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

Extremal $GI/GI/1$ Queues Given Two Moments

Yan Chen
Industrial Engineering and Operations Research, Columbia University, yc3107@columbia.edu
Ward Whitt
Industrial Engineering and Operations Research, Columbia University, ww2040@columbia.edu

This paper studies tight upper and lower bounds for the mean (transient and steady-state) waiting time in the $GI/GI/1$ queue given the first two moments of the interarrival-time and service-time distributions. For distributions with nonempty compact support, we show that these bounds (with one distribution given and overall) are attained at extremal distributions with support on at most three points. The proof exploits theory for the moment problem and penalty functions in addition to standard stochastic theory for the model. We derive an alternative fixed-point characterization for the steady-state mean that is promising for deriving additional structure of the extremal distributions. We then apply relatively tractable numerical algorithms to identify the optimal distributions within the three-point distributions. For the overall upper bound with unbounded support sets, we propose a simple approximation formula and provide a numerical comparison of the approximations and bounds, showing that the new approximate bound is very accurate.

Key words: GI/GI/1 queue, tight bounds, extremal queues, bounds for the mean steady-state mean waiting time, moment problem

History: August 6, 2019

1. Introduction

In this paper we address a long-standing open problem for the classical $GI/GI/1$ queueing model: determining tight bounds for the mean steady-state waiting time, and the distributions that attain them, given the first two moments of the interarrival-time and service-time distributions; see Daley et al. (1992), especially §10, Wolff and Wang (2003) and references therein.

1.1. The $GI/GI/1$ Model

The $GI/GI/1$ single-server queue has unlimited waiting space and the first-come first-served service discipline. There is a sequence of independent and identically distributed (i.i.d.) service times $\{V_n : n \geq 0\}$, each distributed as V with cumulative distribution function (cdf) G, which is independent of a sequence of i.i.d. interarrival times $\{U_n : n \geq 0\}$ each distributed as U with cdf F. With the understanding that a 0^{th} customer arrives at time 0, V_n is the service time of customer n, while U_n is the interarrival time between customers n and $n+1$.

Let U have mean $E[U] \equiv 1$ and squared coefficient of variation (scv, variance divided by the square of the mean) c_a^2; let a service time V have mean $E[V] \equiv \tau \equiv \rho$ and scv c_s^2, where $\rho < 1$, so that the model is stable. (Let $\equiv$ denote equality by definition.)

Let W_n be the waiting time of customer n, i.e., the time from arrival until starting service, assuming that the system starts with an initial workload W_0 having cdf H_0 with a finite mean. The sequence $\{W_n : n \geq 0\}$ is well known to satisfy the Lindley recursion

$$W_n = [W_{n-1} + V_{n-1} - U_{n-1}]^+, \quad n \geq 1, \tag{1}$$

where $x^+ \equiv \max\{x, 0\}$. Let H_n be the cdf of W_n, which is determined by (1). Let $W \equiv W_\infty$ (both used) be the steady-state waiting time, satisfying $W_n \Rightarrow W_\infty$ as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution; see §§X.1-X.2 of Asmussen (2003). The cdf H_∞ of W_∞ is the unique cdf satisfying the stochastic fixed point equation

$$W_\infty \stackrel{\text{d}}{=} (W_\infty + V - U)^+, \tag{2}$$

where $\stackrel{\text{d}}{=}$ denotes equality in distribution. If $P(W_0 = 0) = 1$, then $W_n \stackrel{\text{d}}{=} \max\{S_k : 0 \leq k \leq n\}$ for $n \leq \infty$, $S_0 \equiv 0$, $S_k \equiv X_0 + \cdots + X_{k-1}$ and $X_k \equiv V_k - U_k$, $k \geq 1$. Under the specified finite moment conditions, for $1 \leq n \leq \infty$, W_n is a proper random variable with finite mean, given by

$$E[W_n | W_0 = 0] = \sum_{k=1}^{n} \frac{E[S_k^+]}{k} < \infty, \quad 1 \leq n < \infty, \quad \text{and} \quad E[W_\infty] = \sum_{k=1}^{\infty} \frac{E[S_k^+]}{k} < \infty. \tag{3}$$

1.2. Classical Steady-State Results: Exact, Approximate and Bounds

For the $M/GI/1$ special case, when the interarrival time has an exponential distribution, we have the classical Pollaczek-Khintchine formula

$$E[W] = \frac{\tau\rho(1+c_s^2)}{2(1-\rho)} = \frac{\rho^2(1+c_s^2)}{2(1-\rho)}. \tag{4}$$

A natural commonly used approximation for the $GI/GI/1$ model, inspired by (4), which we call the heavy-traffic approximation, because it is motivated by the early heavy-traffic limit in Kingman (1961), is

$$E[W] \equiv E[W(\rho, c_a^2, c_s^2)] \approx \frac{\rho^2(c_a^2+c_s^2)}{2(1-\rho)}. \tag{5}$$

The heavy traffic limit for the mean states that $(1-\rho)E[W(\rho, c_a^2, c_s^2)] \to (c_a^2+c_s^2)/2$ as $\rho \uparrow 1$.

The most familiar upper bound (UB) on $E[W]$ is the Kingman (1962) bound,

$$E[W] \le \frac{\rho^2([c_a^2/\rho^2]+c_s^2)}{2(1-\rho)}, \tag{6}$$

which also satisfies the same heavy traffic limit.

A better UB depending on these same parameters was obtained by Daley (1977). in particular, the Daley (1977) UB replaces the term c_a^2/ρ^2 by $(2-\rho)c_a^2/\rho$, i.e.,

$$E[W] \le \frac{\rho^2([(2-\rho)c_a^2/\rho]+c_s^2)}{2(1-\rho)}. \tag{7}$$

Note that $(2-\rho)/\rho < 1/\rho^2$ because $\rho(2-\rho) < 1$ for all ρ, $0 < \rho < 1$.

In contrast to the tight UB that we study, the tight lower bound (LB) for the steady-state mean has been known for a long time; see Stoyan and Stoyan (1974), §5.4 of Stoyan (1983), §V of Whitt (1984b), Theorem 3.1 of Daley et al. (1992) and references there. The LB is

$$E[W] \ge \frac{\rho((1+c_s^2)\rho-1)^+}{2(1-\rho)}. \tag{8}$$

The LB is attained asymptotically at a deterministic interarrival time with the specified mean and at any three-point service-time distribution that has all mass on nonnegative-integer multiples of the deterministic interarrival time. The service part follows from Ott (1987). (All service-time distributions satisfying these requirements yield the same mean.)

1.3. Motivation: Approximations for Non-Markovian Open Queueing Networks

One source of motivation for the bounds is provided by parametric-decomposition approximations for non-Markovian open networks of single-server queues, as in Whitt (1983b), where each queue is approximated by a $GI/GI/1$ queue partially characterized by the parameter vector $(\lambda, c_a^2, \tau, c_s^2)$, obtained by solving traffic rate equations for the arrival rate λ at each queue and after solving associated traffic variability equations to generate an approximating scv c_a^2 of the arrival process. Because the internal arrival processes are usually not renewal and the interarrival distribution is not known, there is no concrete $GI/GI/1$ model to analyze. To gain some insight into these approximations (not yet addressing the dependence among interarrival times), It is natural to regard such approximations for the $GI/GI/1$ model as set-valued functions, applying to all models with the same parameter vector $(\lambda, c_a^2, \tau, c_s^2)$.

For the special case of the $GI/M/1$ model with bounded support for the interarrival-time cdf F, the extremal $GI/M/1$ models were studied in Whitt (1984b), where intervals of bounded support were also used together with the theory of Tchebychev systems, as in Karlin and Studden (1966), drawing on Rolski (1972), Holtzman (1973) and Eckberg (1977).(The focus in Whitt (1984b) was on the mean steady state number of customers in the system, but it is easily seen that the extremal interarrival-time distributions are the same for the mean number of customers in the system and the mean steady-state waiting time, because they both depend on the root of the same equation.) For the $GI/M/1$ model, the extremal distributions are two-point distributions.

Let $\mathcal{P}_{2,2}(M) \equiv \mathcal{P}_{2,2}(m_1, c^2, M)$ be the set of all two-point distributions with mean m_1 and second moment $m_2 = m_1^2(c^2+1)$ with support in $[0, m_1 M]$. The set $\mathcal{P}_{2,2}(M)$ is a one-dimensional parametric family. Any element is determined by specifying one mass point. Let $F_b^{(2)}$ have probability mass $c^2/(c^2+(b-1)^2)$ on $m_1 b$, and mass $(b-1)^2/(c^2+(b-1)^2)$ on $m_1(1-c^2/(b-1))$ for $1+c^2 \le b \le M$. The cases $b = 1+c^2$ and $b = M$ constitute the two extremal distributions.

For $GI/M/1$, the interarrival-time cdf achieving the UB with mean m_1 and second moment $m_2 = m_1^2(c_a^2+1)$ with support in $[0, m_1 M_a]$, referred to here as $F_{1+c_a^2}^{(2)}$, arises for $b = 1+c_a^2$. In

particular, $F^{(2)}_{1+c_a^2}$ has probability mass $c_a^2/(1+c_a^2)$ on 0 and probability mass $1/(c_a^2+1)$ on $(m_2/m_1) = m_1(c_a^2+1)$.

The corresponding LB interarrival-time cdf, referred to here as $F^{(2)}_{M_a}$, arises for $b = M_a$. In particular, $F^{(2)}_{M_a}$ has probability mass $c_a^2/(c_a^2+(M_a-1)^2)$ on the upper bound of the support, $m_1 M_a$, and mass $(M_a-1)^2/(c_a^2+(M_a-1)^2)$ on $m_1(1-c_a^2/(M_a-1))$. (For the interarrival time, we scale, i.e., choose measuring units for time, so that $m_1 = 1$.) We use the notation $G^{(2)}_{1+c_s^2}$ and $G^{(2)}_{M_s}$ for the corresponding service-time cdf's G with mean ρ and support $[0, \rho M_s]$.

Since the range of possible values is quite large, while the distributions that attain the bounds are unusual (two-point distributions), the papers Klincewicz and Whitt (1984), Whitt (1984c) and Johnson and Taaffe (1990a) focused on reducing the range by imposing shape constraints. In this paper we do not consider shape constraints.

1.4. Related Literature

The literature on bounds for the $GI/GI/1$ queue is well reviewed in Daley et al. (1992) and Wolff and Wang (2003), so we will be brief. The use of optimization to study the bounding problem for queues seems to have begun with Klincewicz and Whitt (1984) and Johnson and Taaffe (1990b). Bertsimas and Natarajan (2007) provides a tractable semi-definite program as a relaxation model for solving steady-state waiting time of $GI/GI/c$ to derive bounds, while Osogami and Raymond (2013) bounds the transient tail probability of $GI/GI/1$ by a semi-definite program.

Several researchers have studied bounds for the more complex many-server queue. In addition to Bertsimas and Natarajan (2007), Gupta et al. (2010) and Gupta and Osogami (2011) investigate the bounds and approximations of the $M/GI/c$ queue. Gupta et al. (2010) explains why two moment information is insufficient for good accuracy of steady-state approximations of $M/GI/c$. Gupta and Osogami (2011) establishes a tight bound for the $M/GI/K$ in light traffic. Finally, Li and Goldberg (2017) establishes bounds for $GI/GI/c$ intended for the many-server heavy-traffic regime.

1.5. Organization

In §2 we state our main result, Theorem 1, which shows that there exist extremal interarrival-time and service-time cdf's that have support on at most three points when the interarrival-time and service-time cdf's have compact support. In §3 we prove Theorem 1, drawing on the theory for the moment problem as in Smith (1995) plus optimization using a sequence of penalty functions. In §4 we derive a fixed-point characterization of the extremal distributions for the steady-state mean. In §5 we exploit Theorem 1 to develop a multinomial optimization that supports the conjecture that the overall upper bound is attained by two-point distributions, in particular, by the conjectured $F^{(2)}_{1+c_a^2}/G^{(2)}_{M_s}$ model. In §6 we do a simulation study for two-point distributions to expose the form of the upper bound for the mean. In §7 we develop a candidate three-point service cdf for the overall lower bound with finite support. In §8 we present some concluding discussion. We present additional supporting material in the e-companion, starting with a summary of notation in §EC.2.

2. Reduction to Three-Point Distributions

In this section we show that it suffices to consider interarrival-time and service-time cdfs with support on at most three points in our search for bounds on the transient and steady-state mean waiting time $E[W_n]$ for $1 \le n \le \infty$.

Let $\mathcal{P}_n$ be the set of all probability measures on a subset of the positive real line $[0,\infty)$ with specified first n moments. The set $\mathcal{P}_n$ is a convex set, because the convex combination of two probability measures is just the mixture; i.e., for all p, $0 \le p \le 1$, $pP_1 + (1-p)P_2 \in \mathcal{P}_n$ if $P_1 \in \mathcal{P}_n$ and $P_2 \in \mathcal{P}_n$, because the n^{th} moment of the mixture is the mixture of the n^{th} moments, which is just the common value of the components. let $\mathcal{P}_{n,k}$ be the subset of probability measures in $\mathcal{P}_n$ that have support on at most k points.

We use the scv to parameterize, so let $\mathcal{P}_2 \equiv \mathcal{P}_2(m,c^2)$ be the set of all cdf's with mean m and second moment $m^2(c^2+1)$ where $c^2 < \infty$. Let $\mathcal{P}_2(M) \equiv \mathcal{P}_2(m,c^2,M)$ be the subset of all cdf's in $\mathcal{P}_2$ with support in the closed interval $[0,mM]$ having mean m and second moment $m^2(c^2+1)$ where $c^2+1 < M < \infty$. (The last property ensures that the set $\mathcal{P}_2(M)$ is non-empty.) Let subscripts a

and s denote sets for the inter$a$rrival and $s$ervice times, respectively. Let $\mathcal{P}_2^{(c)}(M) \equiv \mathcal{P}_2^{(c)}(m, c^2, M)$ be the subset of $\mathcal{P}_2(m, c^2)$ with support in a compact subset of $[0, M]$, denoted by $\mathcal{C}$, assumed given, fixed and nonempty. (We are usually interested in $\mathcal{C} = [0, M]$, but finite support may be of interest; e.g., that is used in the proof of Theorem 4 here.) Therefore, $\mathcal{P}_{a,2}^{(c)}(M_a)$ is the set of all interarrival-time cdf's F with mean 1, scv c_a^2 and compact support within $[0, M_a]$, while $\mathcal{P}_{s,2,n}^{(c)}(M_s)$ is the set of all service-time cdf's G with mean ρ, scv c_s^2 and compact support within $[0, \rho M_s]$. Then $\mathcal{P}_{a,2,n}^{(c)}(M_a)$ and $\mathcal{P}_{s,2,n}^{(c)}(M_s)$ are the subsets with support on n points.

We are interested in the maps

$$w_n : \mathcal{P}_{a,2}(1, c_a^2) \times \mathcal{P}_{s,2}(\rho, c_s^2) \to \mathbb{R}, \tag{9}$$

where $0 < \rho < 1$ and

$$w_n(F, G) \equiv E[W_n(F, G)], \quad 1 \le n \le \infty, \tag{10}$$

for $W_n \equiv W_n(F, G)]$ in the $GI/GI/1$ queue with interarrival-time cdf $F \in \mathcal{P}_{a,2}$ and service-time cdf $G \in \mathcal{P}_{s,2}$. For $n < \infty$, the distribution of $W_n(F, G)$ also depends on an initial cdf H_0 of W_0, but it is fixed here.

THEOREM 1. (*reduction to a three-point distribution*) *Consider the class of $GI/GI/1$ queues with $E[W_0] < \infty$, $F \in \mathcal{P}_{a,2}(1, c_a^2)$, $G \in \mathcal{P}_{s,2}(\rho, c_s^2)$, $0 < \rho < 1$, where $\mathcal{P}_{a,2}$ and $\mathcal{P}_{s,2}$ are nonempty. For $1 \le n \le \infty$, the functions $w_n : \mathcal{P}_{a,2} \times \mathcal{P}_{s,2} \to \mathbb{R}$ in (9) are continuous. Hence, the following suprema over spaces of probability measures with specified nonempty compact support are attained.*

(*a*) *For each n, $G \in \mathcal{P}_{s,2}$ and $1 + c_a^2 \le M_a < \infty$, there exists $F_n^*(G) \in \mathcal{P}_{a,2,3}^{(c)}(M_a)$ such that*

$$w_{a,n}^{\uparrow}(G) \equiv \sup\{w_n(F, G) : F \in \mathcal{P}_{a,2}^{(c)}(M_a)\} = \sup\{w_n(F, G) : F \in \mathcal{P}_{a,2,3}^{(c)}(M_a)\} = w_n(F_n^*(G), G). \tag{11}$$

(*b*) *For each n, $F \in \mathcal{P}_{a,2}$ and $1 + c_s^2 \le M_s < \infty$, there exists $G_n^*(F) \in \mathcal{P}_{s,2,3}^{(c)}(M_s)$ such that*

$$w_{s,n}^{\uparrow}(F) \equiv \sup\{w_n(F, G) : G \in \mathcal{P}_{s,2}^{(c)}(M_s)\} = \sup\{w_n(F, G) : G \in \mathcal{P}_{s,2,3}^{(c)}(M_s)\} = w_n(F, G_n^*(F)). \tag{12}$$

(c) *For each* n, (M_a, M_s) *with* $1+c_a^2 \le M_a < \infty$ *and* $1+c_s^2 \le M_s < \infty$, *there exists* (F_n^{**}, G_n^{**}) *in* $\mathcal{P}_{a,2,3}^{(c)}(M_a) \times \mathcal{P}_{s,2,3}^{(c)}(M_s)$ *such that*

$$w_n^{\uparrow} \equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2}^{(c)}(M_a), G \in \mathcal{P}_{s,2}^{(c)}(M_s)\} = \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2,3}^{(c)}(M_a), G \in \mathcal{P}_{s,2,3}^{(c)}(M_s)\}$$
$$= w_n(F_n^{**}, G_n^{**}) = w_{a,n}^{\uparrow}(G_n^{**}) = w_{s,n}^{\uparrow}(F_n^{**}). \tag{13}$$

Corresponding results hold for each supremum replaced by an infimum.

REMARK 1. (uniqueness) There is no claim of uniqueness in Theorem 1. Indeed, the $M/GI/1$ steady-state formula in (4) implies that there is no uniqueness in case (b) when F is exponential.

REMARK 2. (extensions) For the transient model, the result remains valid for the mean $E[W_n]$ replaced by $E[(f(W_n)]$ for any continuous bounded real-valued function f. The result also remains valid for the mean $E[W_n]$ replaced by a higher moment $E[W_n^k]$, $k \ge 1$, provided we use the spaces $\mathcal{P}_{a,k+1}$ and $\mathcal{P}_{s,k+1}$, using §X.2 of Asmussen (2003). The result also extends to the time-varying model in which the cdf's F_k and G_k depend on k. Indeed, our proof exploits that model.

3. Proof of Theorem 1

Since the proof draws on results for the moment problem, we first review that.

3.1. The Moment Problem for Distributions with Compact Support

Our problem can be approached via the classical theory for the moment problem, as in Lasserre (2010), Smith (1995) and references therein. Some simplification can be gained by considering continuous functions on a compact metric space domain, so that suprema and infima are attained.

For the general moment problem, let $\mathcal{P}_n(\mathcal{C})$ be the set of all probability measures on a compact subset $\mathcal{C}$ of $\mathbb{R}$ with specified first n moments, where the k^{th} moment of P is defined as $\int x^k \, dP$. Assume that $\mathcal{P}_n(\mathcal{C})$ is not empty and let

$sP_n(\mathcal{C})$ be endowed with the topology of weak convergence, as determined by the Prohorov or Lévy metric, as in §3.2 and §11.3 of Whitt (2002). let $\mathcal{P}_{n,k}(\mathcal{C})$ be the subset of probability measures in $\mathcal{P}_n(\mathcal{C})$ that have support on at most k points in $\mathcal{C}$.

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.

THEOREM 2. (*a version of the classic moment problem*) *Let* $\phi : \mathcal{C} \to \mathbb{R}$ *be a continuous function, where* $\mathcal{C}$ *is a compact subset of* $\mathbb{R}$. *Assume that* $\mathcal{P}_n(\mathcal{C})$ *is not empty. Then there exists* $P^* \in \mathcal{P}_{n,n+1}(\mathcal{C})$ *such that*

$$\sup\{\int_0^M \phi\, dP : P \in \mathcal{P}_n(\mathcal{C})\} = \sup\{\int_0^M \phi\, dP : P \in \mathcal{P}_{n,n+1}(\mathcal{C})\} = \sum_{k=1}^{n+1} \phi(t_k) P^*(\{t_k\}), \tag{14}$$

where $\{t_k : 1 \le k \le n+1\}$ *is the support of* P^*.

.

Proof. First, because the support $\mathcal{C}$ is a compact subset of $\mathbb{R}$ and the set $\mathcal{P}_n(\mathcal{C})$ is not empty by assumption, the space $\mathcal{P}_n(\mathcal{C})$ is a compact metric space with the usual topology of convergence in distribution, as a consequence of Prohorov's theorem; e.g., Theorem 11.6.1 of Whitt (2002). (In general, the set of all probability measures on a compact metric space with the usual topology of weak convergence is itself a compact metric space; see Theorem II.6.4 of Parthasarathy (1967).)

Second, because the function ϕ is continuous, we can apply the continuous mapping theorem as in §3.4 of Whitt (2002) to deduce that the induced map $\phi : \mathcal{P}_n(\mathcal{C}) \to \mathbb{R}$ defined by

$$\phi(P) \equiv \int_0^b \phi\, dP \tag{15}$$

is continuous as well. Hence, the induced map in (15) is a continuous bounded real-valued function on a compact metric space, so that the supremum in (14) is attained. Then the theory for the classical moment problem implies that it is attained in $\mathcal{P}_{n,n+1}(\mathcal{C})$; see §2 of Smith (1995). ∎

REMARK 3. (linear program for finite support) Note that the optimization in (14) reduces to an ordinary linear program (LP) if the compact set $\mathcal{C}$ is a finite set. The decision variables are the probability masses on the specified support set, while the constraints are the specified moments as well as the requirement that the sum of the probability masses is 1.

3.2. The Time-Varying Model and the Lindley Recursion

We start by focusing on part (a). To do so, we start by considering the time-varying $GI_{(k)}/GI/1$ model in which the cdf's F_k of U_k are allowed to depend on k. Afterwards, we force common cdf's

in the limit through a sequence of penalty functions. We work with the Lindley recursion in (1), which remains valid with this generalization. The following is the key lemma. It follows directly from Theorem 2.

LEMMA 1. *For any random variable Y with finite mean, $E[(Y-u)^+]$ is a continuous bounded function of u in $[0,\infty)$. Thus, the supremum*

$$\sup\{\int_0^M E[(Y-u)^+]\,dF(u) : F \in \mathcal{P}_2^{(c)}(m_1,c^2,M)\} \tag{16}$$

is attained by a cdf F^* in $\mathcal{P}_{2,3}^{(c)}(m_1,c^2,M)$.

Lemma 1 immediately applies to show that, for any given H_0 and G, $E[W_1] \equiv E[W_1(H_0,F_0,G)]$ is maximized over $F_0 \in \mathcal{P}_{a,2}^{(c)}(M_a)$ by a cdf $F_0^* \in \mathcal{P}_{a,2,3}^{(c)}(M_a)$.

COROLLARY 1. *Consider the setting of Theorem 1. For any three independent random variable W_0 with cdf H_0 and finite mean, V with cdf $G \in \mathcal{P}_{s,2}$ and U with cdf $F_0 \in \mathcal{P}_{a,2}^{(c)}(M_a)$,*

$$\begin{aligned}\sup\{E[W_1(H_0,F_0,G)] : F_0 \in \mathcal{P}_{a,2}^{(c)}(M_a)\} &\equiv \sup\{E[(W_0+V(G)-U(F_0))^+] : F_0 \in \mathcal{P}_{a,2}^{(c)}\}(M_a)\}, \\ &= \sup\{\int_0^{M_a} E[(W_0+V-u)^+]\,dF_0(u) : F_0 \in \mathcal{P}_{a,2}^{(c)}(M_a)\}.\end{aligned} \tag{17}$$

Proof. Apply Lemma 1 with $Y \equiv W_0 + V$. ∎

Lemma 1 also applies to show that, for any given H_0 and F, $E[W_1(H_0,F,G_0)]$ is maximized over $G \in \mathcal{P}_{s,2}^{(c)}(M_s)$ by a cdf $G_0^* \in \mathcal{P}_{s,2,3}^{(c)}(M_s)$ if we use a reverse-time perspective and look at $\rho M_s - V$. (Recall that the support of V is in $[0,\rho M_s]$.)

3.3. A Penalty Function Method for the Transient Mean

We continue focusing on part (a). We do the proof in detail for $n=2$; general finite n follows by the same argument, as we will explain. Part (b) follows by the same argument applied to $\rho M_s - V$. Part (c) follows from by combining parts (a) and (b); e.g., see Lemma EC.1 in §EC.7 of Whitt and You (2018).

For $n=2$, we introduce some notation. Let $W_2(F_0,F_1) \equiv W_2(H_0,F_0,G,F_1)$ be the waiting time in period 2 as a function of the defining cdf's, in particular, with H_0 being the initial cdf of W_0

having finite mean and G is the common cdf of the service times V_k, assumed to be in $\mathcal{P}_{s,2}$, while F_0 and F_1 are the cdf's of the independent random variables U_0 and U_1, assumed to be in $\mathcal{P}_{a,2}^{(c)}(M_a)$, but without any constraint that $F_0 = F_1$. We use the shorter form because we regard H_0 and G as fixed.

With that notation specified, let

$$\hat{W}_2(F_0, F_1) = W_2(F_0, F_1) - d_M(F_0, F_1), \tag{18}$$

where $d_M(F_0, F_1)$ is a convenient penalty function with the property that $d_M(F_0, F_1) \to \infty$ as $M \to \infty$ if and only if $F_0 \neq F_1$. We will choose a penalty function that will allow us to apply Theorem 2.

To construct a convenient penalty function, we exploit integral probability metrics, as in Example 3.3.6 on p. 50 and §4.4 on p. 89 of Rachev et al. (2013), which draws on Zolotarev (1976, 1983). A probability metric is a semi-metric (does not require the triangle inequality) on a space of probability measures. The space of probability measures we consider are compact subsets of the compact metric space $\mathcal{P}_2(m_1, c^2, M)$ and so themselves are compact metric spaces, so that they have all the desired regularity properties.

In particular, we use the integral probability metric

$$d(F_0, F_1) \equiv \sup\left\{\int_0^{M_a} h(x)\, dF_0(x) - \int_0^{M_a} h(x)\, dF_1(x) : h \in \mathcal{H}\right\}, \tag{19}$$

where $\mathcal{H}$ is a determining class of continuous real-valued functions on $[0, M_a]$ with Lipschitz constant

$$\mathrm{Lip}(h) \equiv \sup\{|h(x) - h(y)|/|x - y| : x \neq y, x, y \in [0, M_a]\} \leq 1. \tag{20}$$

The distance is the dual representation of a Wasserstein distance between F_0 and F_1; see Kantorovich and Rubinstein (1958), Kemperman (1983) and §5.4 of Rachev et al. (2013).

We now return to the main argument. For $n = 2$ in part (a), we do an optimization of the mean $E[\hat{W}_2(F_0, F_1)]$ for $\hat{W}_2(F_0, F_1)$ in (18) over the pair (F_0, F_1) without any direct constraints that $F_0 = F_1$. However, afterwards we let $M \to \infty$, which will force $F_0 = F_1$.

We first construct a deterministic function of the deterministic variables. Let u_0 and u_1 be the first two interarrival times; let v_0 and v_1 be the first two service times; and let w_0 be the initial waiting time. Note that the sample paths satisfy (w.p.1) the continuous relations

$$\begin{aligned} W_1(w_0,u_0,v_0) &= \max\{w_0+v_0-u_0,0\} \quad \text{and} \\ W_2(w_0,u_0,u_1,v_0,v_1) &= \max\{(W_1(w_0,u_0,v_0)+v_1-u_1,0\} \\ &= \max\{\max\{w_0+v_0-u_0,0\}+v_1-u_1,0\}. \end{aligned} \tag{21}$$

Hence, $W_2(w_0,u_0,u_1,v_0,v_1)$ is a continuous function of the five variables. Then write

$$E[\hat{W}_2(F_0,F_1)] = \int_0^{M_a}\int_0^{M_a} \phi(u_0,u_1)\,dF_0(u_0)\,dF_1(u_1) - d_M(F_0,F_1), \tag{22}$$

where $d_M(F_0,F_1) \equiv Md(F_0,F_1)$ for d in (19) and

$$\phi(u_0,u_1) \equiv E[W_2(W_0,u_0,u_1,V_0,V_1)] \tag{23}$$

with

$$E[W_2(W_0,u_0,u_1,V_0,V_1)] = \int_0^{\infty}\int_0^{\infty}\int_0^{\infty} W_2(w_0,u_0,u_1,v_0,v_1)\,dH_0(w_0)\,dG(v_0)\,dG(v_1). \tag{24}$$

Then we have

$$\begin{aligned} w_2^{\uparrow} &= \sup\{E[\hat{W}_2(F_0,F_1)] : F_0 \in \mathcal{P}_{a,2}^{(c)}(M_a), F_1 \in \mathcal{P}_{a,2}^{(c)}(M_a)\} \\ &= \sup\{\int_0^{M_a}\int_0^{M_a} \phi(u_0,u_1)\,dF_0(u_0)\,dF_1(u_1) - d_M(F_0,F_1) : F_0 \in \mathcal{P}_{a,2}^{(c)}(M_a), F_1 \in \mathcal{P}_{a,2}^{(c)}(M_a)\} \end{aligned} \tag{25}$$

Since

$$-d_M(F_0,F_1) = M\inf\left\{\int_0^{M_a} h(x)\,dF_0(x) - \int_0^{M_a} h(x)\,dF_1(x) : h \in \mathcal{H}\right\}, \tag{26}$$

we can rewrite (25) as

$$\sup_{F_0,F_1}\inf_{h\in\mathcal{H}}\left\{\int_0^{M_a}\int_0^{M_a} \phi(u_0,u_1)\,dF_0(u_0)\,dF_1.(u_1) + M\int_0^{M_a} h(x)\,dF_0(x) - M\int_0^{M_a} h(x)\,dF_1(x)\right\} \tag{27}$$

We now apply the minimax theorem to interchange the order of the infimum and supremum in (27) to obtain the alternative version

$$\inf_{h} \sup_{F_0,F_1} \left\{ \int_0^{M_a} \int_0^{M_a} \phi(u_0,u_1)\, dF_0(u_0)\, dF_1(u_1) + M \int_0^{M_a} h(x)\, dF_0(x) - M \int_0^{M_a} h(x)\, dF_1(x) \right\}, \quad (28)$$

where the inner supremum is over $(F_0,F_1) \in \mathcal{P}^{(c)}_{a,2}(M_a) \times \mathcal{P}^{(c)}_{a,2}(M_a)$ and the outer infimum is over $h \in \mathcal{H}$.

We see that the conditions of generalized minimax theorem, extending the classic Von Neumann (1928), are satisfied because the sets $\mathcal{P}^{(c)}_{a,2}(M_a), \mathcal{P}^{(c)}_{s,2}(M_s)$ and $\mathcal{H}$ are compact convex sets, while the objective function is a continuous linear function of F_0, F_1 and h.

Hence we can apply Theorem 2 twice to the inner supremum to deduce that the overall optimum is attained at $(F^*_0(M), F^*_1(M)) \in \mathcal{P}^{(c)}_{a,2,3}(M_a) \times \mathcal{P}^{(c)}_{a,2,3}(M_a)$. We use the fact that the functions to be integrated are continuous in each case. The infimum is attained in the set because it is for a continuous function over a compact metric space.

Finally, we obtain our desired result by letting $M \to \infty$. When we consider a sequence of models in which $M \to \infty$, we attain a sequence of optima, but since this sequence lies in the compact space $\mathcal{P}^{(c)}_{a,2,3}(M_a) \times \mathcal{P}^{(c)}_{a,2,3}(M_a)$, it is tight. Hence, by Prohorov's theorem, there must exist a convergent subsequence; e.g., see §11.6 of Whitt (2002). The limit of any such convergent subsequence is the desired optimum (F^*_0, F^*_1), which must satisfy $F^*_0 = F^*_1 \in \mathcal{P}^{(c)}_{a,2,3}(M_a)$. We draw that final conclusion because $d_M(F_0,F_1) \to \infty$ as $M \to \infty$ whenever $F_0 \neq F_1$. (We use the fact that $E[W_2(F_0,F_1)]$ is bounded by $E[W_0] + 2E[V]$ uniformly in $(F_0,F_1) \in \mathcal{P}^{(c)}_{a,2}(M_a) \times \mathcal{P}^{(c)}_{a,2}(M_a)$.)

To demonstrate optimality, let (F',F') be any alternative $F' \in \mathcal{P}^{(c)}_{a,2}(M_a)$ for the limiting system, which must have the two cdf's coincide. For each M_k in the converging subsequence of optima with associated penalties $\{M_k : k \geq 1\}$, this is a candidate solution, which is dominated by the optimum $(F^*_0(M_k), F^*_1(M_k))$ for each M_k. Since that subsequence converges, that dominance will extend to the limit by continuity.

This same argument applies to any finite n. Instead of (18), we can write

$$\hat{W}_n(F_0,F_1,\ldots,F_{n-1}) = W_n(F_0,F_1,\ldots,F_{n-1}) - \sum_{i=1}^{n-1} d_M(F_{i-1},F_i), \quad (29)$$

In the next subsection we indicate how to get the corresponding result for the steady-state mean $E[W_\infty]$. ∎

REMARK 4. (penalty functions in optimization) Penalty functions are often used in optimization, but they apply in a relatively simple way in the present context, because with the penalty functions, we get a relatively simple iteration of the one-period optimizations without additional constraints, for which Theorem 2 for the moment problem applies directly.

3.4. From the Transient Mean to the Steady-State Mean

We now show that it suffices to consider the transient mean $E[W_n]$ for the three-point distributions and finite n in order to treat $E[W_\infty]$. We can apply the following lemma.

LEMMA 2. (*reduction to the transient mean*) *Consider the $GI/GI/1$ queues in Theorem 1.*

(*a*) *For any specified $G \in \mathcal{P}_{s,2}$, if there exists $F_n^* \in \mathcal{P}_{a,2,3}^{(c)}(M_a)$ such that*

$$w_n(F_n^*, G) = w_{a,n}^{\uparrow}(G) \equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2}^{(c)}(M_a)\} \quad \textit{for all} \quad n \geq 1, \tag{30}$$

then the sequence $\{F_n^* : n \geq 1\}$ is tight, so that there exists a convergent subsequence. Moreover, if F_∞^* is the limit of any convergent subsequence, then F_∞^* is in $\mathcal{P}_{a,2,3}^{(c)}(M_a)$ and F_∞^* is optimal for $E[W_\infty(F,G)]$, i.e., $w_{a,\infty}^{\uparrow}(G) = w(F_\infty^*, G)$ for the steady-state mean.

(*b*) *For any specified $F \in \mathcal{P}_{a,2}$, if there exists $G_n^* \in \mathcal{P}_{s,2,3}^{(c)}(M_s)$ such that*

$$w_n(F, G_n^*) = w_{s,n}^{\uparrow}(F) \equiv \sup\{w_n(F,G) : G \in \mathcal{P}_{s,2}^{(c)}(M_s)\} \quad \textit{for all} \quad n \geq 1, \tag{31}$$

then the sequence $\{G_n^* : n \geq 1\}$ is tight, so that there exists a convergent subsequence. Moreover, if G_∞^* is the limit of any convergent subsequence, then G_∞^* is in $\mathcal{P}_{s,2,3}^{(c)}(M_s)$ and G_∞^* is optimal for $E[W_\infty(F,G)]$, i.e., $w_{s,\infty}^{\uparrow}(F) = w(F, G_\infty^*)$ for the steady-state mean.

(*c*) *If there exists (F_n^*, G_n^*) in $\mathcal{P}_{a,2,3}^{(c)}(M_a) \times \mathcal{P}_{s,2,3}^{(c)}(M_s)$ such that*

$$w_n(F_n^*, G_n^*) = w_n^{\uparrow} \equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2}^{(c)}(M_a), G \in \mathcal{P}_{s,2}^{(c)}(M_s)\} \quad \textit{for all} \quad n \geq 1, \tag{32}$$

then the sequence $\{(F_n^*, G_n^*) : n \geq 1\}$ *is tight, so that there exists a convergent subsequence. Moreover, if* (F_∞^*, G_∞^*) *is the limit of any convergent subsequence, then* (F_∞^*, G_∞^*) *is in* $\mathcal{P}_{a,2,3}^{(c)}(M_a) \times \mathcal{P}_{s,2,3}^{(c)}(M_s)$ *and the pair* (F_∞^*, G_∞^*) *is optimal for* $E[W_\infty]$*, i.e.,* $w_\infty^\uparrow = w_\infty(F_\infty^*, G_\infty^*)$ *for the steady-state mean.*

Moreover, corresponding results hold for supremum replaced by imfimum.

Proof of Lemma 2. We only prove (c), because the others are proved in the same way. As observed before, because the support sets $[0, M_a]$ and $[0, \rho M_s]$ are compact intervals, the spaces $\mathcal{P}_{a,2}^{(c)}(M_a)$, $\mathcal{P}_{s,2}^{(c)}(M_s)$ and their product are compact metric spaces, as are the spaces $\mathcal{P}_{a,2,3}^{(c)}(M_a)$, $\mathcal{P}_{s,2,3}^{(c)}(M_s)$ and their product, because they are closed subsets. Hence the tightness follows, which implies that there exists a convergent subsequence by Prohorov's theorem in §11.6 of Whitt (2002) and the limit (F^*, G^*) of any such subsequence $\{(F_{n_k}^*, G_{n_k}^*) : k \geq 1\}$ must remain in the space $\mathcal{P}_{a,2,3}^{(c)}(M_a) \times \mathcal{P}_{s,2,3}^{(c)}(M_s)$. Moreover, the associated sequence of steady-state waiting times $\{W_\infty(F_{n_k}^*, G_{n_k}^*) : k \geq 1\}$ converges in distribution to $W_\infty(F^*, G^*)$ and the means converge as well, by the continuity results for $GI/GI/1$ in §X.6 of Asmussen (2003).

We conclude by demonstrating optimality. Suppose that (F', G') is another candidate pair of cdf's in $\mathcal{P}_{a,2,3}^{(c)}(M_a) \times \mathcal{P}_{s,2,3}^{(c)}(M_s)$. By the assumed optimality, we must have $w_{n_k}(F_{n_k}^*, G_{n_k}^*) \geq w_{n_k}(F', G')$ for all k. Then, by continuity, using §X.6 of Asmussen (2003), we conclude that $w_\infty^\uparrow = w(F^*, G^*)$ for the steady-state mean. ∎

By the same reasoning, an analog of Lemma 2 holds for two-point distributions. In this case, we assume that the support is the full interval $[0, m_1 M]$.

Corollary 2. *In the setting of Lemma 2, (i) if* $F_n^* \in \mathcal{P}_{a,2,2}(M_a)$ *for all* n *in (a), then* $F_\infty^* \in \mathcal{P}_{a,2,2}(M_a)$*; if* $G_n^* \in \mathcal{P}_{s,2,2}(M_s)$ *for all* n *in (b), then* $G_\infty^* \in \mathcal{P}_{s,2,2}(M_s)$*; if* $(F_n^*, G_n^*) \in \mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$ *for all* n *in (c), then* $(F_\infty^*, G_\infty^*) \in \mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$.

Proof. The same argument applies because $\mathcal{P}_{2,2}(M)$ is a closed subset of $\mathcal{P}_{2,3}(M)$. ∎

4. A Fixed-Point Characterization for the Steady-State Mean

As a basis for further analysis of the extremal distributions for the steady-state waiting time, we develop a fixed-point characterization of the extremal distributions for the steady-state mean. We will be working with one step of the Lindley recursion (1). We obtain a fixed-point equation because, in the case (a), we impose two conditions: (i) the initial cdf H_0 of W_0 is the steady-state cdf of W_∞ for an interarrival-time cdf $\hat{F}$ and (ii) the optimizing cdf F_0^* coincides with $\hat{F}$.

THEOREM 3. (*fixed-point characterization*) *For case* (*a*) *in Theorem 1, let* $G \in \mathcal{P}_{s,2}$ *be given. If there exists a cdf* $\hat{F} \equiv \hat{F}(G) \in \mathcal{P}_{a,2}^{(c)}(M_a)$ *such that, for* $H_0 \stackrel{\rm d}{=} W_\infty(\hat{F}, G)$, *the optimal mean in the first period is attained by* $\hat{F}$ *and equals the initial mean, i.e., if*

$$E[W_1(H_0, F_0^*, G)] \equiv \sup\{E[W_1(H_0, F_0, G)] : F_0 \in \mathcal{P}_{a,2}^{(c)}(M_a)\} = E[W_1(H_0, \hat{F}, G)], \tag{33}$$

then

$$w_{a,n}^{\uparrow}(H_0, G) = w_{a,n}(H_0, \hat{F}, G) \quad \textit{for all} \quad n \geq 1, \tag{34}$$

so that

$$w_{a,\infty}(\hat{F}, G) = w_{a,\infty}^{\uparrow}(G) = \sup\{E[W_\infty(F, G)] : F \in \mathcal{P}_{a,2}^{(c)}(M_a)\}. \tag{35}$$

In addition, if there is a unique optimum $F_0^*(\hat{F})$ *in the optimization* (33) *for* $\hat{F}$, *then* $\hat{F} \in \mathcal{P}_{a,2,3}^{(c)}$. *The analog holds for* (*b*). *Both results hold for supremum replaced by infimum.*

Proof. We focus on part (a); the same argument applies to G by focusing on $\rho M_s - V$. As in the proof of Theorem 1, we consider the generalization to the $GI_{(k)}/GI/1$ model in which the cdf's F_k are allowed to vary with k. We exploit the Markov property and the Lindley recursion (1) in this more general setting to reduce the problem to a one-period problem.

To establish the "if" claim, note that the condition that $F_0^* = \hat{F}$ in (33) implies that the cdf H_0 of $W(\hat{F}, G)$ satisfies the stochastic fixed point equation for the steady-state waiting time in (2) for interarrival-time cdf $\hat{F}$ and service-time cdf G; i.e.,

$$H_1(H_0, F_0^*, G) \equiv H_1(H_0, \hat{F}, G) = H_0 \stackrel{\rm d}{=} W_\infty(\hat{F}, G). \tag{36}$$

The Markov property implies that W_2 depends on F_0 only through H_1.

Hence, under condition (33), for the generalized $GI_{(k)}/GI/1$ model, the optimization problem over F_1 for the second period, given the first period, repeats the initial one-period optimization problem, unconstrained by the choice of F_0. Thus, by mathematical induction, under the given conditions, we have

$$F_n^* = F_0^* = \hat{F} \quad \text{for all} \quad n \geq 1 \quad \text{in the } GI_{(k)}/GI/1 \text{ model.} \tag{37}$$

Hence, the conclusion remains valid for the original $GI/GI/1$ model. To elaborate, consider $n = 2$. For any alternative cdf F, (F, F) is an alternative for $n = 2$ in the $GI_{(k)}/GI/1$ model. However, by our proof, $(\hat{F}, \hat{F})$ dominates (F, F) in the $GI_{(k)}/GI/1$ model, and so in the $GI/GI/1$ model.

Next, by Lemma 2 (a), the optimal transient cdf holds for steady state as well for the original $GI/GI/1$ model, where steady-state holds. Thus, $F_0^* = \hat{F}$ is optimal for $n = \infty$ as well, which is the desired conclusion in (35).

For the reduction to $\mathcal{P}_{a,2,3}^{(c)}(M_a)$, we can apply part (a) of Theorem 1 for $n = 1$ or Corollary 1 to determine that, for each candidate $\hat{F}$, there is $F_0^* \equiv F_0^*(\hat{F}) \in \mathcal{P}_{a,2,3}^{(c)}(M_a)$ for $n = 1$. With uniqueness, $\hat{F}$ itself must be in $\mathcal{P}_{a,2,3}^{(c)}(M_a)$. ∎

Remark 5. (candidate verification) Theorem 3 provides a way to verify that candidate extremal cdf's are in fact extremal. For example, for Theorem 1 (a), let G be specified. To verify that $F_{1+c_a^2}^{(2)}$ is optimal for the steady-state mean, it suffices to verify (33) for $\hat{F} = F_{1+c_a^2}^{(2)}$. Of course, that requires working with the cdf of $W_\infty(F_{1+c_a^2}^{(2)}, G)$.

Our next result, proved in §EC.3 by applying the Kakutani fixed point theorem and an additional asymptotic argument, shows that the conditions in Theorem 3 can be satisfied.

Theorem 4. (*existence*) *For any $G \in \mathcal{P}_{s,2}$ in Theorem 1 (a), there exists a cdf $\hat{F} \equiv \hat{F}(G) \in \mathcal{P}_{a,2}^{(c)}(M_a)$ satisfying the conditions of Theorem 3. The analog holds for $\hat{G}(F)$ in part (b).*

5. A Multinomial Optimization for the Transient Mean $E[W_n]$

In this section we exploit Theorem 1 (c) to formulate an optimization problem for the upper bound of the transient mean based on a multinomial representation. We then can apply Lemma 2 (c) to numerically deduce the form of the overall upper bound in (c) for the steady-state mean. We provide further support with simulations for two-point distributions in §6.

5.1. The Multinomial Representation

We can represent the transient mean starting empty in (3) in terms of two independent multinomial distributions. Let the cdf G in $\mathcal{P}^{(c)}_{s,2,3}(M_s)$ with specified mean ρ and scv c_s^2 be parameterized by the vector of mass points $\mathbf{v} \equiv (v_1, v_2, v_3)$ and the vector of probabilities $\mathbf{p} \equiv (p_1, p_2, p_3)$. For every positive integer k, define a multinomial probability mass function on the vector of nonnegative integers $\mathbf{k} \equiv (k_1, k_2, k_3)$ by

$$P_k(\mathbf{p}) \equiv \frac{k! p_1^{k_1} p_2^{k_2} p_3^{k_3}}{k_1! k_2! k_3!}, \tag{38}$$

where it is understood that $\mathbf{k}\mathbf{e}' \equiv k_1 + k_2 + k_3 = k$. Similarly, let the cdf F in $\mathcal{P}_{a,2,3}$ with specified mean 1 and scv c_a^2 be parameterized by the vector of mass points $\mathbf{u} \equiv (u_1, u_2, u_3)$ and probabilities $\mathbf{q} \equiv (q_1, q_2, q_3)$ on the vector of nonnegative integers $\mathbf{w} \equiv (w_1, w_2, w_3)$, so that

$$Q_k(\mathbf{q}) \equiv \frac{k! q_1^{w_1} q_2^{w_2} q_3^{w_3}}{w_1! w_2! w_3!}, \tag{39}$$

where it is understood that $\mathbf{w}\mathbf{e}' \equiv w_1 + w_2 + w_3 = k$.

Then, from (3),

$$E[W_n | W_0 = 0] = \sum_{k=1}^{n} \frac{1}{k} \sum_{(\mathbf{k},\mathbf{w}) \in \mathcal{I}} \max\{0, \sum_{i=1}^{3} (k_i v_i - w_j u_j)\} P_k(\mathbf{p}) Q_k(\mathbf{q}), \tag{40}$$

where $\mathcal{I}$ is the set of all pairs of vectors $(\mathbf{k}, \mathbf{w})$ with both $\mathbf{k}\mathbf{e}' \equiv k_1 + k_2 + k_3 = k$ and $\mathbf{w}\mathbf{e}' \equiv w_1 + w_2 + w_3 = k$.

For any given n and any given distributions G in $\mathcal{P}^{(c)}_{s,2,3}(M_s)$ parameterized by the pair $(\mathbf{v}, \mathbf{p})$ and F in $\mathcal{P}^{(c)}_{a,2,3}(M_a)$ parameterized by the pair $(\mathbf{u}, \mathbf{q})$, we can calculate the transient mean $E[W_n | W_0 = 0]$

by calculating the sum in (40). We can easily evaluate $E[W_n|W_0=0]$ for candidate cases provided that n is not too large.

Next, for the overall optimization over $\mathcal{P}^{(c)}_{a,2,3}(M_a)\times\mathcal{P}^{(c)}_{s,2,3}(M_s)$, we write

$$\sup\{E[W_n(\mathbf{v},\mathbf{p},\mathbf{u},\mathbf{q})]:((\mathbf{v},\mathbf{p}),(\mathbf{u},\mathbf{q}))\in\mathcal{P}^{(c)}_{a,2,3}(M_a)\times\mathcal{P}^{(c)}_{s,2,3}(M_s)\}, \tag{41}$$

using (40). We now write this optimization problem in a more conventional way, from which we see that the optimization is a form of non-convex nonlinear program. In particular, we write for the means $m_1\equiv E[U]\equiv 1$, $m_2\equiv E[U^2]\equiv m_1^2(c_a^2+1)$, $s_1\equiv E[V]\equiv\rho$ and $s_2\equiv E[V^2]\equiv s_1^2(c_a^2+1)$,

$$\begin{aligned}
\text{maximize}\ & \sum_{k=1}^{n}\frac{1}{k}\sum_{\sum k_i=k,\sum_j w_j=k}\max(\sum_i k_i v_i-\sum_j w_j u_i,0)P(k_1,k_2,k_3)Q(w_1,w_2,w_3)\\
\text{subject to}\ & \sum_{j=1}^{3}u_j q_j=m_1,\quad \sum_{j=1}^{3}u_j^2 q_j=(1+c_a^2)m_1^2,\\
& \sum_{j=1}^{3}v_j p_j=s_1,\quad \sum_{j=1}^{3}v_j^2 p_j=(1+c_s^2)s_1^2,\\
& \sum_{j=1}^{3}p_j=\sum_{k=1}^{3}q_k=1,\\
& M_s\ge v_j\ge 0,\ M_a\ge u_j\ge 0,\ p_j\ge 0,\ q_j\ge 0,\quad 1\le j\le 3.
\end{aligned} \tag{42}$$

5.2. The Numerical Conclusion about the Overall Upper Bound

We solved this non-convex nonlinear program in (42) by applying sequential quadratic programming (SQP) as discussed in Chapter 18 of Nocedal and Wright (1999). In particular, we applied the Matlab variant of SQL, which is a second-order method, implementing Schittkowski's NLPQL Fortran algorithm. This algorithm converges at a local optimum. Since the algorithm is not guaranteed to reach a global optimum, we run the algorithm for a large collection of uniform randomly chosen initial conditions.

We found that the local optimum solution is usually attained at the pair of two-point distributions $(F^{(2)}_{1+c_a^2},G^{(2)}_{b(n)})$, where $b(n)$ depends on n, but $b(n)\to M_s$ as $n\to\infty$; i.e., $G^{(2)}_{b(n)}$ is a two-point distribution that converges to $G^{(2)}_{M_s}$ as $n\to\infty$. In the rare cases that we obtain a different solution,

we found that it is always in $\mathcal{P}^{(c)}_{a,2,2}(M_a) \times \mathcal{P}^{(c)}_{s,2,2}(M_s)$. Moreover, in these cases, we can find a different initial condition for which $(F^{(2)}_{1+c_a^2}, G^{(2)}_{b(n)})$ is the local optimum, and that $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{b(n)})]$ is larger than for other local optima.

From extensive numerical experiments, which draw on our mathematical results, we conclude that the extremal UB interarrival-time cdf $F^{(2)}_{1+c_a^2}$ for $GI/M/1$ also applies to all $GI/GI/1$, but the extremal service-time distribution is more complicated because it depends on both n and M_s. In summary, Theorem 1 and our numerical results support the following conjecture about the overall tight upper bound. For part (b), let $G^{(2)}_{\infty}$ in $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})]$ be shorthand for the limit of $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})]$ as $M_s \to \infty$.

CONJECTURE 1. *(the tight upper bound for $1 \le n \le \infty$ for $W_0 = 0$)*

(a) Given any parameter vector $(1, c_a^2, \rho, c_s^2)$ and a bounded interval $[0, \rho M_s]$ for the service-time cdf G, where $M_s \ge c_s^2 + 1$, the pair $(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})$ attains the tight UB of the steady-state mean $E[W]$, i.e.,

$$E[W(F,G)] \le E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})] \quad \textit{for all} \quad F \in \mathcal{P}_{a,2}(M_a) \quad \textit{and} \quad G \in \mathcal{P}_{s,2}(M_s),$$

while a pair $(F^{(2)}_{1+c_a^2}, G^{(2)}_{b(n)})$ attains the tight UB of the transient mean $E[W_n]$, i.e.,

$$E[W_n(F,G)] \le E[W_n(F^{(2)}_{1+c_a^2}, G^{(2)}_{b(n)}))] \quad \textit{for all} \quad F \in \mathcal{P}_{a,2}(M_a) \quad \textit{and} \quad G \in \mathcal{P}_{s,2}(M_s),$$

where $G^{(2)}_{b(n)}$ is a two-point distribution with $G^{(2)}_{b(n)} \Rightarrow G^{(2)}_{M_s}$ as $n \to \infty$.

(b) When both F and G have unbounded support $[0, \infty)$, the tight UB of $E[W(F,G)]$ is obtained asymptotically in the limit as $M_s \to \infty$ in part (a), i.e.,

$$E[W(F,G)] \le \lim_{M_s \to \infty} E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})] \equiv E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{\infty})] \quad \textit{for all} \quad F \in \mathcal{P}_{a,2} \quad \textit{and} \quad G \in \mathcal{P}_{s,2}.$$

We develop algorithms for computing $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{\infty})]$ in Chen and Whitt (2018). The following is an UB for $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{\infty})]$, assuming Conjecture 1.

THEOREM 5. *(an UB for $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_\infty)]$) under the conjecture) For the $GI/GI/1$ queue with parameter four-tuple $(1, c_a^2, \rho, c_s^2)$, if $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_\infty)]$ is the tight UB as claimed in Conjecture 1, then*

$$E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_\infty)] \le \frac{2(1-\rho)\rho/(1-\delta)c_a^2 + \rho^2 c_s^2}{2(1-\rho)} < \frac{\rho(2-\rho)c_a^2 + \rho^2 c_s^2}{2(1-\rho)}, \tag{43}$$

where $\delta \in (0,1)$ and $\delta = \exp(-(1-\delta)/\rho)$.

Formula (43) relies on Conjecture 1, so it is only verified numerically so far. Formula (43) is based on Conjecture III on p. 211 of Daley et al. (1992). We prove Theorem 5 in §EC.4.

Counterexamples that contradict corresponding conjectures that analogs of Conjecture 1 hold when one distribution is fixed were constructed in §V of Whitt (1984b), drawing on Whitt (1984a), and in §8 of Wolff and Wang (2003).

Tables 1 and 2 compare the numerically computed values of the conjectured tight UB, $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_\infty)]$, drawing on Chen and Whitt (2018), to the heavy-traffic approximation (HTA) in (5), the new conjectured upper bound in (43), the Daley (1977) bound in (7) and the Kingman (1962) bound in (6) over a range of ρ for the scv pairs $(c_a^2, c_s^2) = (4.0, 4.0)$ and $(0.5, 0.5)$. In order to focus on the variability independent of the traffic intensity ρ, we display the scaled mean waiting time values $(1-\rho)E[W]/\rho^2$, which are constant for the heavy-traffic approximation in (5), being equal to $(c_a^2 + c_s^2)/2$. Tables EC.4-EC.7 in the e-companion give results for 12 values of ρ all four cases: $(c_a^2, c_s^2) = (4.0, 4.0)$, $(0.5, 0.5)$, $(4.0, 0.5)$, $(0.5, 4.0)$.

In these tables we also show the value of δ in the new UB (43) and the maximum relative error (MRE) between the UB approximation and the tight UB. The MRE over all four cases was 5.7%. which occurred for $c_a^2 = c_s^2 = 0.5$ and $\rho = 0.5$.

We also display the lower bound (LB) in (8), which is far less than the other values, indicating the wide range of possible values. The extremely low value for the LB occurs because it is associated with the $D/GI/1$ model, which is approached by the $F^{(2)}_{M_a}$ extremal distribution as the support limit $M_a \to \infty$ for any c_a^2. Notice that the LB is actually 0 for many cases with low traffic intensity; that occurs if and only if $P(V \le U) = 1$. Hence, the LB looks especially bad for the case $(c_a^2 =$

4.0, $c_s^2 = 0.5$) in Table EC.6, because it is the same as for the case ($c_a^2 = 0.5, c_s^2 = 0.5$) in Table EC.5 and even for ($c_a^2 = 0.0, c_s^2 = 0.5$) in the $D/GI/1$ model. We discuss the LB in §7.

Table 1 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = c_s^2 = 4.0$.

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.30	0.833	4.000	11.661	11.731	0.041	0.60%	13.333	24.222
0.50	1.500	4.000	6.940	7.020	0.203	1.15%	8.000	10.000
0.70	1.786	4.000	5.168	5.216	0.467	0.93%	5.714	6.082
0.80	1.875	4.000	4.662	4.693	0.629	0.67%	5.000	5.125
0.90	1.944	4.000	4.287	4.302	0.807	0.35%	4.444	4.469
0.95	1.974	4.000	4.134	4.142	0.902	0.18%	4.211	4.216
0.99	1.995	4.000	4.025	4.027	0.980	0.04%	4.040	4.041

Table 2 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = c_s^2 = 0.5$.

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.30	0.000	0.500	1.432	1.466	0.041	2.36%	1.667	3.028
0.50	0.000	0.500	0.827	0.878	0.203	5.72%	1.000	1.250
0.70	0.036	0.500	0.623	0.652	0.467	4.53%	0.714	0.760
0.90	0.194	0.500	0.530	0.538	0.807	1.38%	0.556	0.559
0.95	0.224	0.500	0.514	0.518	0.902	0.65%	0.526	0.527
0.99	0.245	0.500	0.503	0.503	0.980	0.14%	0.505	0.505

From this analysis, we see that conjectured new UB (43) is an excellent approximation for the conjectured UB $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{\infty})]$. Moreover, we see that there is significant improvement going from the Kingman (1962) bound in (6) to the Daley (1977) bound in (7) to the new UB in (43). We also see that the heavy-traffic approximation is consistent with the upper bounds in all cases. The heavy-traffic approximation in (5) tends to be much closer to the UB than the lower bound,

which shows that the overall MRE can be large and that the heavy-traffic approximation tends to be relatively conservative, as usually is desired in applications.

6. A Simulation Study Over All Two-Point Distributions

The optimization in §5 supports Conjecture 1, but not as strongly as we would like. A more convincing conclusion from §5 is that it suffices to reduce the search for an optimum to the smaller subset of two-point distributions, i.e., to the product space $\mathcal{P}^{(c)}_{a,2,2}(M_a) \times \mathcal{P}^{(c)}_{s,2,2}(M_s)$. This space is relatively easy to analyze because each of the sets $\mathcal{P}^{(c)}_{a,2,2}(M_a)$ and $\mathcal{P}^{(c)}_{s,2,2}(M_s)$ is one-dimensional, as indicated in §1.3. The $G^{(2)}_{1+c_s^2}$ counterexample from §8 of Wolff and Wang (2003) also falls in this set.

6.1. Simulation Experiments

To analyze the mean waiting times for the two-point interarrival-time and service-time distributions, we primarily use stochastic simulation. (We also verify for lower traffic intensities by applying the multinomial representation in §5 for finite n.)

We study various simulation approaches in Chen and Whitt (2018). For the transient mean $E[W_n]$, we use direct numerical simulation, but for the steady-state simulations we mostly use the simulation method in Minh and Sorli (1983) that exploits the representation of $E[W]$ in terms of the steady-state idle time I and the random variable I_e that has the associated equilibrium excess distribution, i.e.,

$$E[W] = -\frac{E[X^2]}{2E[X]} - E[I_e] = -\frac{E[X^2]}{2E[X]} - \frac{E[I^2]}{2E[I]} = \frac{\rho^2 c_s^2 + c_a^2 + (1-\rho)^2}{2(1-\rho)} - \frac{E[I^2]}{2E[I]}; \tag{44}$$

which is also used in Wolff and Wang (2003). For each simulation experiment, we perform multiple (usually $20-40$) i.i.d. replications. Within each replication we look at the long-run average after deleting an initial portion to allow the system to approach steady state if deemed helpful. It is well known that obtaining good statistical accuracy is more challenging as ρ increases, e.g., see Whitt (1989), but that challenge is largely avoided by using (44). There is also a well known issue of one long run versus multiple replications, e.g., see Whitt (1991).

We do not report confidence intervals for all the individual results, but we did do a careful study of the statistical precision. To illustrate, Table 3 compares the 95% confidence intervals associated with estimates of the steady-state mean $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})]$ for the parameter triple $(\rho, c_a^2, c_s^2) = (0.5, 4.0, 4.0)$ obtained by making the statistical t test to multiple replications of runs of various length. The table compares the standard simulation for various run lengths N (number of arrivals) and the Minh and Sorli (1983) algorithm for various run lengths T (length of time, over which we average the observed idle periods) and numbers of replications n. (See Chen and Whitt (2018) for more discussion.)

Table 3 Confidence interval halfwidths for estimates of the steady-state mean $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})]$ for the parameter triple $(\rho, c_a^2, c_s^2) = (0.5, 4.0, 4.0)$

	Monte Carlo simulation			Minh and Sorli simulation		
replications	$N = 1 \times 10^5$	$N = 1 \times 10^6$	$N = 1 \times 10^7$	$T = 1 \times 10^5$	$T = 1 \times 10^6$	$T = 1 \times 10^7$
20	6.64E-02	2.45E-02	8.01E-03	1.58E-03	4.81E-04	1.55E-04
40	5.59E-02	1.27E-02	4.22E-03	1.20E-03	3.20E-04	9.89E-05
60	3.69E-02	1.20E-02	4.23E-03	8.44E-04	2.88E-04	8.03E-05
80	3.52E-02	1.17E-02	3.72E-03	7.54E-04	2.27E-04	9.55E-05
100	2.61E-02	9.94E-03	3.13E-03	6.06E-04	2.02E-04	7.20E-05

6.2. The Impact of the Interarrival-Time Distribution

Figure 1 reports simulation results for $E[W_{20}]$ (left) and $E[W]$ (right) in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = M_s = 15$. Recall $b_a \in [1 + c_a^2, M_a]$ and $b_s \in [1 + c_s^2, M_s]$ determine F in $\mathcal{P}^{(c)}_{a,2,2}(M_a)$ and G in $\mathcal{P}^{(c)}_{s,2,2}(M_a)$ respectively, we focus on the impact of b_a (for F) in the permissible range $[5, 15]$ for six values of b_s (for G) ranging from 5 to 15. (Recall that the parameter b was defined in §1.3.)

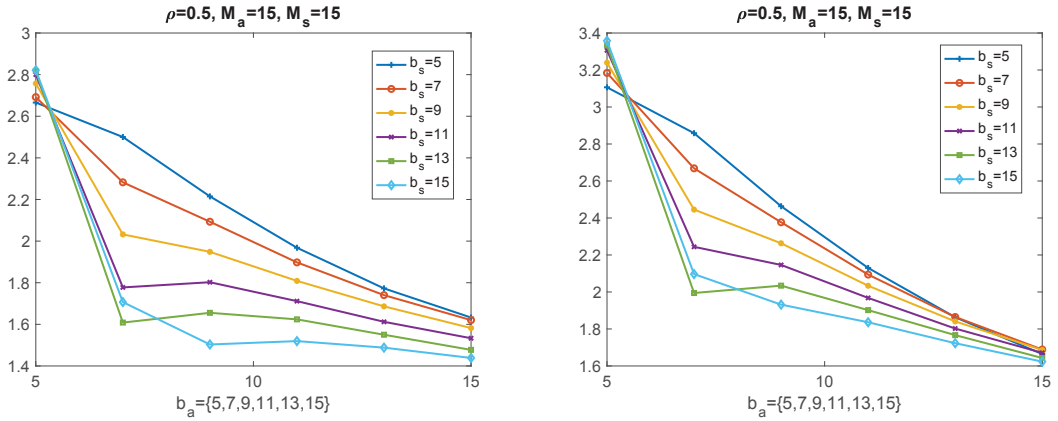

Figure 1 Simulation estimates of the transient mean $E[W_{20}]$ (left) and the steady-state mean $E[W]$ (right) as a function of b_a for six cases of b_s in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = M_s = 15$ under $N = 1 \times 10^7$ and 20 i.i.d replications.

Figure 1 shows that the mean waiting times tend to be much larger at the extreme left, which is associated with $b_a = 5$ or $F^{(2)}_{1+c_a^2}$. Also, the mean is roughly decreasing with b_s increasing except for $b_a = 5$.

On the other hand, a close examination of the extreme case $b_s = 5$ shows that the largest value of b_a does not occur for $b_a = 5$, but in fact occurs at a slightly higher value. That turns out to be the counterexample for the conjecture

$$E[W(F_0, G)] = \sup\{E[W(F, G)] : F \in \mathcal{P}_{a,2}\} \tag{45}$$

for any given $G \in \mathcal{P}_{s,2}$. In particular, Tables 4 present detailed simulation estimates of $E[W]$ and $E[W_{20}]$. In Table 4, we see that the maximum mean waiting time value in the first row, i.e., over b_a when $b_s = 5$ is not attained at $b_a = 5.0$, but is instead attained at $b_a = 5.25$. For emphasis, in each case we highlight both the maximum entry in the first row and the maximum entry in the table. Therefore, for that service-time distribution (which is $G^{(2)}_{1+c_s^2}$), the extremal inter-arrival time is not $F^{(2)}_{1+c_a^2}$.

Table 4 Simulation estimates of $E[W]$ as a function of b_a and b_s when $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = 8 < M_s = 10$ ($N = 1 \times 10^7$ and 20 i.i.d replications).

$b_s \backslash b_a$	5	5.25	5.5	5.75	6	6.25	6.5	6.75	7	7.25	7.5	7.75	8
5	3.11	**3.14**	3.11	3.08	3.04	3.00	2.95	2.90	2.86	2.81	2.76	2.71	2.66
6	3.19	3.06	2.93	2.91	2.90	2.88	2.85	2.81	2.77	2.73	2.69	2.64	2.60
7	3.19	3.07	2.94	2.80	2.75	2.72	2.71	2.70	2.67	2.64	2.60	2.57	2.53
8	3.20	3.06	2.93	2.81	2.66	2.61	2.59	2.57	2.55	2.53	2.51	2.48	2.45
9	3.24	3.09	2.93	2.79	2.67	2.53	2.47	2.46	2.45	2.43	2.41	2.39	2.37
10	**3.28**	3.14	2.98	2.81	2.64	2.51	2.37	2.35	2.35	2.34	2.32	2.31	2.29

Note that $F^{(2)}_{1+c_a^2}$ is not optimal for all other b_s and the difference between $\max\{E[W(F,G_0)] : F\} - E[W(F^{(2)}_{1+c_a^2}, G_0)]$ is very small. Moreover, consistent with Conjecture 1, the overall UB is attained at the pair $(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})$. Finally, note that the difference across each row tends to be greater than the difference across each column.

6.3. The Impact of the Service-Time Distribution

Figure 1 also shows the impact of the service-time distribution, but that impact is more complicated. We see that the curve crosses the other curves in the middle. We now investigate what is the optimal value of b_s over $[1+c_s^2, M_s]$ for $E[W_n]$ and $E[W]$. For that purpose, Figure 2 shows the upper bound of $E[W]$ over $\mathcal{P}^{(c)}_{a,2,2}(M_a) \times \mathcal{P}^{(c)}_{a,2,2}(M_s)$ is attained by $(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})$. Moreover,

$$E[W(F^{(2)}_{b_a}, G^{(2)}_{1+c_s^2})] = \sup\{E[W(F^{(2)}_{b_a}, G)] : G \in \mathcal{P}^{(c)}_{s,2,2}(M_s)\}, \text{ for } \ b_a \in (5,10], M_s = 10. \tag{46}$$

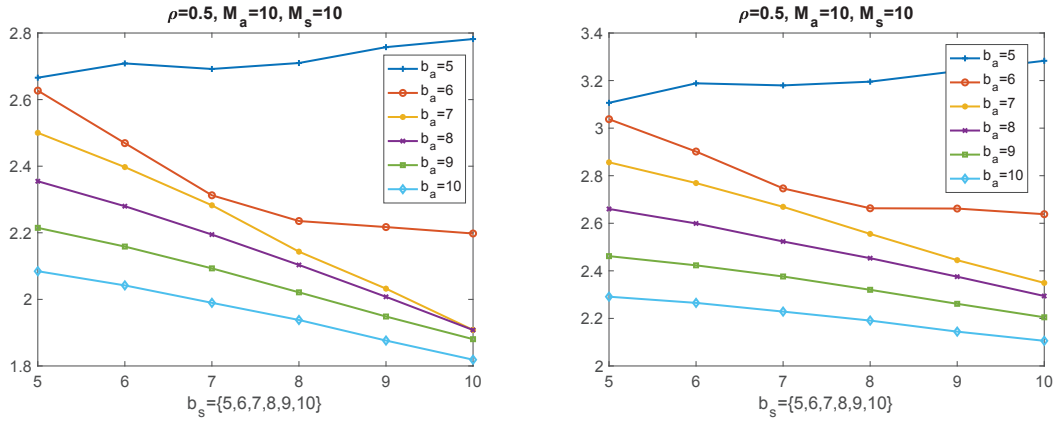

Figure 2 Simulation estimates of the transient mean $E[W_{20}]$ (left) and the steady-state mean $E[W]$ (right) as a function of b_s for six cases of b_a in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = M_s = 10$ under $N = 1 \times 10^7$ and 20 i.i.d replications.

Figure 3 plots the values of $E[W_{10}]$ (left) and $E[W_{20}]$ (right) as a function of b_s in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$, $M_s = 300$ and $b_a = (1 + c_a^2)$. For Figure 3, we use the optimization in §5 with a numerical method to directly compute a good finite truncation of objective in the nonlinear program (42). For these cases, we find $b_s^*(10) = 35.1$ and $b_s^*(20) = 41.1$.

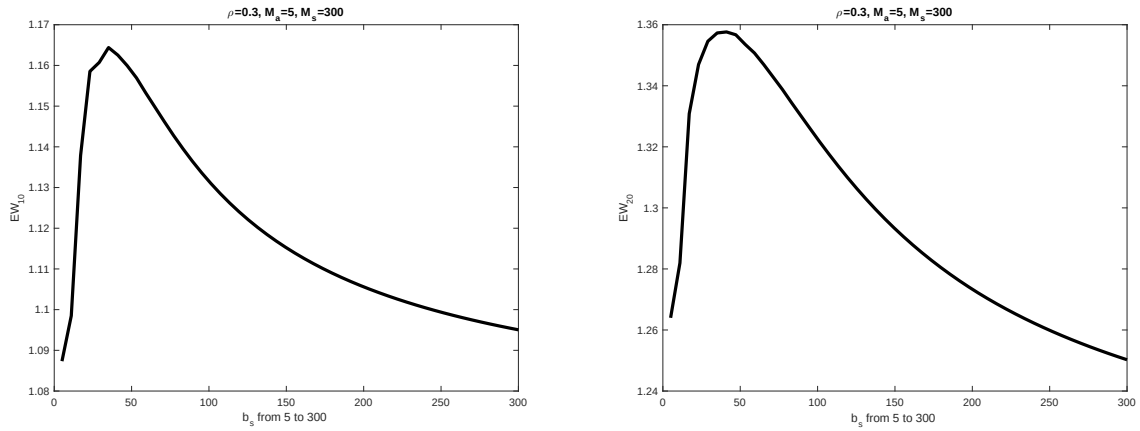

Figure 3 The transient mean waiting time $E[W_n(F^{(2)}_{1+c_a^2}, G)]$ for $n = 10, 20$ as a function of b_s up to $M_s = 300$. $b_s^*(10) = 35.10, b_s^*(20) = 41.12$.

As a function of b_s, our numerical experience shows the transient mean waiting time $E[W_n]$ is approximately first increasing and then decreasing at all traffic levels. Therefore, for each n,

there exists optimal $b(n)$ such that $E[W_n(F^{(2)}_{1+c_a^2}, G^{(2)}_{b(n)})] \geq E[W_n(F^{(2)}_{1+c_a^2}, G); G \in \mathcal{P}^{(c)}_{s,2,2}(M_s)]$ where the $G_{b(n)}$ implies the $u = \rho(1 - c_s^2/(b(n)-1))$.

We next directly examine the steady-state mean waiting time $E[W]$ for set $b_a = (1 + c_a^2)$ and $M_s = 100$. We use Minh and Sorli (1983) method with simulation length over a time interval of length $T = 1 \times 10^7$ and 40 i.i.d. replications. To illustrate, Figure 4 shows the results for the traffic levels $\rho = 0.5$ (left) and $\rho = 0.9$ (right).

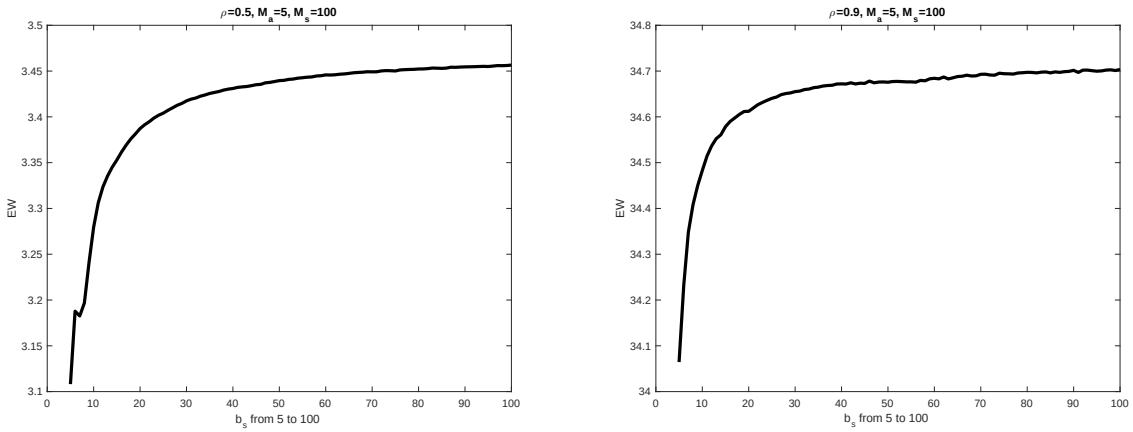

Figure 4 $E[W(F^{(2)}_{1+c_a^2}, G)]$ for $G \in \mathcal{P}^{(c)}_{s,2,2}(M_s)$ as a function of b_s given $b_a = (1 + c_a^2)$ for the case $c_a^2 = c_s^2 = 4$.

Just as in Figure 4 shows that the steady-state mean $E[W]$ is eventually increasing in b_s, given $b_a = (1 + c_a^2)$, strongly supporting the conclusion that the upper bound is attained at $(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})$. Hence, the optimal b_s is M_s. Since $E[W_n] \to E[W]$, we must also have $b(n) \to M_s$ as $n \to \infty$.

7. The Lower Bound with Finite Support

For unbounded support, Ott (1987) showed that the overall LB of $E[W(F, G)]$ for $(F, G) \in \mathcal{P}_{a,2} \times \mathcal{P}_{s,2}$ is attained asymptotically by the $D/G_a^{(3)}/1$ model where the D interarrival time with $c_a^2 = 0$ can be regarded as the limit of $F^{(2)}_{M_a}$ with c_s^2 on $[0, M_a]$ as $M_a \to \infty$ holding the mean fixed at $E[U] = 1$, while the service-time cdf $G_a^{(3)}$ is any three-point distribution in $\mathcal{P}_{s,2}(\rho, c_s^2)$ that has support on integer multiples of the constant interarrival time 1; also see Theorem 3.1 of Daley et al. (1992). It turns out that the mean is insensitive to the service-time cdf provided that all support is on

integer multiples of the interarrival time. Thus, the pure-lattice structure of the $D/G_a^{(3)}/1$ model acts to reduce $E[W]$. The resulting LB has the convenient explicit formula in (8).

However, the overall LB has not yet been established for distributions with finite support. Motivated by the established extremal property of the lattice $D/G_a^{(3)}/1$ model with unbounded support, we investigated a new "nearly-lattice" three-point distribution to use with $F_{M_a}^{(2)}$ called $G_{u,b_su}^{(3)}$. It has support $\{0,u,b_su\}$, where $1<b_s\le M_s$ is an appropriate positive value (see §EC.6 for more explanations.); u is the first point of the cdf $F_{M_a}^{(2)}$ at $u=1-c_a^2/(M_a-1)\in(0,1)$ with $M_a>1+c_a^2$. We provide details of our study in §EC.6.

As expected, for each $(1,c_a^2,\rho,c_s^2,M_a)$ with $M_a>1+c_a^2$, there exists a proper $b_s^*\in(1,\infty)$ such that

$$E[W(D,G_a^{(3)})]\le E[W(F_{M_a}^{(2)},G_{u,b_s^*u}^{(3)})]\le \inf\{E[W(F_{M_a}^{(2)},G_{b_s}^{(2)})]:b_s\in[1+c_s^2,\infty)\}. \tag{47}$$

If $M_a=1+c_a^2$, we have

$$E[W(D,G_a^{(3)})]\le E[W(F_{M_a}^{(2)},G_{1+c_s^2}^{(2)})]\le \inf\{E[W(F_{M_a}^{(2)},G_{b_s}^{(2)})]:b_s\in[1+c_s^2,\infty)\}. \tag{48}$$

8. Conclusions

Theorem 1 showed that the tight upper and lower bounds for both the transient and steady-state mean waiting time, $E[W_n]$, in the $GI/GI/1$ model given interarrival and service times with compact support and specified first two moments are attained at three-point distributions. That result applies both to the overall bound over both distributions as well as the bounds with one of the distributions fixed. Theorems 3 and 4 provided an alternative fixed-point characterization of the extremal distributions for the steady-state mean, which should prove useful for further studies.

In the rest of the paper, we applied numerical methods to further identify the extremal distributions. In §5 we exploited Theorem 1 to construct a multinomial mathematical optimization for the transient mean. In §6 we reported results of extensive simulations over the one-dimensional space of distributions with support on two points. From a practical engineering perspective, these numerical studies answered the important question about the tight upper bound. The combination

of mathematical and numerical results strongly supports Conjecture 1 in §5.2, which states that the overall upper bound is attained by $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{\infty})]$, i.e., at the extremal two-point distributions, modified by a limit, as many have thought. However, because the analysis is partly numerical, it still remains to provide a mathematical proof. We also provided a new upper bound analytical formula (43), which is a valid bound under Conjecture 1. Drawing on algorithms to compute $E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{\infty})]$ in Chen and Whitt (2018), Tables 1, 2 and EC.4-EC.5 illustrate that the new UB formula is quite accurate, providing significantly improvement over previous bounds.

We also conducted a study of the extremal lower bound with finite support in §7 and §EC.6. We have less conclusive results, but we present evidence that it is attained by $E[W(F^{(2)}_{M_a}, G^{(3)}_{u,b_s u})]$, where $F^{(2)}_{M_a}$ is the natural two-point distribution with support on the upper limit of support M_a, while $G^{(3)}_{u,b_s u}$ is a nearly-lattice three-point distribution with mass on the set $\{0, u, k_s u\}$ for an integer k_s, where u is the smaller mass point of $F^{(2)}_{M_a}$ needed to go with the mass point at the upper barrier M_a. This is asymptotically correct as $M_a \to \infty$ because it converges to the known lower bound for unbounded support in (8).

There are many remaining problems for research. In addition to providing a full mathematical proof of Conjecture 1, it remains to identify the extremal distributions with one distribution given. It also remains to establish similar results for other models. The method of proof here can be adapted to other settings.

Acknowledgments

This research was supported by NSF CMMI 1634133.

References

Asmussen, S. 2003. *Applied Probability and Queues*. 2nd ed. Springer, New York.

Berge, C. 1963. *Topological Spaces*. Macmillan, New York. (English translation of the 1959 French edition).

Bertsimas, D., K. Natarajan. 2007. A semidefinite optimization approach to the steady-state analysis of queueing systems. *Queueing Systems* **56** 27–39.

Border, K. C. 1985. *Fixed Point Theorems with Application to Economics and Game Theory*. Cambridge University Press, New York.

Chen, Y., W. Whitt. 2018. Algorithms for the upper bound mean waiting time in the $GI/GI/1$ extremal queue. Submitted for publication, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

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

Daley, D. J., A. Ya. Kreinin, C.D. Trengove. 1992. Inequalities concerning the waiting-time in single-server queues: a survey. U. N. Bhat, I. V. Basawa, eds., *Queueing and Related Models*. Clarendon Press, 177–223.

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

Gupta, V., J. Dai, M. Harchol-Balter, B. Zwart. 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., T. Osogami. 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. 1983. Batch delays versus customer delays. *Bell Laboratories Technical Journal* **62**(7) 2011–2015.

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

Johnson, M. A., M. R. Taaffe. 1990a. Matching moments to phase distributions: Density function shapes. *Stochastic Models* **6**(2) 283–306.

Johnson, M. A., M.R. Taaffe. 1990b. Matching moments to phase distributions: nonlinear programming approaches. *Stochastic Models* **6**(2) 259–281.

Kakutani, S. 1941. A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* **8**(3) 457–459.

Kantorovich, L. V., G. S. Rubinstein. 1958. On a space of completely additive functions. *Vestnik Leningrad. Univ* **13**(7) 52–59.

Karlin, S., W. J. Studden. 1966. *Tchebycheff Systems; With Applications in Analysis and Statistics*, vol. 137. Wiley, New York.

Kemperman, J. H. B. 1983. *On the role of duality in the theory of moments*, *Lecture Notes in Mathematical Economics and Mathematical Systems*, vol. 215. Springer, 63–92.

Kingman, J. F. C. 1961. The single server queue in heavy traffic. *Proc. Camb. Phil. Soc.* **77** 902–904.

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

Klincewicz, J. G., W. Whitt. 1984. On approximations for queues, II: Shape constraints. *AT&T Bell Laboratories Technical Journal* **63**(1) 139–161.

Lasserre, J. B. 2010. *Moments, Positive Polynomials and Their Applications*. Imperial College Press.

Li, Y., D. A. Goldberg. 2017. Simple and explicit bounds for multii-server queues with universal $1/(1-\rho)$ and better scaling. ArXiv:1706.04628v1.

Minh, D. L., R. M. Sorli. 1983. Simulating the $GI/G/1$ queue in heavy traffic. *Operations Research* **31**(5) 966–971.

Nocedal, J., S. J. Wright. 1999. *Numerical Optimization*. Springer, New York.

Osogami, T., R. Raymond. 2013. Analysis of transient queues with semidefinite optimization. *Queueing Systems* **73** 195–234.

Ott, T. J. 1987. Simple inequalties for the $D/G/1$ queue. *Opperations Research* **35**(4) 589–597.

Parthasarathy, K. R. 1967. *Probability Measures on a Metric Space*. Academic Press, New York.

Rachev, S. T., L. B. Klebanov, S. V. Stoyanov, F. J. Fabozzi. 2013. *The Methods of Distances in the Theory of Probability and Statistics*. Springer, New York.

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

Ross, S. M. 1996. *Stochastic Processes*. 2nd ed. Wiley, New York.

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

Stoyan, D. 1983. *Comparison Methods for Queues and Other Stochastic Models*. John Wiley and Sons, New York. Translated and edited from 1977 German Edition by D. J. Daley.

Stoyan, D., H. Stoyan. 1974. Inequalities for the mean waiting time in single-line queueing systems. *Engineering Cybernetics* **12**(6) 79–81.

Von Neumann, J. 1928. Zur theorie der gesellschaftsspiele. *Mathematische Annalen* **100** 295–320.

Whitt, W. 1983a. Comparing batch delays and customer delays. *Bell Laboratories Technical Journal* **62**(7) 2001–2009.

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

Whitt, W. 1984a. Minimizing delays in the $GI/G/1$ queue. *Operations Research* **32**(1) 41–51.

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

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

Whitt, W. 1989. Planning queueing simulations. *Management Science* **35**(11) 1341–1366.

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

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

Whitt, W., W. You. 2018. Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research* **66** 100–120.

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

Zolotarev, V. M. 1976. Metric distances in spaces of random variables and their distributions. *Math. USSR Sbornik* **30** 373–401.

Zolotarev, V. M. 1983. Probability metrics. *Theory of Probability and its Applications* **28**(2) 264–287.

e-Companion to 'Extremal $GI/GI/1$ Queues Given Two Moments" by Y. Chen and W. Whitt

EC.1. Overview

This e-companion contains supplements to the main paper. §EC.2 provides a summary of the notation. §EC.3 provides a proof of Theorem 4. §EC.4 provides a proof of Theorem 5, which justifies the new analytical formula for the overall upper bound, under the assumption that Conjecture 1 is valid. §EC.5 discusses the extension from compact support to unbounded support. §EC.6 provides details of the study of the lower bound with finite support. §EC.7 supplements Table 1 by providing numerical comparisons of the conjectured tight upper bound and known tight lower bound of $E[W(F,G)]$ over the underlying cdf's F and G with specified first two moments, but unbounded support.

EC.2. Summary of the Notation

(i) **acronyms**

(a) UB: upper bound [§1.2]

(b) LB: lower bound [§1.2]

(c) HTA: heavy traffic approximation [§1.2]

(d) LP: linear program [§3.1]

(e) MRE: maximum relative error [Table 1]

(ii) **random variables**

(a) U_k: interarrival time between customers k and $k+1$, $k \geq 0$, having cdf F_k, with generic $U \equiv U(F)$ having cdf F, $E[U] = 1$ and finite scv c_a^2 [§1.1]

(b) V_k: service time of customer k, $k \geq 0$, having cdf G_k, with generic $V \equiv V(G)$ having cdf G and $E[V] = \rho$, $0 < \rho < 1$, and finite scv c_s^2 [§1.1]

(c) $X_k \equiv V_k - U_k$, $k \geq 0$, [§1.1]

(d) $S_k \equiv X_0 + \cdots + X_{k-1}$, $k \geq 1$, with $S_0 \equiv 0$ [§1.1]

(e) W_k: waiting time of customer k, $k \geq 0$, having cdf H_k, with steady-state limit $W \equiv W_\infty$, assuming a customer 0 arrives at time 0 with a waiting time W_0 distributed according to H_0 with finite mean $E[W_0]$ [§1.1]

(f) $W_k(F,G) \equiv W_k(H_0,F,G)$ and $W(F,G) \equiv W_\infty(F,G)$: the random variables W_k and $W \equiv W_\infty$ showing the dependence upon the cdf's H_0, F and G. (The steady-state distribution does not depend on H_0.) [§3.3]

(g) $\hat{W}_2(F_0,F_1)$: modification of $W_2(F_0,F_1)$ by penalty function [§3.3]

(h) Y: a generic random variable [§3.2]

(i) I: a steady-state idle time [§6]

(iii) **special probability distributions**

(a) $F_b^{(2)}$: a two-point interarrival-time distribution with one point at b, given the first two moments; so $F_{1+c_a^2}^{(2)}$ is the natural F for the UB of $E[W]$; [§1.3]

(b) $F_{M_a}^{(2)}$: a two-point interarrival-time distribution with one point at the upper limit M_a given the first two moments; it is a natural candidate for the LB of $E[W]$; [§1.3]

(c) $G_b^{(2)}$: a two-point service-time distribution with one point at b, given the first two moments [§1.3]

(d) $G_{M_s}^{(2)}$: a two-point service-time distribution with one point at the upper limit ρM_s given the first two moments; ; it is a natural candidate for the UB of $E[W]$; [§1.3]

(e) $G_{1+c_s^2}^{(2)}$: a two-point service-time distribution with one point at 0 given the first two moments; it is a natural candidate for the LB of $E[W]$; [§1.3]

(f) $G_a^{(3)}$: a three-point service-time distribution concentrating on multiples of the deterministic interrival time in a $D/GI/1$ model [§1.2 and §7]

(g) $G_{u,b_s u}^{(3)}$: a three-point service-time distribution with support $\{0, u, k_s u\}$ for some integer k, where u is the smaller mass point of the $F_{M_a}^{(2)}$ two-point interarrival-time cdf having higher mass point M_a [§7]

(h) $F^*(G)$: the optimal F as a function of G [Theorem 1]

(i) $G^*(F)$: the optimal G as a function of F [Theorem 1]

(j) (F^{**}, G^{**}): the optimal pair (F, G) [Theorem 1]

(iv) **spaces of probability measures**

(a) $\mathcal{P}$: the space of all probability distributions on $[0, \infty)$ or a subset, with subscripts a and s used to designate the interarrival time or the service time [§2]

(b) $\mathcal{P}_n$: the subset of $\mathcal{P}$ with first n moments specified, [§2]

(c) $\mathcal{P}_2(m_1, m_2) \equiv \mathcal{P}_2(m_1, c^2)$: the space of $\mathcal{P}_2$ with specified first two moments m_1 and $m_2 = m_1^2(c^2 + 1)$ [§2]

(d) $\mathcal{P}_n(M) \equiv \mathcal{P}_n(m_1, c^2, M)$: the subset of $\mathcal{P}_n$ with support on the bounded interval $[0, m_1 M]$ [§2]

(e) $\mathcal{P}_2^{(c)}(M)$: the subset of $\mathcal{P}_2(M)$ with support on a compact subset (denoted by $\mathcal{C}$) of the bounded interval $[0, m_1 M]$ [§2]

(f) $\mathcal{P}_{a,2}(M_a) \equiv \mathcal{P}_{a,2}(1, c_a^2, M_a)$: the space of interarrival-time cdf's F with first two moments $(1, c_a^2 + 1)$ and support $[0, M_a]$ [§2]

(g) $\mathcal{P}_{a,2}^{(c)}(1, c_a^2, M_a)$: the space of interarrival-time cdf's F with first two moments $(1, c_a^2 + 1)$ and support on a compact subset (denoted by $\mathcal{C}$) of the bounded interval $[0, M_a]$ [§2]

(h) $\mathcal{P}_{s,2}(\rho, c_s^2, M_s)$: the space of service-time cdf's G with first two moments $(\rho, \rho^2(c_s^2 + 1))$ and support $[0, \rho M_s]$ [§2]

(i) $\mathcal{P}_{s,2}^{(c)}(1, c_s^2, M_s)$: the space of service-time cdf's G with first two moments $(\rho, \rho^2(c_s^2 + 1))$ and support on a compact subset (denoted by $\mathcal{C}$) of the bounded interval $[0, \rho M_s]$ [§2]

(v) **functions**

(a) $w_n \equiv w_n(F, G)$: Shorthand for the mean, i.e., $w_n(F, G) \equiv E[W(F, G)]$ [§2]

(b) $w_{a,n}^{\uparrow}$: Shorthand for the supremum. [§2]

(c) ϕ: integrand for application of Theorem 2 [§3]

(d) $d_M(F_0, F_1)$: penalty function depending on the parameter M [§3.3]

(vi) **models**

(a) $F/G/1$: a $GI/GI/1$ model with cdf's F and G

(b) $D/G/1$: a $GI/GI/1$ model with deterministic F having unit mass on the mean, which here is 1.

EC.3. Proof of Theorem 4.

In this section we prove Theorem 4 which shows that there exist distributions satisfying the conditions in Theorem 3, which in turn provides a fixed-point characterization of the extremal distributions for the steady-state mean $E[W]$.

Proof. Focusing on (a) with $G \in \mathcal{P}_{s,2}$ fixed, let $\zeta(\hat{F})$ with $\zeta : \mathcal{P}_{a,2}^{(c)}(M_a) \to \mathbb{R}$ be defined by

$$\zeta(\hat{F}) \equiv \sup\{E[(W(\hat{F},G) + V(G) - U(F_0))^+] : F_0 \in \mathcal{P}_{a,2}^{(c)}(M_a)\}, \tag{EC.1}$$

where $W(\hat{F},G)$ is understood to be the steady-state waiting time with the pair $(\hat{F},G)$, $V \equiv V(G)$ has cdf G, $U \equiv U(F_0)$ has cdf F_0 and all three random variables are independent.

Let $\eta(\hat{F})$ be the set of maximizers in (EC.1) when $\hat{F} \in \mathcal{P}_{a,2}^{(c)}(M_a)$. Let $\mathcal{P}_{a,2}^{*}(M_a)$ be the set of all fixed points of the map $\eta : \mathcal{P}_{a,2}^{(c)}(M_a) \to 2^{\mathcal{P}_{a,2}^{(c)}(M_a)}$, i.e.,

$$\mathcal{P}_{a,2}^{*}(M_a) \equiv \{F \in \mathcal{P}_{a,2}^{(c)}(M_a) : F \in \eta(F)\}. \tag{EC.2}$$

To show that $\mathcal{P}_{a,2}^{*}(M_a)$ is nonempty, we apply the Kakutani fixed point theorem; e.g., see Kakutani (1941) and Border (1985), so we state it here.

THEOREM EC.1. (*Kakutani fixed point theorem*) *If S is a non-empty compact and convex subset of some Euclidean space $\mathbb{R}^d$ and $\psi : S \to 2^S$ is a set-valued function with a closed graph such that $\psi(x)$ is non-empty and convex for all $x \in S$, then the map ψ has a fixed point, i.e., there exists $x \in S$ such that $x \in \psi(x)$.*

In order to be able to work within the Euclidean space $\mathbb{R}^d$, we restrict attention to sets of probability measures with finite support in the compact set $\mathcal{C}$ in $[0, M_a]$; each such subset is homeomorphic to a convex compact subset of $\mathbb{R}^n$. We use an asymptotic argument to get the entire set $\mathcal{P}_{a,2}^{(c)}(M_a)$ when the initial support set is infinite. If indeed the initial support set is infinite,

then for $k \geq 3$, let $\mathcal{C}_k$ be a support set with k points within the initial support set $\mathcal{C}$. Let the sets be nested with that $C_k \subset C_{k+1}$ and $\cup_{k=1}^{\infty} \mathcal{C}_k$ being dense in $\mathcal{C}$. Hence, we can apply the Kakutani fixed point theorem to show that the set of fixed points $\mathcal{P}_{a,2}^*(M_a)$ in (EC.2) is nonempty when we restrict F to C_k.

To apply the Kakutani fixed point theorem in Theorem EC.1, we let ψ in Theorem EC.1 be η, where $\eta(\hat{F})$ is the set of all maximizers of $\zeta(\hat{F})$ in (EC.1) restricted to the subset $\mathcal{P}_{a,2}^{(c)}(M_a)$. Thus, we need to show that $\eta(\hat{F})$ has a closed graph and that $\eta(\hat{F})$ is nonempty and convex for each $\hat{F}$. Recall that a set-valued function ψ is said to have a closed graph (or be upper-hemicontinuous) if for all sequences $\{(x_n, y_n) : n \geq 1\}$ such that $y_n \in \psi(x_n)$ for all n, $x_n \to x$ and $y_n \to y$, we also have $y \in \psi(x)$.

To show that η in (EC.1) has a closed graph, we apply the Berge maximum theorem, e.g., Berge (1963), a version of which we state here.

THEOREM EC.2. (*Berge maximum theorem*) *Let S be a compact metric spaces; let $w : S \times S \to \mathbb{R}$ be a continuous function; let $w^{\uparrow}(x_1) \equiv \sup\{w(x_1, x_2) : x_2 \in S\}$; and let $\eta : S \to 2^S$ be the set of $x_2 \in S$ such that $w(x_1, x_2) = w^{\uparrow}(x_1)$. Then η has a closed graph (is upper-hemicontinuous), $\eta(x_1)$ is nonempty, compact and $w^{\uparrow} : S \to \mathbb{R}$ is continuous.*

.

To establish the continuity condition in our context, we use the continuity of the mean steady-state waiting time as a function of the interarrival-time cdf F within the set $\mathcal{P}_{a,2}^{(c)}(M_a)$ with specified finite first two moments, see §X.6 of Asmussen (2003).

It remains to show that $\eta(\hat{F})$ is convex for each $\hat{F}$ when $\eta(\hat{F})$ is the set of all maximizers of $\zeta(\hat{F})$ in (EC.1), but that convexity follows from the linearity in F_0 of the integral in (17). The set $\eta(F)$ is also nonempty because we are maximizing a continuous function over a compact metric space.

To complete the proof, we need to do an asymptotic argument. We need to go beyond the case of finite support. Hence, for each $k \geq 2$, let $\hat{F}(k)$ be a fixed point cdf satisfying the conditions of Theorem 3 with support $\mathcal{C}_k$. Since all these cdf's have common finite first two moments, the

sequence $\{\hat{F}(k) : k \geq 2\}$ is necessarily tight, so that there exists a subsequence $\{\hat{F}(k_j) : j \geq 1\}$ such that $\hat{F}(k_j) \Rightarrow \hat{F}^*$ as $j \to \infty$; see §11.6 of Whitt (2002). Moreover, since the cdf's have finite second moments, we have convergence of the associated steady-state waiting times $W_{k_j} \Rightarrow W^*$ and moments $E[W_{k_j}] \to E[W^*]$ as $j \to \infty$, again by virtue of §X.6 of Asmussen (2003). The limit then yields the desired fixed point in $\mathcal{P}^{(c)}_{a,2}(M_a)$.

EC.4. Proof of Theorem 5

In this section we prove Theorem 5, which provides an UB for $E[W]$ in the conjectured $F^{(2)}_{1+c_a^2}/G^{(2)}_\infty/1$ extremal $GI/GI/1$ queue. The notation $G^{(2)}_\infty$ means the limit of $G^{(2)}_{M_s}$ as $M_s \to \infty$.

Following §10 of Daley et al. (1992), we concentrate on the class $\mathcal{P}_{a,2} \times \mathcal{P}_{s,2}$ and attempt to determine the best choices of functions $a(\rho), b(\rho)$ such that

$$E[W] \leq \frac{a(\rho)c_a^2 + b(\rho)c_s^2}{2(1-\rho)}. \tag{EC.3}$$

We apply Delay's decomposition in the subsequent Theorem EC.3 to $\lim_{M_s\to\infty} E[W(F, G^{(2)}_{M_s})]$ to obtain

$$\lim_{M_s\to\infty} E[W(F, G^{(2)}_{M_s})] = E[W(F, D)] + \lim_{M_s\to\infty} E[W(D, G^{(2)}_{M_s})] = E[W(F, D)] + \frac{c_s^2}{2(1-\rho)}. \tag{EC.4}$$

Consequently, $b(\rho) \geq b_{LB}(\rho) = 1$. From (EC.4), the lower bound of $a(\rho)$ can be given by

$$a(\rho) \geq a_{LB}(\rho) = \inf_{c_a^2>0}\{\frac{2(1-\rho)}{c_a^2} \sup_{F\in\mathcal{P}_{a,2}} E[W(F, D)]\}. \tag{EC.5}$$

The $a_{LB}(\rho)$ is the best choice (if it exists) when set $b(\rho) = 1$. The $a_{LB}(\rho)$ and $b_{LB}(\rho)$ can give a new upper bound for $GI/GI/1$, so that we obtain

$$E[W(F, G)] \leq E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_\infty)] \leq \frac{a_{LB}(\rho)c_a^2 + c_s^2}{2(1-\rho)} \leq \frac{a(\rho)c_a^2 + b(\rho)c_s^2}{2(1-\rho)}. \tag{EC.6}$$

Now we are left to determine the $a_{LB}(\rho)$. At this point we focus on the candidate bounding system $F^{(2)}_{1+c_a^2}/GI/1$, so we obtain a proof only for this case. We obtain an alternative representation in Chen and Whitt (2018), which we state here. In particular, we can convert the queue $F^{(2)}_{1+c_a^2}/GI/1$ into $D/RS(V,p)/1$ where $RS(V,p) = \sum_{k=1}^{N(p)} V_k$ is a random sum of i.i.d. variables distributed as V, $N(p)$ is a geometric random variable on the positive integers having $E[(N(p))] = 1/p$ with $1/p = 1 + c_a^2$. Here is the specific lemma:

LEMMA EC.1. (*Theorem 1 in Chen and Whitt (2018)*) *For the* $F^{(2)}_{1+c_a^2}/GI/1$ *model with service time* V *having mean* ρ *and scv* c_s^2*, the mean steady-state waiting time can be expressed as*

$$\begin{aligned} E[W(F^{(2)}_{1+c_a^2}(p)/GI/1)] &= E[W(D(1/p)/RS(V,p)/1)] + (E[N(p)]-1)E[V] \\ &= E[W(D(1/p)/RS(V,p)/1)] + \rho(1-p)/p \\ &= E[W(D(1/p)/RS(V,p)/1)] + \rho c_a^2. \end{aligned} \tag{EC.7}$$

Proof. The $F^{(2)}_{1+c_a^2}$ interarrival time means that a random number of arrivals, distributed as $N(p)$, arrive at deterministic intervals with deterministic value $1/p = c_a^2 + 1$. So the model has batch arrivals. The result in (EC.7) follows from Halfin (1983) or Theorem 1 of Whitt (1983a), which states that the delay of an arbitrary customer in the batch is distributed the same as the delay of the last customer in the batch when the batch-size distribution is geometric. Because $E[W(D(1/p)/RS(V,p)/1)]$ is the expected delay of the first customer in a batch, we need to add the second term in (EC.7) to get the delay of the last customer in the batch; e.g., see §III of Whitt (1983a). ∎

Hence, we apply Lemma EC.1 to write

$$E[W(F^{(2)}_{1+c_a^2}, G)] = E[W(D, RS(V,p))] + \rho c_a^2. \tag{EC.8}$$

For the rest, we use a stochastic comparison argument involving convex stochastic order, as in §9.5 of Ross (1996) or in §1.7 and Chapter 5 of Stoyan (1983). Let convex order be denoted by $\le_c$. In particular, consider an $F^{(2)}_{1+c_a^2}/GI/1$ system for which $S \le_c S'$ where S' denotes a exponential random variable with mean $E[S]$. Then for two sequences of i.i.d. variables $\{S_n\}$ and $\{S'_n\}$,

$$S_1 + ... + S_{N(p)} \le_c S'_1 + ... + S'_{N(p)}. \tag{EC.9}$$

However, the righthand side is distributed as an exponential random variable with mean $N(p)E[S]$, where $N(p)$ is a geometric random variable with mean $E[N(p)] = 1 + c_a^2$. Hence, we obtain

$$(S_1 + ... + S_{N(p)})/E[N(p)] \le_c S'. \tag{EC.10}$$

Consequently,

$$\begin{aligned}(1+c_a^2)^{-1}W(D,RS(V,p)) &=_d W((1+c_a^2)D, S_1+...+S_{N(p)}) \\ &=_d W(D,(S_1+..+S_{N(p)})/(1+c_a^2)) \\ &\le_c W(D,S') = W(D,M). \end{aligned} \tag{EC.11}$$

Hence,

$$(1+c_a^2)^{-1}E[W(D,RS(V,p))] \le EW[(D,M)] = \delta\rho/(1-\delta). \tag{EC.12}$$

where $\delta = \exp(-(1-\delta)/\rho)$.

Finally, combine (EC.5), (EC.8) and (EC.12) to obtain

$$\begin{aligned} a_{LB}(\rho) &= \inf_{c_a^2>0} \frac{2(1-\rho)\sup_{F\in\mathcal{P}_{a,2}} E[W(F,D)]}{c_a^2} \\ &= \inf_{c_a^2>0} \frac{2(1-\rho)E[W(F^{(2)}_{1+c_a^2},D)]}{c_a^2} \le \inf_{c_a^2>0}\{2\rho(1-\rho) + \frac{(1+c_a^2)\delta\rho/(1-\delta)2(1-\rho)}{c_a^2}\} \\ &\to \frac{\rho(2-2\rho)}{1-\delta} (\text{as } c_a^2\to\infty). \end{aligned} \tag{EC.13}$$

So $a_{LB}(\rho) \le \rho(2-2\rho)/(1-\delta)$ and

$$E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_\infty)] \le \frac{a_{LB}(\rho)c_a^2 + c_s^2}{2(1-\rho)} \le \frac{2(1-\rho)\rho/(1-\delta)c_a^2 + \rho^2 c_s^2}{2(1-\rho)}. \tag{EC.14}$$

■

EC.5. Extension to Unbounded Support

In this section we discuss what happens when we increase the intervals of support $[0, M_a]$ and $[0, \rho M_s]$. Throughout this section we assume that the UB for finite support has been shown to be $(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})$. We ask what happens as we let $M_a \to \infty$ and $M_s \to \infty$.

EC.5.1. Unbounded Support for the Interarrival Time

First, for the interarrival-time cdf F, the cdf $F^{(2)}_{1+c_a^2}$ is optimal for the UB for all M_a, and thus remains optimal as $M_a \to \infty$. In contrast, for the lower bound, which we mostly do not discuss here, the extremal interarrival-time cdf is $F^{(2)}_{M_a}$, which places positive mass on M_a. Then the extremal

interarrival-time cdf $F^{(2)}_{M_a}$ converges to the deterministic distribution with mean 1 as $M_a \to \infty$, which of course has $c_a^2 = 0$, which is likely to be inconsistent with the specified parameter. Nevertheless, the mean waiting time converges to the value $E[W(D,G)]$ of the associated $D/GI/1$ model, as we saw in Tables EC.4-EC.5. Moreover, as discussed in Theorem 3.1 of Daley et al. (1992), that yields the well-known tight LB.

EC.5.2. Unbounded Support for the Service Time

The situation is more complicated when we let $M_s \to \infty$ for the upper bound. Just as for the interarrival-time cdf $F^{(2)}_{M_a}$, the service-time cdf $G^{(2)}_{M_s}$ converges to the deterministic cdf with the mean ρ of $G^{(2)}_{M_s}$ as $M_s \to \infty$. However, the mean waiting time fails to converge to the mean waiting time of the associated $GI/D/1$ queue.

We propose two approaches to this problem. The first way is to exploit the representation in terms of the idle time in (44), as was done in Minh and Sorli (1983) and Wolff and Wang (2003). It turns out that the mean idle time does converge as $M_s \to \infty$. We discuss this approach in Chen and Whitt (2018). The second approach is to exploit the Daley decomposition from §10 of Daley et al. (1992).

EC.5.3. The Daley Decomposition

We now discuss a decomposition for the mean steady-state waiting time $E[W]$ in §10 of Daley et al. (1992). The decomposition appears in equation (10.2) of Daley et al. (1992), where it is attributed to unpublished by D. J. Daley in 1984. We state it in the following theorem. Let $G^{(2)}_\infty$ be shorthand for the limit $E[W(F, G^{(2)}_{M_s})]$ as $M_s \to \infty$ and let D_m denote a deterministic cdf with mass 1 on m.

THEOREM EC.3. (*the Daley decomposition in* (10.2) *of* Daley et al. (1992)) *Consider the* $GI/GI/1$ *model with specified interarrival-time cdf* $F \in \mathcal{P}_{a,2}$. *As* $M_s \to \infty$,

$$\begin{aligned} E[W(F,G^{(2)}_\infty)] \equiv \lim_{M_s\to\infty} E[W(F,G^{(2)}_{M_s})] &= E[W(F,D_\rho)] + E[W(D_1,G^{(2)}_\infty)] \\ &= E[W(F,D_\rho)] + \frac{\rho^2 c_s^2}{2(1-\rho)}. \end{aligned} \tag{EC.15}$$

Proof. We only give a brief overview. We do a regenerative analysis to compute the mean waiting time, looking at successive busy cycles starting empty. We exploit the classic result that the steady-state mean waiting time is the expected sum of the waiting times over one cycle divided by the expected length of one cycle; e.g., see §3.6 and §3.7 of Ross (1996).

As M_s increases, the two-point cdf $G^{(2)}_{M_s}$ necessarily places probability of order $O(1/M_s^2)$ on M_s and the rest of the mass on a point just less than the mean service time, ρ. For very large M_s, there will be only rarely, with probability of order $O(1/M_s^2)$, a large service time of order $O(M_s)$. In the limit, most customers never encounter this large service time, so that we get a contribution to the overall mean $E[W]$ corresponding to $E[W(F, D_\rho)]$ in the first term on the right in (EC.15).

On the other hand, the total impact of the very large waiting time of order M_s is roughly the area of the triangle with height $O(M_s)$ and width $O(M_s)$, which itself is $O(M_s^2)$. When combined with the $O(1/M_s^2)$ probability, this produces an additional $O(1)$ impact on the steady-state mean, which is given by the second term on the right in (EC.15). Moreover, because we can use a law-of-large-numbers argument to treat this large service time, the asymptotic impact of that large service time is independent of the interarrival-time cdf beyond its mean, so we can substitute D_1 for the original interarrival-time cdf F with mean 1 in the second term. ∎

If Conjecture 1 holds, then

$$\sup\{E[W(F,D)] : F \in \mathcal{P}_{a,2}\} = E[W(F^{(2)}_{1+c_a^2}, D)]. \tag{EC.16}$$

Hence, we can apply Theorem EC.3 to show that, if Conjecture 1 (b) holds, then, for all $F \in \mathcal{P}_{a,2}$ and $G \in \mathcal{P}_{s,2}$,

$$\begin{aligned} E[W(F,G)] &\le \lim_{M_s\to\infty} E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{M_s})] \equiv E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{\infty})] \\ &\le E[W(F^{(2)}_{1+c_a^2}, D_\rho)] + \frac{\rho^2 c_s^2}{2(1-\rho)}. \end{aligned} \tag{EC.17}$$

EC.6. More on the Lower Bound with Finite Support

In this section we elaborate on our investigation of the lower bound with finite support, which was briefly discussed in §7.

The new $G^{(3)}_{u,b_s u}$ makes the $F^{(2)}_{M_a}/G^{(3)}_{u,b_s u}/1$ model lattice except for the mass at M_a. If the parameter b_s is chosen as a integer value which is greater than 1, then

$$\lim_{M_a\to\infty} E[W(F^{(2)}_{M_a}, G^{(3)}_{u,b_s u})] = E[W(D, G^{(3)}_a)] \tag{EC.18}$$

which is the tight lower bound of $GI/GI/1$ models over $\mathcal{P}_{a,2}\times\mathcal{P}_{s,2}$.

In previous extensive numerical studies we find that $F^{(2)}_{M_a}$ is good for F, but $G^{(2)}_{1+c_s^2}$ and $G^{(2)}_{M_s}$ might not be nearly optimal for G to minimize the mean waiting time. Moreover, Figure 2 shows$G^{(2)}_{1+c_s^2}$ is the optimal solution to minimize $E[W(F^{(2)}_{M_a}, G)]$ over $\mathcal{P}^{(c)}_{s,2,2}(M_s)$ only for $M_a = 1+c_a^2$. Thus it is interesting to explore better service time distribution when $F = F^{(2)}_{M_a}$ for $M_a > 1+c_a^2$.

EC.6.1. The $G^{(3)}_{u,b_s u}$ Service-Time Distribution

To derive the closed form of $G^{(3)}_{u,b_s u}$, we next solve the moment equations with mass at $x_1 = 0, x_2 = u, x_3 = b_s u$ with $b_s > 1$ and $u > 0$ (recall $u = 1 - c_a^2/(M_a - 1)$),

$$p_1 + p_2 + p_3 = 1, x_1 p_1 + x_2 p_2 + x_3 p_3 = \rho, x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 = (1+c_s^2)\rho^2 \tag{EC.19}$$

to obtain a solution as a function of the single variable b_s. Note the $G^{(3)}_{u,b_s u}$ has no definition for $u = 0$. The probabilities of the points in $\{0, u, b_s u\}$ are then

$$\begin{aligned} p_1 &= \frac{(b_s^2(u^2-\rho u) + b_s(-u^2 + (1+c_s^2)\rho^2) - (1+c_s^2)\rho^2 + u\rho)}{(b_s^2 u^2 - b_s u^2)}, \\ p_2 &= \frac{\rho b_s u - (1+c_s^2)\rho^2)}{b_s u^2 - u^2} \quad \text{and} \quad p_3 = \frac{\rho^2(1+c_s^2) - u\rho}{b_s^2 u^2 - b_s u^2}. \end{aligned} \tag{EC.20}$$

It remains to specify b_s. To do so, we conducted extensive simulation experiments. Based on these experiments, we find that the possible values of b_s depend on $E[V] = \rho$. In particular, if $\rho \in (u/(1+c_s^2), u]$, $b_s \in [(1+c_s^2)\rho/u, \infty)$. When $b_s = (1+c_s^2)\rho/u$, then $G^{(3)}_{u,b_s u} = G^{(2)}_{1+c_s^2}$. If $\rho = u/(1+c_s^2)$, then $G^{(3)}_{u,b_s u}$ is a two-point distribution with mass at $\{0, u\}$. Since inter-arrival time distribution $F^{(2)}_{M_a}$ has mass at $\{u, M_a\}$ and there is no large service time impact, $E[W(F^{(2)}_{M_a}, G^{(3)}_{u,b_s u})] = 0$. If $\rho \in (u, 1)$, then there exists a positive value $\gamma > 0$ which is the largest root of the quadratic equation in b_s

$$b_s^2(u^2 - \rho u) + b_s(-u^2 + (1+c_s^2)\rho^2) - (1+c_s^2)\rho^2 + u\rho = 0, \tag{EC.21}$$

such that $b_s \in [(1+c_s^2)\rho/u, \gamma)$. Therefore, the possible range of b_s depends on ρ. In general,

$$b_s \in \left[\frac{(1+c_s^2)\rho}{u}, \mathbf{1}_{\{\rho\in(u/(1+c_s^2),u]\}}\infty + \mathbf{1}_{\{\rho>u\}}\gamma\right). \tag{EC.22}$$

EC.6.2. The Impact of Service Time in $F_{M_a}^{(2)}/G_{u,b_su}^{(3)}/1$

We study the impact of b_s to $E[W(F_{M_a}^{(2)}, G_{u,b_su}^{(3)})]$ and seek for optimal b_s^* in (EC.22) to minimize $E[W(F_{M_a}^{(2)}, G_{u,b_su}^{(3)}]$ by Minh and Sorli (1983) simulation with $T = 1\times 10^7$ and 20 i.i.d replications. Following the range of b_s in EC.22, we simulate the model under $M_a = 6, 8, 10$ and various settings of b_s ($\gamma- = \gamma - 0.0001$. For example, γ^- is 19.167 when $M_a = 6$ by some simple calculation.).

Table EC.1 Simulation estimates of $E[W(F_{M_a}^{(2)}, G_{u,b_su}^{(3)})]$ under the case $c_a^2 = c_s^2 = 4, \rho = 0.5$

b_s	13	14	15	16	17	18	19	γ^-	γ^-	γ^-	γ^-
$M_a=6$	3.01	2.95	2.89	2.82	2.76	2.72	2.67	2.66	2.66	2.66	2.66
b_s	10	12	14	16	18	20	22	24	26	28	30
$M_a=8$	2.36	2.22	2.10	1.98	1.85	1.73	1.69	1.68	1.65	1.61	1.58
b_s	10	12	14	16	18	20	22	24	26	28	30
$M_a=10$	1.97	1.87	1.78	1.70	1.61	1.53	1.48	1.44	1.41	1.39	1.37

From the above simulation, we see the $E[W(F_{M_a}^{(2)}, G_{u,b_su}^{(3)})]$ is monotone decreasing with b_s increasing. Thus the optimal $b_s^* = \gamma^-$ when $\rho > u$ or $b_s^* = \infty$ when $\rho \in (u/(1+c_s^2), u]$.

EC.6.3. Simulation Comparisons

From extensive simulation experiments, we conclude that the LB for $E[W]$ is attained, at least approximately, by the $F_{M_a}^{(2)}/G_{u,b_su}^{(3)}/1$ model. Following from Figure 1 and 2, we see there exists an optimal $b_s^*(b_a)$ such that the lower bound of $E[W]$ is attained by $E[W(F_{M_a}^{(2)}, G_{b_s}^{(2)})]$ over $\mathcal{P}_{a,2,2}^{(c)}(M_a)\times \mathcal{P}_{s,2,2}^{(c)}(M_s)$. Since the mean of $F_{M_a}^{(2)}/G_{u,b_su}^{(3)}/1$ is monotone decreasing as b_s increases, we set b_s sufficiently large for $F_{M_a}^{(2)}/G_{u,b_su}^{(3)}/1$ and set the optimal $b_s^*(b_a)$ for $F_{M_a}^{(2)}/G_{b_s}^{(2)}/1$ to make a careful simulation comparison under the case $c_a^2 = c_s^2 = 4$ under different settings of b_a.

Table EC.6.3 shows the results for the $E[W(F^{(2)}_{M_a}, G^{(2)}_{b_s})]$ under optimal b^*_s within $[0, M_s]$ ($M_s = 1000$). We compare it to Ott's lower bound, the HTA and conjectured UB and UB Approx.

Table EC.2 Simulation performance of lower bound with different settings of M_a for the model $F^{(2)}_{M_a}/G^{(2)}_{b_s}/1$ ($T = 5 \times 10^8$ and 20 i.i.d replications)

ρ	Ott LB	$M_a = 20$	$M_a = 10$	$M_a = 8$	$M_a = 6$	HTA	Tight UB	UB Approx
0.30	0.107	0.261	0.262	0.307	0.815	0.514	1.50	1.51
0.50	0.750	1.01	1.02	1.70	2.68	2.00	3.47	3.51
0.70	2.92	3.33	6.34	6.95	7.76	6.53	8.44	8.52
0.90	15.8	29.1	33.0	33.5	34.1	72.2	74.6	74.8

We study the simulation performance of $E[W(F^{(2)}_{M_a}, G^{(3)}_{u,b_s u})]$ under optimal $b^*_s = \min\{1000, \gamma - 0.0001\}$ by Minh and Sorli (1983) algorithm with simulation length $T = 5 \times 10^8$ and 20 independent repetitive experiments.

Table EC.3 Simulation performance of lower bound with different settings of M_a for the model $F^{(2)}_{M_a}/G^{(3)}_{u,b_s u}/1$ ($T = 1 \times 10^7$ and 20 i.i.d replications)

ρ	Ott LB	$M_a = 20$	$M_a = 10$	$M_a = 8$	$M_a = 6$	HTA	Tight UB	UB Approx
0.30	0.107	0.151	0.203	0.230	0.685	0.514	1.50	1.51
0.50	0.750	0.857	0.973	1.50	2.66	2.00	3.47	3.51
0.70	2.92	3.17	5.56	6.33	7.56	6.53	8.44	8.52
0.90	15.8	27.2	31.8	32.7	33.7	72.2	74.6	74.8

Therefore, we conclude by stating a conjecture associated with lower bound.

CONJECTURE EC.1. *Given any parameter vector* $(1, c_a^2, \rho, c_s^2)$ *and a bounded interval* $[0, M_a]$ *for the interarrival-time cdf* F*, the pair* $(F^{(2)}_{M_a}, G^{(3)}_{u,b_s u})$ *attains the tight LB of the steady-state mean* $E[W]$ *for* $M_a > 1 + c_a^2$*, i.e.,*

$$E[W(F,G)] \geq E[W(F^{(2)}_{M_a}, G^{(3)}_{u,b_s u})] \quad \textit{for all} \quad F \in \mathcal{P}^{(c)}_{a,2}(M_a) \quad \textit{and} \quad G \in \mathcal{P}_{s,2}. \tag{EC.23}$$

If $M_a = 1 + c_a^2$, *the pair* $(F^{(2)}_{1+c_a^2}, G^{(2)}_{1+c_s^2})$ *attains the tight LB of the steady-state mean* $E[W]$, *i.e.,*

$$E[W(F,G)] \geq E[W(F^{(2)}_{1+c_a^2}, G^{(2)}_{1+c_s^2})] \quad \textit{for all} \quad F \in \mathcal{P}^{(c)}_{a,2}(M_a) \quad \textit{and} \quad G \in \mathcal{P}_{s,2}. \tag{EC.24}$$

EC.7. Numerical Comparison of the Bounds and Approximations

We now supplement Tables 1 and 2 by making numerical comparisons for both the scaled means $(1-\rho)E[W]/\rho^2$ and the unscaled means $E[W]$ for 12 values of ρ in the four cases: $(c_a^2, c_s^2) = (4.0, 4.0)$, $(0.5, 0.5)$, $(4.0, 0.5)$, $(0.5, 4.0)$. Tables EC.4-EC.7 present the scaled values, while Tables EC.8-EC.11 then present the corresponding unscaled values.

Table EC.4 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = c_s^2 = 4.0$.

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.10	0.000	4.000	38.001	38.002	0.000	0.00%	40.000	202.000
0.20	0.000	4.000	18.078	18.112	0.007	0.19%	20.000	52.000
0.30	0.833	4.000	11.661	11.731	0.041	0.60%	13.333	24.222
0.40	1.250	4.000	8.640	8.722	0.107	0.94%	10.000	14.500
0.50	1.500	4.000	6.940	7.020	0.203	1.15%	8.000	10.000
0.60	1.667	4.000	5.883	5.946	0.324	1.07%	6.667	7.556
0.70	1.786	4.000	5.168	5.216	0.467	0.93%	5.714	6.082
0.80	1.875	4.000	4.662	4.693	0.629	0.67%	5.000	5.125
0.90	1.944	4.000	4.287	4.302	0.807	0.35%	4.444	4.469
0.95	1.974	4.000	4.134	4.142	0.902	0.18%	4.211	4.216
0.98	1.990	4.000	4.052	4.055	0.960	0.07%	4.082	4.082
0.99	1.995	4.000	4.025	4.027	0.980	0.04%	4.040	4.041

Table EC.5 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = c_s^2 = 0.5$.

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.10	0.000	0.500	4.750	4.750	0.000	0.00%	5.000	25.250
0.20	0.000	0.500	2.252	2.264	0.007	0.54%	2.500	6.500
0.30	0.000	0.500	1.432	1.466	0.041	2.36%	1.667	3.028
0.40	0.000	0.500	1.049	1.090	0.107	3.82%	1.250	1.813
0.50	0.000	0.500	0.827	0.878	0.203	5.72%	1.000	1.250
0.60	0.000	0.500	0.708	0.743	0.324	4.71%	0.833	0.944
0.70	0.036	0.500	0.623	0.652	0.467	4.53%	0.714	0.760
0.80	0.125	0.500	0.569	0.587	0.629	2.95%	0.625	0.641
0.90	0.194	0.500	0.530	0.538	0.807	1.38%	0.556	0.559
0.95	0.224	0.500	0.514	0.518	0.902	0.65%	0.526	0.527
0.98	0.240	0.500	0.505	0.507	0.960	0.27%	0.510	0.510
0.99	0.245	0.500	0.503	0.503	0.980	0.14%	0.505	0.505

Table EC.6 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = 4.0$ and $c_s^2 = 0.5$

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.10	0.000	2.250	36.251	36.252	0.000	0.00%	38.250	200.250
0.20	0.000	2.250	16.328	16.362	0.007	0.21%	18.250	50.250
0.30	0.000	2.250	9.911	9.981	0.041	0.71%	11.583	22.472
0.40	0.000	2.250	6.890	6.972	0.107	1.16%	8.250	12.750
0.50	0.000	2.250	5.190	5.270	0.203	1.51%	6.250	8.250
0.60	0.000	2.250	4.133	4.196	0.324	1.50%	4.917	5.806
0.70	0.036	2.250	3.418	3.466	0.467	1.39%	3.964	4.332
0.80	0.125	2.250	2.912	2.943	0.629	1.06%	3.250	3.375
0.90	0.194	2.250	2.537	2.552	0.807	0.59%	2.694	2.719
0.95	0.224	2.250	2.384	2.392	0.902	0.31%	2.461	2.466
0.98	0.240	2.250	2.301	2.305	0.960	0.17%	2.332	2.332
0.99	0.245	2.250	2.275	2.277	0.980	0.09%	2.290	2.291

Table EC.7 A comparison of the bounds and approximations for the scaled steady-state mean $(1-\rho)E[W]/\rho^2$ in the $GI/GI/1$ model as a function of ρ for the case $c_a^2 = 0.5$ and $c_s^2 = 4.0$

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.10	0.000	2.250	6.500	6.500	0.000	0.00%	6.750	27.000
0.20	0.000	2.250	4.002	4.014	0.007	0.30%	4.250	8.250
0.30	0.833	2.250	3.182	3.216	0.041	1.08%	3.417	4.778
0.40	1.250	2.250	2.799	2.840	0.107	1.47%	3.000	3.563
0.50	1.500	2.250	2.577	2.628	0.203	1.91%	2.750	3.000
0.60	1.667	2.250	2.458	2.493	0.324	1.40%	2.583	2.694
0.70	1.786	2.250	2.373	2.402	0.467	1.23%	2.464	2.510
0.80	1.875	2.250	2.319	2.337	0.629	0.74%	2.375	2.391
0.90	1.944	2.250	2.280	2.288	0.807	0.32%	2.306	2.309
0.95	1.974	2.250	2.264	2.268	0.902	0.15%	2.276	2.277
0.98	1.990	2.250	2.255	2.257	0.960	0.06%	2.260	2.260
0.99	1.995	2.250	2.253	2.253	0.980	0.03%	2.255	2.255

Table EC.8 A comparison of the unscaled bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = c_s^2 = 4.0$

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.10	0.000	0.044	0.422	0.422	0.000	0.00%	0.444	2.244
0.20	0.000	0.200	0.904	0.906	0.007	0.19%	1.000	2.600
0.30	0.107	0.514	1.499	1.508	0.041	0.60%	1.714	3.114
0.40	0.333	1.067	2.304	2.326	0.107	0.94%	2.667	3.867
0.50	0.750	2.000	3.470	3.510	0.203	1.15%	4.000	5.000
0.60	1.500	3.600	5.295	5.352	0.324	1.07%	6.000	6.800
0.70	2.917	6.533	8.441	8.520	0.467	0.93%	9.333	9.933
0.80	6.000	12.800	14.917	15.017	0.629	0.67%	16.000	16.400
0.90	15.750	32.400	34.721	34.843	0.807	0.35%	36.000	36.200
0.95	35.625	72.200	74.621	74.755	0.902	0.18%	76.000	76.100
0.98	95.550	192.080	194.557	194.702	0.960	0.07%	196.000	196.040
0.99	195.525	392.040	394.533	394.684	0.980	0.04%	396.000	396.020

Table EC.9 A comparison of the unscaled bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = c_s^2 = 0.5$

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.10	0.000	0.006	0.053	0.053	0.000	0.00%	0.056	0.281
0.20	0.000	0.025	0.113	0.113	0.007	0.54%	0.125	0.325
0.30	0.000	0.064	0.184	0.189	0.041	2.36%	0.214	0.389
0.40	0.000	0.133	0.280	0.291	0.107	3.82%	0.333	0.483
0.50	0.000	0.250	0.414	0.439	0.203	5.72%	0.500	0.625
0.60	0.000	0.450	0.637	0.669	0.324	4.71%	0.750	0.850
0.70	0.058	0.817	1.017	1.065	0.467	4.53%	1.167	1.242
0.80	0.400	1.600	1.822	1.877	0.629	2.95%	2.000	2.050
0.90	1.575	4.050	4.295	4.355	0.807	1.38%	4.500	4.525
0.95	4.037	9.025	9.284	9.344	0.902	0.65%	9.500	9.512
0.98	11.515	24.010	24.271	24.338	0.960	0.27%	24.500	24.505
0.99	24.008	49.005	49.265	49.336	0.980	0.14%	49.500	49.503

Table EC.10 A comparison of the unscaled bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = 4.0$ and $c_s^2 = 0.5$

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.10	0.000	0.025	0.403	0.403	0.000	0.00%	0.425	2.225
0.20	0.000	0.113	0.816	0.818	0.007	0.21%	0.913	2.513
0.30	0.000	0.289	1.274	1.283	0.041	0.71%	1.489	2.889
0.40	0.000	0.600	1.837	1.859	0.107	1.16%	2.200	3.400
0.50	0.000	1.125	2.595	2.635	0.203	1.51%	3.125	4.125
0.60	0.000	2.025	3.720	3.777	0.324	1.50%	4.425	5.225
0.70	0.058	3.675	5.583	5.662	0.467	1.39%	6.475	7.075
0.80	0.400	7.200	9.317	9.417	0.629	1.06%	10.400	10.800
0.90	1.575	18.225	20.546	20.668	0.807	0.59%	21.825	22.025
0.95	4.037	40.613	43.033	43.168	0.902	0.31%	44.413	44.513
0.98	11.515	108.045	110.479	110.667	0.960	0.17%	111.965	112.005
0.99	24.008	220.523	222.971	223.167	0.980	0.09%	224.483	224.503

Table EC.11 A comparison of the unscaled bounds and approximations for the steady-state mean $E[W]$ as a function of ρ for the case $c_a^2 = 0.5$ and $c_s^2 = 4.0$

ρ	Tight LB (8)	HTA (5)	Tight UB $F^{(2)}_{1+c_a^2}/G^{(2)}_{\infty}$	conj UB (43)	δ (43)	MRE	Daley (7)	Kingman (6)
0.10	0.000	0.025	0.072	0.072	0.000	0.00%	0.075	0.300
0.20	0.000	0.113	0.200	0.201	0.007	0.30%	0.213	0.413
0.30	0.107	0.289	0.409	0.414	0.041	1.08%	0.439	0.614
0.40	0.333	0.600	0.746	0.757	0.107	1.47%	0.800	0.950
0.50	0.750	1.125	1.289	1.314	0.203	1.91%	1.375	1.500
0.60	1.500	2.025	2.212	2.244	0.324	1.40%	2.325	2.425
0.70	2.917	3.675	3.875	3.923	0.467	1.23%	4.025	4.100
0.80	6.000	7.200	7.422	7.477	0.629	0.74%	7.600	7.650
0.90	15.750	18.225	18.470	18.530	0.807	0.32%	18.675	18.700
0.95	35.625	40.613	40.871	40.932	0.902	0.15%	41.088	41.100
0.98	95.550	108.045	108.307	108.373	0.960	0.06%	108.535	108.540
0.99	195.525	220.523	220.783	220.853	0.980	0.03%	221.018	221.020