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

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 exposes some important open problems in queueing theory. We use simulation and optimization to evaluate the tight upper and lower bounds for the transient and steady-state mean waiting time in the $GI/GI/1$ queue when the interarrival times and service times are partially specified by their first two moments. For the special case in which the interarrival-time and service-time distributions are two-point distributions with bounded support, we apply simulation to provide evidence to support the conjecture that the tight upper bound overall is attained at two-point distributions where the interarrival-time distribution has one mass point at 0, while the service-time distribution has one mass at the upper limit. To extend that conclusion, we apply optimization to provide evidence that the tight upper bound over three-point distributions are attained at two-point distributions. We also obtain results for the tight upper bound of the supremum over one distribution when the other is specified. We also study the tight lower bounds.

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

History: April 12, 2021

1. Introduction

This paper contributes to the theory of extremal $GI/GI/1$ queues. We are interested in tight upper and lower bounds for the transient mean waiting time $E[W_n]$ and steady-state mean waiting time $E[W] \equiv E[W_\infty]$ when the interarrival-time cdf F has mean 1 and scv (squared coefficient of variation, variance divided by the square of the mean) $c_a^2 < \infty$ and the service-time cdf G has mean ρ and scv $c_s^2 < \infty$, so that the traffic intensity is ρ, $0 < \rho < 1$.

We now formulate a version of the problem for distributions with bounded support. For that purpose, let $\mathcal{P}_{a,2}(1, c_a^2, M_a)$ be the set of all interarrival-time cdf's F with mean 1, scv c_a^2 and support

M_a, where $M_a \geq 1 + c_a^2$ to be feasible. Similarly, let $\mathcal{P}_{s,2}(\rho, c_s^2, M_s)$ be the set of all service-time cdf's G with mean ρ, scv c_s^2 and support ρM_s, where $M_s \geq 1 + c_s^2$ to be feasible.

For the upper bound, we are interested in three problems:

$$\begin{aligned}
&(a) \quad \sup\{E[W_n(F,G)] : F \in \mathcal{P}_{a,2}(1, c_a^2, M_a)\} \quad \text{for given} \quad G \in \mathcal{P}_{s,2}(\rho, c_s^2, M_s), \\
&(b) \quad \sup\{E[W_n(F,G)] : G \in \mathcal{P}_{s,2}(1, c_s^2, M_s)\} \quad \text{for given} \quad F \in \mathcal{P}_{a,2}(1, c_a^2, M_a) \quad \text{and} \\
&(c) \quad \sup\{E[W_n(F,G)] : F \in \mathcal{P}_{a,2}(1, c_a^2, M_a), G \in \mathcal{P}_{s,2}(\rho, c_s^2, M_s)\}. \qquad (1)
\end{aligned}$$

We wish to identify the distributions that attain these suprema. Since we are optimizing a continuous funcion over a compact metric space, the optimum values are always attained.

Let $\mathcal{P}_{a,2,2}(1, c_a^2, M_a)$ and $\mathcal{P}_{s,2,2}(\rho, c_s^2, M_s)$ be the subsets of two-point distributions in $\mathcal{P}_{a,2}(1, c_a^2, M_a)$ and $\mathcal{P}_{s,2}(\rho, c_s^2, M_s)$, respectively. Each of these-sets is a one-parameter family, which can be indexed by either the lower or upper mass point. Let F_0 be the two-point distribution in $\mathcal{P}_{a,2,2}(1, c_a^2, M_a)$ with one mass at 0; let G_u be the two-point distribution in $\mathcal{P}_{s,2,2}(1, c_s^2, M_s)$ with one mass at the upper limit of support, ρM_s.

The case of greatest interest in (1) is no doubt the overall upper bound in case (c). Our numerical results support the following conjecture about the overall tight upper bound.

CONJECTURE 1. (*the tight upper bound for* $1 \leq n \leq \infty$)

(*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* $M_s \geq c_s^2 + 1$*, the pair* (F_0, G_u) *attains the tight upper bound of the steady-state mean* $E[W]$*, i.e.,*

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

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

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

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 upper bound of $E[W(F,G)]$ is obtained asymptotically in the limit as $M_s \to \infty$ in part (a), i.e.,*

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

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 1 (b). In Chen and Whitt (2020) we obtained an UB for $E[W(F_0,G_{u^*})]$, which is remarkably accurate. (See Theorem 1 and Tables 1 and 2 in §2.3 below.)

In §3 we report results of a simulation study to investigate the special case of two-point distributions. In particular, we experimentally (numerically) verify Conjecture 1 for the restriction to two-point distributions. We also consider cases (a) and (b) in (1) in the special case of two-point distributions. These results supplement our recent papers Chen and Whitt (2020, 2021b,a), which contribute to the substantial literature, reviewed in Daley et al. (1992), especially §10, and Wolff and Wang (2003). To gain simulation efficiency for the steady-state mean, we exploit the representation of $E[W]$ in terms of the idle-time distribution proposed by by Minh and Sorli (1983).

For the maximum over F for specified service-time cdf G in case (a) of (1), by applying Tchebycheff systems, Theorem 2 (a) of Chen and Whitt (2021b) establishes that F_0 is optimal if G is completely monotone, i.e., can be represented as a mixture of exponentials. Here we experimentally show that F_0 is either optimal or nearly optimal for all two-point G.

On the other hand, for the maximum over G for specified F in case (b) of (1), Theorem 2 (b) of Chen and Whitt (2021b) tells a more complicated story, consistent with §V of Whitt (1984b) and §8 of Wolff and Wang (2003). In particular, it concludes that the maximum is attained at G_0 (G_u) when the cdf F is strictly concave (strictly convex). (An explanation in terms of singularities of Laplace transforms is given in Whitt (1984a).) The cdf F_0 has neither of these properties. It can be well approximated by a cdf that is strictly increasing right after 0 and right before M_a, but flat or relatively flat in between. The complexity of the problems in cases (a) and (b) at least partly explains the difficulty in establishing Conjecture 1.

This paper has been extracted from an earlier unpublished paper that (unsuccessfully) tried to prove the following conjecture (which we still think is true).

CONJECTURE 2. (*three-point extremal distributions*) *All the tight upper bounds in* (1) *and the corresponding tight lower bounds are attained by three-point distributions.*

Motivated by Conjecture 2, in §4 we develop a multinomial representation for three-point distributions and apply it to formulate a non-convex nonlinear program (NLP) for the overall upper bound over three-point distributions, 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. In our experiments, we found that all local optima for the overall upper bound are two-point distributions and that the best local optimum always has interarrival-time cdf with one mass at 0.

As reviewed in §2.4.1 of Chen and Whitt (2020), the overall tight lower bound is known, and it is known to be attained at a three-point distribution, which is not a two-point distribution. Otherwise, Conjectures 1 and 2 remain to be proved or refuted. In §5 we study the lower bound with bounded support, which was not considered previously. Finally, in §6 we draw conclusions.

2. Background

We briefly review the $GI/GI/1$ model in §2.1, the established bounds in §2.2, the two-point distributions in §2.3. and related literature in §2.4.

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, 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 $E[U] \equiv \lambda^{-1} \equiv 1$ and squared coefficient of variation (scv, variance divided by the square of the mean) $c_a^2 < \infty$; let a service time V have mean $E[V] \equiv \tau \equiv \rho$ and scv $c_s^2 < \infty$, 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 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

$$W_n = [W_{n-1} + V_{n-1} - U_{n-1}]^+, \quad n \geq 1, \tag{2}$$

where $x^+ \equiv \max\{x, 0\}$. Let W be the steady-state waiting time, satisfying $W_n \Rightarrow W$ as $n \to \infty$, where $\Rightarrow$ denotes convergence in distribution for any proper cdf H_0. It is well known that the cdf H of W is the unique cdf satisfying the stochastic fixed point equation

$$W \stackrel{\rm d}{=} (W + V - U)^+, \tag{3}$$

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

$$E[W_n] \equiv E[W_n | W_0 = 0] = \sum_{k=1}^{n} \frac{E[S_k^+]}{k} < \infty, \quad 1 \leq n < \infty, \quad \text{and} \quad E[W] = \sum_{k=1}^{\infty} \frac{E[S_k^+]}{k} < \infty; \tag{4}$$

see §§X.1-X.2 of Asmussen (2003) or (13) in §8.5 of Chung (2001). We will exploit the formula for the transient mean in (4) in our analysis.

2.2. Background on the Bounds and Approximations

The familiar heavy-traffic approximation for the mean steady-state waiting time is

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

Formula (5) combines the heavy-traffic limit in Kingman (1961) with the exact Pollaczek-Khintchine formula when the arrival process is a Poisson process, so that $c_a^2 = 1$.

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

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

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

A better upper bound 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.,

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

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

In Chen and Whitt (2020) we developed algorithms to compute the conjectured tight upper bound. There we showed that it provides significant improvement to (6) and (7) away from heavy traffic. 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 1 (b). We also obtained an upper bound for $E[W(F_0, G_{u^*})]$, which is remarkably accurate.

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

$$E[W(F_0, G_{u^*})] \le \frac{2(1-\rho)\rho/(1-\delta)c_a^2 + \rho^2 c_s^2}{2(1-\rho)} < \frac{\rho(2-\rho)c_a^2 + \rho^2 c_s^2}{2(1-\rho)}, \tag{8}$$

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

In contrast to the tight upper bound that we primarily study, the tight lower bound 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:

$$E[W] \ge \frac{\rho((1+c_s^2)\rho - 1)^+}{2(1-\rho)}. \tag{9}$$

The lower bound in (9) 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.)

Tables 1 and 2 compare the numerically computed values of the conjectured tight upper bound, $E[W(F_0, G_{u^*})]$, comparing it to the heavy-traffic approximation (HTA) in (5), the new upper bound

in (8), 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 these tables we also show the value of δ in the new upper bound (8) (UB) and the maximum relative error (MRE) between the upper bound approximation and the tight upper bound. 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 (9), 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 $P(V \leq U) = 1$. 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 (5)	Tight UB	UB Approx (8)	δ	MRE	Daley (7)	Kingman (6)
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 (5)	Tight UB	UB Approx (8)	δ	MRE	Daley (7)	Kingman (6)
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 (8) 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 (6) to the Daley (1977) bound in (7) to the new UB in (8). We also see that the heavy-traffic approximation is consistent with the upper bounds in all cases. Moreover, all the upper bound approximations are asymptotically correct as $\rho \uparrow 1$. 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. The very wide range (UB - LB) caused by the lower bound was a major incentive for adding extra information about the underlying distributions in order to obtain useful set-valued approximations in Chen and Whitt (2021c). We discuss the lower bound for the transient mean here in §5.

2.3. 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 $\mathcal{P}_{2,2}(M) \equiv \mathcal{P}_{2,2}(m_1, c^2, M)$ is a one-dimensional parametric family. Typically, we use $\mathcal{P}_{a,2,2}(M_a)$ and $\mathcal{P}_{s,2,2}(M_s)$ with notations a and s to denote the sets of probability measures for inter-arrival time and service time. 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 $F_0 \equiv F^{(2)}_{1+c^2}$ and $F_u \equiv F_b^{(2)}$ for some $b \in (1+c^2, M)$ denote the extremal distributions with masses at the end points $m_1(1+c^2)$ and $m_1 b$, as in §1.

2.4. Related Literature

Optimization has been used previously to study the bounding problem for queues, beginning with Klincewicz and Whitt (1984) and Johnson and Taaffe (1990). Due to intractability(e.g., lack of convexity), new approaches have been proposed to simplify the problem, e.g, reformulating the problem into tractable relaxed convex programs, imposing extra conditions and limitations; see Bertsimas and Natarajan (2007) and Gupta and Osogami (2011)). Optimal solutions are not difficult to obtain, but it is difficult to assess the approximation error.

In addition, several researchers have studied bounds for the more complex many-server queue. 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. Osogami and Raymond (2013) bounds the transient tail probability of $GI/GI/1$ by a semi-definite program. Li and Goldberg (2017) establishes bounds for $GI/GI/c$ intended for the many-server heavy-traffic regime. van Eekelen et al. (2019) address the classical extremal queueing problem by measuring dispersion in terms of Mean Absolute Deviation (MAD) instead of variance. Finally, we mention that optimization also plays a critical role in recent work on robust queueing, as in Bandi et al. (2015) and Whitt and You (2018, 2019).

3. The Simulation Experiments

To analyze the mean waiting times for the two-point interarrival-time and service-time distributions, we use stochastic simulation.

3.1. The Simulation Methodology

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

$$E[W] = -\frac{E[X^2]}{2E[X]} - E[I_e] = -\frac{E[X^2]}{2E[X]} - \frac{E[I^2]}{2E[I]} = \frac{\rho^2 c_s^2 + c_a^2 + (1-\rho)^2}{2(1-\rho)} - \frac{E[I^2]}{2E[I]}; \tag{10}$$

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 (10).

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 $E[W(F_0, G_u)]$ 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 (2020) for more discussion.)

3.2. 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 $M_a = M_s = 30$. (The maximum 95% confidence interval was less than 10^{-4}.) We focus on 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 §2.3.)

Table 3 **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	$N = 10^5$	$N = 10^6$	$N = 10^7$	$T = 10^5$	$T = 10^6$	$T = 10^7$
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

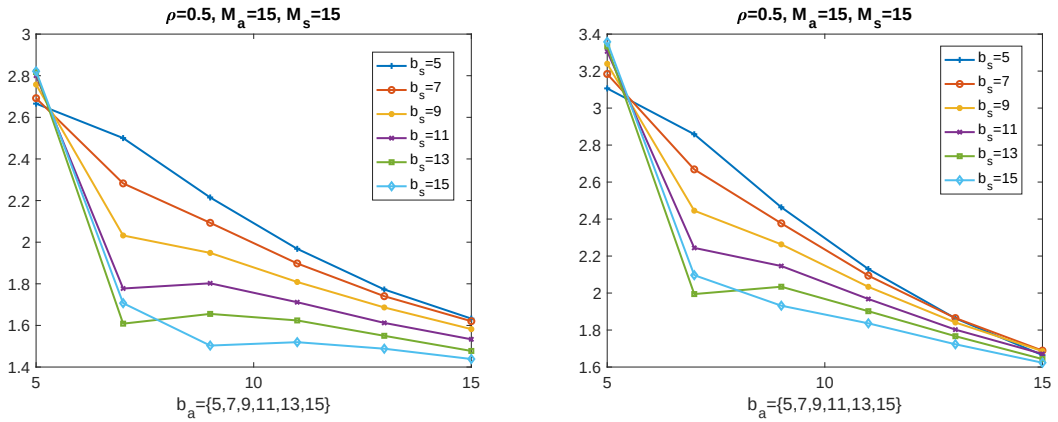

Figure 1 **Simulation estimates of the transient mean $E[W_{20}]$ (left) and the steady-state mean $E[W]$ (right) as a function of b_a for six cases of b_s the in the case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$ and $M_a = M_s = 30$.**

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 4 and 5 present detailed simulation estimates of $E[W]$ and $E[W_{20}]$. In both Tables 4 and 5 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 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 $M_a = 7 < M_s = 10$.**

$b_s \backslash b_a$	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

Table 5 **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

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 1, 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.

3.3. The Impact of the Service-Time Distribution

Figure 1 also shows the impact of the service-time distribution, but that impact is more complicated. 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 case $\rho = 0.5$, $c_a^2 = c_s^2 = 4.0$, $M_s = 300$ and $b_a = (1 + c_a^2)$. For these cases, we find $b_s(10) = 35.1$ and $b_s(20) = 41.1$.

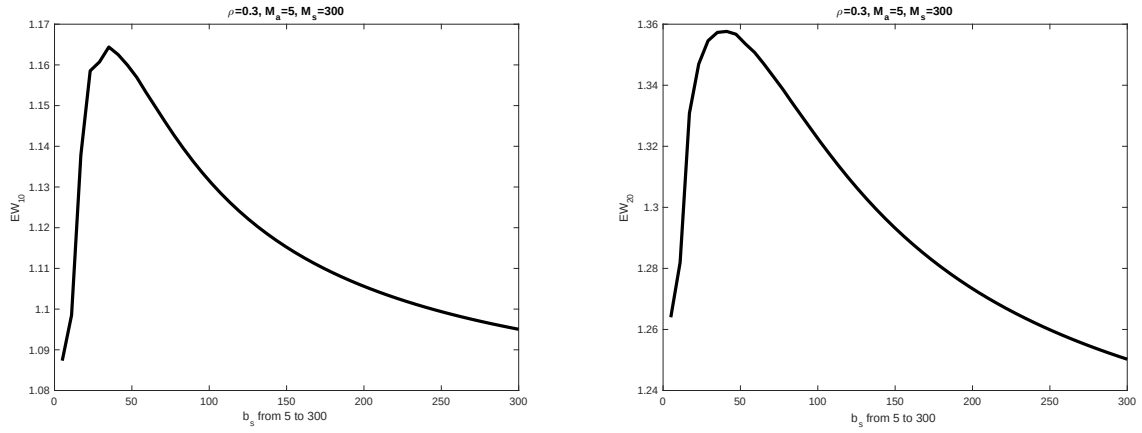

Figure 2 **The transient mean waiting time $E[W_n]$ for $n = 10, 20$ as a function of b_s up to $M_s = 300$. $b_s(10) = 35.1, b_s(20) = 41.1$.**

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 $n = 15$. We define $G_{u,n}$ is a two-point distribution with $b_s(n) \in (1 + c_s^2, M_s)$ that converges to G_u as $n \to \infty$.

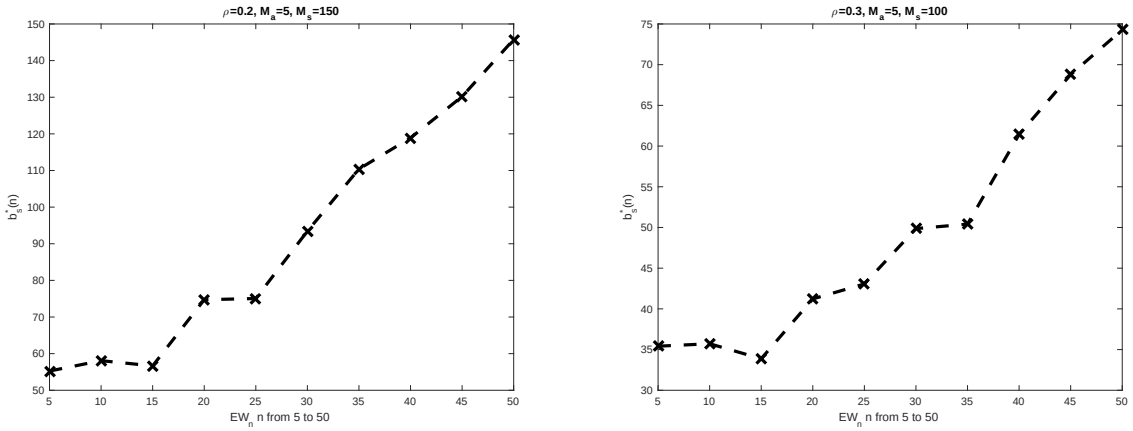

Figure 3 **Performance of $b_s(n)$ associated with $E[W_n(F_0, G_{u,n})]$ for $5 \le n \le 50$.**

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).

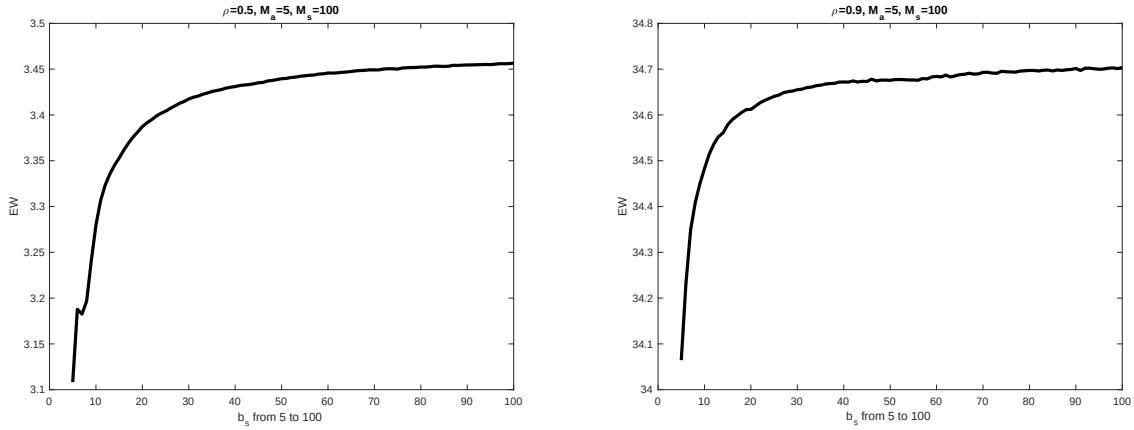

Figure 4 **$E[W(F_0, G)]$ for $G \in \mathcal{P}_{s,2,2}(M_s)$ as a function of b_s given $b_a = (1 + c_a^2)$.**

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 $n \to \infty$.

3.4. 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 4 and 5. 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: $b_s^*(10) = 5.0, b_s^*(15) = 8, b_s^*(20) = 11, b_s^*(25) = 18, b_s^*(30) = 20$.

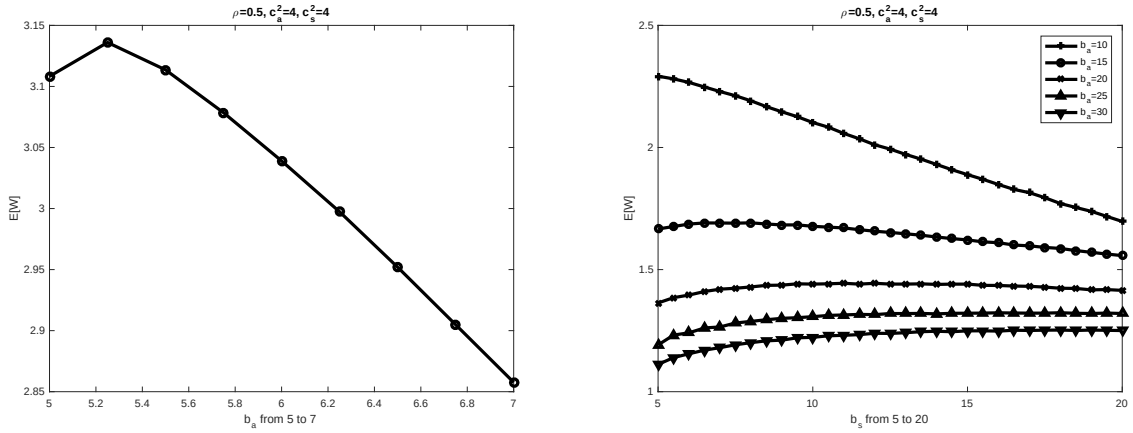

Figure 5 **Simulation estimates of the steady-state mean $E[W]$ as a function of b_a in $[(1 + c_a^2), M_a] = [5, 7]$ for $b_s = 5$, i.e., for G_0 (left) and as a function of b_s in $[(1 + c_s^2), M_s] = [5, 20]$ for various b_a (right).**

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).

4. The Nonlinear Program

In this section we apply numerical optimization for the transient mean $E[W_n]$ to deduce the form of the extremal distributions for the overall upper bound among three-point distributions. For that purpose, let $\mathcal{P}_{a,2,3}$ and $\mathcal{P}_{s,2,3}$ be the sets of three-point distributions, paralleling $\mathcal{P}_{a,2,2}$ and $\mathcal{P}_{s,2,2}$ in §2.3. 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.

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

We can represent the transient mean in (4) 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

$$P_k(\mathbf{p}) \equiv \frac{k! p_1^{k_1} p_2^{k_2} p_3^{k_3}}{k_1! k_2! k_3!}, \tag{11}$$

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

$$Q_k(\mathbf{q}) \equiv \frac{k! q_1^{w_1} q_2^{w_2} q_3^{w_3}}{w_1! w_2! w_3!}, \tag{12}$$

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

Then, from (4),

$$E[W_n] = \sum_{k=1}^{n} \frac{1}{k} \sum_{(\mathbf{k},\mathbf{w})\in\mathcal{I}} \max\{0, \sum_{i=1}^{3}(k_i v_i - w_j u_j)\} P_k(\mathbf{p}) Q_k(\mathbf{q}), \tag{13}$$

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 (13). 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

$$\sup\{E[W_n(\mathbf{v}, \mathbf{p}, \mathbf{u}, \mathbf{q})] : ((\mathbf{v}, \mathbf{p}), (\mathbf{u}, \mathbf{q})) \in \mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)\}, \tag{14}$$

using (13). 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

$$\begin{aligned}
\text{maximize} & \sum_{k=1}^{n} \frac{1}{k} \sum_{\sum k_i = k, \sum_j w_j = k} \max(\sum_i k_i v_i - \sum_j w_j u_i, 0) P(k_1, k_2, k_3) Q(w_1, w_2, w_3) \\
\text{subject to} & \sum_{j=1}^{3} u_j q_j = m_1, \quad \sum_{j=1}^{3} u_j^2 q_j = (1 + c_a^2) m_1^2, \\
& \sum_{j=1}^{3} v_j p_j = s_1, \quad \sum_{j=1}^{3} v_j^2 p_j = (1 + c_s^2) s_1^2, \\
& \sum_{j=1}^{3} p_j = \sum_{k=1}^{3} q_k = 1, \\
& M_s \geq v_j \geq 0, \; M_a \geq u_j \geq 0, \; p_j \geq 0, \; q_j \geq 0, \quad 1 \leq j \leq 3.
\end{aligned} \tag{15}$$

We solved this non-convex NLP in (15) 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 $(F_0, G_{u,n})$. 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 §3) to carefully identify the optimal values among these candidate $G_{u,n}$. Tables 6-9 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 6 **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 7 **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 8 **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 9 **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 6-9 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 6 and 9 above to Tables 1 and 2, 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 6 and 9 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 6 and 9 shows that the finite support bound M_s makes a big difference for higher traffic intensities.

4.3. When One Distribution is Deterministic

Tables 10 and 11 also show optimization results for $E[W_n]$ from (15) 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 6-9, we perform an additional search to identify the optimal distribution $G_{u,n}$ for each n.

Table 10 **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 11 **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

Tables 10 and 11. showed the extremal transient mean waiting times $E[W_n]$ as a function of n and ρ. For all those cases, the transient mean was maximized at $(F_0, G_{u,n})$. We now consider the

steady-state mean $E[W]$, applying simulation as in §3. 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 shows that the upper bounds 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 $E[W_n]$, we find 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 upper bound is attained when b_a^* is $(1+c_a^2)$ and $b_s^* = M_s$. Hence, the upper bound for the steady-state mean waiting time is attained at (F_0, G_u).

5. The Lower Bound with Finite Support

For unbounded support, Ott (1987) showed that the overall lower bound of $E[W(F,G)]$ for $(F,G) \in \mathcal{P}_{a,2}(\infty) \times \mathcal{P}_{s,2}(\infty)$ 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 lower bound has the convenient explicit formula in (9).

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

$$\lim_{M_a \to \infty} E[W(F_u, G_{u,b_s u})] = E[W(D, A_3)] \tag{16}$$

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 $M_a > 1 + c_a^2$.

5.1. 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 $u = 1 - c_a^2/(M_a - 1)$),

$$p_1 + p_2 + p_3 = 1, x_1 p_1 + x_2 p_2 + x_3 p_3 = \rho, x_1^2 p_1 + x_2^2 p_2 + x_3^2 p_3 = (1 + c_s^2)\rho^2 \tag{17}$$

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

$$\begin{aligned} p_1 &= \frac{(b_s^2(u^2 - \rho u) + b_s(-u^2 + (1 + c_s^2)\rho^2) - (1 + c_s^2)\rho^2 + u\rho)}{(b_s^2 u^2 - b_s u^2)}, \\ p_2 &= \frac{\rho b_s u - (1 + c_s^2)\rho^2)}{b_s u^2 - u^2} \quad \text{and} \quad p_3 = \frac{\rho^2(1 + c_s^2) - u\rho}{b_s^2 u^2 - b_s u^2}. \end{aligned} \tag{18}$$

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

$$b_s^2(u^2 - \rho u) + b_s(-u^2 + (1 + c_s^2)\rho^2) - (1 + c_s^2)\rho^2 + u\rho = 0, \tag{19}$$

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

$$b_s \in \left[\frac{(1 + c_s^2)\rho}{u}, \mathbf{1}_{\{\rho \in (u/(1+c_s^2), u]\}} \infty + \mathbf{1}_{\{\rho > u\}} \gamma \right). \tag{20}$$

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

$$b_s \in \arg\min_b E[W(F_u, G_{u,bu})]. \tag{21}$$

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.,*

$$E[W(F,G)] \geq E[W(F_u, G_{u,b_s u})] \quad \textit{for all} \quad F \in \mathcal{P}_{a,2}(M_a) \quad \textit{and} \quad G \in \mathcal{P}_{s,2}. \tag{22}$$

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.,*

$$E[W(F,G)] \geq E[W(F_0, G_0)] \quad \textit{for all} \quad F \in \mathcal{P}_{a,2}(M_a) \quad \textit{and} \quad G \in \mathcal{P}_{s,2}. \tag{23}$$

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

$$E[W(D, A_3)] \leq E[W(F_u, G_{u,b_s u})] \leq \inf\{E[W(F_u, G_u)] : b \in [1 + c_s^2, \infty)\}. \tag{24}$$

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

$$E[W(D, A_3)] \leq E[W(F_0, G_0)] \leq \inf\{E[W(F_0, G_u)] : b \in [1 + c_s^2, \infty)\}. \tag{25}$$

5.2. 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 (20) 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 (20), 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 $\rho \in (u/(1 + c_s^2), u]$.

5.3. 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 5.3 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

6. Conclusions

We have studied tight upper and lower bounds for the mean steady-state waiting time $E[W]$ and the mean transient waiting time $E[W_n]$ 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. Overall, we have exposed important open problems and constructed and applied numerical algorithms to support Conjectures 1-3.

In §3 we applied simulation to study the special case of two-point interarrival-time and service-time distributions. For this special case, we have experimentally verified the main outstanding Conjecture 1, 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, it still remains to consider a broader range of alternatives and, even for this restricted case, to provide a mathematical proof.

We have also studied cases (a) and (b) of (1). For (a), we have confirmed and elaborated on the counterexample to the optimality of F_0 from §8 of Wolff and Wang (2003) involving G_0, but mostly found that F_0 tends to be at least nearly optimal. For (b), we see that the upper bound for G either is G_u or approaches it for all two-point F. This misses the story in §V of Whitt (1984b) and Theorem 2 of Chen and Whitt (2021b), which shows that G_0 attains the upper bound when F is completely monotone or strictly concave, whereas G_u attains the upper bound when F is strictly convex with finite support. That complexity evidently does not affect the overall tight upper bound in Conjecture 1 because F_0 is far from being strictly concave or strictly concave. In summary, our accumulated

experience is that F_0 tends to be an upper bound for all G without mass on 0, whereas the upper bound for G clearly depends strongly on the structure of F. That complexity is evidently avoided in Conjecture 1 because we focus on F_0 rather than any F.

In §4 we constructed a multinomial representation of the transient mean in the case of three-point distributions and applied it together with nonlinear programming to numerically conclude that the overall upper bound over F and G is always attained at two-point distributions, further supporting Conjecture 1.

In §5 we applied simulation to study the tight lower bound of the transient mean. There we found that a modification of the three-point lower-bound distribution identified by Ott (1987) holds for the transient mean.

There are many remaining problems for research. In addition to providing full mathematical proofs of Conjectures 1-3 or refuting them, it remains to identify the extremal distributions with one distribution given, as in parts (a) and (b) of (1). It also remains to establish similar results for other stochastic models. Hopefully this paper will help advance those goals.

Acknowledgments

This research was supported by NSF CMMI 1634133. Conflicts of Interest: none

References

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

Bandi, C., D. Bertsimas, N. Youssef. 2015. Robust queueing theory. *Operations Research* **63**(3) 676–700.

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. 2020. Algorithms for the upper bound mean waiting time in the $GI/GI/1$ queue. *Queueing Systems* **94** 327–356.

Chen, Y., W. Whitt. 2021a. Applying optimization to study extremal $GI/GI/1$ transient mean waiting times. Submitted for publication, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

Chen, Y., W. Whitt. 2021b. Extremal $GI/GI/1$ queues given two moments: Exploiting Tchebycheff systems. *Queueing Systems* **97** 101–124.

Chen, Y., W. Whitt. 2021c. Set-valued performance approximations for the $GI/GI/K$ queue given partial information. *Probability in the Engineering and Informational Sciences* **35**(2) xxx–yyy.

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.

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

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.

Li, Y., D. A. Goldberg. 2017. Simple and explicit bounds for multi-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.

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.

van Eekelen, W., D. J. den Hartog, J. S. H. van Leeuwaarden. 2019. Mad dispersion measure makes extremal queue anlysis simple. Working paper.

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: Extremal distributions. *AT&T Bell Laboratories Technical Journal* **63**(1) 115–137.

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

Whitt, W., W. You. 2018. Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research* **66**(1) 100–120.

Whitt, W., W. You. 2019. Time-varying robust queueing. *Operations Research* **67**(6) 1766–1782.

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.