W(F_L, G) \le \theta_W(F, G) \tag{25}$$

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

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

$$\theta_W(F, G_L) \ge \theta_W(F, G) \tag{26}$$

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

(c) For all (F, G) such that Assumption 2 holds, the decay rate $\theta_W(F, G)$ is well defined as the unique positive solution of (2) and (24) holds.

Acknowledgments

We thank Richard Zalik of Auburn University, Auburn, AL, USA, for assistance with T systems and NSF for support through CMMI 1634133.

References

[1] J. Abate, G.L. Choudhury, W. Whitt, Exponential approximations for tail probabilities in queues, I: waiting times, Oper. Res. 43 (5) (1995) 885–901.
[2] S. Asmussen, Applied Probability and Queues, second ed., Springer, New York, 2003.
[3] Y. Chen, W. Whitt, Algorithms for the upper bound mean waiting time in the $GI/GI/1$ queue, Queueing Syst. 94 (2020) 327–356.
[4] Y. Chen, W. Whitt, Set-valued performance approximations for the $GI/GI/K$ queue given partial information, Prob Eng. Inf. Sci. 34 (2020).

[5] A.E. Eckberg, Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems, Math. Oper. Res. 2 (2) (1977) 135–142.
[6] P.W. Glynn, W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, J. Appl. Probab. 31 (1994) 131–156.
[7] V. Gupta, T. Osogami, On Markov-Krein characterization of the mean waiting time in $M/G/K$ and other queueing systems, Queueing Syst. 68 (2011) 339–352.
[8] J.M. Holtzman, The accuracy of the equivalent random method with renewal inouts, Bell Syst. Tech. J. 52 (9) (1973) 1673–1679.
[9] M.A. Johnson, M.R. Taaffe, Tchebycheff systems for probability analysis, Amer. J. Math. Management Sci. 13 (1–2) (1993) 83–111.
[10] S. Karlin, W.J. Studden, Tchebycheff Systems; With Applications in Analysis and Statistics, Vol. 137, Wiley, New York, 1966.
[11] M.F. Neuts, Y. Takahashi, Asymptotic behavior of stationary distributions in the $GI/PH/C$ queue with heterogeneous servers, Z. Wahrscheinlichkeitstheor. Verwandte Geb. 57 (1981) 441–452.
[12] T. Rolski, Some inequalities for $GI/M/n$ queues, Zast. Mat. 13 (1) (1972) 43–47.
[13] T. Rolski, Order relations in the set of probability distribution functions and their applications in queueing theory, Dissertationes Math. Polish Acad. Sci. 132 (1) (1976) 3–47.
[14] L.P. Seelen, H.C. Tijms, M.H. van Hoorn, Tables for Multi-Server Queues, North-Holland, Amsterdam, 1985.
[15] J. Smith, Generalized Chebychev inequalities: Theory and application in decision analysis, Oper. Res. 43 (1995) 807–825.
[16] W. Whitt, Approximating a point process by a renewal process: two basic methods, Oper. Res. 30 (1982) 125–147.
[17] R.A. Zalik, Chebychev and weak Chebychev systems, in: M. Gasca, A.A. Miccelli (Eds.), Total Positivity and Its Applications, in: MAIA, vol. 359, Kluwer Academic Publishers, Dordrecht, 1996, pp. 301–332.
[18] R.A. Zalik, Some properties of Chebycheff systems, J. Comput. Anal. Appl. 13 (1) (2011) 20–26.