Extremal $GI/GI/1$ Queues Given Two Moments: Exploiting Three-Point Distributions Numerically

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

The tight upper bound for the steady-state mean waiting time $E[W]$ in the $GI/GI/1$ queue given the first two moments of the interarrival-time and service-time distributions has recently been shown to be attained by three-point distributions when those distributions have bounded support. In this paper we exploit the three-point representation to develop and apply numerical algorithms to provide evidence that the tight upper bound is attained within this class by special two point distributions. For the conjectured overall upper bound, the two-point distribution for the interarrival time has one mass point at 0, while the two-point distribution for the service-time has one mass at the upper limit. With unbounded support, the service-time distribution involves a limit; there is one mass point at a high value, but that upper mass point must increase to infinity while the probability on that point must decrease to 0 appropriately.

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

History: May 13, 2020

1. Introduction

For the classical $GI/GI/1$ queueing model, we can understand how the level of congestion depends on the model data by looking at the heavy-traffic approximation for the mean steady-state waiting time

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

where the interarrival time has mean 1 and scv (squared coefficient of variation, variance divided by the square of the mean) c_a^2 and the service time has mean ρ and scv c_s^2, so that the traffic intensity is ρ, $0 < \rho < 1$. Formula (1) combines the heavy-traffic limit in Kingman (1961) with the exact Pollaczek-Khintchine formula when the arrival process is a Poisson process, so that $c_a^2 = 1$.

However, it is natural to wonder how accurate is formula (1) given the partial information provided by the parameter vector (ρ, c_a^2, c_s^2). This issue was partly addressed by Kingman (1962) when he established an upper bound given this information (see (4) in §2.2), but that upper bound is not tight. Continued interest in this issue has led to considerable further research over the years; see Bertsimas and Natarajan (2007), Daley et al. (1992), Gupta et al. (2010), Gupta and Osogami (2011) and Wolff and Wang (2003).

It has long been conjectured that the tight upper bound for $E[W]$ in the $GI/GI/1$ queue given the first two moments of the interarrival-time and service-time distributions (which is equivalent to the parameters above) is attained asymptotically by two-point distributions; see Daley et al. (1992), especially §10. and references therein. The conjectured two-point distribution for the interarrival time has one mass point at 0, but the conjectured two-point distribution for the service time involves a limit; there is one mass point at a high value, but that upper mass point must increase to infinity while the probability on that point must decrease to 0 appropriately.

The present paper is a sequel to Chen and Whitt (2020b), in which we obtained a partial result; in particular, we proved that the upper and lower bounds for the steady-state mean $E[W]$ and the associated transient mean $E[W_n|W_0 = 0]$ over distributions with support on bounded intervals are attained by three-point interarrival-time and service-time distributions; see Theorem 1 here. In this paper we exploit the three-point representation to provide numerical evidence supporting the conjecture itself. We develop and apply a nonconvex nonlinear program based on a multinomial representation and simulation algorithms.

In Chen and Whitt (2020a) we also studied the conjectured extremal model with unbounded support. In particular, Theorem 3.1 there established an overall representation in terms of the mean waiting time in the extremal model in terms of a $D/RS(D)/1$ discrete-time model involving a geometric random sum of deterministic random variables (the $RS(D)$), where the two deterministic random variables in the model may have different values, so that the extremal steady-state waiting time need not have a lattice distribution, while Theorem 3.2 established a stochastic upper bound

for the steady-state waiting time (involving convex stochastic order) that yields a remarkably good approximation for the conjectured upper bound for the mean waiting time; see Theorem 2 here. Moreover, we developed algorithms to numerically compute and simulate the conjectured tight upper bound. We applied these algorithms to show that the conjectured upper bound provides a significant improvement over the classic upper bounds by Kingman (1962) and Daley (1977).

Here is how the rest of the paper is organized. In §2 we provide technical background. In §2.3 we state the main reduction result from Chen and Whitt (2020b). In §3 we state the conjectures based on our new numerical results. In §4 we develop a multinomial representation and apply it to formulate a non-convex nonlinear program (NLP) for the overall UB, which we solve by applying sequential quadratic programming (SQP) as discussed in Ch. 18 of Nocedal and Wright (1999). The SQP algorithm converges at a local optimum, so we apply it with randomly selected initial conditions. We found that all local optima for the overall UB are two-point distributions and that the best local optimum always has interarrival-time cdf with one mass at 0. In §5 we do a systematic simulation study of models with two-point distributions. For efficiency, we exploit the representation of $E[W]$ in terms of the idle-time distribution proposed by by Minh and Sorli (1983). These simulations exhibit interesting structure and provide strong support for the conjecture. In §6 we study the lower bound with bounded support, which was not considered previously. Finally, in §7 we draw conclusions.

2. Background

In §2.1 we specify the model and the quantities of interest. In §2.2 we review the classic bounds and approximations. In §2.3 we review the three-point result from Chen and Whitt (2020b).

2.1. 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 to find an empty system, V_n is the service time of customer n, whle U_n is the interarrival time between customers n and $n+1$.

Let U have mean $\mathbb{E}[U] \equiv \lambda^{-1} \equiv 1$ and squared coefficient of variation (scv, variance divided by the square of the mean) c_a^2; let a service time V have mean $\mathbb{E}[V] \equiv \tau \equiv \rho$ and scv c_s^2, where $\rho \equiv \lambda\tau < 1$, 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 empty with $W_0 \equiv 0$. The sequence $\{W_n : n \geq 0\}$ is well known to satisfy the Lindley recursion

$$W_{n+1} = [W_n + V_n - U_n]^+, \quad n \geq 0, \tag{2}$$

where $x^+ \equiv \max\{x, 0\}$. Let W be the steady-state waiting time. It is also well known that $W_n \stackrel{\text{d}}{=} \max\{S_k : 0 \leq k \leq n\}$ and $W \stackrel{\text{d}}{=} \max\{S_k : k \geq 0\}$, where $\stackrel{\text{d}}{=}$ denotes equality in distribution, $S_0 \equiv 0$, $S_k \equiv X_0 + \cdots + X_{k-1}$, $k \geq 1$, and $X_k \equiv V_k - U_k$, $k \geq 0$; e.g., see §§X.1-X.2 of Asmussen (2003) or (13) in §8.5 of Chung (2001). It is also known that, under the specified finite moment conditions, W_n and W are proper random variables with finite means, given by

$$E[W_n] = \sum_{k=1}^{n} \frac{\mathbb{E}[S_k^+]}{k} < \infty \quad \text{and} \quad E[W] = \sum_{k=1}^{\infty} \frac{\mathbb{E}[S_k^+]}{k} < \infty. \tag{3}$$

2.2. Classical Results: Exact, Approximate and Bounds

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

$$E[W] \leq \frac{\rho^2([c_a^2/\rho^2] + c_s^2)}{2(1-\rho)}, \tag{4}$$

which is known to be asymptotically correct in heavy traffic (as $\rho \to 1$).

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

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

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 (1984), Theorem 3.1 of Daley et al. (1992) and references there:

$$E[W(LB)] = \frac{\rho((1+c_s^2)\rho - 1)^+}{2(1-\rho)}. \tag{6}$$

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.)

2.3. Reduction to Three-Point Distributions

We now review Theorem 1 of Chen and Whitt (2020b), which establishes the reduction to three-point distributions. For that purpose, let $\mathcal{P}_n$ be the set of all probability measures on a subset of $\mathbb{R}$ with specified first n moments. 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 interarrival and service times, respectively. If a support bound M is omitted, i.e., if we write $\mathcal{P}_{a,2} \equiv \mathcal{P}_{a,2}(1, c_a^2)$, then the support is understood to be $[0, \infty)$. Let $\mathcal{P}_{a,2,k}(M)$ denote the subset with support on at most k points within bounded $[0, M]$. We use this definition for both the cdf's we consider: F of U and G of V, but recall that our parameter specification with $\mathbb{E}[U] = 1$ makes the support of F be $[0, M_a]$, while the support of G is $[0, \rho M_s]$.

We consider the mean waiting time $E[W_n]$ for $1 \le n \le \infty$ expressed as a mapping of the underlying distributions; i.e., let

$$w_n : \mathcal{P}_{a,2}(1, c_a^2) \times \mathcal{P}_{s,2}(\rho, c_s^2) \to [0, \infty), \tag{7}$$

where $0 < \rho < 1$ and

$$w_n(F, G) \equiv E[W_n(F, G)], \quad 1 \le n \le \infty, \tag{8}$$

in the $GI/GI/1$ queue with interarrival-time cdf $F \in \mathcal{P}_{a,2}(1, c_a^2)$ and service-time cdf $G \in \mathcal{P}_{s,2}(\rho, c_s^2)$, as given explicitly in (3).

THEOREM 1. (*reduction to a three-point distribution (Theorem 1 of Chen and Whitt (2020b))*) *Consider the class of $GI/GI/1$ queues with $W_0 = 0$, $F \in \mathcal{P}_{a,2} \equiv \mathcal{P}_{a,2}(1, c_a^2)$, $G \in \mathcal{P}_{s,2} \equiv \mathcal{P}_{s,2}(\rho, c_s^2)$, $0 < \rho < 1$, 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 (7) 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 $F_n^*(G) \in \mathcal{P}_{a,2,3}(M_a)$ such that

$$w_{a,n}^{\uparrow}(G) \equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2}(M_a)\} = \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2,3}(M_a)\} = w_n(F_n^*(G), G). \quad (9)$$

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

$$w_{s,n}^{\uparrow}(F) \equiv \sup\{w_n(F,G) : G \in \mathcal{P}_{s,2}(M_s)\} = \sup\{w_n(F,G) : G \in \mathcal{P}_{s,2,3}(M_s)\} = w_n(F, G_n^*(F)). \quad (10)$$

(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 $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$ such that

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

Corresponding results hold for each supremum replaced by an infimum.

3. The Numerical Conclusion about the Overall Upper Bound

We now give an overview of the conclusions we draw from the numerical algorithms and associated experiments to be presented in the following sections.

3.1. Reduction to Two-Point Distributions

From our NLP in §4 we make the following initial conjecture.

CONJECTURE 1. (*further reduction for the UB*) *Within the class of three-point distributions, the tight upper bounds in Theorem 1 are actually attained at two-point distributions.*

We now turn to a stronger conjecture for the overall UB.

3.2. The Extremal Two-Point Distributions

From all our numerical studies, we conclude that the tight upper bound for distributions with bounded support is attained at special two-point distributions with mass at one of the endpoints. Given the support $[0, M_a]$ for F and $[0, \rho M_s]$ for G, let these extremal two-point distributions be deonted by

- $F_0 : c_a^2/(1+c_a^2)$ on 0, $1/(1+c_a^2)$ on $1+c_a^2$;
- F_u: $(M_a-1)^2/(c_a^2+(M_a-1)^2)$ on $1-c_a^2/(M_a-1)$, $c_a^2/(c_a^2+(M_a-1)^2)$ on M_a;
- $G_0 : c_s^2/(1+c_s^2)$ on 0, $1/(1+c_s^2)$ on $\rho(1+c_s^2)$;
- G_u: $(M_s-1)^2/(c_s^2+(M_s-1)^2)$ on $\rho(1-c_a^2/(M_s-1))$, $c_s^2/(c_s^2+(M_s-1)^2)$ on ρM_s.

From extensive numerical experiments, which draw on our mathematical results, we conclude that the extremal UB interarrival-time cdf is the two-point distribution with one mass at 0, denoted by F_0, but the extremal service-time distribution is more complicated because it depends on both n and M_s. In summary, Theorem 1 above and our numerical results support the following conjecture about the overall tight upper bound.

CONJECTURE 2. (*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_0, G_u) *attains the tight UB of the steady-state mean* $E[W]$, *i.e.,*

$$E[W(F,G)] \le E[W(F_0,G_u)] \quad \textit{for all} \quad F \in \mathcal{P}_{a,2}(M_a) \quad \textit{and} \quad G \in \mathcal{P}_{s,2}(M_s),$$

while a pair $(F_0, G_{u,n})$ *attains the tight UB of the transient mean* $E[W_n]$, *i.e.,*

$$E[W_n(F,G)] \le E[W_n(F_0,G_{u,n}))] \quad \textit{for all} \quad F \in \mathcal{P}_{a,2}(M_a) \quad \textit{and} \quad G \in \mathcal{P}_{s,2}(M_s),$$

where $G_{u,n}$ *is a two-point distribution with* $G_{u,n} \Rightarrow G_u$ *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.,*

$$E[W(F,G)] \le \lim_{M_s \to \infty} E[W(F_0,G_u)] \equiv E[W(F_0,G_{u^*})] \quad \textit{for all} \quad F \in \mathcal{P}_{a,2} \quad \textit{and} \quad G \in \mathcal{P}_{s,2}.$$

Let G_{u^*} in $E[W(F,G_{u^*})]$ be shorthand for the limit of $E[W(F,G_u)]$ as $M_s \to \infty$ as in Conjecture 2 (b). In Chen and Whitt (2020a) we obtained an UB for $E[W(F_0,G_{u^*})]$, which is remarkably accurate.

THEOREM 2. (*an UB for* $E[W(F_0,G_{u^*})]$, *Theorem 3.2 of Chen and Whitt (2020a)*) *For the* $GI/GI/1$ *queue with parameter four-tuple* $(1,c_a^2,\rho,c_s^2)$, *if* $E[W(F_0,G_{u^*})]$ *is the tight UB as claimed in Conjecture 2, then*

$$E[W(F_0,G_{u^*})] \le \frac{2(1-\rho)\rho/(1-\delta)c_a^2+\rho^2c_s^2}{2(1-\rho)} < \frac{\rho(2-\rho)c_a^2+\rho^2c_s^2}{2(1-\rho)}, \tag{12}$$

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

REMARK 1. (when one distribution is specified) Counterexamples were constructed in §V of Whitt (1984)and in §8 of Wolff and Wang (2003) that contradict corresponding conjectures that analogs of Conjecture 2 hold when one distribution is fixed.

Tables 1 and 2 compare the numerically computed values of the conjectured tight UB, $E[W(F_0,G_{u^*})]$, drawing on Chen and Whitt (2020b), to the heavy-traffic approximation (HTA) in (1), the new upper bound in (12), the Daley (1977) bound in (5) and the Kingman (1962) bound in (4) 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 these tables we also show the value of δ in the new UB (12) 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 (6), which is far less than the other values, indicating the wide range of possible values. The extremely low LB occurs because it is associated with the $D/GI/1$ model, which is approached by the F_u 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 $(c_a^2 = 4.0, c_s^2 = 0.5)$, because it is the same as for the case $(c_a^2 = 0.5, c_s^2 = 0.5)$ in Table 2 and even for $(c_a^2 = 0.0, c_s^2 = 0.5)$ in the $D/GI/1$ model.

Table 1 **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 $c_s^2 = 4.0$**

ρ	Tight LB	HTA (1)	Tight UB	UB Approx (12)	δ	MRE	Daley (5)	Kingman (4)
0.10	0.000	0.044	0.422	0.422	0.000	0.00%	0.444	2.244
0.15	0.000	0.106	0.653	0.654	0.001	0.05%	0.706	2.406
0.20	0.000	0.200	0.904	0.906	0.007	0.19%	1.000	2.600
0.25	0.042	0.333	1.182	1.187	0.020	0.40%	1.333	2.833
0.30	0.107	0.514	1.499	1.508	0.041	0.60%	1.714	3.114
0.35	0.202	0.754	1.868	1.883	0.070	0.79%	2.154	3.454
0.40	0.333	1.067	2.304	2.326	0.107	0.94%	2.667	3.867
0.45	0.511	1.473	2.829	2.859	0.152	1.06%	3.273	4.373
0.50	0.750	2.000	3.470	3.510	0.203	1.15%	4.000	5.000
0.55	1.069	2.689	4.272	4.321	0.261	1.13%	4.889	5.789
0.60	1.500	3.600	5.295	5.352	0.324	1.07%	6.000	6.800
0.65	2.089	4.829	6.632	6.698	0.393	1.00%	7.429	8.129
0.70	2.917	6.533	8.441	8.520	0.467	0.93%	9.333	9.933
0.75	4.125	9.000	11.014	11.102	0.546	0.80%	12.000	12.500
0.80	6.000	12.800	14.917	15.017	0.629	0.67%	16.000	16.400
0.85	9.208	19.267	21.484	21.597	0.716	0.53%	22.667	22.967
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 2 **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 $c_s^2 = 0.5$**

ρ	Tight LB	HTA (1)	Tight UB	UB Approx (12)	δ	MRE	Daley (5)	Kingman (4)
0.10	0.000	0.006	0.053	0.053	0.000	0.00%	0.056	0.281
0.15	0.000	0.013	0.082	0.082	0.001	0.11%	0.088	0.301
0.20	0.000	0.025	0.113	0.113	0.007	0.54%	0.125	0.325
0.25	0.000	0.042	0.146	0.148	0.020	1.35%	0.167	0.354
0.30	0.000	0.064	0.184	0.189	0.041	2.36%	0.214	0.389
0.35	0.000	0.094	0.228	0.235	0.070	3.16%	0.269	0.432
0.40	0.000	0.133	0.280	0.291	0.107	3.82%	0.333	0.483
0.45	0.000	0.184	0.342	0.357	0.152	4.43%	0.409	0.547
0.50	0.000	0.250	0.414	0.439	0.203	5.72%	0.500	0.625
0.55	0.000	0.336	0.515	0.540	0.261	4.62%	0.611	0.724
0.60	0.000	0.450	0.637	0.669	0.324	4.71%	0.750	0.850
0.65	0.000	0.604	0.800	0.837	0.393	4.45%	0.929	1.016
0.70	0.058	0.817	1.017	1.065	0.467	4.53%	1.167	1.242
0.75	0.188	1.125	1.312	1.388	0.546	5.42%	1.500	1.563
0.80	0.400	1.600	1.822	1.877	0.629	2.95%	2.000	2.050
0.85	0.779	2.408	2.646	2.700	0.716	1.99%	2.833	2.871
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

From this analysis, we see that conjectured new UB (12) is an excellent approximation for the conjectured UB $E[W(F_0, G_{u^*})]$. Moreover, we see that there is significant improvement going from the Kingman (1962) bound in (4) to the Daley (1977) bound in (5) to the new UB in (12). We also see that the heavy-traffic approximation is consistent with the UBs in all cases. Moreover, all the UB approximations are asymptotically correct as $\rho \uparrow 1$. The heavy-traffic approximation in (1) 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. However, the LB falls far below the UB, showing a wide range. We think that the UB is much more representative of typical cases than the LB.

4. The Nonlinear Program

In this section we combine Theorem 1 (c) above with numerical optimization for the transient mean $E[W_n]$ to deduce the form of the extremal distributions for the overall upper bound. Theorem 6 in §7.3 of Chen and Whitt (2020b) shows that results for the transient mean $E[W_n]$ for all n imply corresponding results for the steady-state mean. In §4.1 we formulate an optimization problem for the transient mean based on a multinomial representation. We follow in §4.2 by presenting numerical examples applying the algorithm. We provide further support with simulations for two-point distributions in §5.

.

4.1. The Multinomial Representation for the Transient Mean $E[W_n]$

We can represent the transient mean in (3) in terms of two independent multinomial distributions. Let the cdf G in $\mathcal{P}_{s,2,3}$ 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

$$P_k(\mathbf{p}) \equiv \frac{k! p_1^{k_1} p_2^{k_2} p_3^{k_3}}{k_1! k_2! k_3!}, \tag{13}$$

where it is understood that $\mathbf{ke}' \equiv k_1 + k_2 + k_3 = k$. 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

$$Q_k(\mathbf{q}) \equiv \frac{k! q_1^{w_1} q_2^{w_2} q_3^{w_3}}{w_1! w_2! w_3!}, \tag{14}$$

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

Then, from (3),

$$E[W_n] = \sum_{k=1}^{n} \frac{1}{k} \sum_{(\mathbf{k},\mathbf{w}) \in \mathcal{I}} \max\{0, \sum_{i=1}^{3} (k_i v_i - w_j u_j)\} P_k(\mathbf{p}) Q_k(\mathbf{q}), \tag{15}$$

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

For any given n and any given distributions G in $\mathcal{P}_{s,2,3}$ parameterized by the pair $(\mathbf{v}, \mathbf{p})$ and F in $\mathcal{P}_{a,2,3}$ parameterized by the pair $(\mathbf{u}, \mathbf{q})$, we can calculate the transient mean $E[W_n]$ by calculating the sum in (15). We can easily evaluate $E[W_n]$ for candidate cases provided that n is not too large.

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

$$\sup\{E[W_n(\mathbf{v}, \mathbf{p}, \mathbf{u}, \mathbf{q})] : ((\mathbf{v}, \mathbf{p}), (\mathbf{u}, \mathbf{q})) \in \mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)\}, \tag{16}$$

using (15). 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 (NLP). In particular, for the moments we write $m_1 \equiv E[U] \equiv 1$, $m_2 \equiv E[U^2] \equiv m_1^2(c_a^2 + 1)$, $s_1 \equiv E[V] \equiv \rho$ and $s_2 \equiv E[V^2] \equiv s_1^2(c_a^2 + 1)$. Then the NLP for the UB is

$$\begin{aligned}
\text{maximize} & \sum_{k=1}^{n} \frac{1}{k} \sum_{\sum k_i = k, \sum_j w_j = k} \max(\sum_i k_i v_i - \sum_j w_j u_i, 0) P(k_1, k_2, k_3) Q(w_1, w_2, w_3) \\
\text{subject to} & \sum_{j=1}^{3} u_j q_j = m_1, \quad \sum_{j=1}^{3} u_j^2 q_j = (1 + c_a^2) m_1^2, \\
& \sum_{j=1}^{3} v_j p_j = s_1, \quad \sum_{j=1}^{3} v_j^2 p_j = (1 + c_s^2) s_1^2, \\
& \sum_{j=1}^{3} p_j = \sum_{k=1}^{3} q_k = 1, \\
& M_s \ge v_j \ge 0, \, M_a \ge u_j \ge 0, \, p_j \ge 0, \, q_j \ge 0, \quad 1 \le j \le 3.
\end{aligned} \tag{17}$$

We solved this non-convex NLP in (17) 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 $(F_0, G_{u,n})$, where $G_{u,n}$ is a two-point distribution with $b_s(n) \in (1 + c_s^2, M_s)$ that converges to G_u as $n \to \infty$. In the rare cases that we obtain a different solution, we found that it is always in $\mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$. Moreover, in these cases, we can find a different initial condition for which $(F_0, G_{u,n})$ is the local optimum, and that $E[W(F_0, G_{u,n})]$ is larger than for other local optima.

4.2. Numerical Results from the Optimization and Numerical Search

To illustrate our results, we report results from a further experiment in which we performed a numerical search over the candidate two-point service-time distributions $G_{u,n}$ for the mean waiting time $E[W_n(F_0, G_{u,n})]$ as a function of n using the multinomial exact representation in §4.1 for a class of models ($\rho \in \{0.1, ..., 0.9\}$, $c_a^2 \in \{0.5, 4.0\}, c_s^2 \in \{0.5, 4.0\}, M_a = M_s = 10$), and $n = 1, 5, ..., 50$. For all these cases, we first found by the optimization that the local optimum was obtained at $(F_0, G_{u,n})$. We then conducted the search (using simulation, see §5) to carefully identify the optimal values among these candidate $G_{u,n}$. Tables 3-6 present numerical results for the cases $(c_a^2, c_s^2) = (4.0, 4.0)$, $(4.0, 0.5)$, $(0.5, 4.0)$ and $(0.5, 0.5)$) for a range of n and ρ..

Table 3 **Numerical values of $E[W_n(F_0, G_{u,n})]$ from the optimization for $c_a^2 = c_s^2 = 4.0$ and $M_a = M_s = 10$**

n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
1	0.080	0.160	0.240	0.320	0.400	0.489	0.579	0.668	0.758
5	0.269	0.538	0.813	1.095	1.414	1.777	2.140	2.505	2.882
10	0.357	0.716	1.102	1.525	2.056	2.634	3.228	3.869	4.555
15	0.386	0.778	1.220	1.744	2.410	3.137	3.949	4.832	5.776
20	0.395	0.804	1.281	1.871	2.626	3.508	4.499	5.602	6.808
25	0.399	0.814	1.313	1.948	2.781	3.782	4.933	6.242	7.693
30	0.400	0.820	1.332	1.999	2.896	3.992	5.291	6.794	8.508
35	0.400	0.822	1.343	2.032	2.979	4.163	5.590	7.270	9.185
40	0.400	0.824	1.349	2.056	3.040	4.299	5.846	7.696	9.858
45	0.400	0.824	1.354	2.072	3.088	4.411	6.067	8.075	10.423
50	0.400	0.825	1.356	2.084	3.126	4.505	6.260	8.421	11.002

Table 4 **Numerical values of $E[W_n(F_0, G_{u,n})]$ from the optimization for $c_a^2 = 4.0$ and $c_s^2 = 0.5$ and $M_a = M_s = 10$**

n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
1	0.080	0.160	0.240	0.320	0.400	0.481	0.563	0.644	0.725
5	0.269	0.538	0.807	1.078	1.356	1.638	1.920	2.207	2.499
10	0.357	0.714	1.073	1.447	1.831	2.241	2.702	3.203	3.740
15	0.386	0.772	1.167	1.590	2.074	2.621	3.225	3.902	4.660
20	0.395	0.792	1.206	1.679	2.228	2.860	3.603	4.449	5.411
25	0.399	0.799	1.230	1.730	2.324	3.039	3.888	4.893	6.053
30	0.400	0.803	1.242	1.759	2.393	3.169	4.118	5.262	6.615
35	0.400	0.805	1.248	1.779	2.439	3.268	4.306	5.579	7.114
40	0.400	0.805	1.252	1.791	2.474	3.347	4.460	5.857	7.567
45	0.400	0.806	1.254	1.800	2.498	3.408	4.591	6.102	7.982
50	0.400	0.806	1.256	1.806	2.517	3.458	4.702	6.319	8.364

Table 5 **Numerical values of $E[W_n(F_0, G_{u,n})]$ from the optimization for $c_a^2 = 0.5, c_s^2 = 4.0$ and $M_a = M_s = 10$**

n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
1	0.033	0.082	0.147	0.220	0.305	0.400	0.500	0.600	0.700
5	0.051	0.147	0.303	0.515	0.780	1.097	1.465	1.874	2.301
10	0.051	0.151	0.331	0.607	0.982	1.458	2.043	2.723	3.477
15	0.051	0.152	0.335	0.636	1.075	1.654	2.400	3.301	4.338
20	0.051	0.152	0.337	0.647	1.122	1.779	2.648	3.744	5.033
25	0.051	0.152	0.337	0.652	1.148	1.864	2.836	4.097	5.624
30	0.051	0.152	0.337	0.653	1.163	1.923	2.981	4.392	6.141
35	0.051	0.152	0.337	0.654	1.172	1.965	3.096	4.642	6.600
40	0.051	0.152	0.337	0.655	1.177	1.995	3.190	4.857	7.015
45	0.051	0.152	0.337	0.655	1.181	2.018	3.268	5.046	7.395
50	0.051	0.152	0.337	0.655	1.183	2.034	3.333	5.214	7.744

Table 6 **Numerical values of $E[W_n(F_0, G_{u,n})]$ from the optimization for $c_a^2 = 0.5, c_s^2 = 0.5$ and $M_a = M_s = 10$**

n	$\rho = 0.1$	$\rho = 0.2$	$\rho = 0.3$	$\rho = 0.4$	$\rho = 0.5$	$\rho = 0.6$	$\rho = 0.7$	$\rho = 0.8$	$\rho = 0.9$
1	0.033	0.069	0.106	0.145	0.187	0.230	0.274	0.317	0.361
5	0.050	0.106	0.171	0.248	0.347	0.472	0.626	0.802	1.008
10	0.050	0.107	0.176	0.265	0.386	0.557	0.793	1.096	1.483
15	0.050	0.107	0.176	0.268	0.398	0.590	0.872	1.271	1.813
20	0.050	0.107	0.176	0.268	0.402	0.606	0.917	1.388	2.067
25	0.050	0.107	0.176	0.268	0.404	0.615	0.943	1.471	2.273
30	0.050	0.107	0.176	0.268	0.404	0.619	0.961	1.533	2.446
35	0.050	0.107	0.176	0.268	0.405	0.622	0.973	1.580	2.593
40	0.050	0.107	0.176	0.268	0.405	0.623	0.982	1.616	2.722
45	0.050	0.107	0.176	0.268	0.405	0.624	0.988	1.645	2.834
50	0.050	0.107	0.176	0.268	0.405	0.624	0.993	1.668	2.935

Tables 3-6 illustrate the well known property that $E[W_n]$ is increasing in n, c_a^2 and c_s^2. We also see that $E[W_n]$ tends to be slightly smaller for the pair $(0.5, 4.0)$ than for the pair $(4.0, 0.5)$, but these are similar, as suggested by the HT limit. In support of the corresponding result for $E[W]$, we see convergence well before the final $n = 50$ for the lower traffic intensities.

It is interesting to compare Tables 3 and 6 above to Tables 1 and 2 in §3 which considers the limiting case of $M_s \to \infty$ for same traffic intensities in the cases $c_a^2 = c_s^2 = 4.0$ and $c_a^2 = c_s^2 = 0.5$. The values in Tables 3 and 6 here are consistently lower, significantly so for the larger traffic intensities.

That can be explained by the finite support bound $M_s = 10$ here as opposed to the limiting case as $M_s \to \infty$ in Table 1. Tables 3 and 6 shows that the finite support bound M_s makes a big difference for higher traffic intensities.

Tables 7 and 8 also show optimization results for $E[W_n]$ from (17) for the special cases of the $GI/D/1$ and $D/GI/1$ models with $(c_a^2 = 4.0, M_a = 100)$ and $(c_s^2 = 4.0, M_s = 100)$, respectively. For the $GI/D/1$ model, the optimization terminates with the same extremal two-point cdf F_0. For the $D/GI/1$ model, as in Tables 3-6, we perform an additional search to identify the optimal distribution $G_{u,n}$ for each n.

Table 7 **Numerical values of $E[W_n]$ in the extremal $GI/D/1$ model with $M_a = 100$, $c_a^2 = 4.0$ and $c_s^2 = 0.0$**

$\rho \backslash n$	10	15	20	25	30	35	40	45	50
0.10	0.357	0.386	0.395	0.398	0.400	0.400	0.400	0.400	0.400
0.15	0.536	0.579	0.593	0.598	0.599	0.600	0.600	0.600	0.600
0.20	0.714	0.772	0.791	0.797	0.800	0.802	0.803	0.804	0.804
0.25	0.893	0.965	0.988	1.001	1.009	1.012	1.013	1.014	1.015
0.30	1.071	1.158	1.194	1.217	1.228	1.234	1.237	1.239	1.240
0.35	1.250	1.353	1.413	1.447	1.463	1.474	1.480	1.484	1.486
0.40	1.428	1.562	1.648	1.691	1.719	1.737	1.748	1.756	1.760
0.45	1.607	1.785	1.896	1.958	2.002	2.028	2.047	2.060	2.069
0.50	1.785	2.022	2.159	2.251	2.310	2.353	2.383	2.405	2.421
0.55	1.977	2.274	2.447	2.572	2.656	2.720	2.765	2.800	2.827
0.60	2.183	2.539	2.762	2.922	3.042	3.129	3.200	3.253	3.296
0.65	2.398	2.814	3.100	3.305	3.466	3.590	3.689	3.770	3.836
0.70	2.622	3.106	3.461	3.724	3.931	4.102	4.242	4.358	4.456
0.75	2.859	3.423	3.847	4.182	4.451	4.674	4.865	5.029	5.171
0.80	3.101	3.757	4.262	4.673	5.017	5.309	5.562	5.784	5.982
0.85	3.350	4.108	4.707	5.205	5.631	6.005	6.336	6.632	6.900
0.90	3.611	4.481	5.186	5.784	6.306	6.773	7.194	7.579	7.933

Table 8 **Numerical values of $E[W_n]$ in the extremal $D/GI/1$ model with $M_s = 10$, $c_a^2 = 0.0$ and $c_s^2 = 4.0$**

$\rho \backslash n$	10	15	20	25	30	35	40	45	50
0.10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.15	0.012	0.025	0.025	0.012	0.012	0.012	0.012	0.012	0.025
0.20	0.048	0.058	0.058	0.048	0.048	0.048	0.048	0.048	0.058
0.25	0.091	0.115	0.115	0.091	0.091	0.091	0.091	0.091	0.115
0.30	0.174	0.195	0.195	0.174	0.174	0.174	0.174	0.174	0.195
0.35	0.272	0.300	0.301	0.274	0.274	0.274	0.274	0.274	0.301
0.40	0.407	0.441	0.445	0.418	0.419	0.419	0.419	0.419	0.447
0.45	0.568	0.620	0.631	0.601	0.602	0.603	0.603	0.603	0.640
0.50	0.764	0.833	0.862	0.844	0.848	0.851	0.852	0.853	0.892
0.55	0.985	1.086	1.142	1.139	1.154	1.162	1.168	1.171	1.219
0.60	1.241	1.382	1.472	1.514	1.547	1.569	1.585	1.595	1.642
0.65	1.520	1.728	1.860	1.951	2.017	2.064	2.099	2.125	2.176
0.70	1.837	2.121	2.319	2.462	2.574	2.659	2.728	2.783	2.840
0.75	2.183	2.563	2.843	3.035	3.223	3.362	3.477	3.575	3.658
0.80	2.536	3.038	3.422	3.673	3.978	4.186	4.365	4.520	4.657
0.85	2.924	3.568	4.068	4.371	4.826	5.128	5.394	5.632	5.844
0.90	3.317	4.110	4.747	5.120	5.755	6.171	6.545	6.886	7.200

5. A Systematic Study Over All Two-Point Distributions

The optimization in §4 supports Conjecture 2, but not as strongly as we would like. A more convincing conclusion from §4 is that the optimum is attained by some two-point distribution, as stated in Conjecture 1. Given that weaker conclusion, it suffices to reduce the search for an optimum to the smaller subset of two-point distributions.

5.1. The Two-Point Distributions

Let $\mathcal{P}_{2,2}(m_1, c^2, M)$ 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 $\mathcal{P}_{2,2}(m_1, c^2, M)$ is a one-dimensional parametric family. Any element is determined by specifying one mass point. Let $F_b^{(2)}$ be the cdf that has 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. Let the extremal distributions with mass at the end points be as in §3.2.

Given Conjecture 1 based on the NLP, we next, perform a search for an optimum over the smaller subset of two-point distributions, i.e., to the product space $\mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$. (The G_0 counterexample from §8 of Wolff and Wang (2003) also falls in this set.)

5.2. 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 §4.1 for finite n.)

We study various simulation approaches in Chen and Whitt (2020a). 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.,

$$E[W] = -\frac{E[X^2]}{2E[X]} - E[I_e] = -\frac{E[X^2]}{2E[X]} - \frac{E[I^2]}{2E[I]} = \frac{\rho^2 c_s^2 + c_a^2 + (1-\rho)^2}{2(1-\rho)} - \frac{E[I^2]}{2E[I]}; \tag{18}$$

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, but that challenge is largely avoided by using (18).

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 9 compares the 95% confidence intervals associated with estimates of the steady-state mean $E[W(F_0, G_u)]$ 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 (2020a) for more discussion.)

5.3. 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 $M_a = M_s = 30$. (The maximum 95% confidence interval was less than 10^{-4}.) We focus on

Table 9 **Confidence interval halfwidths for estimates of the steady-state mean $E[W(F_0, G_u)]$ 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	$N = 1E+05$	$N = 1E+06$	$N = 1E+07$	$T = 1E+05$	$T = 1E+06$	$T = 1E+07$
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

the impact of b_a (for F) in the permissible range $[5, 30]$ for six values of b_s (for G) ranging from 5 to 30. (Recall that the parameter b was defined in §5.1.)

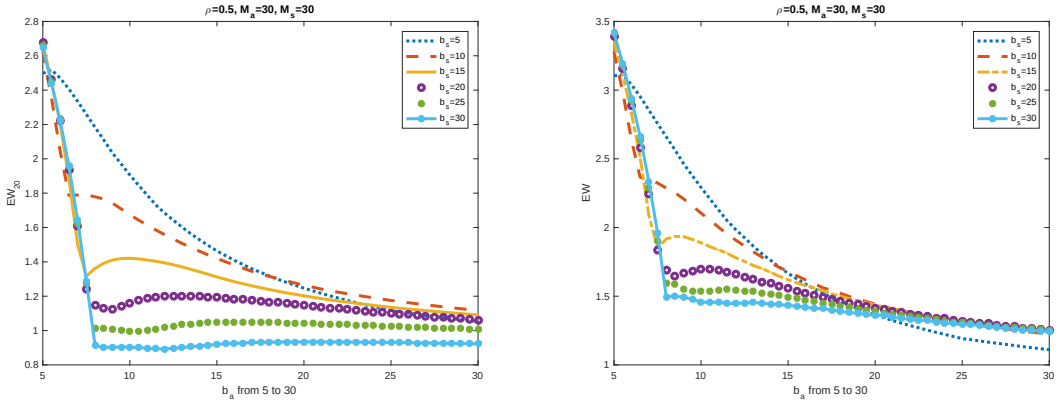

Figure 1 **Simulation estimates of the transient mean $E[W_{20}]$ (left) and the steady-state mean $E[W]$ (right) as a function of b_a for six cases of b_s the in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = M_s = 30$.**

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_0. However, we see some subtle behavior. For example, for $b_s = 20$, we clearly see that the mean is not monotonically decreasing in b_a, but nevertheless, F_0 is clearly optimal.

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. In particular, Tables 10 and 11 present detailed simulation estimates of $E[W]$ and $E[W_{20}]$. In both Tables 10 and 11 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_0), the extremal inter-arrival time is not F_0.

Table 10 **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 $M_a = 7 < M_s = 10$.**

$b_s \backslash b_a$	5.00	5.25	5.50	5.75	6.00	6.25	6.50	6.75	7.0
5.0	3.110	**3.134**	3.117	3.083	3.040	2.997	2.950	2.910	2.863
5.5	3.179	3.026	3.019	3.009	2.975	2.938	2.901	2.860	2.823
6.0	3.191	3.065	2.932	2.907	2.905	2.876	2.844	2.809	2.767
7.0	3.181	3.067	2.942	2.797	2.748	2.720	2.713	2.691	2.670
8.0	3.195	3.056	2.934	2.810	2.664	2.611	2.591	2.564	2.553
9.0	3.239	3.092	2.931	2.792	2.663	2.525	2.472	2.467	2.449
10.0	**3.282**	3.142	2.986	2.812	2.640	2.507	2.367	2.350	2.349

Note that F_0 is optimal for all other b_s and the difference between $\max\{E[W(F, G_0)] : F\} - E[W(F_0, G_0)]$ is very small. Moreover, consistent with Conjecture 2, the overall UB is attained at the pair (F_0, G_u). Finally, note that the difference across each row tends to be greater than the difference across each column.

5.4. The Impact of the Service-Time Distribution

Figure 1 also shows the impact of the service-time distribution, but that impact is more complicated. For $E[W]$ with $b_a = 5.5$, we see that the curve crosses the other curves in the middle. We now investigate what the optimal value of b_s will be over $[1 + c_s^2, M_s]$ for $E[W_n]$ and $E[W]$. For that purpose, Figure 2 plots the values of $E[W_{10}]$ (left) and $E[W_{20}]$ (right) as a function of b_s in the

Table 11 **Simulation estimates of $E[W_{20}]$ as a function of b_a and b_s when $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = 7 < M_s = 10$.**

$b_s \backslash b_a$	5.00	5.25	5.50	5.75	6.00	6.25	6.50	6.75	7.00
5.0	2.497	**2.530**	2.518	2.497	2.469	2.439	2.406	2.371	2.335
5.5	2.557	2.414	2.420	2.422	2.402	2.378	2.351	2.320	2.288
6.0	2.561	2.447	2.328	2.318	2.328	2.312	2.290	2.266	2.239
7.0	2.549	2.447	2.331	2.204	2.165	2.149	2.154	2.150	2.132
8.0	2.556	2.430	2.319	2.208	2.074	2.029	2.021	2.010	2.007
9.0	2.598	2.456	2.310	2.183	2.068	1.937	1.895	1.903	1.898
10.0	**2.626**	2.506	2.353	2.188	2.043	1.921	1.786	1.779	1.789

case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$, $M_s = 300$ and $b_a = (1 + c_a^2)$. For Figure 2, we use the optimization in §4 with a numerical method to directly compute a good finite truncation of objective in the nonlinear program (17). For these cases, we find $b_s(10) = 35.1$ and $b_s(20) = 41.1$.

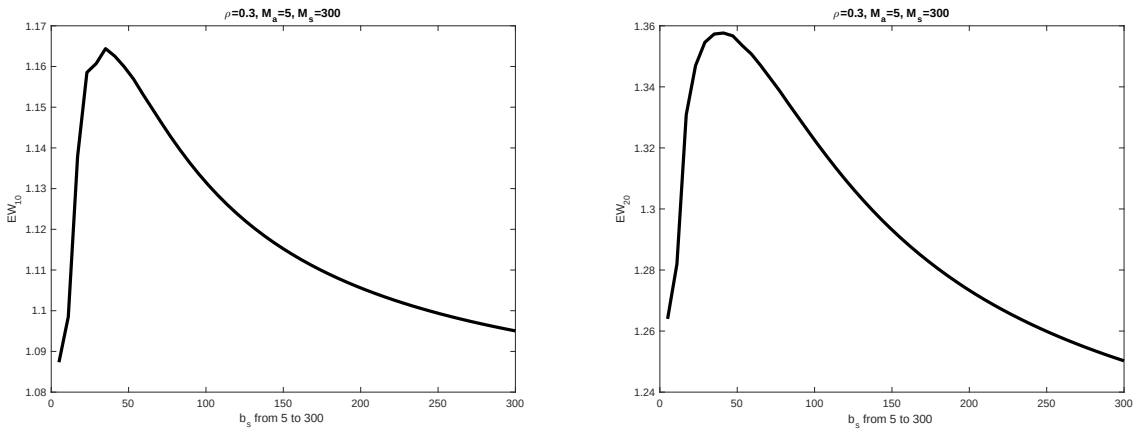

Figure 2 **The transient mean waiting time $E[W_n]$ for $n = 10, 20$ as a function of b_s up to $M_s = 300$. $b_s(10) = 35.1, b_s(20) = 41.1$.**

As a function of b_s, 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 $b_s(n)$ such that

$E[W_n(F_0, G_u; b_s(n))] \geq \{E[W_n(F_0, G_u; b_s)] : b_s \in [1+c_s^2, M_s]\}$. Another important observation is that $b_s(n)$ is a function of n and $b_s(20) > b_s(10)$ under traffic level $\rho = 0.3$.

Now we investigate the extremal $b_s(n)$ as a function of n. Figure 3 shows $E[W_n]$ as a function of n for the light traffic $\rho = 0.2$ (left) and $\rho = 0.3$ (right). Figure 3 shows that $b_s(n)$ tends to be increasing with n given $b_a = (1+c_a^2)$, but is not uniformly so. In particular, for $\rho = 0.3$ on the right, we see a dip at $n = 15$.

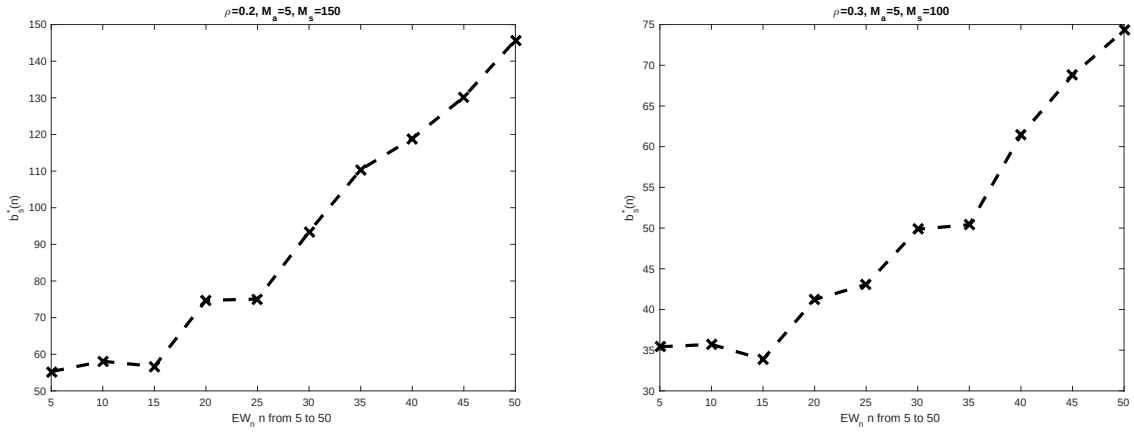

Figure 3 **Performance of $b_s(n)$ associated with $E[W_n(F_0, G_{u,n}]$ for $5 \leq n \leq 50$.**

Nevertheless, the upper bound queue over $\mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$ for transient mean waiting time $E[W_n]$ is $F_0/G_{u,n}/1$ with $b_s(n)$ primarily increasing with n.

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 1×10^7 and 40 i.i.d. replications. (The maximum 95% confidence interval was again less than 10^{-4}.) To illustrate, Figure 4 shows the results for the traffic levels $\rho = 0.3$ (left) and $\rho = 0.9$ (right).

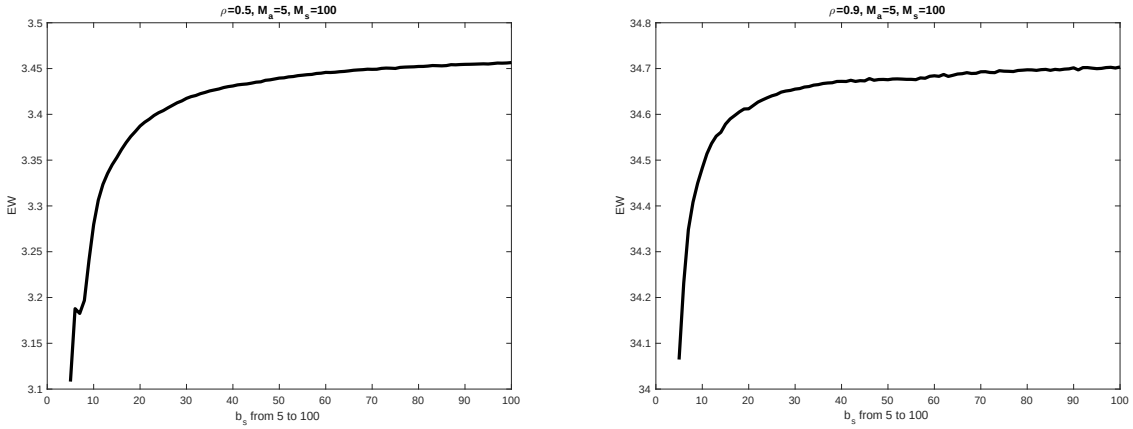

Figure 4 $E[W(F_0, G)]$ **for** $G \in \mathcal{P}_{s,2,2}(M_s)$ **as a function of** b_s **given** $b_a = (1 + c_a^2)$**.**

Just as in Figure 3, 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_0, G_u). Hence, the optimal b_s is M_s. Since $E[W_n] \to E[W]$, we must also have $b_s(n) \to M_s$ as $n \to \infty$.

5.5. Additional Counterexamples When One Distribution is Given

In this section we report additional experiments to provide more counterexamples when one distribution is given. Recall that strong evidence has already been given in Tables 10 and 11. For the steady-state mean $E[W]$, we use simulation method in Minh and Sorli (1983) with simulation length $T^* = 1 \times 10^7$ and 20 i.i.d. replications to compute $E[W]$ for the case $\rho = 0.5$, $c_a^2 = 4$, and $c_s^2 = 4$ with $b_a \in [1 + c_a^2, M_a]$ (LHS of the following Figure 5). For the RHS of Figure 5, we use Monte Carlo simulation method with $N = 1 \times 10^7$ and report average results based on 20 identical independent replications for studying the effects of b_s on $E[W]$ for different cases of b_a. It is already known that when $b_a = (1 + c_a^2)$, the $E[W]$ is increasing with b_s.

Figure 5 shows simulation estimates of the steady-state mean $E[W]$ as a function of b_a in $[(1 + c_a^2), M_a = 7]$ for $b_s = 5$, i.e., for G_0 (left) and as a function of b_s in $[(1 + c_s^2), M_s = 20]$ for various b_a (right). The optimal values of b_s as a function of b_a, denoted by $b_s^*(b_a)$, are: $b_s^*(10) = 5.0, b_s^*(15) = 8, b_s^*(20) = 11, b_s^*(25) = 18, b_s^*(30) = 20$.

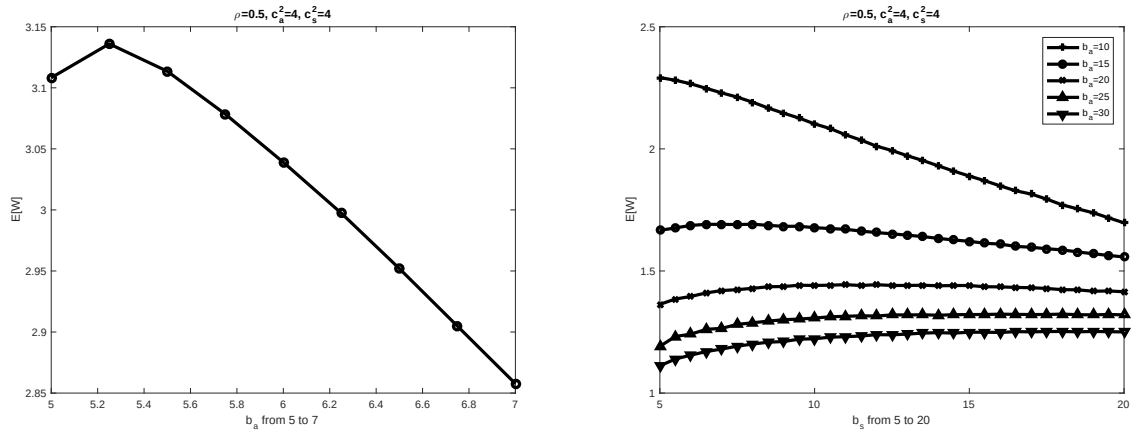

Figure 5 **Simulation estimates of the steady-state mean $E[W]$ as a function of b_a in $[(1+c_a^2), M_a] = [5, 7]$ for $b_s = 5$, i.e., for G_0 (left) and as a function of b_s in $[(1+c_s^2), M_s] = [5, 20]$ for various b_a (right).**

The plot on the left in Figure 5 dramatically shows the counterexample from Wolff and Wang (2003)); it shows that the maximum is not attained at F_0 when the service-time cdf is G_0. The plot on the right shows the more complex behavior that is possible for b_s (the service-time cdf G) as a function of b_a (the interarrival-time cdf F). When $b_a = 5$ (F_0), we see that the mean is increasing in b_s, but when $b_a > 5$, we see more complicated behavior. For the three cases $b_a = 15, 20, 25$, there exists $b_s^*(b_a) \in (1+c_s^2, M_s)$ such that the extremal service-time cdf is neither associated with b_s on the left (G_0) nor with b_s on the right (G_u).

5.6. When One Distribution is Deterministic

We have already looked at the $GI/D/1$ and $D/GI/1$ models in Tables 7 and 8. They showed the transient mean waiting times $E[W_n]$ as a function of n and ρ resulting from the optimization in §4. For all those cases, the transient mean was maximized at $(F_0, G_{u,n})$. We now consider the steady-state mean $E[W]$.

For $D/GI/1$ and $GI/D/1$, we implement the same simulation search for different cases of b_a, b_s throughout traffic level from $\rho = 0.1$ to $\rho = 0.9$. We use Monte Carlo simulation method with $N = 1 \times 10^7$ and report average of 20 identical independent replications. Tables 12 and 13 present results that are consistent with optimization results for transient mean waiting time that the upper bounds of $D/GI/1$ and $GI/D/1$ of steady-state mean and transient mean are attained by G_u and F_0.

Table 12 **Simulation search for $GI/D/1$ over b_a with mean 1 arrival**

$b_a \backslash \rho$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
5.0	0.400	0.804	1.242	1.770	2.469	3.496	5.171	8.50	18.41
5.5	0.000	0.450	0.964	1.536	2.262	3.307	5.006	8.34	18.30
6.0	0.000	0.000	0.626	1.271	2.040	3.102	4.812	8.19	18.26
6.5	0.000	0.000	0.206	0.965	1.795	2.896	4.627	8.02	18.01
7.0	0.000	0.000	0.000	0.600	1.526	2.674	4.436	7.83	17.95
7.5	0.000	0.000	0.000	0.163	1.224	2.436	4.232	7.65	17.71
8.0	0.000	0.000	0.000	0.000	0.875	2.182	4.017	7.46	17.50
8.5	0.000	0.000	0.000	0.000	0.468	1.909	3.802	7.26	17.49
9.0	0.000	0.000	0.000	0.000	0.000	1.612	3.573	7.09	17.19
9.5	0.000	0.000	0.000	0.000	0.000	1.277	3.337	6.88	17.05
10.0	0.000	0.000	0.000	0.000	0.000	0.899	3.084	6.68	16.83

Table 13 **Simulation search for $D/GI/1$ over b_s with mean 1 arrival**

$b_s \backslash \rho$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
10	0.000	0.058	0.195	0.447	0.893	1.670	3.114	6.23	16.00
11	0.004	0.064	0.200	0.457	0.903	1.682	3.129	6.24	16.02
12	0.007	0.067	0.205	0.462	0.911	1.691	3.141	6.26	16.04
13	0.008	0.068	0.210	0.469	0.918	1.702	3.151	6.27	16.05
14	0.009	0.070	0.211	0.474	0.924	1.709	3.160	6.28	16.06
15	0.010	0.073	0.216	0.476	0.929	1.714	3.167	6.29	16.07
16	0.011	0.075	0.218	0.481	0.934	1.721	3.174	6.29	16.08
17	0.011	0.076	0.221	0.484	0.938	1.726	3.179	6.30	16.09
18	0.011	0.077	0.223	0.487	0.941	1.730	3.184	6.31	16.10
19	0.011	0.079	0.224	0.490	0.945	1.734	3.189	6.31	16.10
20	0.012	0.080	0.227	0.492	0.948	1.737	3.193	6.32	16.11

To sum up, for the transient mean waiting time $E[W_n]$, the numerical experiments show that there exists $b_a^* = (1 + c_a^2)$ and $b_s(n)$ such that the $\sup\{E[W_n(F,G)] : F, G \in \mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)\}$ is attained. We find that $b_s(n)$ is not strictly increasing, but that there exists an n_0 after which it is increasing In all cases, we find that $G_{u,n} \Rightarrow G_u$ as $n \to \infty$. For the steady-state mean waiting time $E[W]$, the UB is attained when b_a^* is $(1 + c_a^2)$ and $b_s^* = M_s$. Hence, the UB for the steady-state mean waiting time is attained at (F_0, G_u).

6. 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 $D/A_3/1$ model where the D interarrival time with $c_a^2 = 0$ can be regarded as the limit of F_u 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 A_3 is any three-point distribution in $\mathcal{P}_{s,2}$ 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 $D/A_3/1$ model acts to reduce $E[W]$. The resulting LB has the convenient explicit formula in (6).

However, the overall LB has not yet been established for distributions with finite support. Motivated by the established extremal property of the lattice $D/A_3/1$ model with unbounded support, we investigate a new "nearly-lattice" three-point distribution to use with F_u called $G_{u,b_s u}$. It has support $\{0, u, b_s u\}$, where $1 < b_s \le M_s$ is an appropriate positive value and u is the first point of the cdf F_u at $u = 1 - c_a^2/(M_a - 1) \in (0,1)$ with $M_a > 1 + c_a^2$.

The new $G_{u,b_s u}$ makes the $F_u/G_{u,b_s u}/1$ 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

$$\lim_{M_a \to \infty} E[W(F_u, G_{u,b_s u})] = E[W(D, A_3)] \tag{19}$$

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_u is good for F, but G_0 and G_u might not be nearly optimal for G to minimize the mean waiting time. Moreover, Figure 3 shows G_0 is the optimal solution to minimize $E[W(F_0, G)]$ over $\mathcal{P}_{s,2,2}(M_s)$ only for $M_a = 1 + c_a^2$. Thus it is interesting to explore better service time distribution when $F = F_u$ for $M_a > 1 + c_a^2$.

6.1. The $G_{u,b_s u}$ Service-Time Distribution

To derive the closed form of $G_{u,b_s u}$, 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 $u = 1 - c_a^2/(M_a - 1)$),

$$p_1 + p_2 + p_3 = 1, x_1 p_1 + x_2 p_2 + x_3 p_3 = \rho, x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 = (1 + c_s^2)\rho^2 \tag{20}$$

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

$$p_1 = \frac{(b_s^2(u^2-\rho u)+b_s(-u^2+(1+c_s^2)\rho^2)-(1+c_s^2)\rho^2+u\rho)}{(b_s^2u^2-b_s u^2)},$$
$$p_2 = \frac{\rho b_s u-(1+c_s^2)\rho^2)}{b_s u^2-u^2} \quad \text{and} \quad p_3 = \frac{\rho^2(1+c_s^2)-u\rho}{b_s^2u^2-b_s u^2}. \tag{21}$$

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 $G_{u,b_s u}=G_0$. If $\rho=u/(1+c_s^2)$, then $G_{u,b_s u}$ is a two-point distribution with mass at $\{0,u\}$. Since inter-arrival time distribution F_u has mass at $\{u,M_a\}$ and there is no large service time impact, $E[W(F_u,G_{u,b_s u})]=0$. 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

$$b_s^2(u^2-\rho u)+b_s(-u^2+(1+c_s^2)\rho^2)-(1+c_s^2)\rho^2+u\rho=0, \tag{22}$$

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

$$b_s\in\left[\frac{(1+c_s^2)\rho}{u},\mathbf{1}_{\{\rho\in(u/(1+c_s^2),u]\}}\infty+\mathbf{1}_{\{\rho>u\}}\gamma\right). \tag{23}$$

To sum up, the b_s is determined optimally within its valid range via solving

$$b_s\in\arg\min_b E[W(F_u,G_{u,bu})]. \tag{24}$$

Numerically, the b_s can be decided by a simulation search.

Conjecture 3. *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 $(F_u,G_{u,b_s u})$ attains the tight LB of the steady-state mean $E[W]$ for $M_a>1+c_a^2$, i.e.,*

$$E[W(F,G)]\geq E[W(F_u,G_{u,b_s u})] \quad \textit{for all} \quad F\in\mathcal{P}_{a,2}(M_a) \quad \textit{and} \quad G\in\mathcal{P}_{s,2}. \tag{25}$$

If $M_a=1+c_a^2$, the pair (F_0,G_0) attains the tight LB of the steady-state mean $E[W]$, i.e.,

$$E[W(F,G)]\geq E[W(F_0,G_0)] \quad \textit{for all} \quad F\in\mathcal{P}_{a,2}(M_a) \quad \textit{and} \quad G\in\mathcal{P}_{s,2}. \tag{26}$$

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

$$E[W(D, A_3)] \le E[W(F_u, G_{u,b_s u})] \le \inf\{E[W(F_u, G_u)] : b \in [1 + c_s^2, \infty)\}. \tag{27}$$

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

$$E[W(D, A_3)] \le E[W(F_0, G_0)] \le \inf\{E[W(F_0, G_u)] : b \in [1 + c_s^2, \infty)\}. \tag{28}$$

6.2. The Impact of Service Time in $F_u/G_{u,b_s u}/1$

We study the impact of b to $E[W(F_u, G_{u,bu})]$ and determine the optimal b_s in (23) to minimize $E[W(F_u, G_{u,bu})]$ 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 (23), we simulate the model under $M_a = 6, 8, 10$ and various settings of b_s ($\gamma- \equiv \gamma - 0.0001$. For example, γ^- is 19.2 when $M_a = 6$.)

Table 14 **Simulation estimates of $E[W(F_u, G_{u,bu})]$ under the case $c_a^2 = c_s^2 = 4, \rho = 0.5$**

b	13	14	15	16	17	18	19	γ^-	γ^-	γ^-	γ^-
$M_a = 6 (u = 0.20)$	3.01	2.95	2.89	2.82	2.76	2.72	2.67	2.66	2.66	2.66	2.66
b	10	12	14	16	18	20	22	24	26	28	30
$M_a = 8 (u = 0.42)$	2.36	2.22	2.10	1.98	1.85	1.73	1.69	1.68	1.65	1.61	1.58
b	10	12	14	16	18	20	22	24	26	28	30
$M_a = 10 (0.55)$	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 $E[W(F_u, G_{u,b_s u})]$ is monotone decreasing as b_s increases. Thus the optimal b_s is γ^- when $\rho > u$ or ∞ when $\rho \in (u/(1 + c_s^2), u]$.

6.3. Simulation Comparisons

From extensive simulation experiments, we conclude that the LB for $E[W]$ is attained, at least approximately, by the $F_u/G_{u,b_s u}/1$ model. Following from Figure 1 and 3, we see there exists an optimal $b_s^*(b_a)$ such that the lower bound of $E[W]$ is attained by $E[W(F_u, G_u)]$ over $\mathcal{P}_{a,2,2}(M_a) \times$

$\mathcal{P}_{s,2,2}(M_s)$. Since the mean of $F_u/G_{u,b_s u}/1$ is monotone decreasing as b_s increases, we set b_s sufficiently large for $F_u/G_{u,b_s u}/1$ and set the optimal $b_s^*(b_a)$ for $F_u/G_u/1$ to make a careful simulation comparison under the case $c_a^2 = c_s^2 = 4$ under different settings of b_a.

Table 6.3 shows the results for the $E[W(F_u, G_u)]$ 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 15 Simulation performance of lower bound with different settings of M_a for the model $F_u/G_u/1$ ($T = 5 \times 10^8$ 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 $E[W(F_u, G_{u,b_s u})]$ 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 16 Simulation performance of lower bound with different settings of M_a for the model $F_u/G_{u,b_s u}/1$ ($T = 1 \times 10^7$ 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

7. Conclusions

We have studied tight upper and lower bounds for the mean steady-state waiting time in the $GI/GI/1$ model given the first two moments of the interarrival time and service time, specified by the parameter vector $(1, c_a^2, \rho, c_s^2)$, when the underlying distributions have bounded support. Theorem 1 from Chen and Whitt (2020b) shows that the upper and lower bounds (for the transient mean $E[W_n]$ as well as the steady-state mean $E[W]$ in (3), overall and with one distribution fixed) are attained at distributions with support on at most three points. In this paper we applied numerical methods to further identify the extremal distributions.

From a practical engineering perspective, we have addressed the important question about the tight upper bound. The combination of mathematical and numerical results strongly supports Conjecture 2 in §3, which states that the overall upper bound is attained by $E[W(F_0, G_{u^*})]$, i.e., at the extremal two-point distributions, modified by a limit, as some have thought. However, because the analysis is partly numerical, it still remains to provide a mathematical proof. A convenient explicit formula for an upper bound for the conjectured tight upper bound appears in formula (12). Tables 1 and 2 show that the new UB formula (12) is quite accurate, providing significantly improvement over previous bounds. We also obtained numercial results for the lower bound with bounded support in §6, leading to Conjecture 3.

There are many remaining problems for research. In addition to providing a full mathematical proof of Conjecture 2, it remains to identify the extremal distributions with one distribution given, as in parts (a) and (b) of Theorem 1. It also remains to establish similar results for other models.

Acknowledgments

This research was supported by NSF CMMI 1634133.

References

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

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

Chen, Y., W. Whitt. 2020a. Algorithms for the upper bound mean waiting time in the $GI/GI/1$ queue. *Queueing Systems* **94** 327–356.

Chen, Y., W. Whitt. 2020b. Extremal $GI/GI/1$ queues given two moments: Three-point distributions. Revision under review for Operations Research, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Chung, K. L. 2001. *A Course in Probability Theory*. 3rd ed. Academic Press, New York.

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.

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.

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.

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.

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

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.

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

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.