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

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

In this paper we show that the upper bound for the mean steady-state waiting time in the classical $GI/GI/1$ queue, given the first two moments of the interarrival-time and service-time distributions, is attained by probability distributions with support on at most three points. We start by focusing on the extremal problem for the transient mean with one of the underlying distributions fixed. We then restrict attention to distributions with finite support. In that context, we apply basic optimization theory to formulate the extremal problem as a non-convex nonlinear program with linear constraints. We show that any local optimum must be a fixed point involving a linear program. We then show that the linear program must have a unique solution, implying that the local optimum must correspond to an extreme point of the linear program, and thus must be a three-point distribution. Finally, we apply asymptotics to obtain corresponding results for the other cases.

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

History: April 23, 2020

1. Introduction

In this paper we address a long-standing open problem for the $GI/GI/1$ queueing model: determining a tight upper bound for the mean steady-state waiting time, and the interarrival-time and service-time distributions that attain it, given the first two moments of these underlying distributions; see Daley et al. (1992), especially §10, Wolff and Wang (2003) and references therein. In this paper we obtain in important partial result: We show that the upper bound is attained by probability distributions with support on at most three points.

We approach this problem by considering two more general problems of independent interest. First, we consider the corresponding two problems with one of the underlying distributions fixed (with finite first two moments) and specified. Second, we consider those problems also for the transient expected waiting time. For the transient mean, we exploit the known explicit formula for the mean. For the case of finite support, we apply basic optimization theory for a non-convex nonlinear program with linear constraints to show that all local optima are attained at three-point distributions. Finally, we apply asymptotics to connect those results to the general cases.

2. Background

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; let a service time V have mean $E[V] \equiv \tau \equiv \rho$ 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 with an initial workload W_0 having cdf H_0 with a finite mean. The sequence $\{W_n : n \geq 0\}$ is well known to satisfy the Lindley recursion

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

where $x^+ \equiv \max\{x, 0\}$. Let 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

$$W \overset{\mathrm{d}}{=} (W + V - U)^+, \tag{2}$$

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

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

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 (3) in our analysis.

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

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

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

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

2.3. Related Literature

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

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

Since the first version of this paper was completed, we have subsequently completed other related papers. In Chen and Whitt (2020) we developed some effective algorithms for the widely conjectured upper bound model, which involves only two-point distributions. There we showed that its mean value is a significant improvement over established bounds, such as in Kingman (1962) and Daley (1977). We also developed an explicit formula for an upper bound for the conjectured tight upper bound, which is very accurate. In Chen and Whitt (2019) we developed sufficient conditions for special two-point extremal distributions by applying the theory of Tchebycheff systems.

2.4. Organization

In §3 we state our main contribution, Theorem 1, which shows that the extremal distributions have support on at most three points. We also outline the proof there. In §4 we prove the part of the main theorem with the service-time distribution fixed. In §5 we prove the corresponding part of the theorem with the interarrival-time distribution fixed. It is shorter because much of the same argument can be used again. In §6 we show that the results in §§4 and 5 can be combined to quickly provide a proof for the case neither distribution is fixed. In §7 we present three asymptotic results needed to complete the overall proof. Finally, in §8 we draw conclusions.

3. The Main Result

In this section we state our main result, Theorem 1, and outline the proof. We start in §3.1 by introducing the notation we will use.

3.1. Notation

Let $\mathcal{P}_n$ be the set of all probability measures on a subset of $\mathbb{R}$ with specified first n moments. The set $\mathcal{P}_n$ is a convex set, because the convex combination of two probability measures is just the mixture; i.e., for all p, $0 \le p \le 1$, $pP_1 + (1-p)P_2 \in \mathcal{P}_n$ if $P_1 \in \mathcal{P}_n$ and $P_2 \in \mathcal{P}_n$, because the n^{th} moment of the mixture is the mixture of the n^{th} moments, which is just the common value of the components.

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. For inter-arrival time, we let $\mathcal{P}_{a,2}(S) \equiv P_{a,2}(1, c_a^2, S)$ be the set of probability measures for inter-arrival time distribution F on $[0, \infty)$ with two moments specified, as determined by the parameter pair $(1, c_a^2)$, and support in the set S. Let $\mathcal{P}_{s,2}(S) \equiv P_{s,2}(\rho, c_s^2, S)$ denote the corresponding set of probability measures for the service time distribution G. For example, if $S = [0, M_a]$, then we write $P_{a,2}(M_a) \equiv P_{a,2}(1, c_a^2, [0, M_a])$; if $S = \mathcal{F}$ where $\mathcal{F}$ is a finite set including ending points in $\{0, M_a\}$, then we write $\mathcal{P}_{a,2}(\mathcal{F}) \equiv \mathcal{P}_{a,2}(1, c_a^2, \mathcal{F})$. If S 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 $\mathcal{P}_{a,2,k}(S)$ denote the subset with support on at most k points within S for various S as above.

We introduce some further simplified notation in our proof. In particular, see the beginning of §4.1 and §5.

3.2. Theorem Statement

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

$$w_n : \mathcal{P}_{a,2}(1, c_a^2) \times \mathcal{P}_{s,2}(\rho, c_s^2) \to [0, \infty), \tag{4}$$

where $0 < \rho < 1$ and

$$w_n(F, G) \equiv E[W_n(F, G)], \qquad 1 \le n \le \infty, \tag{5}$$

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*) *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 (4) 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*

$$w_{a,n}^{\uparrow}(G) \equiv \sup\{w_n(F, G) : F \in \mathcal{P}_{a,2}(M_a)\} = \sup\{w_n(F, G) : F \in \mathcal{P}_{a,2,3}(M_a)\} = w_n(F_n^*(G), G). \tag{6}$$

(*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*

$$w_{s,n}^{\uparrow}(F) \equiv \sup\{w_n(F,G) : G \in \mathcal{P}_{s,2}(M_s)\} = \sup\{w_n(F,G) : G \in \mathcal{P}_{s,2,3}(M_s)\} = w_n(F, G_n^*(F)). \qquad (7)$$

(*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*

$$\begin{aligned} w_n^{\uparrow} &\equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2}(M_a), G \in \mathcal{P}_{s,2}(M_s)\} = \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2,3}(M_a), G \in \mathcal{P}_{s,2,3}(M_s)\} \\ &= w_n(F_n^{**}, G_n^{**}) = w_{a,n}^{\uparrow}(G_n^{**}) = w_{s,n}^{\uparrow}(F_n^{**}). \qquad (8) \end{aligned}$$

Corresponding results hold for each supremum replaced by an infimum.

3.3. Outline of the Proof

We first prove part (a) of Theorem 1. We start with the transient mean $E[W_n]$ with $n < \infty$ and obtain results for the steady-state mean afterwards by an asymptotic argument. For the transient mean, we first do the proof for the special case of finite support, and then treat the general case by a second asymptotic argument. For part (a) given finite support, we do the proof for the case of the fixed service-time cdf G having a positive probability density function (pdf), and then treat the general case by a third asymptotic argument. As required by this logic, at the end in §7 we perform the three asymptotic arguments in reverse order.

For the special case of (a) with finite support, we analyze the optimization problem as a nonlinear program with a smooth non-convex objective function and linear constraints. The basic nonlinear program theory as in Bertsekas (2016) then allows us to show that any local optimum must be the fixed point of a linear program (LP). We then apply duality theory to show that the optimal solution of the LP must be unique, and thus the local optimum must correspond to an extreme point of the linear program, which is a three-point distribution. Starting from a local optimum, the LP yields that given local optimum as the unique solution of the LP, which must be an extreme-point and thus a three-point distribution.

4. Proof of Theorem 1 (a) for $n < \infty$

As indicated above, we start by proving part (a) for the transient waiting time $E[W_n]$ with $n < \infty$ as given in (3). For $n = 1$, the conclusion follows from the classic moment problem reviewed in Smith (1995) because we can express $E[W_1]$ as the integral of a continuous real-valued function over the bounded interval $[0, M_a]$ via

$$E[W_1] = \int_0^{M_a} \phi(u)\, dF(u), \tag{9}$$

where $\phi(u) \equiv E[(V - u)^+]$. Unfortunately, the partial sums lead to convolution, which prevent such a representation for $n \geq 2$. We circumvent that difficulty to some extent in Chen and Whitt (2019) by focusing on the more general $GI_{(n)}/GI/1$ model that allows a different interarrival-time cdf each time period, but we only obtain partial results by that approach.

4.1. Gradient of the Mean Transient Waiting Time

To treat finite n with $n \geq 2$, we consider finite support $\mathcal{F}$ in $\mathcal{P}_{a,2}(M_a)$, i.e., $\mathcal{P}_{a,2}(\mathcal{F})$. Let the elements of $\mathcal{F}$ be $0 = u_1 < u_2 < \ldots < u_m = M_a$ with $m \equiv |\mathcal{F}| \geq 3$. With this assumption, we will simplify the notation. In particular, we will suppress the fixed service-time cdf G and we will replace F by its probability mass function $p \equiv (p_1, \ldots, p_m)$.

With these notation conventions, the optimization in part (a) becomes

$$\begin{aligned}
&\max\{w_n(p) \equiv w_n(F,G) \equiv E[W_n(F,G)] : F \in \mathcal{P}_{a,2}(\mathcal{F})\} \\
&\quad = \max\{\sum_{k=1}^{n} \frac{1}{k} E[S_k^+(p)]\} \\
&\text{such that} \quad \sum_{i=1}^{m} p_i = 1, \quad \sum_{i=1}^{m} u_i p_i = 1, \quad \sum_{i=1}^{m} u_i^2 p_i = (1 + c_a^2), \quad \text{and} \quad p_i \geq 0.
\end{aligned} \tag{10}$$

We now show that the function $w_n(p)$ is a smooth function of $p \equiv (p_1, \ldots, p_m)$. In particular, we show that the gradient is well defined. We do that by showing that the Frechet derivative is well defined. For that purpose, let $\|p\|$ be the l_1 norm in $\mathbb{R}^m$, i.e.,

$$\|p\| \equiv \sum_{i=1}^{m} |p_i|. \tag{11}$$

The function $w_n(p)$ is said to be Frechet differentiable at $\hat{p}$ if the following limit is well defined:

$$\lim_{\|p-\hat{p}\|\to 0} \frac{\left|w_n(p) - w_n(\hat{p}) - \nabla w_n(\hat{p})^t \cdot (p-\hat{p})\right|}{\|p-\hat{p}\|} = 0, \tag{12}$$

where where $\nabla w_n(\hat{p})$ is the gradient, which we regard as an $m \times 1$ column vector,

$$\nabla w_n(\hat{p}) \equiv \left(\left(\frac{\partial w_n}{\partial p_1}(\hat{p}) \right), \ldots, \left(\frac{\partial w_n}{\partial p_m}(\hat{p}) \right) \right)^t \tag{13}$$

with t denoting the transpose of vector in $\mathbb{R}^m$. The gradient is associated with the local linear approximation of $w_n(p)$ at some $\hat{p} \in \mathbb{R}^m$, using the dot product, as

$$w_n(p) \approx w_n(\hat{p}) + \nabla w_n(\hat{p})^t \cdot (p - \hat{p}). \tag{14}$$

Remark 1. (extension) The Frechet derivative can be generalized to Banach spaces using the total variation metric, which in our setting is just $d_{TV}(p,\hat{p}) = (1/2)\|p-\hat{p}\|$; see Ch. 6 of Serfling (1980) and Wang (1993). For example, the following result also holds if the cdf F has a pdf f over $\mathbb{R}$ instead of having finite support. Then $d_{TV}(F_1,F_2) \equiv \int_0^\infty |f_1(x) - f_2(x)|dx$.

Theorem 2. (*Frechet derivative*) *In the finite support setting above, the function $w_n(p)$ is Frechet differentiable with partial derivatives at $\hat{p}$ given by*

$$\frac{\partial w_n}{\partial p_i}(\hat{p}) = \sum_{j=1}^{n} E[(\sum_{k=1}^{j} V_k(G) - \sum_{k=1}^{j-1} U_k(\hat{p}) - u_i)^+], \tag{15}$$

so that

$$\nabla w_n(\hat{p})^t \cdot (p-\hat{p}) = \sum_{i=1}^{m} \frac{\partial w_n}{\partial p_i}(\hat{p})(p_i - \hat{p}_i). \tag{16}$$

Proof. We do the proof for $n = 2$; the argument for higher n is analogous. For any real-valued functions $f(x)$ and $g(x)$, let $f(x) = \Theta(g(x))$ denote that there exists $m, M > 0$ such that $mg(x) \leq |f(x)| \leq Mg(x)$ for all x. Then, adding and subtracting by $\hat{p}_i$ and $\hat{p}_j$ inside the expression for $w_2(p)$, we get

$$w_2(p) = \sum_i E[(V_1 - u_i)^+]p_i + \frac{1}{2}\sum_{i,j} E[(V_1 + V_2 - u_i - u_j)^+]p_i p_j$$

$$
\begin{aligned}
&= \sum_i E[(V_1-u_i)^+](p_i-\hat{p}_i+\hat{p}_i) \\
&\quad + \frac{1}{2}\sum_{i,j} E[(V_1+V_2-u_i-u_j)^+](p_i-\hat{p}_i+\hat{p}_i)(p_j-\hat{p}_j+\hat{p}_j) \\
&= \sum_i E[(V_1-u_i)^+]\hat{p}_i + \frac{1}{2}\sum_{i,j} E[(V_1+V_2-u_i-u_j)^+]\hat{p}_i\hat{p}_j \\
&\quad + \sum_i E[(V_1-u_i)^+](p_i-\hat{p}_i) + \sum_i E[(V_1+V_2-U_1(\hat{F})-u)^+(p_i-\hat{p}_i) + \Theta(\|p-\hat{p}\|) \\
&= w_2(\hat{p}) + \sum_i \nabla w_2(\hat{p})^t(p_i-\hat{p}_i) + \Theta(\|p-\hat{p}\|)^2),
\end{aligned}
\tag{17}
$$

where

$$
\frac{\partial w_2}{\partial p_i}(\hat{p}) = \sum_{j=1}^{2} E[(\sum_{k=1}^{j} V_k(\hat{G}) - \sum_{k=1}^{j-1} U_k(F) - u_i)^+]. \tag{18}
$$

To justify the conclusion in (17), we observe that there exists a constant C such that $E[(V_1 + V_2 - u_i - u_j)^+] \le C < \infty$ for all i and j. Consequently, the second term in the second line of (17) associated with the second order of $(p_i - \hat{p}_i)$ can be bounded by the square of the norm, in particular,

$$
\begin{aligned}
&\Big|\frac{1}{2}\sum_{i,j} E[(V_1+V_2-u_i-u_j)^+](p_i-\hat{p}_i)(p_j-\hat{p}_j)\Big| \le C\sum_{i,j}\Big|(p_i-\hat{p}_i)(p_j-\hat{p}_j)\Big| \\
&\le C\sum_{i,j}\big|(p_i-\hat{p}_i)\big|\big|(p_j-\hat{p}_j)\big| = C\|p-\hat{p}\|^2.
\end{aligned}
\tag{19}
$$

Therefore, as $\|p-\hat{p}\| \to 0$,

$$
\frac{\big|w_2(p) - w_2(\hat{p}) - \sum_i \frac{\partial w_2}{\partial p_i}(\hat{p})(p_i-\hat{p}_i)\big|}{\|p-\hat{p}\|} \le C\frac{\|p-\hat{p}\|^2}{\|p-\hat{p}\|} = C\|p-\hat{p}\| \to 0. \tag{20}
$$

Hence, we have shown that $w_n(p)$ is Frechet differentiable. ▪

4.2. Local Optimality

There exists a global optimum because we are maximizing a continuous function over a compact subset of $\mathbb{R}^m$. Recall that a point $\hat{p}$ is a local optimum for (10) if there exists $\delta > 0$ such that

$$
w_n(p) \le w_n(\hat{p}) \quad \text{for all} \quad p \quad \text{such that} \quad \|p-\hat{p}\| < \delta. \tag{21}
$$

Clearly, there exists at least one local optimum because the global optimum is necessarily a local optimum. We apply the following necessary condition for a local optimum from Proposition 3.1.1 of Bertsekas (2016).

PROPOSITION 1. (*necessary condition for a local optimum, Proposition 3.1.1 of Bertsekas (2016)*) *If $\hat{p}$ is a local optimum of $w_n(p)$, then*

$$\nabla w_n(\hat{p})^t \cdot (p - \hat{p}) \leq 0 \quad \textit{for all} \quad p \in \mathcal{P}_{a,2}(\mathcal{F}). \tag{22}$$

Given the special linear form of the constraints, we obtain the following corollary.

COROLLARY 1. (*LP fixed point property*) *If $\hat{p}$ is a local optimal solution for $w_n(p)$ in (10), then $p = \hat{p}$ is an optimal solution of the following linear program*

$$\sup\{\nabla w_n(\hat{p})^t \cdot p \equiv \sum_{i=1}^{m} \frac{\partial w}{\partial p_i}(\hat{p}) p_i : p \in \mathcal{P}_{a,2}(\mathcal{F})\}. \tag{23}$$

Proof. Clearly $p = \hat{p}$ in (22) yields equality. Then (23) is obtained directly from (22) by subtracting the constant term. ■

We will prove that, if $\hat{p}$ is a local optimum of $w_n(p)$ in $\mathcal{P}_{a,2}(\mathcal{F})$, then $\hat{p}$ is a three-point distribution. To do so, we will show that, for any $\hat{p}$, the LP under $\hat{p}$ in Corollary 1 necessarily has a unique optimal solution. Thus the unique optimal solution, which coincides with $\hat{p}$, must be an extreme point, i.e., in $\mathcal{P}_{a,2,3}(\mathcal{F})$. For that purpose, we impose regularity conditions on the fixed service-time cdf G.

4.3. Uniqueness in the Linear Program

We now prove that the LP in (23) has a unique solution, assuming that the fixed service-time cdf G has a positive pdf g over $[0, \infty)$. That condition is satisfied if G is a phase-type distribution or has a rational Laplace transform; see Asmussen (2003) or §II.5.10 of Cohen (1982). We will later relax this condition in our asymptotic argument in §7.

THEOREM 3. (*uniqueness in the LP for local optimality*) *Suppose that G has a positive pdf g over $[0, \infty)$. If $\hat{p}$ is a local optimal solution of (10) for $2 \leq n < \infty$, then the LP given $\hat{p}$ in (23) has a unique optimal solution, which must be $\hat{p}$. Thus $\hat{p}$ must be an extreme point, so that $\hat{p} \in \mathcal{P}_{a,2,3}(\mathcal{F})$.*

Proof. We apply duality theory for the LP in (23). From basic LP duality theory as in Ch. 4 of Bertsimas and Tsitsiklis (1997), the dual problem associated with the LP in (23) is to find the vector $\lambda^* \equiv (\lambda_0^*, \lambda_1^*, \lambda_2^*)$ that attains the minimum

$$\min\{\lambda_0 + \lambda_1 + \lambda_2(1 + c_a^2)\}$$

$$\text{such that} \quad \psi(u_i) \equiv \lambda_0 + \lambda_1 u_i + \lambda_2 u_i^2 \ge \phi_a(u_i) \quad \text{for all} \quad i, \quad 1 \le i \le m. \tag{24}$$

where the decision variables λ_i are unconstrained and

$$\phi_a(u) \equiv \sum_{i=1}^{n} E[(\sum_{k=1}^{i} V_k(G) - \sum_{k=1}^{i-1} U_k(\hat{p}) - u)^+], \quad u \ge 0. \tag{25}$$

We apply the following lemma; e.g., see pp. 1128-9 of Appa (2002).

Lemma 1. (*non-degeneracy and uniqueness in LP*) *A standard LP has a unique optimal solution if and only if its dual has a non-degenerate optimal solution.*

From (25), we see that the constraints produce the quadratic function $\psi(u)$ that is required to dominate $\phi_a(u)$ for all $u \in \mathcal{F}$. We now use the following lemma.

Lemma 2. (*structure of the objective function*) *If the fixed cdf G of V has a positive pdf g over $[0, \infty)$, then the random variable $Y_i \equiv \sum_{k=1}^{i} V_k - \sum_{k=1}^{i-1} U_k$ has a cdf Γ_i with support in $[-(i-1)M_a, \infty)$ which has a positive pdf γ_i over $[0, \infty)$ for each i, $1 \le i \le m$. Hence, for $x > 0$, the cdf of Y_i can be expressed by*

$$\Gamma_i(x) = \Gamma_i(0) + \int_0^x \gamma_i(y)\, dy \quad \textit{for} \quad x \ge 0, \tag{26}$$

so that the function ϕ_a in (25) can be expressed as

$$\phi_a(u) \equiv \frac{\partial w_n}{\partial p}(\hat{p}) = \sum_{i=1}^{n} \int_0^{\infty} (x - u)^+ \gamma_i(x)\, dx > 0, \quad u \ge 0. \tag{27}$$

Hence, $\phi_a(u) > 0$ and the first two derivatives of ϕ_a in (25) exist for $u > 0$ and satisfy

$$\dot{\phi_a}(u) = \sum_{i=1}^{n} (\Gamma_i(u) - 1) < 0, \qquad \ddot{\phi_a} = \sum_{i=1}^{n} \gamma_i(u) > 0, \quad u \ge 0. \tag{28}$$

Thus, ϕ_a is continuous, strictly decreasing and strictly convex on $[0, M_a]$.

Proof. For the initial pdf property, see §V.4 of Feller (1971). To calculate the derivatives, we apply the Leibniz integral rule for differentiation of integrals of integrable functions that are differentiable almost everywhere. Observe that the derivative of $(x-u)^{+}\gamma_i(x)$ with respect to u is $-\gamma_i(x)$ for $u<x$. That implies that

$$\dot{\phi}_a(u) = -\sum_{i=1}^{n}\int_{u}^{\infty}\gamma_i(x)\,dx = \sum_{i=1}^{n}(\Gamma_i(u)-1). \tag{29}$$

The rest follows directly. ■

To continue the proof, we now show that the dual problem has a non-degenerate optimal solution. To do so, we exploit the structure of the function ϕ_a in (25) established in Lemma 2. Under the condition, ϕ_a is continuous, strictly positive, strictly decreasing and strictly convex. Recall that we are working with standard LP's, where the cdf F has finite support set $\mathcal{F}$, but the support set $\mathcal{F}$ always contains the two endpoints 0 and M_a.

The inequality constraints in (24) are only required to hold at the finitely many point in the support set $\mathcal{F}$. Even though we exploit the structure of continuous functions, the following argument applies to any finite support set.

If $M_a = m_2$, then the primal has the unique feasible, and thus optimal, two-point feasible distribution with masses on 0 and m_2. So henceforth assume that $M_a > m_2$ as well. We start knowing that both the dual LP (24) and the primal LP (23) have feasible solutions and the feasible region of the primal LP (23) is compact, thus they both have at least one optimal solution. We will show that the primal LP (23) has a unique solution by applying Lemma 1 and showing that no optimal solution of the dual (24) can be degenerate. That implies that the dual has at least one non-degenerate optimal solution. Hence, we will show that we cannot have the optimal λ_i^* be 0 for any i.

First, we must have $\lambda_0 \geq \phi_a(0) > 0$, so we cannot have $\lambda_0^* = 0$. Next, suppose that $\lambda_1 = 0$. In this setting, with $\lambda_0^* > 0$ and $\lambda_1^* = 0$, if $\lambda_2^* \geq 0$, then ψ can intersect ϕ_a only at 0, which cannot correspond to a feasible solution of the primal. (We exploit complementary slackness here and in the following.) On the other hand, if $\lambda_2^* < 0$, then ϕ_a can only intersect ψ at the two endpoints, without violating the conditions at the endpoints, but that does not correspond to a feasible solution of the primal, assuming that $M_a > m_2$. Hence, we cannot have a degenerate optimal solution with $\lambda_1^* = 0$. Finally, suppose that $\lambda_2^* = 0$, which makes ψ linear. If $\lambda_0 = \phi_a(0) > 0$, then again ψ can only meet ϕ_a at the two endpoints without violating the conditions at the endpoints, but that does not correspond to a feasible solution of the primal, assuming that $M_a > m_2$. Otherwise, ψ can only have one intersection point with ϕ_a (as we have done). ■

4.4. Uniqueness for Minimization

The proof for the corresponding minimization problem of (23) is very similar, but is somewhat more complex because it requires more care in treating the underlying finite support set $\mathcal{F}$ of F. We now assume that $1 \in \mathcal{F}$ (so that the fixed mean is in $\mathcal{F}$) as well as the two endpoints 0 and M_a.

As before, we apply Lemma 1 to shows uniqueness, which leads to the maximization version of the dual LP in (24), i.e.,

$$\begin{aligned} &\max\{\lambda_0 + \lambda_1 + \lambda_2(1 + c_a^2)\} \\ &\text{such that} \quad \psi(u_i) \equiv \lambda_0 + \lambda_1 u_i + \lambda_2 u_i^2 \le \phi_a(u_i) \quad \text{for all} \quad i, \quad 1 \le i \le m. \end{aligned} \tag{30}$$

where the decision variables λ_i are again unconstrained.

Next, as before, we show that the dual LP does not have any degenerate solution. Thus, suppose that $(\lambda_0^*, \lambda_1^*, \lambda_2^*)$ is an optimal solution for the dual. First, if $\lambda_2^* = 0$, then $\psi(u)$ must be linear, so that the intersections of $\psi(u)$ and $\phi_a(u)$ can only occur at two adjacent points, so that it cannot correspond to a feasible solution. Since $1 \in \mathcal{F}$, these points both must be ≥ 1 or both must be ≤ 1, but neither pair of points corresponds to a feasible solution.

Second, we consider two cases to analyze λ_1^*. If $\lambda_0^* = \phi_a(0) > 0$, by considering a Taylor series at the origin, then we must have $\lambda_1^* \le \dot{\phi}_a(0) < 0$. On the other hand, if $\lambda_1^* = 0$ and $\lambda_0^* < \phi_a(0)$, then we must have $\lambda_2^* < 0$ to have any intersection. However, if $\lambda_2^* < 0$, then there can be an intersection of $\psi(u)$ and $\phi_a(u)$ only at two adjacent points in the support set $\mathcal{F}$. So again that cannot correspond to a feasible solution.

Finally, we rule out the case $\lambda_0^* = 0$. Under the condition $\lambda_0^* = 0$, the function ψ either first decreases and then increases or it first increases and then decreases. If it first decreases, then it can only have an intersection at the right endpoint. Hence, it cannot correspond to a feasible solution. Finally, if it first increases, then it can again can only have an intersection at two adjacent points in $\mathcal{F}$, so that it cannot correspond to a feasible solution. it cannot correspond to a feasible solution.

■

5. Proof of Theorem 1 (b)

The proof is mostly the same as for part (a), so we will be brief. Paralleling §4.1, to treat finite n with $n \ge 2$, we consider finite support $\mathcal{G}$ in $\mathcal{P}_{s,2}(M_s)$, i.e., $\mathcal{P}_{s,2}(\mathcal{G})$. Let the elements of $\mathcal{G}$ be $0 = v_1 < v_2 < \ldots < v_m = \rho M_s$ with $m \equiv |\mathcal{G}| \ge 3$. As before, we will simplify the notation. In particular, we will suppress the fixed interarrival-time cdf F and we will replace G by its probability mass function $q \equiv (q_1, \ldots, q_m)$.

With these notation conventions, paralleling (10), the optimization in part (b) becomes

$$\begin{aligned}
&\max\{w_n(q) \equiv w_n(F,G) \equiv E[W_n(F,G)] : G \in \mathcal{P}_{s,2}(\mathcal{G})\}\\
&\quad = \max\{\sum_{k=1}^{n} \frac{1}{k} E[S_k^+(q)]\}\\
&\text{such that} \quad \sum_{i=1}^{m} q_i = 1, \quad \sum_{i=1}^{m} v_i q_i = 1, \quad \sum_{i=1}^{m} v_i^2 q_i = \rho^2(1+c_s^2), \quad \text{and} \quad q_i \ge 0.
\end{aligned} \tag{31}$$

We then have the following analog of the differentiability result in Theorem 2. We omit the identical proof.

Theorem 4. (*Frechet derivative for (b)*) *In the finite support setting above, the function $w_n(q)$ is Frechet differentiable with partial derivatives at $\hat{q}$ given by*

$$\frac{\partial w_n}{\partial q_i}(\hat{q}) = \sum_{j=1}^{n} E[(\sum_{k=1}^{j-1} V_k(\hat{q}) - \sum_{k=1}^{j} U_k(F) + v_i)^+], \tag{32}$$

so that

$$\nabla w_n(\hat{q})^t \cdot (q - \hat{q}) = \sum_{i=1}^{m} \frac{\partial w_n}{\partial q_i}(\hat{q})(q_i - \hat{q}_i). \tag{33}$$

Turning to local optimality, we have the following fixed point property, just as in Corollary 1.

Corollary 2. (*LP fixed point property for (b)*) *If $\hat{q}$ is a local optimal solution for $w_n(q)$ in* (31), *then $q = \hat{q}$ is an optimal solution of the following linear program*

$$\sup\{\nabla w_n(\hat{q})^t \cdot q \equiv \sum_{i=1}^{m} \frac{\partial w}{\partial q_i}(\hat{q}) q_i : q \in \mathcal{P}_{s,2}(\mathcal{G})\}. \tag{34}$$

Turning to uniqueness in the LP (34), we have the analog of Theorem 3.

Theorem 5. (*uniqueness in the LP for local optimality in (b)*) *Suppose that F has a positive pdf f over $[0,\infty)$. If $\hat{q}$ is a local optimal solution of* (31) *for $2 \le n < \infty$, then the LP given $\hat{q}$ in* (34) *has a unique optimal solution, which must be $\hat{q}$. Thus $\hat{q}$ must be an extreme point, so that $\hat{q} \in \mathcal{P}_{s,2,3}(\mathcal{G})$.*

Just as for Theorem 3, we exploit duality to prove Theorem 5. However, to carry out this step, it is convenient to perform a change of variables so that we can directly apply the detailed argument used for part (a). In particular, we replace v by

$$\tilde{v} \equiv \rho M_s - v. \tag{35}$$

With this change, we can write the LP in (34) as

$$\begin{aligned} &\max\{w_n(q) \equiv w_n(F,G) \equiv E[W_n(F,G)] : G \in \mathcal{P}_{s,2}(\mathcal{G})\} \\ &\quad = \max\{\sum_{i=1}^{m} \tilde{c}_i q_i\} \\ &\text{such that} \quad \sum_{i=1}^{m} q_i = 1, \quad \sum_{i=1}^{m} \tilde{v}_i q_i = \tilde{m}_1, \quad \sum_{i=1}^{m} \tilde{v}_i^2 q_i = \tilde{m}_2, \quad \text{and} \quad q_i \ge 0, \end{aligned} \tag{36}$$

where, from Theorem 4,

$$\tilde{c}_i \equiv \sum_{j=1}^{n} E[(\sum_{k=1}^{j-1} V_k(\hat{q}) - \sum_{k=1}^{j} U_k(F) + \rho M_s - \tilde{v}_i)^+] \tag{37}$$

and

$$\tilde{m}_k \equiv E[(\rho M_s - V)^k], \quad k = 1, 2. \tag{38}$$

.

For the optimization problem in (36), we have the associated dual

$$\begin{aligned}&\min\{\lambda_0+\lambda_1\tilde{m}_1+\lambda_2\tilde{m}_2\}\\ &\text{such that}\quad \psi(\tilde{v}_i)\equiv\lambda_0+\lambda_1\tilde{v}_i+\lambda_2\tilde{v}_i^2\ge\phi_s(\tilde{v}_i)\quad\text{for all}\quad i,\quad 1\le i\le m,\end{aligned}\tag{39}$$

where the decision variables λ_i are unconstrained and

$$\phi_s(\tilde{v})\equiv\sum_{i=1}^{n}E[(\sum_{k=1}^{i-1}V_k(\hat{q})-\sum_{k=1}^{i}U_k(F)+\rho M_s-\tilde{v})^+].\tag{40}$$

We now have an analog of Lemma 2. Let $x^-\equiv\min\{x,0\}$ and recall that $x=x^++x^-$.

Lemma 3. (*structure of the objective function for (b)*) *If the fixed cdf F of U has a positive pdf f over $[0,\infty)$, then $Z_i\equiv\sum_{k=1}^{i-1}V_k(\hat{q})-\sum_{k=1}^{i}U_k(F)+\rho M_s$ has support in $(-\infty,\rho M_s+(i-1)a]$, where $a>0$ is the upper limit of the support of V. Thus Z_i has a positive pdf θ_i over $(-\infty,\rho M_s]$ for each i, $1\le i\le m$. Hence,*

$$\begin{aligned}\phi_s(\tilde{v})&=\sum_{i=1}^{n}E[(Z_i-\tilde{v})^+]=\sum_{i=1}^{n}\left(E[Z_i-\tilde{v}]-E[(Z_i-\tilde{v})^-]\right),\\ &=\sum_{i=1}^{n}\left(E[Z_i-\tilde{v}]-\int_{-\infty}^{\rho M_s}(x-\tilde{v})^-\,d\Theta_i(x)\right),\end{aligned}\tag{41}$$

where

$$\Theta_i(x)=\int_{-\infty}^{x}\theta_i(y)\,dy\quad\text{for}\quad x\le\rho M_s,\tag{42}$$

so that, paralleling Lemma 2, the first two derivatives of $\phi_s(\tilde{v})$ are

$$\dot{\phi}_s(\tilde{v})=\sum_{i=1}^{n}(\Theta_i(\tilde{v})-1))<0\quad\text{and}\quad\ddot{\phi}_s(\tilde{v})=\sum_{i=1}^{n}\theta_i(\tilde{v})>0,\quad\tilde{v}\in[0,\rho M_s].\tag{43}$$

Thus $\phi_s(\tilde{v})$ in (40) is a continuous, strictly positive, strictly decreasing and strictly convex function on $[0,\rho M_s]$.

Proof. Just as in Lemma 2, we differentiate the integral to go from (41) to (43). For each term in the sum for $\dot{\phi}_s(\tilde{v})$, we get -1 from the first term in (41) and $\Theta_i(\tilde{v})$ from the second. ■

With Lemma 3, the rest of the proof for (b) can use the same detailed argument used for part (a). However, we must recall the change of variables made in (35). For example, 0 appears in the extremal cdf for F if and only if ρM_s appears in the extremal cdf for G.

6. Proof of Theorem 1 (c)

Part (c) of Theorem 1 follows directly from the conclusions of parts (a) and (b) because the two iterated suprema coincide with the joint supremum; e.g., see Lemma EC.1 of Whitt and You (2018).

7. Three Remaining Asymptotic Arguments

We now complete the proof of Theorem 1 (a) by providing three asymptotic arguments. In §7.1 we relax the condition that G have a positive pdf in Theorem 3. In §7.2 we relax the condition that F have finite support, introduced in §4.1. Finally, in §7.3 we extend the result from the transient mean to the steady-state mean. These asymptotic arguments apply to (b) and (c) as well.

All three asymptotic arguments are based on continuity and compactness. There are four key facts: (i) the bounded interval $[0, M_a]$ is a compact metric space; (ii) the space of probability measures on a compact metric space is itself a compact metric space; (iii) any sequence from a compact metric space has a convergent subsequence with the limits of all convergent subsequences being in that compact metric space; and (iv) the mean waiting time functions in (4) and (5) are continuous functions. To go beyond the compact setting, we apply Prohorov's theorem and the notion of tightness; e.g., see §11.6 of Whitt (2002). Let $\Rightarrow$ denote convergence in distribution.

7.1. Relaxing the Positive Density Condition

In our proof of Theorem 1 (a), we applied Theorem 3, which required that the service-time cdf G have a positive density. We now relax that condition. We observe that the class of cdf's G with positive densities in $\mathcal{P}_{s,2}$ is dense in the class of all cdf's in $\mathcal{P}_{s,2}$.

To give a concrete demonstration, consider the subclass of $\mathcal{K} \bigcap \mathcal{P}_{s,2}$, where $\mathcal{K}$ is the set of all distributions with rational Laplace transform, as in §II.5.10 of Cohen (1982), i.e., the union of $\mathcal{K}_n$ where n is the order of the polynomial in the denominator.

Lemma 4. (*a dense subset*) *The subset* $\mathcal{K} \bigcap \mathcal{P}_{s,2}$ *is a dense subset of* $\mathcal{P}_{s,2}$.

Proof. Observe that any point mass on the positive halfline can be expressed as the limit of Erlang E_n distributions, which are in $\mathcal{K}_n$, with fixed mean and variance approaching 0 as $n \to \infty$. Thus, any distribution with finite support is the limit of finite mixtures of E_n distributions (which also are in $\mathcal{K}_n$). Since arbitrary distributions can be expressed as limits of distributions with finite support, we see that the conclusion holds. ■

For any fixed $G \in \mathcal{P}_{s,2}$ and $n \geq 1$, let G_n be a cdf in $\mathcal{K}_n \bigcap \mathcal{P}_{s,2}$ such that $G_n \Rightarrow G$ as $n \to \infty$. Let F_n^* be a extremal distribution in $\mathcal{P}_{a,2,3}(M_a)$ associated with G_n for each n from Theorem 3. Since the set $\mathcal{P}_{a,2,3}(M_a)$ is compact, the sequence $\{F_n^* : n \geq 1\}$ is tight; e.g., see §11.6 of Whitt (2002). Hence, it contains a convergent subsequence with limit F^*, which is in $\mathcal{P}_{a,2,3}(M_a)$ because it is compact. Finally, because the mean transient waiting time is a continuous function, the limit F^* is an extremal distribution associated with the limiting G. To verify the extremal property, suppose that $\hat{F}$ is an alternative cdf in $\mathcal{P}_{a,2}(M_a)$ such that $w_n(\hat{F}, G) > w_n(F^*, G)$. That would then require that $w_n(\hat{F}, G_n) > w_n(F^*, G_n)$ for sufficiently large n, but that cannot occur. ■

7.2. Relaxing the Finite Support Condition

We now show how to relax the finite support condition introduced for part (a) in §4.1. For that purpose, we will consider a sequence of nested support sets. We say that a sequence of support sets $\{\mathcal{F}_k : k \geq 1\}$ is nested if $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all $n \geq 1$. We say that $\mathcal{F}_k \to [0, M_a]$ as $k \to \infty$ for a nested sequence of support sets if each $x \in [0, M_a]$ can be expressed as

$$x = \lim_{k \to \infty} \{x_k : x_k \in \mathcal{F}_k\}. \tag{44}$$

We have the following approximation lemma.

Lemma 5. (*approximation lemma*) *If $\mathcal{F}_k \to [0, M_a]$ as $k \to \infty$ for a nested sequence of support sets, then Any cdf $F \in \mathcal{P}_{a,2}([0, M_a])$ can be expressed as the limit of cdf's $F_k \in \mathcal{P}_{a,2}([0, \mathcal{F}_k])$.*

Lemma 6. (*extremal cdf for support* $[0, M_a]$) *Assume that $\mathcal{F}_k \to [0, M_a]$ as $k \to \infty$ for a nested sequence of support sets. If $F_k^* \in \mathcal{P}_{a,2,3}(\mathcal{F}_k)$ is the optimal cdf for support set $\mathcal{F}_k$, then there exists a convergent subsequence of $\{F_k^* : k \geq 1\}$ with limiting cdf $F^* \in \mathcal{P}_{a,2,3}([0, M_a])$ and the cdf F^* is an optimal cdf in $\mathcal{P}_{a,2,3}([0, M_a])$.*

Proof. The key fact is that $\mathcal{P}_{a,2,3}([0,M_a])$ is a compact subset of $\mathcal{P}_{a,2}([0,M_a])$. That implies the existence of the convergent subsequence with a limit in the same space. Then the continuity implies the extremal property in the limit. ■

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

So far, we have established the results for $w_n(F,G)$ in (4) and (5) for $n<\infty$. We now show that these results extend to the steady-state mean with $n=\infty$.

THEOREM 6. (*reduction to the transient mean*) *Consider the $GI/GI/1$ queues in Theorem 1.*

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

$$w_n(F_n,G) = w^{\uparrow}_{a,n}(G) \equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2,3}(M_a)\} \quad \textit{for all} \quad n \geq 1, \tag{45}$$

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

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

$$w_n(F,G_n) = w^{\uparrow}_{s,n}(F) \equiv \sup\{w_n(F,G) : G \in \mathcal{P}_{s,2,3}(M_s)\} \quad \textit{for all} \quad n \geq 1, \tag{46}$$

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

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

$$w_n(F_n,G_n) = w^{\uparrow}_{n} \equiv \sup\{w_n(F,G) : F \in \mathcal{P}_{a,2,3}(M_a), G \in \mathcal{P}_{s,2,3}(M_s)\} \quad \textit{for all} \quad n \geq 1, \tag{47}$$

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

Proof. We only prove (c), because the others are proved in the same way. As observed before, because the support sets $[0, M_a]$ and $[0, \rho M_s]$ are compact intervals, the spaces $\mathcal{P}_{a,2}(M_a)$, $\mathcal{P}_{s,2}(M_s)$ and their product are compact metric spaces, as are the spaces $\mathcal{P}_{a,2,3}(M_a)$, $\mathcal{P}_{s,2,3}(M_s)$ and their product, because they are closed subsets. Hence the tightness follows, which implies that there exists a convergent subsequence by Prohorov's theorem in §11.6 of Whitt (2002) and the limit (F,G) of any such subsequence $\{(F_{n_k}, G_{n_k}) : k \geq 1\}$ must remain in the space $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$. Suppose that (F', G') is another candidate pair of cdf's in $\mathcal{P}_{a,2,3}(M_a) \times \mathcal{P}_{s,2,3}(M_s)$. By the assumed optimality, we must have $w_{n_k}(F_{n_k}, G_{n_k}) \geq w_{n_k}(F', G')$ for all k. Then, by continuity, using §X.6 of Asmussen (2003) again, we conclude that $w^{\uparrow} = w(F,G)$ for the steady-state mean. ■

By the same reasoning, an analog of Theorem 6 holds for two-point distributions.

Corollary 3. *In the setting of Theorem 6, (i) if $F_n \in \mathcal{P}_{a,2,2}(M_a)$ for all n in (a), then $F \in \mathcal{P}_{a,2,2}(M_a)$; if $G_n \in \mathcal{P}_{s,2,2}(M_s)$ for all n in (b), then $G \in \mathcal{P}_{s,2,2}(M_s)$; if $(F_n, G_n) \in \mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$ for all n in (c), then $(F,G) \in \mathcal{P}_{a,2,2}(M_a) \times \mathcal{P}_{s,2,2}(M_s)$.*

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

8. Conclusions

We have proved Theorem 1, establishing that the upper bounds of the transient and steady-state mean waiting time, given two specified moments of the underlying interarrival-time and/or service-time distribution, are attained by three-point distributions. We provided a detailed proof of Theorem 1 (a), after which the other cases (b) and (c) follow relatively easily. Our proof of part (a) in §4 is based on first treating the transient mean with finite support. In that setting we applied basic optimization theory to the explicit formula for the transient mean in (3). In that setting, we obtain the non-convex nonlinear program with linear constraints shown in (10). Corollary 1 concludes that any local optimum must be the fixed point of a linear program. Under an extra density condition for the fixed service-time distribution, Theorem 3 concludes that the linear program has a unique solution, implying that the local optimum must be an extreme point, which is a three-point distribution. In §7 we applied asymptotic arguments to extend the results to the general setting.

There is much yet to do. In particular, it remains to identify the actual tight bounds and the extremal distributions that attain them. The answers are known in special cases, as discussed in Chen and Whitt (2019). The conjectured overall optimum in (c) is discussed extensively in Chen and Whitt (2020), where algorithms are developed. The results here can be the basis for numerical algorithms to explore these problems further. In that regard, the LP in Corollary 1 can be a useful tool to verify candidate optima as well as generate counterexamples. A complication, though, is determining the form of the objective function in that LP. So far, it seems necessary to resort to simulation for that purpose. Hopefully the methods here will be useful for other problems and stimulate more research.

Acknowledgments

This research was supported by NSF CMMI 1634133.

References

Appa, G. 2002. On the uniqueness of solutions to linear programs. *Journal of the Operational Research Society* **53** 1127–1132.

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

Bertsekas, D.P. 2016. *Nonlinear Programming*. 3rd ed. Athena Scientific.

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

Bertsimas, D., J. N. Tsitsiklis. 1997. *Introduction to Linear Optimization*. Athena, Belmont, MA.

Chen, Y., W. Whitt. 2019. Extremal $GI/GI/1$ queues given two moments: Exploiting Tchebycheff systems. Submitted for publication, Columbia University, http://www.columbia.edu/∼ww2040/allpapers.html.

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

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

Cohen, J. W. 1982. *The Single Server Queue*. 2nd ed. North-Holland, Amsterdam.

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

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

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

Feller, W. 1971. *An Introduction to Probability Theory and its Applications*. Second edition ed. John Wiley, New York.

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.

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

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

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

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

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 multii-server queues with universal $1/(1-\rho)$ and better scaling. ArXiv:1706.04628v1.

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

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

Serfling, R.J. 1980. *Approximations Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

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

Wang, T. 1993. l_p-frechet differentiable preference and local utility analysis. *Journal of Economic Theory* **61** 139–159.

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

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

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

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

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

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