Extremal Models for the $GI/GI/K$ Waiting-Time Tail-Probability Decay Rate

Yan Chen and Ward Whitt

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

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.

Keywords: queues, tail probabilities, Tchebycheff systems, bounds

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 (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 focus on in 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, $\hat{f}(s) \equiv \int_0^\infty e^{-st}\, dF(t)$. In particular, the equation for the decay rate is

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

In this light-tailed setting, we show that the theory of Tchebycheff (T) systems from [2], as used in [3, 4, 5, 6], 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 [7] we provide evidence that this heuristic approach is effective for the steady-state mean $E[W]$. Here we start in §2 by giving background on T systems. In §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 §2.1 by reviewing the classical moment problem, as in [8]. 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 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:

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

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 [2].)

Theorem 2.1. (*a version of the classic moment problem, §2.1 of [8]*) *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*

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

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 $\sigma(P_U^*) \leq n+1$ and $\sigma(P_L^*) \leq n+1$.

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^*; for (4); see the seminal book [2] and the review papers [5, 9].

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 $2k \geq n+1$. When n is odd, we must have $\sigma(P) \geq (n+1)/2$. When n is even, we must have $\sigma(P) \geq 1+(n/2)$. The final story under the T-system assumption is different in these two cases. It is summarized in (5) of [4] and on p. 342 of [6].

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 [2]), 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 [2], 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 $\{u_i(t) : 0 \leq i \leq n\}$ 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 [2]) 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 [2]*) *In the setting of Theorem 2.1 extended by requiring that the moment vector is in the interior of the moment*

space, if $\{u_0, ..., u_n\}$ *and* $\{u_0, ..., u_n, \phi\}$ *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* 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*

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

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 [2]*) *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

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

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 [2] is somewhat complicated. Thus, we give an equivalent definition

Definition 4. (*extended T system, §I.2 of [2] and Theorem 1 of [10]*) *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 [2]) 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 [2], Theorem 5 and Corollary 1 of [9], and Theorem 29 of [5].

Theorem 2.3. (*Wronskians and ECT systems, p. 376 of [2]*) *Under the smoothness condition, the system of* $n+1$ *functions* $\{u_i : 0 \leq i \leq n\}$ *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*,* $0 \leq k \leq n$*.*

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* $u_i(t) \equiv t^i$, $0 \leq i \leq n$*, 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 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 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* $[0, M]$*:* $\mathcal{A}_1(m) \equiv \{t^k : 0 \le k \le m\}$, $\mathcal{A}_2 \equiv \{(-1)^{m+1}e^{-s_i t} : s_i > s_{i+1} > 0 \quad \text{for all} \quad i\}$ *and* $\mathcal{A}_3 \equiv \{e^{z_i t} : 0 < z_i < z_{i+1} \quad \text{for all} \quad i\}$*. 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 [2] or Example 6 of [9].

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 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 [2] and Example 6 of [9]. ∎

3. Extremal Models for the Asymptotic Decay Rate

In §3.1, we provide technical background on the decay rate for $K = 1$; in §3.2 we show that this approach also applies to the $GI/GI/K$ model for $K > 1$. In §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 §3.4 we

obtain more useful (as shown in [7]) three-point extremal distributions when we are also given the third moment and values of the Laplace transform of U and V.

3.1. Theory for the Asymptotic Decay Rate for $K = 1$

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 [11], defined by

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

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 $E[e^{sV}]$ 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 [11], large deviation limits in Corollary 1 in §1.2 of [12] and approximations in [1].

Part of the appeal of this approach is that it extends directly to $K > 1$, as we show in §3.2. Moreover, it has been observed that the approximation $P(W > t|W > 0) \approx e^{-\theta_W t}$ is quite good for $K \geq 1$; see [13]. 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, e.g, $\hat{f}(s) \equiv \int_0^\infty e^{-st}\, dF(t)$. Equivalently, as in §XIII.1 of [11], $\kappa_F(s) + \kappa_G(-s) = 0$, 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 [14] shows

that the distribution of W can be characterized by all complex roots of equations related to (2).

Given the simple structure in (2), 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.

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 (8) be well defined, as can be seen from Corollary 1 and Proposition 2 in [12] and Theorems 2.1, 5.5 and 5.3 in Chapter XIII in [11]. 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 $E[e^{tV}] \leq e^{tM_s}$ for all $t > 0$, so that $s^* = \infty$ 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 [11].) 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 $[0, \rho K M_s]$, the p^{th} moment of $E[V(K)^p] = K^p E[V^p]$ and the laplace transforms are related by $\hat{g}_{V(K)}(s) = \hat{g}_V(Ks)$. 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 $V^*(K) = KV^*$.

In [1], it was observed that the extension to $K > 1$ in (5) there was proved for the $GI/PH/K$ by [15]. 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 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*

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

Proof. Fix the interarrival-time cdf F and consider a sequence of phase-type service-time $\{G_n : n \geq 1\}$ 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 [15], (9) 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, (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 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 $1 + c^2 \leq b \leq M$.

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

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

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 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 (2) for the $F_u/G/1$ model (with service cdf G) such that $0 < \bar{s} < s^*$, where s^* is defined in Assumption 1,*

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

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 $\bar{s}$ 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 (2) to see that order for the Laplace transforms translates into order for θ_W, recalling that (i) (2) 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

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

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

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

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

To justify all the inequalities, we can apply the T-system theory working with bounded support sets, as in §2.2 and §2 of [4]. To treat F, we apply Lemma 2.2 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 2.2 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 [4]. ∎

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 [12]. 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 [16]*

3.4. 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 (2) 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 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 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, *where equation* (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 $\theta_W < \mu_s < s^*$, 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:

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

(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 §2.2.1 and (5) of [4]. 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 2.2. 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 $\{1, t, t^2, t^3, e^{-\mu_a t}, e^{-st}\}$. 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 (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 $\{1, t, t^2, t^3, e^{\mu_s t}, e^{st}\}$. 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 $\hat{g}(-s)$. Then combine the conclusions above.

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

Acknowledgement. We thank Richard Zalik of Auburn University 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, Operations Research 43 (5) (1995) 885–901.

[2] S. Karlin, W. J. Studden, Tchebycheff Systems; With Applications in Analysis and Statistics, Vol. 137, Wiley, New York, 1966.

[3] T. Rolski, Some inequalities for $GI/M/n$ queues, Zast. Mat. 13 (1) (1972) 43–47.

[4] A. E. Eckberg, Sharp bounds on Laplace-Stieltjes transforms, with applications to various queueing problems, Mathematics of Operations Research 2 (2) (1977) 135–142.

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

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

[7] Y. Chen, W. Whitt, Set-valued queueing approximations for the $GI/GI/K$ queue given partial information, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html (2020).

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

[9] R. A. Zalik, Chebychev and weak Chebychev systems, in: M. Gasca, A. A. Miccelli (Eds.), Total Positivity and its Applications, MAIA Volume 359, Kluwer Academic Publishers, Dordrecht, 1996, pp. 301–332.

[10] R. A. Zalik, Some properties of Chebycheff systems, J. Comput. Anal. Appl. 13 (1) (2011) 20–26.

[11] S. Asmussen, Applied Probability and Queues, 2nd Edition, Springer, New York, 2003.

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

[13] L. P. Seelen, H. C. Tijms, M. H. van Hoorn, Tables for Multi-Server Queues, North-Holland, Amsterdam, 1985.

[14] J. W. Cohen, The Single Server Queue, 2nd Edition, North-Holland, Amsterdam, 1982.

[15] M. F. Neuts, Y. Takahashi, Asymptotic behavior of stationary distributions in thee $GI/PH/C$ queue with hterogeneous servers, Z. Wahrscheinlichkeiteth. 57 (1981) 441–452.

[16] Y. Chen, W. Whitt, Algorithms for the upper bound mean waiting time in the $GI/GI/1$ queue, Queueing Systems 94 (2020) 327–356.