12-2016

Appointment systems under service level constraints

David CHEN

Rowan WANG
Singapore Management University, ROWANWANG@smu.edu.sg

Zhenzhen YAN

Saif BENJAAFAR

Oualid JOUINI

Citation

CHEN, David; WANG, Rowan; YAN, Zhenzhen; BENJAAFAR, Saif; and JOUINI, Oualid. Appointment systems under service level constraints. (2016).
Available at: https://ink.library.smu.edu.sg/lkcsb_research/5302

Appointment Systems under Service Level Constraints

David Chen[1] • **Rowan Wang**[2] • **Zhenzhen Yan**[3] • **Saif Benjaafar**[4] • **Oualid Jouini**[5]

[1] Rotman School of Management, University of Toronto, Toronto, Ontario, Canada M5S 3E6

[2] Lee Kong Chian School of Business, Singapore Management University, Singapore 178899

[3] Department of Decision Sciences, National University of Singapore, Singapore 119245

[4] Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, Minnesota 55455

[5] Laboratoire Génie Industriel, Ecole Centrale Paris, Châtenay-Malabry, France 92290

Rui.Chen@Rotman.Utoronto.Ca • rowanwang@smu.edu.sg • a0109727@nus.edu.sg • saif@umn.edu • oualid.jouini@ecp.fr

July 15, 2016

Abstract

We consider a new model of appointment scheduling where customers are given the earliest possible appointment times under the service level constraint that the expected waiting time of each individual customer cannot exceed a given threshold. We apply the theory of majorization to analytically characterize the structure of the optimal appointment schedule. We show that, the optimal inter-appointment times increase with the order of arrivals. That is, the optimal inter-arrival time between two customers later in the arrival process is longer than that between two customers earlier in the arrival process. We study the limiting behavior of our system, and prove that, when customer service times follow an exponential distribution, our system converges asymptotically to the D/M/1 queueing system as the number of arrivals approaches infinity. We also extend our analysis to systems with multiple servers.

Keywords: Appointment scheduling; service level constraint; waiting time; majorization

1 Introduction

In this paper, we tackle the classical appointment scheduling problem from a completely new angle. We study an appointment system where a finite number of customers are scheduled to arrive in such a way that (1) the expected waiting time of each individual customer cannot exceed a given threshold, and (2) the appointment times are set as early as possible (without breaking the waiting time constraint). Using a transient queueing analysis approach, we analytically characterize the structure of the optimal appointment schedule and prove the limiting behavior of our system. Compared with the literature, our paper brings unique features in both modeling perspectives and analysis methods. We discuss in detail these new features in the following subsections.

1.1 Modeling Perspective

The fundamental principle of appointment scheduling is on the balance between servers' idling (when appointments are scheduled far from each other) and customers' waiting (when appointments are scheduled close to each other). For decades, appointment scheduling has drawn significant attention in the queueing, optimization, operations management, and health care research communities; see Cayirli and Veral (2003) and Mondschein and Weintraub (2003) for comprehensive reviews of the literature. As pointed out by Cayirli and Veral (2003), the overwhelming majority of the studies assign unit costs (weights) to servers' idling and customers' waiting and then search appointment schedules that minimize the expected total system cost which is a linear combination of servers' idling time and customers' waiting time. Mondschein and Weintraub (2003) notice that other objective function forms used in the literature (including those with servers' overtime cost) are equivalent to the one above.

Despite the fruitful results available in academia, the implementation or guidance of appointment scheduling in practice is still very limited. Many service firms are still using simple rules of thumb. The authors have discussed with practitioners in different service industries and found out, among others, four main concerns that obstruct the application of results from academic literature to industry.

First, as the optimal appointment schedules are found through minimizing the sum of servers' idling cost and customers' waiting cost, it is obviously true that the resulting schedules depend critically on the relative costs of servers' idling and customers' waiting. Therefore, obtaining accurate cost parameters becomes a crucial issue in the application of theoretical results. However, from personal communication with practitioners, there is a lack of methods or guidelines for estimating customers' waiting cost. Fries and Marathe (1981) relate the difficulty in estimating waiting cost to the connection between customers' waiting and the issues of goodwill as well as the cost of society. We also notice that long waiting times would lead to reneging and negative word of month, which further complicates the estimation of the cost.

Second, most of the literature models customers' waiting cost as a linear function of waiting time. However, in reality, the magnitude of customers' annoyance from waiting may not be proportional to the length of waiting time. From recent empirical studies (see, e.g., Baron et al. 2016), in various service encounters, customers' perception of waiting reveals a threshold type behavior in time. That is, customers are generally satisfied with their waiting experience if they wait no more than a certain time length (e.g. 20 minutes), and their patience declines rapidly when their waiting time exceeds that threshold. In many service industries, this acceptance threshold can be obtained from customer satisfaction or complaint surveys (see, e.g., Baron et al. 2016). The firms usually consider the acceptance thresholds as their performance targets.

Third, and very importantly, a schedule that minimizes the total system cost may not lead to equal waiting experiences for each individual customer. Hassin and Mendel (2008) provide numerical results showing that for both the dome-shape system (appointment intervals initially increase and then decrease) and the equal-space system (appointment intervals stay fixed), customers who are scheduled to come later wait longer than those who are scheduled to come earlier. Cayirli and Veral (2003) highlight that the increasing waiting trend is observed under most commonly studied appointment systems. The inequity in waiting time among customers certainly leads to fairness issues, which would clearly create problems in practice.

Fourth, besides the usual concept of waiting time which describes the duration from the time when a customer arrives and joins the queue to the time when she starts her service, there is another important measure which captures the duration from the time a customer requests service to the time when she arrives (i.e., her appointment time). This can be viewed as the indirect waiting time. Indirect waiting time is often ignored in the literature. This is because the indirect waiting cost is considered to be much lower than the direct waiting cost; for example, customers are less inconvenienced waiting at home before arriving to service systems. However, longer indirect waiting time could in fact lead to higher probability of no-show. From personal communication with practitioners, no-show is more frequently seen at the end of a work day than at the beginning of a work day. That is, customers who are given later appointment times and therefore with longer indirect waiting time are less likely to show up. When there are alternative service providers available, indirect waiting time is quite often a major selection criterion for customers.

Unlike a traditional appointment system that minimizes the sum of servers' idling cost and customers' waiting cost, in this paper, we study an appointment system under a specific service level constraint, that is, the expected waiting time of each individual customer must be less than a certain value. Customers are then given the earliest possible appointment times without breaking the service level constraint. Our model resolves the above four concerns simultaneously. (1) Our model only deals with time, and cost is never involved. Thus, our results can be applied in practice without any cost estimation. This resolves the first concern. (2) A unique feature of our model is the service level constraint which gives the upper limit of the expected waiting time of each individual customer. As a result, our appointment schedule ensures fairness among customers. None of the customers wait longer than the acceptance threshold in expectation. This resolves the second and third concerns. (3) Since each individual customer is scheduled to arrive as early as possible, her indirect waiting time is minimized. This resolves the fourth concern. In addition to these, our model has many other advantages. (4) When customers are given the earliest possible appointment times, the servers' idling time and overtime are automatically minimized (without breaking the service level constraint). (5) Our model can be viewed as both prospective scheduling (while the appointment times of all the customers are decided together at once) and sequential scheduling (while the appointment time of each customer is set one after another at the time when service is requested). The interpretation of our problem in the prospective scheduling setting is to find the earliest possible appointment times for all the customers such that the service level constraint is fulfilled, while the interpretation in the sequential scheduling setting is, given that all the previous inter-appointment times are minimized while keeping the service level constraint valid, we need to find the shortest inter-appointment time for the next customer such that the service level constraint is still valid. It is easy to see but worth mentioning that, under our model, for two systems, one with m customers and the other with n customers ($m < n$), with the same service level constraint, the optimal appointment times of the first m customers in the two systems coincide.

Our modeling immediately raises several interesting and important questions: (1) What is the structure of the optimal inter-appointment times? Are they constant, increasing, or decreasing with the order of arrivals? (2) Should the length of the optimal inter-appointment times be equal to the length of the expected service time of a customer? (3) If the optimal inter-appointment times are not constant, are there any simple upper and lower bounds? (4)

How are the optimal inter-appointment times affected by the service level constraint?

1.2 Analysis Method

In the past few years, there has been a growing body of literature on appointment scheduling from the optimization community (see, e.g., Kong et al. 2013 and the references therein). The studies there mainly focus on applying optimization techniques (e.g. robust optimization) to develop computationally tractable programming models (or approximations) for searching the optimal appointment schedules. On the other hand, appointment scheduling has received relatively less recent attention in the queueing community. This is, in part, due to the nature of appointment systems that (1) there is only a finite number of arrivals; and (2) the inter-arrival times between customers may not be equal. These features create difficulties in applying standard queueing methodology which relies on steady state analysis (and therefore assumes infinite arrivals) and requires homogeneous inter-arrival times. As a matter of fact, very few analytical results exist on the structural properties of optimal schedules.

In this paper, we take the queueing approach to explore the structure of optimal appointment schedules. We study a system with a single server and a finite number of customers to schedule. Customer service times are i.i.d. and follow an Erlang distribution. Note that compared with the exponential service time distribution which is assumed in most literature, the Erlang service time distribution, while still holding the Markov property in someway that helps mathematical tractability, is a tremendously relaxed assumption. It largely increases the applicability of the results and managerial insights obtained from our study. Wang et al. (2014) use an embedded Markov chain approach to study a queueing system with finite arrivals where customer inter-arrival times are stochastic and heterogeneous. They characterize performance measures such as the average expected waiting time and examine the effect of heterogeneity in inter-arrival and service times. We follow their embedded Markov chain approach to obtain the waiting time distribution for each individual customer. We then apply the theory of majorization to analyze structural properties of the optimal schedule (which, to the best of our knowledge, is the first of its kind in the literature). We prove that, to keep the expected waiting time of each individual customer less than a certain threshold, the minimum inter-appointment times required increase with the order of arrivals. That is, the inter-appointment time between the m^{th} and $(m+1)^{th}$ arrivals is no less than the inter-appointment time between the $(m-1)^{th}$ and m^{th} customers. We also identify several additional properties of the optimal schedule. For example, other than for the first few arrivals, the expected service time of a customer is a lower bound of the optimal inter-appointment times; and later arrivals have higher chances to see an empty system. For the case where service time is exponentially distributed, we prove the convergence of our system to the D/M/1 queueing system as the number of arrivals approaches infinity. We also discuss the extension of our results to systems with multiple servers.

To help understanding and for notational convenience, we start the analysis with the case where service time is exponentially distributed. We then prove, later in the text, that the main results also hold for Erlang service time distributions. Throughout the paper, "increase/decrease" means "nondecrease/nonincrease". The rest of the paper is organized as follows. In Section 2, we describe the model with exponential service time distributions and analyze the structure of the optimal appointment schedule. In Section 3, we extend the analysis to the case with Erlang service time distributions and discuss the robustness of our results in systems with multiple servers. In Section 4, we provide concluding comments.

2 Problem Description and Analysis

We consider a service system with a single server and M customers to be scheduled to come over time. We index customers by the order of their appointments, so that customer m, for $m = 1, ..., M$, is the m^{th} customer to arrive. All customers show up punctually. We denote by A_m, for $m = 1, ..., M$, the appointment time of customer m. The server begins service (i.e., the system starts) at time 0, and we have $0 \leq A_{m-1} \leq A_m$. Customer service times are i.i.d. and follow an exponential distribution with a finite mean $\frac{1}{\mu}$. Upon arrival, a customer starts her service immediately if the server is available. If not, the customer joins the queue and waits. Customers waiting in queue are served on a first-come, first-served basis (i.e., the same order of their appointment times). There is a service level constraint on the waiting time of each individual customer; namely, the expected waiting time of each customer must be less than or equal to a certain value s. The system provides each customer the earliest possible appointment time fulfilling the waiting time constraint. In Table 1, we summarize the main notations used in the analysis.

2.1 Preliminary Results

Let A^*_m denote the optimal appointment time of customer m. It is easy to see that the optimal appointment time of the first customer is $A^*_1 = 0$. We denote by T_m, for $m = 2, ..., M$, the inter-appointment time between customers $m-1$ and m. That is, $T_m = A_m - A_{m-1}$. Let $\mathbf{T}^* = (T^*_2, ..., T^*_M)$ denote the optimal schedule (shortest inter-appointment times satisfying the

A_m	Appointment time of customer m
A_m^*	Optimal appointment time of customer m
T_m	Inter-appointment time between customers $m-1$ and m
T_m^*	Optimal inter-appointment time between customers $m-1$ and m
$W_m(x)$	Expected waiting time in queue of customer m, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, ..., m-1$
W_m^*	$W_m(T_m^*)$
$R_m(x)$	Number of customers found in system by customer m, upon her arrival, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, ..., m-1$
R_m^*	$R_m(T_m^*)$
$p_{m,i}(x)$	$\Pr\{R_m(x) = i\}$
$p_{m,i}^*$	$\Pr\{R_m^* = i\}$
$P_{m,n}(x)$	$\sum\limits_{i=n}^{m-1} p_{m,i}(x)$
$P_{m,n}^*$	$\sum\limits_{i=n}^{m-1} p_{m,i}^*$
$c_i(x)$	$\frac{(\mu x)^i}{i!} e^{-\mu x}$
$c_{m,i}^*$	$\frac{(\mu T_m^*)^i}{i!} e^{-\mu T_m^*}$

Table 1: Notations

service level constraint). That is, for any schedule $\mathbf{T} = (T_2, ..., T_M)$, if $\sum\limits_{i=2}^{m} T_i < \sum\limits_{i=2}^{m} T_i^*$ (i.e., $A_m < A_m^*$) for some $m = 2, ..., M$, then the expected waiting time of some customer n, for $n = 2, ...m$, must be greater than s (the service level constraint is broken). We now analyze the properties of $(T_2^*, ..., T_M^*)$.

Define $W_m(x)$ to be the expected waiting time in queue of customer m, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, ..., m-1$. We also define $W_m^* = W_m(T_m^*)$. Since $A_1^* = 0$, we have $W_1^* = 0$. From the service level constraint, $W_m^* \leq s$ for $m = 2, ..., M$.

Lemma 1. *$W_m(x)$ is decreasing in x, and T_m^* is decreasing in s.*

First, it is trivial to show that, with fixed schedule of previous customers, for the next customer, the later she comes, the less she waits.

Next, if customers $m-1$ and m are scheduled to arrive together, then the expected waiting time of customer m equals the expected waiting time of customer $m-1$ plus the expected service time of customer $m-1$. Therefore, if customers $1, 2, ..., m$ are scheduled to arrive together at time 0, then the expected waiting time of customer m equals $\frac{m-1}{\mu}$. Let $\lfloor x \rfloor$ denote the largest integer not greater than x.

Lemma 2. *$T_m^* = 0$ for $m = 2, ..., \lfloor \mu s \rfloor + 1$, and $W_m^* = s$ for $m = \lfloor \mu s \rfloor + 2, ..., M$.*

Proof: For each customer m, we are searching for the smallest x such that $W_m(x) \leq s$. Since $\frac{(\lfloor \mu s \rfloor + 1) - 1}{\mu} \leq s$ and $\frac{\lfloor \mu s \rfloor + 1}{\mu} > s$, it is optimal to schedule customers $1, 2, ..., \lfloor \mu s \rfloor + 1$ together at time 0. From customer $\lfloor \mu s \rfloor + 2$, since $W_m(x)$ is decreasing in x (Lemma 1), T_m^* is such that $W_m^* = s$. □

Lemma 2 says that it is optimal to schedule the first $\lfloor \mu s \rfloor + 1$ customers together at time 0. The rest would have expected waiting time equal to s. Since the service time of each customer is exponentially distributed with mean $\frac{1}{\mu}$, the expected waiting time of a customer depends on the number of customers found in system (in queue or in service) upon her arrival. We denote $R_m(x)$ as the random variable describing the number of customers found in system by customer m, upon her arrival, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, ..., m-1$, and let $E[R_m(x)]$ be its expected value. The total number of customers in system immediately after A_m is $R_m(x) + 1$. Now, for $i = 0, ..., m-1$ and $m = 1, ..., M$, let $p_{m,i}(x) = \Pr\{R_m(x) = i\}$ refer to the probability that the m^{th} customer finds, upon arrival, i customers in system, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, ..., m-1$. We also define $R_m^* = R_m(T_m^*)$ and $p_{m,i}^* = \Pr\{R_m^* = i\}$.

When a customer finds i customers in system upon arrival, her expected waiting time is equal to $\frac{i}{\mu}$. Therefore, we have $W_m(x) = \sum_{i=1}^{m-1} p_{m,i}(x)\frac{i}{\mu} = E[R_m(x)]\frac{1}{\mu}$ and $W_m^* = \sum_{i=1}^{m-1} p_{m,i}^* \frac{i}{\mu} = E(R_m^*)\frac{1}{\mu}$. The service level constraint on waiting time ($W_m^* \leq s$) can then be interpreted as $E(R_m^*) \leq \mu s$. That is, the expected number of customers found in system upon each arrival is not greater than μs. From Lemma 2, except for the first $\lfloor \mu s \rfloor + 1$ customers who are scheduled to arrive together at time 0, the expected number of customers found in system upon each arrival equals μs. Suppose now customers $1, ..., m-1$ are scheduled optimally and R_{m-1}^* is equal to μs. The goal is to find the smallest inter-appointment time T_m^* such that $E[R_m(T_m^*)]$ is also equal to μs. Notice that the earliest available appointment time of customer m is the appointment time of customer $m-1$, that is $T_m = 0$. If customer m arrives together with customer $m-1$, then $E[R_m(0)] = E(R_{m-1}^*) + 1 = \mu s + 1 > \mu s$. The constraint is broken. Therefore, we need customer m to arrive later, not together with customer $m-1$.

Lemma 3. *The expected number of service completions during the time interval (A_{m-1}^*, A_m^*) equals 1 for $m = \lfloor \mu s \rfloor + 3, ..., M$.*

From Lemma 3, the optimal inter-appointment time T_m is such that, the expected number of customers who complete the service and leave the system during T_m exactly equals 1. This is because, the expected number of customers in system immediately after A_{m-1}^* is $\mu s + 1$. After the system completes 1 customer, the number of customers in system will return to μs again. So, searching the optimal inter-appointment time is equivalent to asking how long it takes the system to complete one service. A specious guess to this question could be $\frac{1}{\mu}$ (i.e., the expected service time of a customer).

Lemma 4. *$T_m^* \geq \frac{1}{\mu}$ for $m = \lfloor \mu s \rfloor + 3, ..., M$.*

Lemma 4 states that except for the first $\lfloor \mu s \rfloor + 2$ customers, the optimal inter-appointment times have a lower bound $\frac{1}{\mu}$. The reason is that, for a system with exponential service rate μ, if the server is always busy during a time interval with length $\frac{1}{\mu}$, then the expected number of service completions is equal to 1. However, if the server is not always busy, the expected number of service completions is less than 1.

To further analyze the properties of $(T_2^*, ..., T_M^*)$, we need to find the relationship between T_m^* and T_{m-1}^*. Conditioning on the number of customers found, upon arrival, by customer $m-1$, we obtain

$$p_{m,i}(x) = \sum_{j=i-1}^{m-2} p_{m-1,j}^* \Pr\{R_m(x) = i \mid R_{m-1}^* = j\} \tag{1}$$

for $1 \leq i \leq m-1$, and

$$p_{m,0}(x) = 1 - \sum_{i=1}^{m-1} p_{m,i}(x)$$

for $2 \leq m \leq M$. Similarly,

$$p_{m,i}^* = \sum_{j=i-1}^{m-2} p_{m-1,j}^* \Pr\{R_m^* = i \mid R_{m-1}^* = j\}, \tag{2}$$

and

$$p_{m,0}^* = 1 - \sum_{i=1}^{m-1} p_{m,i}^*.$$

For the m^{th} customer to find i customers given that the $(m-1)^{th}$ customer finds j, there must be exactly $j-i+1$ service completions during the time interval (A_{m-1}^*, A_m) with length x. Since service time is exponentially distributed with rate μ, the number of service completions during a time interval with length x is Poisson distributed with rate μx. Define

$$c_i(x) = \frac{(\mu x)^i}{i!} e^{-\mu x}$$

and

$$c_{m,i}^* = \frac{(\mu T_m^*)^i}{i!} e^{-\mu T_m^*}.$$

Then,

$$\Pr\{R_m(x) = i \mid R_{m-1}^* = j\} = c_{j-i+1}(x)$$

and

$$\Pr\{R_m^* = i \mid R_{m-1}^* = j\} = c_{m,j-i+1}^*.$$

Equations (1) and (2) become

$$p_{m,i}(x) = \sum_{j=i-1}^{m-2} p^*_{m-1,j} c_{j-i+1}(x) \tag{3}$$

and

$$p^*_{m,i} = \sum_{j=i-1}^{m-2} p^*_{m-1,j} c^*_{m,j-i+1}.$$

Now, define

$$P_{m,n}(x) = \sum_{i=n}^{m-1} p_{m,i}(x)$$

and

$$P^*_{m,n} = \sum_{i=n}^{m-1} p^*_{m,i}$$

for $n = 0, ..., m-1$. Then, we have

$$W_m(x) = \frac{1}{\mu} \sum_{i=1}^{m-1} P_{m,i}(x)$$

and

$$W^*_m = \frac{1}{\mu} \sum_{i=1}^{m-1} P^*_{m,i}.$$

From Equation (3), we have $P_{m,n}(x) = \sum\limits_{i=n}^{m-1} \sum\limits_{j=i-1}^{m-2} p^*_{m-1,j} c_{j-i+1}(x) = \sum\limits_{i=n-1}^{m-2} c_{i-n+1}(x) \sum\limits_{j=i}^{m-2} p^*_{m-1,j}$. That is,

$$P_{m,n}(x) = \sum_{i=n-1}^{m-2} P^*_{m-1,i} c_{i-n+1}(x). \tag{4}$$

Similarly,

$$P^*_{m,n} = \sum_{i=n-1}^{m-2} P^*_{m-1,i} c^*_{m,i-n+1}. \tag{5}$$

Next, we apply the theory of majorization to show the relationship between T^*_m and T^*_{m-1}. We firstly introduce the concepts of majorization.

2.2 Majorization

For an n-dimensional vector $\mathbf{x} = (x_1, ..., x_n)$, we denote by $(x_{(1)}, ..., x_{(n)})$ the vector with the same components but sorted in increasing order (i.e., $x_{(1)} \leq ... \leq x_{(n)}$), and by $(x_{[1]}, ..., x_{[n]})$ the vector with the same components but sorted in decreasing order (i.e., $x_{[1]} \geq ... \geq x_{[n]}$).

Definition 1. *For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\mathbf{x}$ is majorized by $\mathbf{y}$ ($\mathbf{y}$ majorizes $\mathbf{x}$), denoted by $\mathbf{x} \prec \mathbf{y}$, if*

$$\begin{cases} \sum\limits_{i=1}^{j} x_{[i]} \leq \sum\limits_{i=1}^{j} y_{[i]} & \text{for } j = 1, ..., n-1, \\ \sum\limits_{i=1}^{n} x_{[i]} = \sum\limits_{i=1}^{n} y_{[i]}, \end{cases}$$

or, equivalently,

$$\begin{cases} \sum\limits_{i=1}^{j} x_{(i)} \geq \sum\limits_{i=1}^{j} y_{(i)} & \textit{for } j = 1, ..., n-1, \\ \sum\limits_{i=1}^{n} x_{(i)} = \sum\limits_{i=1}^{n} y_{(i)}. \end{cases}$$

Definition 2. *A square matrix $D = (d_{ij})$ is a doubly stochastic matrix if $d_{ij} \geq 0$ and $\sum\limits_{i} d_{ij} = \sum\limits_{j} d_{ij} = 1$, $\forall i, j$.*

Lemma 5. *For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the following conditions are equivalent*

(1) $\mathbf{x} \prec \mathbf{y}$,

(2) $\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i$ *and* $\sum\limits_{i=1}^{n} (x_i - z)^+ \leq \sum\limits_{i=1}^{n} (y_i - z)^+$, $\forall z \in \mathbb{R}$,

(3) $\sum\limits_{i=1}^{n} |x_i - z| \leq \sum\limits_{i=1}^{n} |y_i - z|$, $\forall z \in \mathbb{R}$,

(4) $\mathbf{x} = D\mathbf{y}$ *for some doubly stochastic matrix D.*

The proof of Lemma 5 and more properties of majorization can be found in Marshall et al. (2011).

Proposition 1. $(P^*_{m,m-1}, P^*_{m,m-2}, ..., P^*_{m,0}) \prec (0, P^*_{m-1,m-2}, ..., P^*_{m-1,0})$ *for $m = \lfloor \mu s \rfloor + 3, ..., M$.*

Proof: Let $\mathbf{u} = (P^*_{m,m-1}, P^*_{m,m-2}, ..., P^*_{m,0})^T$ and $\mathbf{v} = (0, P^*_{m-1,m-2}, ..., P^*_{m-1,0})^T$, and define

$$D = \begin{pmatrix} 1 - c^*_{m,0} & c^*_{m,0} & 0 & ... & 0 \\ 1 - \sum\limits_{i=0}^{1} c^*_{m,i} & c^*_{m,1} & c^*_{m,0} & ... & 0 \\ 1 - \sum\limits_{i=0}^{2} c^*_{m,i} & c^*_{m,2} & c^*_{m,1} & ... & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ d & 1 - \sum\limits_{i=0}^{m-2} c^*_{m,i} & 1 - \sum\limits_{i=0}^{m-3} c^*_{m,i} & ... & 1 - c^*_{m,0} \end{pmatrix}$$

where d satisfies $d + (1 - \sum\limits_{i=0}^{m-2} c^*_{m,i}) + (1 - \sum\limits_{i=0}^{m-3} c^*_{m,i}) + ... + (1 - c^*_{m,0}) = 1$.

First, note that $P^*_{m,0} = P^*_{m-1,0} = 1$, $\sum\limits_{i=1}^{m-1} P^*_{m,i} = \sum\limits_{i=1}^{m-2} P^*_{m-1,i} = \mu s$, and

$$\begin{aligned} &(1 - \sum_{i=0}^{m-2} c^*_{m,i}) P^*_{m-1,m-2} + (1 - \sum_{i=0}^{m-3} c^*_{m,i}) P^*_{m-1,m-3} + ... + (1 - c^*_{m,0}) P^*_{m-1,0} \\ =& \sum_{i=0}^{m-2} P^*_{m-1,i} - \sum_{i=0}^{m-2} P^*_{m-1,i} (\sum_{j=0}^{i} c^*_{m,i-j}) \\ =& 1 + \mu s - \mu s = 1. \end{aligned}$$

From Equation (5), we have $\mathbf{u} = D\mathbf{v}$.

Now, notice that $D = (d_{ij})$ satisfies $\sum\limits_{i} d_{ij} = \sum\limits_{j} d_{ij} = 1$. If $d \geq 0$, then D is a doubly stochastic matrix. By Lemma 5, we have $\mathbf{u} \prec \mathbf{v}$.

If $d < 0$, then D is not a doubly stochastic matrix. Since $\sum_{i=1}^{m} u_i = \sum_{i=1}^{m} v_i = 1 + \mu s$, we prove $\mathbf{u} \prec \mathbf{v}$ by showing that $\sum_{i=1}^{m}(u_i - z)^+ \leq \sum_{i=1}^{m}(v_i - z)^+$, $\forall z \in \mathbb{R}$ (Lemma 5).

First, for any $z \leq 0$, we have $\sum_{i=1}^{m}(u_i - z)^+ = \sum_{i=1}^{m}(u_i - z) = \sum_{i=1}^{m} u_i + mz = \sum_{i=1}^{m} v_i + mz = \sum_{i=1}^{m}(u_i - z) = \sum_{i=1}^{m}(u_i - z)^+$. For any $z > 0$, since $u_m = v_m = 1$, we only need to show that $\sum_{i=1}^{m-1}(u_i - z)^+ \leq \sum_{i=1}^{m-1}(v_i - z)^+$.

Now, for matrix D, since $\sum_j d_{ij} = 1$, we have $D\mathbf{e}_m = \mathbf{e}_m$, where $\mathbf{e}_m = (1, ..., 1)^T$ is the m-dimensional identity column vector. Define D_{m-1} as the matrix that consists of the first $m-1$ rows of D. Since $(u_i - z)^+ = (\mathbf{u} - z\mathbf{e}_m)_i^+ = (D\mathbf{v} - zD\mathbf{e}_m)_i^+ \leq [D(\mathbf{v} - z\mathbf{e}_m)^+]_i$, we have $\sum_{i=1}^{m-1}(u_i - z)^+ \leq \sum_{i=1}^{m-1}[D(\mathbf{v} - z\mathbf{e}_m)^+]_i = \mathbf{e}_{m-1}^T D_{m-1}(\mathbf{v} - z\mathbf{e}_m)^+ = \sum_{j=1}^{m-1}(\mathbf{e}_{m-1}^T D_{m-1})_j (v_j - z)^+ = \sum_{j=2}^{m-1}(\mathbf{e}_{m-1}^T D_{m-1})_j (v_j - z)^+$, where the last equality is due to the fact that $v_1 = 0$ and therefore $(v_1 - z)^+ = 0$, $\forall z > 0$. Notice that $(\mathbf{e}_{m-1}^T D_{m-1})_j = \sum_{i=1}^{m-1} d_{ij} \leq 1$ for $j = 2, ..., m-1$, we have $\sum_{i=1}^{m-1}(u_i - z)^+ \leq \sum_{j=2}^{m-1}(\mathbf{e}_{m-1}^T D_{m-1})_j (v_j - z)^+ \leq \sum_{j=2}^{m-1}(v_j - z)^+ = \sum_{j=1}^{m-1}(v_j - z)^+$. This completes the proof. □

2.3 Structure of Optimal Inter-Appointment Times

Now, we are ready to show the main result regarding the structure of the optimal inter-appointment times.

Theorem 1. $T_{m+1}^* \geq T_m^*$.

Proof: From Lemma 2, $T_m^* = 0$ for $m = 2, ..., \lfloor \mu s \rfloor + 1$; and from Lemma 1, $W_m(x)$ is decreasing in x. Therefore, to prove $T_{m+1}^* \geq T_m^*$, we only need to show that $W_{m+1}(T_m^*) \geq W_m(T_m^*) = W_m^* = s$ for $m = \lfloor \mu s \rfloor + 2, ..., M$.

From Equation (4), we have

$$\begin{aligned}
W_{m+1}(T_m^*) &= \frac{1}{\mu}\sum_{i=1}^{m} P_{m+1,i}(T_m^*) \\
&= \frac{1}{\mu}\sum_{i=1}^{m}\sum_{j=i-1}^{m-1} P_{m,j}^* c_{j-i+1}(T_m^*) \\
&= \frac{1}{\mu}\sum_{i=1}^{m}\sum_{j=i-1}^{m-1} P_{m,j}^* c_{m,j-i+1}^* \\
&= \frac{1}{\mu}\sum_{i=0}^{m-1} c_{m,i}^* \sum_{j=i}^{m-1} P_{m,j}^*
\end{aligned}$$

$$= \frac{1}{\mu}(\sum_{i=0}^{m-2} c^*_{m,i} \sum_{j=i}^{m-1} P^*_{m,j} + c^*_{m,m-1} P^*_{m,m-1}),$$

and $W^*_m = \frac{1}{\mu} \sum\limits_{i=0}^{m-2} c^*_{m,i} \sum\limits_{j=i}^{m-2} P^*_{m-1,j}$. Comparing $W_{m+1}(T^*_m)$ with W^*_m, we see that it is sufficient to prove $\sum\limits_{j=i}^{m-1} P^*_{m,j} \geq \sum\limits_{j=i}^{m-2} P^*_{m-1,j}$ for $i = 0, ..., m-2$ and $m = \lfloor \mu s \rfloor + 2, ..., M$.

For $m = \lfloor \mu s \rfloor + 3, ..., M$, from Proposition 1, $(P^*_{m,m-1}, P^*_{m,m-2}, ..., P^*_{m,0}) \prec (0, P^*_{m-1,m-2}, ..., P^*_{m-1,0})$. Therefore, by Definition 1, we have $\sum\limits_{j=i}^{m-1} P^*_{m,j} \geq \sum\limits_{j=i}^{m-2} P^*_{m-1,j}$ for $i = 0, ..., m-2$.

Last, for $m = \lfloor \mu s \rfloor + 2$, we have $m - 1 = \lfloor \mu s \rfloor + 1$ and $A^*_{m-1} = 0$. Then, $p^*_{m-1,m-2} = 1$, and $p^*_{m-1,i} = 0$ for $i = 0, ..., m-3$. Therefore, $P^*_{m-1,n} = 1$ for $n = 0, ..., m-2$, and $\sum\limits_{j=i}^{m-2} P^*_{m-1,j} = m - 1 - i$. We now prove $\sum\limits_{j=i}^{m-1} P^*_{m,j} \geq m - 1 - i$ for $i = 0, ..., m-2$ by induction. First, for $i = 0$, $\sum\limits_{i=0}^{m-1} P^*_{m,i} = \sum\limits_{i=1}^{m-1} P^*_{m,i} + P^*_{m,0} = \mu s + 1 \geq \lfloor \mu s \rfloor + 1 = m - 1$. Now, suppose $\sum\limits_{j=i}^{m-1} P^*_{m,j} \geq m - 1 - i$ for $i = 0, ..., n$, where $n \leq m - 1$. Then, for $i = n + 1$, $\sum\limits_{j=n+1}^{m-1} P^*_{m,j} = \sum\limits_{j=n}^{m-1} P^*_{m,j} - P^*_{m,n} \geq \sum\limits_{j=n}^{m-1} P^*_{m,j} - 1 \geq m - 1 - n - 1 = m - 1 - (n + 1)$. This completes the proof. □

From Lemma 2 and Theorem 1, we see that, the optimal appointment schedule has the structure that (1) the first $\lfloor \mu s \rfloor + 1$ customers are scheduled to come together at time 0; and (2) from customer $\lfloor \mu s \rfloor + 2$, the inter-appointment time increases. The optimal schedule has some other interesting properties. Denote w^*_m as the random variable describing the waiting time in queue of customer m under the optimal schedule ($W^*_m = E(w^*_m)$).

Corollary 1. *For $m = \lfloor \mu s \rfloor + 2, ..., M$, the following hold*

(a) $p^*_{m+1,0} \geq p^*_{m,0}$,

(b) $Pr\{w^*_{m+1} \leq \frac{1}{\mu}\} \geq Pr\{w^*_m \leq \frac{1}{\mu}\}$.

Proof: (a) Since $(P^*_{m+1,m}, P^*_{m+1,m-1}, ..., P^*_{m+1,0}) \prec (0, P^*_{m,m-1}, ..., P^*_{m,0})$, by Definition 1, we have $P^*_{m+1,0} + P^*_{m+1,1} \leq P^*_{m,0} + P^*_{m,1}$. That is, $1 + 1 - p^*_{m+1,0} \leq 1 + 1 - p^*_{m,0}$, or $p^*_{m+1,0} \geq p^*_{m,0}$.

(b) If the m^{th} customer finds i customers ($i \geq 1$) in system upon arrival, her waiting time in queue is Erlang distributed with shape i and rate μ, and therefore, $\Pr\{w^*_m \leq \frac{1}{\mu} | R^*_m = i\} = 1 - \sum\limits_{j=0}^{i-1} \frac{1}{j!} e^{-1} = 1 - \sum\limits_{j=0}^{i-1} C_j$, where $C_j = \frac{1}{j!} e^{-1}$. Thus,

$$\begin{aligned}
\Pr\{w^*_m \leq \frac{1}{\mu}\} =& p^*_{m,0} + \sum_{i=1}^{m-1} p^*_{m,i} (1 - \sum_{j=0}^{i-1} C_j) \\
=& 1 - \sum_{i=1}^{m-1} p^*_{m,i} \sum_{j=0}^{i-1} C_j \\
=& 1 - (C_0 P^*_{m,1} + C_1 P^*_{m,2} + ... + C_{m-2} P^*_{m,m-1})
\end{aligned}$$

$$
\begin{aligned}
&=1-(C_0\sum_{j=1}^{1}P^*_{m,j}-C_1\sum_{j=1}^{1}P^*_{m,j}+C_1\sum_{j=1}^{2}P^*_{m,j}-C_2\sum_{j=1}^{2}P^*_{m,j}\\
&\quad+...+C_{m-3}\sum_{j=1}^{m-2}P^*_{m,j}-C_{m-2}\sum_{j=1}^{m-2}P^*_{m,j}+C_{m-2}\sum_{j=1}^{m-1}P^*_{m,j})\\
&=1-[(C_0-C_1)\sum_{j=1}^{1}P^*_{m,j}+(C_1-C_2)\sum_{j=1}^{2}P^*_{m,j}+...+(C_{m-3}-C_{m-2})\sum_{j=1}^{m-2}P^*_{m,j}\\
&\quad+C_{m-2}\sum_{j=1}^{m-1}P^*_{m,j}]\\
&=1-[\sum_{i=1}^{m-2}(C_{i-1}-C_i)\sum_{j=1}^{i}P^*_{m,j}+C_{m-2}\sum_{j=1}^{m-1}P^*_{m,j}].
\end{aligned}
$$

It is easy to see that C_j is decreasing in j. That is, $C_i \geq C_{i+1}$. Now, $(P^*_{m+1,m}, P^*_{m+1,m-1}, ..., P^*_{m+1,0}) \prec (0, P^*_{m,m-1}, ..., P^*_{m,0})$, so $\sum_{j=0}^{i} P^*_{m+1,j} \leq \sum_{j=0}^{i} P^*_{m,j}$ for $i = 0, ..., m-1$, and $\sum_{j=0}^{m} P^*_{m+1,j} \leq \sum_{j=0}^{m-1} P^*_{m,j}$. Since $P^*_{m+1,0} = P^*_{m,0} = 1$, we have $\sum_{j=1}^{i} P^*_{m+1,j} \leq \sum_{j=1}^{i} P^*_{m,j}$ for $i = 1, ..., m-1$, and $\sum_{j=1}^{m} P^*_{m+1,j} \leq \sum_{j=1}^{m-1} P^*_{m,j}$. This implies

$$
\begin{aligned}
\Pr\{w^*_{m+1}\leq\frac{1}{\mu}\}&=1-[\sum_{i=1}^{m-1}(C_{i-1}-C_i)\sum_{j=1}^{i}P^*_{m+1,j}+C_{m-1}\sum_{j=1}^{m}P^*_{m+1,j}]\\
&\geq 1-[\sum_{i=1}^{m-1}(C_{i-1}-C_i)\sum_{j=1}^{i}P^*_{m,j}+C_{m-1}\sum_{j=1}^{m-1}P^*_{m,j}]\\
&=1-[\sum_{i=1}^{m-2}(C_{i-1}-C_i)\sum_{j=1}^{i}P^*_{m,j}+(C_{m-2}-C_{m-1})\sum_{j=1}^{m-1}P^*_{m,j}+C_{m-1}\sum_{j=1}^{m-1}P^*_{m,j}]\\
&=1-[\sum_{i=1}^{m-2}(C_{i-1}-C_i)\sum_{j=1}^{i}P^*_{m,j}+C_{m-2}\sum_{j=1}^{m-1}P^*_{m,j}]\\
&=\Pr\{w^*_m\leq\frac{1}{\mu}\}.\quad\square
\end{aligned}
$$

Corollary 1 states that, upon arrival, while seeing the equal expected number of customers, a later arrival has a higher chance to find an empty system and is more likely to wait shorter than the duration of her expected service time. These results are intuitively true noticing that a later arrival also has a higher chance to see a longer queue (e.g. the 10^{th} arrival could see 9 customers in system while the 5^{th} arrival could see 4 at most).

It is worth highlighting here the fact that, $p^*_{m,0}$ increases with m can be viewed as the reason why T^*_m increases with m. As we explained earlier after Lemma 4, during a time interval with fixed length, the expected number of service completions depends on the proportion of time while the server is busy (working). As m increases, $p^*_{m,0}$ increases, that is, the proportion of server-busy time decreases, and therefore it takes longer to complete 1 service.

2.4 Asymptotic Analysis

In this section, we study the limiting behavior of our system. We prove that our system converges to the D/M/1 queueing system as the number of arrivals approaches infinity.

First, since $T^*_{m+1} \geq T^*_m$, $\lim_{M\to\infty} T^*_M$ exists (can be infinity). Let $T^* = \lim_{M\to\infty} T^*_M$, and $c^*_i = \lim_{M\to\infty} c^*_{M,i} = \lim_{M\to\infty} c_i(T^*_M) = c_i(\lim_{M\to\infty} T^*_M)$. Define $p^*_{m,i} = P^*_{m,i} = 0$ for $i \geq m$. We now prove that $\lim_{M\to\infty} p^*_{M,i}$ exists.

Lemma 6. *$\lim_{M\to\infty} p^*_{M,i}$ exists for $i = 0, 1, ...$*

Proof: We prove this by induction on i. First, notice that $p^*_{M,i} \leq 1$, $\forall i \ \forall M$. For $i = 0$, from Corollary 1, $p^*_{M+1,0} \geq p^*_{M,0}$ for $M \geq \lfloor \mu s \rfloor + 2$. Since $p^*_{M,0} \leq 1$, $\forall M$, we conclude that $\lim_{M\to\infty} p^*_{M,0}$ exists. Now, suppose $\lim_{M\to\infty} p^*_{M,i}$ exists for $i = 0, 1, ..., n$. Then, for $i = n+1$, since $p^*_{M,i} = 0$ for $i \geq M$, we have $P^*_{M,j} = \sum_{i=j}^{M-1} p^*_{M,i} = \sum_{i=j}^{\infty} p^*_{M,i}$ for $j = 0, ..., M-1$; and since $P^*_{M,i} = 0$ for $i \geq M$, we have $P^*_{M,j} = \sum_{i=j}^{\infty} p^*_{M,i}$ for $j = 0, 1, ...$ This implies that $(P^*_{M,j}, P^*_{M,j-1}, ..., P^*_{M,0}) \prec (P^*_{M-1,j}, P^*_{M-1,j-1}, ..., P^*_{M-1,0})$ for $j \geq M$ and $M \geq \lfloor \mu s \rfloor + 3$. As a result, $\sum_{j=0}^{i} P^*_{M,j} \leq \sum_{j=0}^{i} P^*_{M-1,j}$ for all $i \geq 0$ and $M \geq \lfloor \mu s \rfloor + 3$. Let $i = n+2$, we have $\sum_{j=0}^{n+2} P^*_{M,j} \leq \sum_{j=0}^{n+2} P^*_{M-1,j}$, that is,

$$P^*_{M,0} + P^*_{M,1} + ... + P^*_{M,n+2} \leq P^*_{M-1,0} + P^*_{M-1,1} + ... + P^*_{M-1,n+2},$$

or

$$\sum_{i=0}^{M-1} p^*_{M,i} + \sum_{i=1}^{M-1} p^*_{M,i} + ... + \sum_{i=n+2}^{M-1} p^*_{M,i} \leq \sum_{i=0}^{M-2} p^*_{M-1,i} + \sum_{i=1}^{M-2} p^*_{M-1,i} + ... + \sum_{i=n+2}^{M-2} p^*_{M-1,i}.$$

This leads to

$$1 + (1 - p^*_{M,0}) + ... + (1 - \sum_{i=0}^{n+1} p^*_{M,i}) \leq 1 + (1 - p^*_{M-1,0}) + ... + (1 - \sum_{i=0}^{n+1} p^*_{M-1,i}),$$

or $-\sum_{i=0}^{n+1}(n+2-i)p^*_{M,i} \leq -\sum_{i=0}^{n+1}(n+2-i)p^*_{M-1,i}$.

Define $Q^*_{M,n} = \sum_{i=0}^{n+1}(n+2-i)p^*_{M,i}$. We have $Q^*_{M,n} \geq Q^*_{M-1,n}$ (for $M \geq \lfloor \mu s \rfloor + 3$). Since $Q^*_{M,n} \leq \sum_{i=0}^{n+1}(n+2-i) = \sum_{i=1}^{n+2} i = \frac{(n+2)(n+3)}{2}$, we conclude that $\lim_{M\to\infty} Q^*_{M,n}$ exists. Now, notice that $p^*_{M,n+1} = Q^*_{M,n} - \sum_{i=0}^{n}(n+2-i)p^*_{M,i}$, and therefore

$$\lim_{M\to\infty} p^*_{M,n+1} = \lim_{M\to\infty}[Q^*_{M,n} - \sum_{i=0}^{n}(n+2-i)p^*_{M,i}] = \lim_{M\to\infty} Q^*_{M,n} - \sum_{i=0}^{n}(n+2-i) \lim_{M\to\infty} p^*_{M,i}.$$

Since $\lim_{M\to\infty} Q^*_{M,n}$, $\lim_{M\to\infty} p^*_{M,0}$, $\lim_{M\to\infty} p^*_{M,1}$,..., $\lim_{M\to\infty} p^*_{M,n}$ all exist and are all finite, we conclude that $\lim_{M\to\infty} p^*_{M,n+1}$ exists. This completes the induction and the proof. □

Theorem 2. *$T^* = \lim_{M\to\infty} T^*_M = s(1 + \frac{1}{\mu s})\ln(1 + \frac{1}{\mu s})$, and $p^*_i = \lim_{M\to\infty} p^*_{M,i} = \frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^i$ for $i = 0, 1, ...$*

Proof: Recall that $p^*_{M,i} = \sum\limits_{j=i-1}^{M-2} p^*_{M-1,j} c^*_{M,j-i+1}$ for $1 \le i \le M-1$, and $p^*_{M,0} = 1 - \sum\limits_{i=1}^{M-1} p^*_{M,i}$. Since $p^*_{M,i} = 0$ for $i \ge M$, we have $p^*_{M,i} = \sum\limits_{j=i-1}^{\infty} p^*_{M-1,j} c^*_{M,j-i+1}$ for $i = 1, 2, ...,$ and $p^*_{M,0} = 1 - \sum\limits_{i=1}^{\infty} p^*_{M,i}$. Let M goes to infinity, we have

$$p^*_i = \lim_{M\to\infty} \sum_{j=i-1}^{\infty} p^*_{M-1,j} c^*_{M,j-i+1} = \sum_{j=i-1}^{\infty} p^*_j c^*_{j-i+1} = \sum_{j=0}^{\infty} p^*_{j+i-1} c^*_j = \sum_{j=0}^{\infty} p^*_{j+i-1} \frac{(\mu T^*)^j}{j!} e^{-\mu T^*},$$

for $i = 1, 2, ...,$ and $p^*_0 = 1 - \sum\limits_{i=1}^{\infty} p^*_i$.

Now, for $M \ge \lfloor \mu s \rfloor + 2$, we have $\sum\limits_{i=1}^{\infty} i p^*_{M,i} = \sum\limits_{i=1}^{\infty} P^*_{M,i} = \sum\limits_{i=1}^{m-1} P^*_{M,i} = \mu s$. Again, let M goes to infinity, we get $\sum\limits_{i=1}^{\infty} i p^*_i = \mu s$.

Thus, we have

$$\begin{cases} p^*_0 = 1 - \sum\limits_{i=1}^{\infty} p^*_i, \\ \sum\limits_{i=1}^{\infty} i p^*_i = \mu s, \\ p^*_i = \sum\limits_{j=0}^{\infty} p^*_{j+i-1} \frac{(\mu T^*)^j}{j!} e^{-\mu T^*} \text{ for } i = 1, 2, ... \end{cases}$$

We now verify that $T^* = s(1 + \frac{1}{\mu s}) \ln(1 + \frac{1}{\mu s})$ and $p^*_i = \frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^i$ for $i = 0, 1, ...$ is the solution of the above system of equations.

For the first equation, LHS (left hand side) $= p^*_0 = \frac{1}{1+\mu s}$. RHS (right hand side) $= 1 - \sum\limits_{i=1}^{\infty} p^*_i = 1 - \sum\limits_{i=1}^{\infty} \frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^i = 1 - \frac{1}{1+\mu s} \sum\limits_{i=1}^{\infty} (\frac{\mu s}{1+\mu s})^i = 1 - \frac{1}{1+\mu s} \frac{\frac{\mu s}{1+\mu s}}{1 - \frac{\mu s}{1+\mu s}} = 1 - \frac{\mu s}{1+\mu s} = \frac{1}{1+\mu s}$. LHS = RHS.

For the second equation, define $\alpha = \sum\limits_{i=1}^{\infty} i p^*_i$. LHS $= \sum\limits_{i=1}^{\infty} i p^*_i = \sum\limits_{i=1}^{\infty} i \frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^i =$ $\frac{1}{1+\mu s} \sum\limits_{i=1}^{\infty} i(\frac{\mu s}{1+\mu s})^i$. That is,

$$\alpha = \frac{1}{1+\mu s} \sum_{i=1}^{\infty} i(\frac{\mu s}{1+\mu s})^i. \tag{6}$$

Then, $\frac{\mu s}{1+\mu s}\alpha = \frac{\mu s}{1+\mu s}\frac{1}{1+\mu s} \sum\limits_{i=1}^{\infty} i(\frac{\mu s}{1+\mu s})^i$. That is,

$$\frac{\mu s}{1+\mu s}\alpha = \frac{1}{1+\mu s} \sum_{i=1}^{\infty} i(\frac{\mu s}{1+\mu s})^{i+1}. \tag{7}$$

From (6) and (7), we obtain $\frac{1}{1+\mu s}\alpha = \frac{1}{1+\mu s} \sum\limits_{i=1}^{\infty} (\frac{\mu s}{1+\mu s})^i = \frac{\mu s}{1+\mu s}$, and therefore $\alpha = \mu s$. LHS = RHS.

For the third equation, $\frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^i = \sum\limits_{j=0}^{\infty} \frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^{j+i-1} \frac{(\mu T^*)^j}{j!} e^{-\mu T^*}$. So, $\frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^i =$ $\frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^i \sum\limits_{j=0}^{\infty} (\frac{\mu s}{1+\mu s})^{j-1} \frac{(\mu T^*)^j}{j!} e^{-\mu T^*}$, and therefore $\sum\limits_{j=0}^{\infty} (\frac{\mu s}{1+\mu s})^{j-1} \frac{(\mu T^*)^j}{j!} e^{-\mu T^*} = 1$. That is, $\sum\limits_{j=0}^{\infty} (\frac{\mu s}{1+\mu s})^j \frac{(\mu T^*)^j}{j!} = \frac{\mu s}{1+\mu s} e^{\mu T^*}$, or $\sum\limits_{j=0}^{\infty} \frac{(\frac{\mu s \mu T^*}{1+\mu s})^j}{j!} = \frac{\mu s}{1+\mu s} e^{\mu T^*}$. Since $e^x = \sum\limits_{j=0}^{\infty} \frac{x^j}{j!}$, we have $e^{\frac{\mu s \mu T^*}{1+\mu s}} =$

$\frac{\mu s}{1+\mu s}e^{\mu T^*}$. This implies that $\ln(e^{\frac{\mu s \mu T^*}{1+\mu s}}) = \ln(\frac{\mu s}{1+\mu s}e^{\mu T^*})$, or $\frac{\mu s \mu T^*}{1+\mu s} = \mu T^* + \ln(\frac{\mu s}{1+\mu s})$. Thus, we have $\frac{\mu}{1+\mu s}T^* = -\ln(\frac{\mu s}{1+\mu s})$, and so $T^* = -\frac{1+\mu s}{\mu}\ln(\frac{\mu s}{1+\mu s}) = \frac{1+\mu s}{\mu}\ln[(\frac{\mu s}{1+\mu s})^{-1}] = (s + \frac{1}{\mu})\ln(\frac{1+\mu s}{\mu s}) = s(1+\frac{1}{\mu s})\ln(1+\frac{1}{\mu s})$.

Now, we conclude that $T^* = \lim_{M\to\infty} T^*_M = s(1+\frac{1}{\mu s})\ln(1+\frac{1}{\mu s})$, and $p^*_i = \lim_{M\to\infty} p^*_{M,i} = \frac{1}{1+\mu s}(\frac{\mu s}{1+\mu s})^i$ for $i = 0, 1, ...$ □

Theorem 2 shows that as the number of arrivals approaches infinity, our system converges asymptotically to a D/M/1 queueing system having deterministic inter-arrival times with length $s(1+\frac{1}{\mu s})\ln(1+\frac{1}{\mu s})$ and exponential service times with rate μ.

As a result of Theorem 1 together with Theorem 2, we see that T^*_m has an upper bound that is equal to $s(1+\frac{1}{\mu s})\ln(1+\frac{1}{\mu s})$, for $m = 1, ..., M$. Recall Lemma 4 that T^*_m has a lower bound that is equal to $\frac{1}{\mu}$, for $m = \lfloor \mu s \rfloor + 3, ..., M$. Therefore, we have obtained both upper and lower bounds of T^*_m for $m = \lfloor \mu s \rfloor + 3, ..., M$, in explicit forms.

3 Extension

In this section, we study two extensions of our system, one with Erlang service time distributions and the other with multiple servers.

3.1 Erlang Service Time Distribution

When customer service times follow an Erlang distribution with shape I (a finite positive integer) and rate μ (a finite positive real number), each customer has I phases of service, and the duration for each phase of service follows an exponential distribution with rate μ. Now, the expected waiting time of a customer depends on the number of phases found in system upon her arrival, instead of the number of customers. First, it is easy to see that the corresponding Lemma 1-4 still hold for the Erlang case.

Lemma 7. *For systems with Erlang service time distributions, the following hold*
(1) $W_m(x)$ is decreasing in x, and T^*_m is decreasing in s,
(2) $T^*_m = 0$ for $m = 2, ..., \lfloor \frac{\mu s}{I} \rfloor + 1$, and $W^*_m = s$ for $m = \lfloor \frac{\mu s}{I} \rfloor + 2, ..., M$,
(3) The expected number of phase completions during the time interval (A^*_{m-1}, A^*_m) equals I for $m = \lfloor \frac{\mu s}{I} \rfloor + 3, ..., M$,
(4) $T^*_m \geq \frac{I}{\mu}$ for $m = \lfloor \frac{\mu s}{I} \rfloor + 3, ..., M$.

We keep the same notations but switching from the "the number of customers" to "the number of phases". Namely, we denote $R_m(x)$ as the random variable describing the number of phases found in system by customer m, upon her arrival, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, ..., m-1$. The total number of phases in system immediately after A_m is $R_m + I$. Then, for $i = 0, ..., (m-1)I$ and $m = 1, ..., M$, let $p_{m,i}(x) = \Pr\{R_m(x) = i\}$ refer to the probability that the m^{th} customer finds, upon arrival, i phases in system, with $T_m = x$ and $T_n = T_n^*$ for $n = 2, ..., m-1$. We also define $R_m^* = R_m(T_m^*)$ and $p_{m,i}^* = \Pr\{R_m^* = i\}$.

When a customer finds i phases in system upon arrival, her expected waiting time is equal to $\frac{i}{\mu}$. Therefore, we have $W_m(x) = \sum\limits_{i=1}^{m-1} p_{m,i}(x)\frac{i}{\mu} = E[R_m(x)]\frac{1}{\mu}$ and $W_m^* = \sum\limits_{i=1}^{m-1} p_{m,i}^* \frac{i}{\mu} = E(R_m^*)\frac{1}{\mu}$. The service level constraint on waiting time ($W_m^* \leq s$) can then be interpreted as $E(R_m^*) \leq \mu s$. That is, the expected number of phases found in system upon each arrival is no more than μs.

To compute $p_{m,i}(x)$, we separate into two cases, $I \leq i \leq (m-1)I$ and $1 \leq i < I$. If the m^{th} customer finds i phases, for $I \leq i \leq (m-1)I$, then the $(m-1)^{th}$ customer must at least find $i - I$ phases. On the other hand, if the m^{th} customer finds i phases, for $1 \leq i < I$, then the $(m-1)^{th}$ customer could find 0 phases. Under both cases, for the m^{th} customer to find i customers given that the $(m-1)^{th}$ customer finds j, there must be exactly $j - i + I$ phase completions during the time interval (A_{m-1}^*, A_m) with length x. Since service time is Erlang distributed with shape I and rate μ, the number of phase completions during a time interval with length x is Poisson distributed with rate μx. Thus, we have

$$p_{m,i}(x) = \sum_{j=i-I}^{(m-2)I} p_{m-1,j}^* c_{j-i+I}(x) \tag{8}$$

for $I \leq i \leq (m-1)I$,

$$p_{m,i}(x) = \sum_{j=0}^{(m-2)I} p_{m-1,j}^* c_{j-i+I}(x) \tag{9}$$

for $1 \leq i < I$, and

$$p_{m,0}(x) = 1 - \sum_{i=1}^{(m-1)I} p_{m,i}(x).$$

Similarly,

$$p_{m,i}^* = \sum_{j=i-I}^{(m-2)I} p_{m-1,j}^* c_{m,j-i+I}^*$$

for $I \leq i \leq (m-1)I$,

$$p_{m,i}^* = \sum_{j=0}^{(m-2)I} p_{m-1,j}^* c_{m,j-i+I}^*$$

for $1 \leq i < I$, and

$$p_{m,0}^* = 1 - \sum_{i=1}^{(m-1)I} p_{m,i}^*.$$

Now, define

$$P_{m,n}(x) = \sum_{i=n}^{(m-1)I} p_{m,i}(x)$$

and

$$P^*_{m,n} = \sum_{i=n}^{(m-1)I} p^*_{m,i}$$

for $n = 0, ..., (m-1)I$. Then, we have

$$W_m(x) = \frac{1}{\mu} \sum_{i=1}^{(m-1)I} P_{m,i}(x)$$

and

$$W^*_m = \frac{1}{\mu} \sum_{i=1}^{(m-1)I} P^*_{m,i}.$$

When $n \geq I$, from Equation (8), we have

$$P_{m,n}(x) = \sum_{i=n}^{(m-1)I} \sum_{j=i-I}^{(m-2)I} p^*_{m-1,j} c_{j-i+I}(x) = \sum_{i=n-I}^{(m-2)I} c_{i-n+I}(x) \sum_{j=i}^{(m-2)I} p^*_{m-1,j}.$$

That is,

$$P_{m,n}(x) = \sum_{i=n-I}^{(m-2)I} P^*_{m-1,i} c_{i-n+I}(x).$$

Similarly,

$$P^*_{m,n} = \sum_{i=n-I}^{(m-2)I} P^*_{m-1,i} c^*_{m,i-n+I}. \tag{10}$$

When $n < I$, from Equation (9), we have

$$P_{m,n}(x) = \sum_{i=n}^{I-1} \sum_{j=0}^{(m-2)I} p^*_{m-1,j} c_{j-i+I}(x) + \sum_{i=I}^{(m-1)I} \sum_{j=i-I}^{(m-2)I} p^*_{m-1,j} c_{j-i+I}(x).$$

Similarly

$$P^*_{m,n} = \sum_{i=n}^{I-1} \sum_{j=0}^{(m-2)I} p^*_{m-1,j} c^*_{m,j-i+I} + \sum_{i=I}^{(m-1)I} \sum_{j=i-I}^{(m-2)I} p^*_{m-1,j} c^*_{m,j-i+I},$$

Note that the first term equals

$$\begin{aligned}
&\sum_{i=n}^{I-1} \sum_{j=0}^{(m-2)I} p^*_{m-1,j} c^*_{m,j-i+I} \\
=& c^*_{m,1} p^*_{m-1,0} + c^*_{m,2}(p^*_{m-1,1} + p^*_{m-1,0}) + ... + c^*_{m,I-n}(p^*_{m-1,I-n-1} + ... + p^*_{m-1,0}) + c^*_{m,I-n+1} \\
&(p^*_{m-1,I-n} + ... + p^*_{m-1,1}) + ... + c^*_{m,(m-1)I-(I-1)}(p^*_{m-1,(m-2)I} + ... + p^*_{m-1,(m-3)I+n+1}) \\
&+ c^*_{m,(m-1)I-I+2}(p^*_{m-1,(m-2)I} + ... + p^*_{m-1,(m-3)I+n+2}) + ... + c^*_{m,(m-1)I-n-1}(p^*_{m-1,(m-2)I} \\
&+ p^*_{m-1,(m-2)I-1}) + c^*_{m,(m-1)I-n} p^*_{m-1,(m-2)I} \\
=& c^*_{m,1}(P^*_{m-1,0} - P^*_{m-1,1}) + c^*_{m,2}(P^*_{m-1,0} - P^*_{m-1,2}) + ... + c^*_{m,I-n}(P^*_{m-1,0} - P^*_{m-1,I-n})
\end{aligned}$$

$$+ c^*_{m,I-n+1}(P^*_{m-1,1} - P^*_{m-1,I-n+1}) + c^*_{m,I-n+2}(P^*_{m-1,2} - P^*_{m-1,I-n+2}) + ... + c^*_{m,(m-1)I-(I-1)}$$
$$(P^*_{m-1,(m-3)I+n+1} - 0) + c^*_{m,(m-1)I-I+2}P^*_{m-1,(m-3)I+n+2} + ... + c^*_{m,(m-1)I-n-1}P^*_{m-1,(m-2)I-1}$$
$$+ c^*_{m,(m-1)I-n}P^*_{m-1,(m-2)I}$$
$$= \sum_{i=0}^{(m-2)I} c^*_{m,i-n+I}P^*_{m-1,i} + P^*_{m-1,0}\sum_{i=1}^{I-n-1} c^*_{m,i} - \sum_{i=1}^{(m-2)I} c^*_{m,i}P^*_{m-1,i},$$

and the second term equals $\sum\limits_{i=I}^{(m-1)I}\sum\limits_{j=i-I}^{(m-2)I} p^*_{m-1,j}c^*_{m,j-i+I} = P^*_{m,I} = \sum\limits_{i=0}^{(m-2)I} P^*_{m-1,i}c^*_{m,i}$. Thus, we have for $n < I$,

$$P^*_{m,n} = \sum_{i=0}^{(m-2)I} P^*_{m-1,i}c^*_{m,i-n+I} + P^*_{m-1,0}\sum_{i=0}^{I-n-1} c^*_{m,i}.$$

Define $\tilde{I} = \frac{I(I-1)}{2}$, then we have the following proposition which corresponds to Proposition 1 for the case with exponential service time distributions.

Proposition 2. *$(P^*_{m,(m-1)I}, P^*_{m,(m-1)I-1}, ..., P^*_{m,(m-2)I+1}, P^*_{m,(m-2)I}, ..., P^*_{m,1}, \tilde{I}P^*_{m,0}) \prec (0, 0, ..., 0, P^*_{m-1,(m-2)I}, ..., P^*_{m-1,1}, \tilde{I}P^*_{m-1,0})$ for $m = \lfloor \mu s \rfloor + 3, ..., M$.*

Proof: Let $\mathbf{u} = (P^*_{m,(m-1)I}, P^*_{m,(m-1)I-1}, ..., P^*_{m,(m-2)I+1}, P^*_{m,(m-2)I}, ..., P^*_{m,1}, \tilde{I}P^*_{m,0})^T$ and $\mathbf{v} = (0, 0, ..., 0, P^*_{m-1,(m-2)I}, ..., P^*_{m-1,1}, \tilde{I}P^*_{m-1,0})^T$, and define

$$D = \begin{pmatrix}
1 - c^*_{m,0} & 0,...,0 & c^*_{m,0} & 0 & ... & 0 & 0 \\
1 - \sum_{i=0}^{1} c^*_{m,i} & 0,...,0 & c^*_{m,1} & c^*_{m,0} & ... & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
1 - \sum_{i=0}^{(m-2)I-1} c^*_{m,i} & 0,...,0 & c^*_{m,(m-2)I-1} & c^*_{m,(m-2)I-2} & ... & c^*_{m,0} & 0 \\
1 - \sum_{i=1}^{(m-2)I} c^*_{m,i} - \frac{1}{\tilde{I}}c^*_{m,0} & 0,...,0 & c^*_{m,(m-2)I} & c^*_{m,(m-2)I-1} & ... & c^*_{m,1} & \frac{1}{\tilde{I}}c^*_{m,0} \\
1 - \sum_{i=2}^{(m-2)I+1} c^*_{m,i} - \frac{1}{\tilde{I}}\sum_{i=0}^{1} c^*_{m,i} & 0,...,0 & c^*_{m,(m-2)I+1} & c^*_{m,(m-2)I} & ... & c^*_{m,2} & \frac{1}{\tilde{I}}\sum_{i=0}^{1} c^*_{m,i} \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
1 - \sum_{i=I}^{(m-1)I-1} c^*_{m,i} - \frac{1}{\tilde{I}}\sum_{i=0}^{I-1} c^*_{m,i} & 0,...,0 & c^*_{m,(m-1)I-1} & c^*_{m,(m-1)I-2} & ... & c^*_{m,I} & \frac{1}{\tilde{I}}\sum_{i=0}^{I-1} c^*_{m,i} \\
d & 1,...,1 & 1 - \sum_{i=0}^{(m-1)I-1} c^*_{m,i} & 1 - \sum_{i=0}^{(m-1)I-2} c^*_{m,i} & ... & 1 - \sum_{i=0}^{I} c^*_{m,i} & 1 - \frac{1}{\tilde{I}}(\sum_{i=0}^{I-1} c^*_{m,i} + ... + \sum_{i=0}^{1} c^*_{m,i} + c^*_{m,0})
\end{pmatrix}$$

where d satisfies $d + \sum\limits_{i=I}^{(m-1)I-1}(1 - \sum\limits_{j=0}^{i} c^*_{m,i}) + I - \frac{1}{\tilde{I}}\sum\limits_{i=0}^{I-1}\sum\limits_{j=0}^{i} c^*_{m,i} = 1$.

First, note that $P^*_{m,0} = P^*_{m-1,0} = 1$, and $\sum\limits_{i=1}^{(m-1)I} P^*_{m,i} = \sum\limits_{i=1}^{(m-2)I} P^*_{m-1,i} = \mu s$. Now, consider $(1 - \sum\limits_{i=0}^{(m-1)I-1} c^*_{m,i})P^*_{m-1,(m-2)I} + (1 - \sum\limits_{i=0}^{(m-1)I-2} c^*_{m,i})P^*_{m-1,(m-2)I-1} + ... + (1 - \sum\limits_{i=0}^{I} c^*_{m,i})P^*_{m-1,1} + (1 - \frac{1}{\tilde{I}}\sum\limits_{i=0}^{I-1} c^*_{m,i} - \frac{1}{\tilde{I}}\sum\limits_{i=0}^{I-2} c^*_{m,i} - ... - \frac{1}{\tilde{I}}\sum\limits_{i=0}^{1} c^*_{m,i} - \frac{1}{\tilde{I}}c^*_{m,0})\tilde{I}P^*_{m-1,0}$. The sum of the negative terms equals

$$[(\sum_{i=0}^{I-1} c^*_{m,i})P^*_{m-1,0} + c^*_{m,I}P^*_{m-1,1} + c^*_{m,I+1}P^*_{m-1,2} + ... + c^*_{m,(m-1)I-1}P^*_{m-1,(m-2)I}]$$

$$+[(\sum_{i=0}^{I-2} c^*_{m,i})P^*_{m-1,0} + c^*_{m,I-1}P^*_{m-1,1} + c^*_{m,I}P^*_{m-1,2} + ... + c^*_{m,(m-1)I-2}P^*_{m-1,(m-2)I}]$$

$$+... + [c^*_{m,0}P^*_{m-1,1} + c^*_{m,1}P^*_{m-1,2} + ... + c^*_{m,(m-2)I-1}P^*_{m-1,(m-2)I}]$$

$$=P^*_{m,1} + P^*_{m,2} + ... + P^*_{m,(m-1)I} = \sum_{i=1}^{(m-1)I} P^*_{m,i} = \mu s.$$

Thus,

$$(1 - \sum_{i=0}^{(m-1)I-1} c^*_{m,i})P^*_{m-1,(m-2)I} + (1 - \sum_{i=0}^{(m-1)I-2} c^*_{m,i})P^*_{m-1,(m-2)I-1} + ... + (1 - \sum_{i=0}^{I} c^*_{m,i})P^*_{m-1,1}$$

$$+(1 - \frac{1}{\tilde{I}}\sum_{i=0}^{I-1} c^*_{m,i} - \frac{1}{\tilde{I}}\sum_{i=0}^{I-2} c^*_{m,i} - ... - \frac{1}{\tilde{I}}\sum_{i=0}^{1} c^*_{m,i} - \frac{1}{\tilde{I}}c^*_{m,0})\tilde{I}P^*_{m-1,0}$$

$$= \sum_{i=1}^{(m-2)I} P^*_{m-1,i} + \tilde{I}P^*_{m-1,0} - \mu s = \mu s + \tilde{I} - \mu s = \tilde{I} = \tilde{I}P^*_{m,0}.$$

Then, from Equation (10), we have $\mathbf{u} = D\mathbf{v}$.

The rest of the proof is the same as that for Proposition 1, and is therefore omitted for the sake of brevity. □

Following Proposition 2, we obtain results correspond to those for the case with exponential service time distributions.

Theorem 3. *$T^*_{m+1} \geq T^*_m$ for systems with Erlang service time distributions.*

Proposition 3. *For systems with Erlang service time distributions, the following hold*

(a) $p^*_{m+1,0} \geq p^*_{m,0}$ for $m = \lfloor \frac{\mu s}{I} \rfloor + 2, ..., M$,

(b) $Pr\{w^*_{m+1} \leq \frac{I}{\mu}\} \geq Pr\{w^*_m \leq \frac{I}{\mu}\}$ for $m = \lfloor \frac{\mu s}{I} \rfloor + 2, ..., M$,

(c) $\lim_{M\to\infty} p^*_{M,i}$ exists for $i = 0, 1, ...$

The proof for Theorem 3 and Proposition 3 is similar to the proof for Theorem 1, Corollary 1, and Lemma 6, and is therefore omitted for the sake of brevity. We see that all the results hold for the case with Erlang service time distributions, and the system converges asymptotically to a D/Er/1 queueing system as the number of arrivals approaches infinity.

3.2 Multiple Servers

Consider a system with N (a finite positive integer) parallel and identical servers. Customer service times are i.i.d. and follow an exponential distribution with rate μ (a finite positive real number). An arriving customer starts service immediately if there is an available server. Otherwise, she waits in the queue and will be served by the first available server. We continue to use similar notations.

Since there are N servers, it is optimal to schedule the first N customers to arrive together at time 0, and their waiting time equals 0. Notice that, when n servers are occupied (n customers are under service) simultaneously, the time it takes to complete one service is exponentially distributed with rate μn. So, if customers $1, 2, ..., m$ are scheduled to arrive together at time 0, the expected waiting time of customer m equals $\frac{(m-N)^+}{\mu N}$. The following Lemma corresponds to Lemma 1-4 for the single server case.

Lemma 8. *For systems with N servers, the following hold*
(1) $W_m(x)$ is decreasing in x, and T_m^* is decreasing in s,
(2) $T_m^* = 0$ for $m = 2, ..., \lfloor \mu N s \rfloor + N$, and $W_m^* = s$ for $m = \lfloor \mu N s \rfloor + N + 1, ..., M$,
(3) The expected number of service completions during the time interval (A_{m-1}^*, A_m^*) equals 1 for $m = \lfloor \mu N s \rfloor + N + 2, ..., M$,
(4) $T_m^* \geq \frac{1}{\mu N}$ for $m = \lfloor \mu N s \rfloor + N + 2, ..., M$.

As for the single server case, we first characterize the probability $\Pr\{R_m^* = i \mid R_{m-1}^* = j\}$ for $\lfloor \mu N s \rfloor + N + 1 \leq m \leq M$, $1 \leq i \leq m-1$, and $i - 1 \leq j \leq m-2$. For the m^{th} customer to find i customers given that the $(m-1)^{th}$ customer finds j, there must be exactly $j - i + 1$ service completions during the time interval (A_{m-1}, A_m). We distinguish the following three cases.

Case 1, $N \leq i \leq j+1$: In this case, all the servers are busy during the time interval (A_{m-1}, A_m). Therefore, the departure process is Poisson with rate μN. So,

$$\Pr\{R_m^* = i \mid R_{m-1}^* = j\} = \frac{(\mu N T_m^*)^{j-i+1}}{(j-i+1)!} e^{-\mu N T_m^*}.$$

Case 2, $1 \leq i \leq j + 1 \leq N$: In this case, both the m^{th} and the $(m+1)^{th}$ customers start service immediately upon arrival, and $\Pr\{R_m^* = i \mid R_{m-1}^* = j\}$ corresponds to the probability that exactly $j - i + 1$ among $j + 1$ customers complete their service during the time interval (A_{m-1}, A_m) (and the other i customers do not complete). The service time of each customer is exponentially distributed with rate μ. Noticing that $\binom{j+1}{j-i+1} = \frac{(j+1)!}{i!(j-i+1)!}$, we have

$$\Pr\{R_m^* = i \mid R_{m-1}^* = j\} = \frac{(j+1)!}{i!(j-i+1)!} (1 - e^{-\mu T_m^*})^{j-i+1} (e^{-\mu T_m^*})^i.$$

Case 3, $1 \leq i < N < j + 1$: In this case, all the servers are busy immediately after A_{m-1} and there are $j - N + 1$ queued customers. However, some servers become idle and the queue is empty at A_m. To have $j - i + 1$ service completions, we need the following two events to happen during the time interval (A_{m-1}, A_m); (1) the first $j - N + 1$ queued customers leave the queue and enter service, which implies that $j - N + 1$ customers complete their service; (2) $N - i$ customers complete their service afterwards. The departure process during (1) is the same as that in Case 1. The time it takes to complete the first $j-N+1$ services (when all the servers are busy) is Erlang distributed with shape $j-N+1$ and rate μN. The departure process during (2) is same as that in Case 2. Exactly $N-i$ among N customers complete their service. Suppose the time duration for (1) equals t for $t \in (0, T_m^*)$, then the time duration for (2) equals $T_m^* - t$. Thus,

$$\Pr\{R_m^* = i \mid R_{m-1}^* = j\} = \int\limits_0^{T_m^*} \left[\frac{N!}{i!(N-i)!}(1-e^{-\mu(T_m^*-t)})^{N-i} e^{-\mu(T_m^*-t)i}\right] \frac{(\mu N)^{j-N+1} t^{j-N} e^{-\mu N t}}{(j-N)!} dt.$$

As a result, we have

$$p_{m,i}^* = \sum_{j=i-1}^{N-1} p_{m-1,j}^* \frac{(j+1)!}{i!(j-i+1)!}(1-e^{-\mu T_m^*})^{j-i+1}(e^{-\mu T_m^*})^i$$
$$+ \sum_{j=N}^{m-2} p_{m-1,j}^* \int\limits_0^{T_m^*} \left[\frac{N!}{i!(N-i)!}(1-e^{-\mu(T_m^*-t)})^{N-i} e^{-\mu(T_m^*-t)i}\right] \frac{(\mu N)^{j-N+1} t^{j-N} e^{-\mu N t}}{(j-N)!} dt \quad (11)$$

for $i = 1, ..., N-1$, and

$$p_{m,i}^* = \sum_{j=i-1}^{m-2} p_{m-1,j}^* \frac{(\mu N T_m^*)^{j-i+1}}{(j-i+1)!} e^{-\mu N T_m^*} \quad (12)$$

for $i = N, ..., m-1$. As for the single server case, $p_{m,0}^* = 1 - \sum_{i=1}^{m-1} p_{m,i}^*$ for $m = \lfloor \mu N s \rfloor + N + 1, ..., M$. Notice also that $p_{\lfloor \mu N s \rfloor + N, \lfloor \mu N s \rfloor + N - 1} = 1$, and $p_{\lfloor \mu N s \rfloor + N, i} = 0$ for $i \neq \lfloor \mu N s \rfloor + N - 1$.

Now, $W_m^* = \frac{(m-N)^+}{\mu N}$ for $m = 1, ..., \lfloor \mu N s \rfloor + N$. For $m = \lfloor \mu N s \rfloor + N + 1, ..., M$, when a customer finds i customers in system upon arrival, her expected waiting time equals 0 if $i < N$, and $\frac{i-N+1}{\mu N}$ if $i \geq N$. Therefore, we have $W_m^* = \sum_{i=N}^{m-1} p_{m,i}^* \frac{i-N+1}{\mu N}$.

Due to the complexity of Equations (11) and (12), it is difficult to apply the theory of majorization here. However, through extensive numerical experiments (see Figure 1 for an illustration of a system with $M = 30$, $N = 4$, $\mu = 0.5$, and $s = 10$), we see that the main results (e.g. T_m^* increases and converges) also hold for the case with multiple servers.

4 Concluding Comments

We summarize the contributions of our paper as follows:

- We bring a new modeling perspective to appointment scheduling by considering a system that minimizes servers' idling time under a constraint on customers's waiting time. Our model avoids the well-known difficulty in estimating customers' waiting cost.

- Our waiting time constraint captures the threshold type behavior in customers' perception of waiting (patience to wait), which is empirically observed.

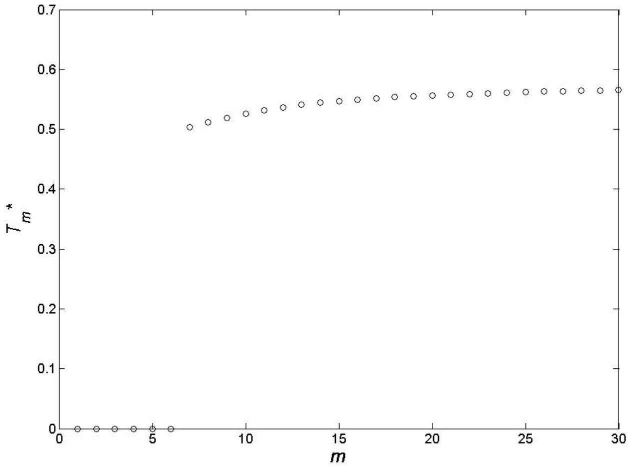

Figure 1: Optimal Inter-Appointment Times for Systems with Multiple Servers

- We incorporate the fairness concern in our modeling. Our resulting schedule leads to the same expected waiting time among all customers (other than the first few ones).
- Our system minimizes customers' indirect waiting time, which is ignored in most literature.
- We assume Erlang service time distributions, which increases the applicability of our model (compared with exponential service time distributions used in most of the existing literature).
- We apply the theory of majorization to analytically characterize the structure of the optimal appointment schedule. To the best of our knowledge, this is the first of its kind in the appointment scheduling literature.
- We study the limiting behavior of our system and prove the convergence to the D/M/1 queueing system for the case with exponential service time distributions.
- We confirm the robustness of our results in systems with multiple servers, which is seldom treated in the literature.

For future research, first, it will be useful to consider other types of service level constraints such as one on the probability of long waiting (e.g. $\Pr\{W_m^* > s\} \leq a$, $\forall m$). However, it is not easy to apply the theory of majorization and prove the structural results in that case. It will also be interesting to consider the possibility of customer non-punctuality and no-shows and see how the optimal appointment schedules are affected.

References

Baron O, Berman O, Krass D, Wang J (2016). Strategic Idling and Dynamic Scheduling in an Open-Shop Service Network: Case Study and Analysis. *Working Paper, University of Toronto.*

Cayirli T, Veral E (2003). Outpatient Scheduling in Health Care: a Review of Literature. *Production and Operations Management.* 12(4):519-549.

Fries BE, Marathe VP (1981). Determination of Optimal Variable-Sized Multiple-Block Appointment Systems. *Operations Research.* 29(2):324-345.

Hassin R, Mendel S (2008). Scheduling Arrivals to Queues: a Single-Server Model with No-Shows. *Management Science.* 54(3):565-572.

Kong Q, Lee CY, Teo CP, Zheng Z (2013). Scheduling Arrivals to a Stochastic Service Delivery System Using Copositive Cones. *Operations Research.* 61(3):711-726.

Marshall AW, Olkin I, Arnold BC (2009). *Inequalities: Theory of Majorization and Its Applications.* Springer.

Mondschein SV, Weintraub GY (2003). Appointment Policies in Service Operations: a Critical Analysis of the Economic Framework. *Production and Operations Management.* 12(2):266-286.

Wang R, Jouini O, Benjaafar S (2014). Service Systems with Finite and Heterogeneous Customer Arrivals. *Manufacturing & Service Operations Management.* 16(3):365-380.