A STOCHASTIC APPROACH TO APPOINTMENT SEQUENCING

AHMAD REZA POURGHADERI

(M.Sc. Industrial Engineering, University of Tehran)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF INDUSTRIAL & SYSTEMS ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2014

Declaration

I hereby declare that the thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

AHMAD REZA POURGHADERI

17 Aug 2014

To my dear prophet Muhammad (S.A.W.)

To my lovely wife Fatemeh

Acknowledgments

First and foremost, I would like to express my sincerest gratitude to my supervisor, Dr. Boray Huang, for his constant support, motivation, guidance and training without which this dissertation would not have been possible. He guided me through difficult times in research and also in life, and has been a wonderful friend and mentor who has always amazed me with his intelligence and creativity and above all compassion and patience.

I would like to especially thank my thesis committee members, Professor Kien Ming Ng and Dr. Shuangchi He, for their invaluable advice on improving my thesis.

I am grateful to the wonderful faculty members in our department. I would like to particularly thank the head of our department, Professor Loon Ching Tang for his encouragement when I just embarked on my Ph.D. journey and the consistent support throughout it. I would also like to extend my appreciation to the staff of our department office, especially Ms. Lai Chuan for her commitment and support.

Gratitude also goes to Professor David Yao from Colombia University and Professor Mark Van Oyen from University of Michigan for their constructive comments on my research.

During my Ph.D., I am very fortunate to have the opportunities to

experience various teaching duties and learn from many excellent educators, including Professor Szu Hui Ng, Professor Kien Ming Ng, and Dr. Chin Hon Tan. I am grateful to their generous support and guidance in this early stage of my teaching journey.

A great deal of appreciation goes to my parents for their unconditional love and support, as always. I would also like to extend my very great appriciation to my parents in law, Mr. Jamshidian and Mrs. Nikyar, for being a constant source of support and encouragement over the past seven years.

Words cannot express my gratitude and love to my wife, *Fatemeh Jamshidian*. Her faith in me is my greatest source of inspiration and motivation. My life will not be so complete and meaningful without her and our little baby, *Raja*.

17 August 2014 **Ahmad Reza Pourghaderi**

Singapore

The financial support for this work was provided by the Academic Research Fund of the National University of Singapore under NUS Research Scholarship (No.: C-266-000-207-532). It was also partially supported by the Singapore Lee Foundation.

Contents

Summary

Efficiently regulating the arrival of customers through a well-designed appointment system is a critical factor to the performance of many service delivery systems. Among various applications, perhaps the most important application of appointment systems is in healthcare for out-patient and elective surgery scheduling. In this thesis, a useful managerial insight is obtained which could improve the performance of appointment systems in terms of the customers' waiting times that is a main concern for most healthcare providers.

We study a single server appointment-based queueing system with two classes of customers, *regular* and *fast*. The excess service time of a fast customer is stochastically less than that of a regular customer where the excess service time for each customer is defined to be the difference between the service duration and the corresponding job allowance (the length of the appointment slot allocated to the customer). The majority of the appointment scheduling research focus on finding the optimal schedule (appointment times) for either homogeneous customers or heterogeneous customers in a predetermined sequence. Very little is known about the structure of the optimal arrival sequence for various objective functions. In contrast, we focus on finding the optimal arrival sequence to minimize the customer's waiting time.

We first consider customers with exponential service durations includ-

ing only one fast customer to provide counter-examples to challenge the Smallest Variance first (SV) and the Shortest Expected Processing Time first (SEPT) rules which are widely conjectured to minimize the customer's waiting times in the literature. We also provide a sufficient condition to guarantee that SEPT/SV is not optimal as well as a reasonable explanation for this counter intuitive observation by introducing a new concept, *voucher effect*, in appointment systems.

Moreover, we have observed that the optimal slot for the fast customer is not necessarily the first one, but it is always in the first half of the sequence. Based on this interesting observation, a useful insight is obtained which implies that each fast customer must be scheduled in a position that is in the first half of the positions after the previous fast customer, the First Half Rule (FHR). This sequencing rule is established under the likelihood ratio ordering assumption of the excess service times. In addition, a simple and effective FHR-based heuristic algorithm to completely characterise the optimal sequence is proposed which shows an impressive performance over the test problems.

While the application of the FHR is not limited to appointment systems with constant job allowance, it could be applied to any system with equally spaced appointment times and two classes of customers from a same service distribution family with Monotone Likelihood Ratio Property, for example exponential, beta, Weibull, normal with known variance, uniform, gamma,

Poisson, geometric and binomial distributions.

Eventually, we extend our results to address two important practical issues: the server unpunctuality and the customer no-shows. Our results also could be applied to schedule the breaks in an appointment system with equally spaced appointment times as well.

List of Tables

List of Figures

Chapter 1

Introduction

Efficiently regulating the arrival of customers through a well-designed appointment system is a critical factor to the performance of many service delivery systems. Among various applications of appointment systems, e.g. in accounting services, professional consultants, legal services, barber shops/beauty salons, visa services, container vessel and terminal operations, airport gates and runway schedules, perhaps the most important application is in healthcare industry.

Nowadays, healthcare service providers are mostly under a great deal of pressure to improve efficiency. Besides the high cost of medical resources, another reason for this pressure is that the world is rapidly ageing. According to the World Health Organization (WHO)[1], the world's population of people 60 years of age and older has doubled since 1980 and is forecast

[1] `http://www.who.int/features/factfiles/ageing`

to reach 2 billion (21%) by 2050. United Nations also reports that the population ageing is taking place in nearly all the countries of the world [2]. Consequently, the healthcare demand is rapidly growing which emphasizes the need for improving appointment scheduling systems to increase the utilization of expensive personnel and equipment-based medical resources as well as reducing waiting times for patients.

Over the past six decades, the problem of designing appointment systems has been studied widely in the operations research and medical literature. However, the current literature is still unable to introduce a general appointment sequencing policy to minimize the customers' waiting time. It motivates us to study the appointment sequencing problem.

1.1 Research scope and objective

Appointment scheduling problem is mostly studied in the context of outpatient and elective surgery scheduling. In the current literature, most studies consider appointment systems with no customer classification, assuming customers are homogeneous and do not address the sequencing aspect of the problem. Among studies considering heterogeneous customers, the majority focus on finding the optimal schedule (appointment times) for a given sequence of customers. For homogeneous customers or heteroge-

[2]For more information about population ageing please refer to *World Population Ageing 2013* report by United Nations, Department of Economic and Social Affairs.

neous customers with a given sequence, the problem can be solved efficiently mainly based on sub-modularity and convexity properties of the objective function (see e.g. Begen and Queyranne 2011, Denton and Gupta 2003, Ge et al. 2014, Kaandorp and Koole 2007, and Wang 1993). Even without complete information about the service time distributions, the problem of scheduling arrivals of a fixed sequence can be formulated as a convex conic optimization problem with a tractable semi-definite relaxation (Kong et al., 2013).

In contrast, the literature on the appointment sequencing problem is relatively limited. Very little is known about the structure of the optimal sequence of customers for various objective functions. While there is no analytical result for sequencing more than three customers, the optimality of the smallest variance first rule (SV) is widely conjectured for various measurements. The reason for this conjecture is that the customers' waiting time is usually an important component of the objective, thereby implying that scheduling a customer with a lower variance of service duration earlier in the schedule would decrease the expected waiting time of the customers arriving later.

The appointment scheduling and sequencing problem is extremely difficult. Since the transient performance measures of the system are analyzed, standard queueing theory does not apply. The technical complexity originates from the fact that for calculation of the expected waiting time of an

arrival we have to consider the impact of all earlier events. Even under a deterministic model where the time interval allocated to each customer is set to a constant, and the service duration for each customer is known in advance, the sequencing problem is still NP-hard in the strong sense (Kong et al., 2014).

We study an appointment system with two distinct classes of customers with random required service durations. Customers within one class are homogeneous and have the same service time distributions. The problem is static in the sense that the number of customers of each class and their service time distribution are known in advance. We investigate the impact of the sequencing policy on customers' waiting times when there is a predetermined scheduling policy. In other words, we focus on finding the optimal sequence of customers when the time allowance that should be allocated to each customer is already decided.

1.2 Summary of research contributions

We start from the simplest case where there are $N-1$ identical *regular* customers and only one special *fast* customer to be sequenced in a system with equally spaced appointment times. The service times are exponentially distributed and the service rate is higher for the fast customer. We propose a method to exactly compute the expected customers' waiting times and provide counter intuitive examples to show the optimal slot for the fast

customer is not necessarily the first one. The results challenge the Smallest Variance (SV) first rule and the Shortest Expected Processing Time (SEPT) first rule, which is another famous sequencing policy expected to be optimal in this case. Then we introduce a new concept, called *voucher effect*, in appointment systems to explain this counter intuitive observation.

Moreover, it is observed that the optimal slot for the fast customer is always within the first half of the sequence. Based on this observation, a new sequencing rule, called the First Half Rule (FHR), is proposed. We relax the constant appointment interval and the exponential distribution assumptions and study a more general appointment system. We establish the FHR assuming that the excess service time of the fast customer is smaller than that of a regular customer in likelihood ratio order where the excess service time of a customer is the difference between his/her service duration and allocated time interval.

The next contribution of this work is the extension of the applicability of the FHR to an appointment system with a late server. It could be helpful especially for the surgical scheduling where the operating room is still occupied by the previous surgery team at the beginning of the new session.

After considering an unpunctual server, the results have been extended to multiple fast customers which makes the FHR a powerful sequencing rule for appointment systems with two classes of customers. To the best of our

knowledge, the FHR is the only appointment sequencing rule analytically established for sequencing of more than three arrivals with known service distributions.

A simple and effective appointment sequencing heuristic, based on the FHR, is proposed which shows an impressive performance to find the optimal sequence.

Finally, we address the no-show phenomenon which is nowadays a main concern for many service providers especially in the healthcare industry. The applicability of the FHR is shown where the fast customer is a customer with a smaller probability of showing up for service. The significance of this result is that it holds for any service distribution. This result also can be used to schedule break times in static appointment scheduling problem with homogeneous customers.

1.3 Notations

We consider the problem of sequencing N punctual customers to a single server queue. The operational target is to minimize the total expected customers' waiting time. We assume the customers can be divided into two classes based on their excess service times: *fast* (with an excess service time distribution F) and *regular* (with an excess service time distribution G), where F is stochastically smaller than G in *likelihood ratio order*. The excess service time for each customer is defined to be the difference between

his/her service duration and the corresponding job allowance.

Inspired by Kendall's queueing notation (Kendall et al., 1953), Pegden and Rosenshine (1990) used the notation $S(n)/M/1$ to denote a queueing system in which n identical customers are to be scheduled to a single exponential server. In a later study, Hassin and Mendel (2008) consider a showing up probability of p for each customer and denote the system by $S(n,p)/M/1$. Following a similar definition, we denote the system as an $S(M, N-M)/(F,G)/1$ queue where there are M fast and $N-M$ regular customers to be sequenced.

We denote the system as $S(M, N-M)/(SM, SM')/1$ where the excess service times follow shifted exponential distributions with rates μ and μ', both shifted by a constant $x \geq 0$. Moreover, under deterministic service time assumption, the problem is denoted by $S(M, N-M)/(D, D')/1$.

In the next section, we provide some background information on appointment scheduling and stochastic ordering.

1.4 Background

In this section, we first briefly review some important aspects of the appointment scheduling problem and then give an overview of some fundamentals of the usual and likelihood ratio stochastic orders.

1.4.1 Appointment scheduling systems overview

We borrow some perspectives from Gupta and Denton (2008). Studies on appointment scheduling are classified into two categories: *static* and *dynamic*. In the static appointment scheduling, all decisions must be made prior to the beginning of a service session, while in the dynamic case, the schedule of future arrivals is revised continuously based on the current state of the system over the course of the day.

A scheduled customer usually faces two types of access delays: *indirect* and *direct*. Indirect waiting time is the difference between the time that s/he requests an appointment and the time of that appointment. Direct waiting time is the difference between the customer's appointment time (or his/her arrival time if the customer is tardy) and the time when s/he is actually served by the service provider.

To evaluate an appointment system, a variety of performance criteria can be used such as cost-based, fairness and congestion measures. The cost-based measure is the most common criteria used in the literature. Waiting time or flow time (the total time a customer spends in the service centre) of customers and available time, overtime or idle time of server can be used for calculating the total system cost. The mean and variance of queue sizes are two examples for congestion and fairness measures respectively.

Arrival process is a key factor that affects the performance of appointment systems. Presence of late cancellations, unpunctual patients, no-

shows (late cancellations that cannot be replaced) and walk-ins (unscheduled arrivals that may be urgent) are the most important factors of the arrival characteristics of customers that make the appointment scheduling more complicated. Unpunctuality and no-shows of doctors are also challenging factors for appointment scheduling in healthcare.

Service time variability is another important factor influencing the performance of appointment systems. According to Gupta and Denton (2008), in the healthcare primary care setting, the vast majority of patients require services that can be performed within a fixed time length while in specialty care clinics the patients' service times tend to vary more depending on the patients' diagnoses and other characteristics. Patient classification can be used to sequence patients as well as to adjust the appointment intervals based on the distinct service time characteristics of different patient types. Although there is some evidence that classifying the patients may be advantageous, the majority of the studies assume patients are homogeneous, and use independently and identically distributed service times for all patients (Cayirli and Veral, 2003).

A variety of probability distributions are chosen for customer service times in the literature. Some studies used empirical data to show that the frequency distributions of observed service times are uni-modal and right-skewed (e.g. Meza 1998). Many analytical studies use exponential service times for mathematical tractability purposes.

In this thesis, we consider the static appointment sequencing problem to minimize the direct waiting time of punctual customers of two classes. We first study a system with exponential service times, punctual server, and without no-shows. Then, we relax these assumptions to address more practical situations.

1.4.2 Fundamentals of stochastic ordering

In probability theory, a stochastic order is a partial order which quantifies the comparison of random variables. Many different orders exist with various applications. In this section, we briefly give an overview of two important stochastic orders that compare the *location* (*magnitude*) of random variables: the usual stochastic order and the likelihood ratio order.

Usual Stochastic Order: A random variable X is smaller than a random variable Y in the *usual stochastic order*, denoted by $X \leq_{st} Y$, if for all $t \in (-\infty, +\infty)$,

$$\mathbb{P}(X > t) \leq \mathbb{P}(Y > t).$$

It follows that $X \leq_{st} Y$ if, and only if, $\mathbb{E}[\phi(X)] \leq \mathbb{E}[\phi(Y)]$ for all non-decreasing function ϕ for which the expectations exist.

Likelihood Ratio Order: A random variable X is said to be smaller

than a random variable Y in the *likelihood ratio order*, denoted by $X \leq_{lr} Y$, if $\frac{f(t)}{g(t)}$ is non-decreasing in t over the union of the supports of X and Y where f and g are density functions (mass functions for discrete variables) of X and Y respectively ($a/0$ is taken to be $+\infty$ for $a > 0$).

Let $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ and $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ be two vectors in $\mathbb{R}^n$. We denote $\mathbf{x} \leq \mathbf{y}$ if $x_i \leq y_i$ for $i = 1, 2, ..., n$. Let ϕ be a multivariate function with domain in $\mathbb{R}^n$. If $\phi(\mathbf{x}) \leq [\geq] \phi(\mathbf{y})$ whenever $\mathbf{x} \leq \mathbf{y}$, then we say that the function ϕ is non-decreasing [non-increasing]. A set $U \subseteq \mathbb{R}^n$ is called *upper* if $\mathbf{y} \in U$ whenever $\mathbf{y} \geq \mathbf{x}$ and $\mathbf{x} \in U$. Now, the multivariate extension of the usual stochastic order can be described as follows.

Usual Multivariate Stochastic Order: A random vector $\mathbf{X}$ is smaller than a random vector $\mathbf{Y}$ in the *usual stochastic order*, denoted by $\mathbf{X} \leq_{st} \mathbf{Y}$, if for all upper sets $U \subseteq \mathbb{R}^n$,

$$\mathbb{P}\{\mathbf{X} \in U\} \leq \mathbb{P}\{\mathbf{Y} \in U\}.$$

It follows that $\mathbf{X} \leq_{st} \mathbf{Y}$ if, and only if, $\mathbb{E}[\phi(\mathbf{X})] \leq \mathbb{E}[\phi(\mathbf{Y})]$ for all non-decreasing function ϕ for which the expectations exist.

We also use "$=_{st}$" to denote equality in law whenever two random variables or vectors have the same distribution.

Some useful properties of stochastic orders which are used in this thesis have been represented in Appendix A.

The most relevant literature to our research is reviewed in the next section.

1.5 Literature review

Starting with the pioneering work of Bailey (1952) and Welch and Bailey (1952), the problem of designing appointment systems has been studied extensively during the past six decades. In this chapter, we briefly review the most relevant findings to our paper and refer the readers to Cayirli and Veral (2003), Erdogan et al. (2010), Gupta (2007), and Gupta and Denton (2008) for comprehensive literature reviews.

In many service delivery systems, the customer population can be distinctly classified into different groups based on service time characteristics. For instance, in ambulatory care, some variables used for classifying patients include major problem, acute problem, acute problem follow-up and chronic problem (Arbitman, 1986), or for CAT scans, patients could be classified by procedure type (such as head, spine, brain, chest), or by age, if pediatric and geriatric patients typically have longer consultation time compared to adult patients (Cayirli et al., 2006). Another common situation in healthcare is the new/repeat classification of patients (see Cayirli et al. 2008 and Kong et al. 2013). The mean and variance of the consul-

tation times of the new patients are usually higher than those of repeat patients, because the conditions of the new patients are unknown prior to the first visit. According to Cayirli et al. (2008), a higher percentage of more complicated cases (e.g., new patients) normally translates into higher variability in the system performance, and thus proper sequencing of patients becomes more valuable. In another study, Cayirli et al. (2006) found that the impact of sequencing on the performance measure can be more important than the impact of the appointment scheduling rule, and the panel characteristics such as walk-ins, no-shows, punctuality and overall session volume, influence the effectiveness of appointment systems.

There are two sequencing policies mostly recommended by researchers for various objectives in the appointment scheduling literature: the Smallest Variance first rule (SV), and the Shortest Expected Processing Time first rule (SEPT).

Smallest Variance first rule (SV) was initially proposed by Weiss (1990). He can be considered as the first to address the problem of jointly determining the optimal order of customers and the optimal appointment times. He showed that sequencing lower-variance procedure first is optimal in two-customer case under exponential or uniform service duration where the objective function includes the customer waiting cost and the server idle cost. He also conjectured that the SV rule could be optimal in more complicated systems. Many researchers subsequently recommended the SV rule

for various appointment scheduling settings (e.g., Klassen and Rohleder 1996, Wang 1999, Rohleder and Klassen 2000, Dexter and Ledolter 2005, Denton et al. 2007, Pinedo (2009), and Choi and Wilhelm 2012).

Klassen and Rohleder (1996) consider the problem of sequencing patients of two classes, *low* and *high* variance patients, when appointment intervals are constant. They use simulation to compare alternative ways of sequencing and find that low-variance patients should be scheduled at the beginning of the session (called the LVBEG rule) when the objective is to minimize a linear combination of the patient waiting time and the operating room idle time costs. In a later study, Rohleder and Klassen (2000) consider the possibility that the scheduler cannot sequence patients perfectly when some patients insist on particular slots. In addition, they consider the possibility that the scheduler can make an error when classifying patients and find that the LVBEG rule still performs well under these assumptions.

Wang (1999) investigates the sequencing and scheduling problem of a set of customers with different exponential service rates. He proposes recursive expressions to find the customers' flow time distributions as well as an efficient method to evaluate an objective function that includes the customers' flow times and the server's completion time. Wang explicitly states that the optimal sequencing policy is in order of increasing variance. However, no analytical proof is provided for the case of sequencing more

than two customers. We will challenge this result in the next chapter.

Dexter and Ledolter (2005) study prediction bounds for operating room times (the time we can start surgery for each patient) and consider the effect of sequencing on the mean tardiness. They point out that sequencing less uncertain cases earlier reduces the patients' waiting times.

Denton et al. (2007) formulate the appointment sequencing and scheduling problem as a two-stage stochastic mixed-integer programming model and incorporate the service time uncertainty into the model using a sample approximation approach. They have shown that the scheduling problem of the start times for a given sequence is a linear stochastic program which can be solved efficiently by the L-shaped algorithm described in Denton and Gupta (2003). They have considered several simple sequencing heuristic rules and found that SV can provide the best results among the proposed heuristics. However, it is concluded that finding a general optimal sequencing policy is very complicated.

Pinedo (2009) considers the sequencing of two surgeries with independent uniform durations. He found that the variance of the surgery duration has a much stronger influence on the optimal schedule than does the mean. He also conjectured the optimality of the SV rule for more than two patients.

Recently, Choi and Wilhelm (2012) study the problem of sequencing two or three surgeries with durations that follow the lognormal, gamma

or normal distributions. The time allocated to each surgery is its mean duration time. They consider patient waiting, surgeon idle and operating room overtime cost, and have shown the optimality of SV rule for the two-patient case with normal service times. For other cases, they have numerically confirmed the efficiency of the SV rule.

Shortest Expected Processing Time first rule (SEPT) is optimal for various machine scheduling models in manufacturing systems. In the single machine setting, when all jobs are available at time zero, the SEPT has been shown to minimize the total completion time and the average number of jobs waiting for processing (Pinedo, 2012). The appointment scheduling model approaches to the classic single machine scheduling model when the appointment intervals approach to zero, i.e. all customers arrive at the beginning of the service session. The SEPT rule therefore minimizes the total expected customer waiting time under this extreme assumption. Moreover, in general, the calculation of the waiting times in an appointment system appears to be similar to the calculation of the job tardiness times in the counterpart manufacturing system. The SEPT has been shown to minimize the total expected tardiness under some considerations (Pinedo, 2012). Based on the above discussion, one may expect SEPT to be optimal for some appointment scheduling measurements.

Lehaney et al. (1999) implemented an appointment system that sorts patients in ascending order of mean service duration in a National Health

Service hospital in the US. It is found that SEPT could significantly improve the performance of the clinic in terms of the patient waiting times. Lebowitz (2003) also reports that scheduling short procedures first improved operating room efficiency in another hospital and provided a better operating room schedule by decreasing staff member overtime expense without reducing surgical throughput.

Marcon and Dexter (2006) consider the impact of sequencing on post anesthesia care unit staffing. They compare seven sequencing rules using discrete event simulation and find that the longest case first rule which is usually used in practice, performs poorly from a staffing perspective. They recommend the shortest case first rule for a number of decision rules.

Gul et al. (2011) evaluates how 12 different sequencing and scheduling heuristics perform with respect to the expected patient waiting time and expected surgical suite overtime. It is found that among the sequencing heuristics, SEPT performs best.

Besides the papers capturing special patterns for the structure of the optimal sequence, there are researchers who found that **it is very difficult to generalize any sequencing results** (e.g. Bosch and Dietz 2000, 2001, Jebali et al. 2006, and Mancilla and Storer 2012).

In Bosch and Dietz (2000, 2001), the authors assume that patients may only be scheduled at regularly-spaced times (every 10 minutes), called lattice program, with the objective of minimizing a weighted sum of waiting

time and overtime costs. They considered three classes of patients with phase type and log-normal distributions. Given a sequence, an efficient gradient-based algorithm is proposed to find the optimal schedule of starting times based on sub-modularity properties of the objective function. Then, they proposed a pairwise interchange heuristic to sequence patients of different classes based on sub-modularity and convexity of the objective function with respect to the arrival time vector. They found that it is very difficult to generalize any appointment sequencing rule.

Jebali et al. (2006) develop a two-step approach to solve the surgery assignment and sequencing problem. Firstly, operations are assigned to operating rooms with regards to the operating room under-time and overtime costs. Secondly, optimal sequences are found to minimize the total overtime cost for each operating room. No special pattern is captured for the optimal sequence.

Mancilla and Storer (2012) formulate a sample approximation two-stage stochastic programming model for the appointment sequencing and scheduling problem which is quite similar to the model proposed by Denton et al. (2007). The master problem is used to find sequences and the sub-problems are scheduling problems (stochastic linear programs). They propose a heuristic solution approach based on Benders' decomposition. It is realized that the master problem becomes extremely hard to solve as cuts are added. They consider waiting time, idle time and overtime costs

and have shown that the problem is NP-complete even with two scenarios. The proposed heuristic could provide better results than just sorting by variance over the test problems where it utilized much more computing time than SV rule. No special pattern has been captured for the structure of the optimal sequence.

In summary, very little is known about the structural properties of the optimal sequence for various appointment scheduling objective functions. We look at the problem from a new perspective assuming that the job allowances are predetermined and focusing on the structure of the optimal sequence to minimize the customers' waiting time. By developing an exact waiting time calculation method, we are able to investigate the performance of the SEPT and SV rules. After showing that they are not necessarily optimal, we introduce a new sequencing rule, the First Half Rule (FHR) which works for sequencing a finite number of customers of two classes under a mild assumption.

1.6 Organization of the thesis

The structure of this thesis is as follows. In §2, an appointment sequencing problem with exponential service time and one fast customer is investigated. Counter-examples for the optimality of SEPT/SV are presented in §2.2.2, followed by an explanation which introduce the voucher effect to justify scheduling a customer with a higher variance and mean service

duration first in §2.3. In §3, we establish the FHR based on the likelihood ratio order assumption of the excess service times. Then, we present two important FHR extensions: late server case in §3.2, and multiple fast customers in §3.3. A effective FHR-based appointment sequencing heuristic algorithm is developed in §3.4. Later, through incorporating no-shows in §4, another practical application of the FHR is addressed. Finally, we conclude in §5 and present some potential directions for future research.

Chapter 2

Sequencing customers of two classes with exponential service durations

We first study an appointment system with one fast customer and $N-1$, $N \geq 2$, regular customers. The service durations are exponentially distributed with rate μ for the fast customer and μ' for the regular customers, $\mu > \mu'$. The job allowances are set to a predetermined constant $x \geq 0$. That is the customers will arrive punctually at times $0, x, ..., (N-1)x$.

The question we are interested to answer is *in order to minimize the total expected waiting time, should we schedule the fast customer first?*

As discussed in §1.5, according to both SV and SEPT rules, the answer is expected to be yes. Wang (1999) considers appointment sequencing

of customers with exponential service times and explicitly expresses that SV is optimal for minimizing the weighted average of the total customers' flow time and the server completion time where the flow time for each customer is summation of the waiting time and the service duration (see *Preposition 4* in Wang 1999). Considering the fact that the total expected service duration is constant, independent of the sequence, and assuming the server availability cost is zero, our objective function exactly matches the Wang's objective. Also, Pinedo (2009), after showing the optimality of SV in sequencing two surgeries, expressed that "*showing that SV is optimal in an environment with n surgeries is considerably harder*". We first in §2.1, show that the conjecture of optimality of SV and SEPT is true for the three-customer case and then in §2.2 provide some counter-examples which show that this is not true in general. Later in §2.3, we give an explanation for this counter-intuitive observation.

Let *sequence* m, be the arrival sequence in which the fast customer is sequenced in the m-th position, i.e. the fast customer is scheduled at the m-th slot and will arrive at time $(m-1)x$. Let $\mathbb{E}W_n^m$ denote the expected waiting time of the n-th arrival of sequence m. Also, let $\mathbb{E}W^m$ denote the total expected waiting time of sequence m, i.e. $\mathbb{E}W^m = \sum_{n=1}^{N} \mathbb{E}W_n^m$.

2.1 $S(1,2)/(SM, SM')/1$

When there are one fast and two regular customers to be sequenced, the total expected waiting time for different sequences can be calculated by the following expressions. Without loss of generality, we can assume that $\mu' = 1$.

$$\mathbb{E}W^1 = e^{-x} + \frac{1}{\mu}e^{-\mu x} + \frac{1}{\mu - 1}e^{-(\mu+1)x} + \frac{1}{\mu(1-\mu)}e^{-2\mu x}$$

$$\mathbb{E}W^2 = e^{-x} + \frac{1}{\mu}e^{-\mu x} + \frac{1}{1 - \mu}e^{-(\mu+1)x} + \frac{\mu}{\mu - 1}e^{-2x}$$

$$\mathbb{E}W^3 = 2e^{-x} + (x+1)e^{-2x}$$

The calculation details can be found in Appendix B.1. The following result shows that SEPT/SV is optimal for $S(1,2)/(SM, SM')/1$.

Lemma 2.1. *The optimal sequence to minimize the total expected waiting time of three customers with exponential service durations in an equally spaced appointment system is the descending order of the customer service rates.*

Proof.

$$\begin{aligned}\mathbb{E}W^2 - \mathbb{E}W^1 &= \frac{\mu}{\mu-1}e^{-2x} - \frac{2}{\mu-1}e^{-(\mu+1)x} + \frac{1}{\mu(\mu-1)}e^{-2\mu x} \\ &= \frac{[(\mu - e^{-(\mu-1)})e^{-x}]^2}{\mu(\mu-1)}\end{aligned}$$

It is obviously positive, since $\mu > 1$.

In addition, the last arrival's service duration does not affect the waiting time of any customer. Therefore, the slowest customer should be scheduled last. □

In the next section, we increase the number of regular customers to more than 2 and violate the optimality of SV/SEPT.

2.2 $S(1, N-1)/(SM, SM')/1$, $N > 3$

For more than three customers, the first challenge is to compute the expected waiting times. The waiting time of the n-th arrival is a function of the service time of all the first $n-1$ arrivals and can be computed recursively by

$$W_n^m = [W_{n-1}^m + S_{n-1}^m - x_{n-1}^m]^+$$

where $W_1^m = 0$, and for $n = 2, 3, ..., N$, W_n^m,S_n^m and x_n^m are the waiting time, the service time and the job allowance of the nth arrival of sequence m respectively and $[.]^+ = max\{0, .\}$. Define the *excess service time* of the nth arrival $Z_n^m = S_n^m - x_n^m$. The total waiting time calculation can be represented as a Maximum Cost Flow problem and W_n^m can be calculated using the following *Max-Flow* expression (Kong et al., 2013).

$$W_n^m = \max\{0, Z_{n-1}^m, Z_{n-1}^m + Z_{n-2}^m, ..., \sum_{i=2}^{n-1} Z_i^m, \sum_{i=1}^{n-1} Z_i^m\} \tag{2.1}$$

We have benefited from the memoryless property of the exponential distribution and recursively computed the probability that there are j, $j = 1, 2, ..., n-1$, customers in the system upon the nth arrival. We then have used these probabilities to find the expected waiting times. The calculation is tedious and there is no closed form formula for the expected waiting times. Our proposed method to compute the expected waiting times for $S(1, N-1)/(SM, SM')/1$ is presented in the next section.

2.2.1 Expected waiting time calculation

Consider an $S(1, N-1)/(SM, SM')/1$ queue with constant job allowance of x and exponential service rates of μ and μ' for fast and regular customers respectively. Let t_i denote the i-th appointment time in the schedule, i.e. $t_i = (i-1)x$ where $i = 1, 2, ..., N$. Denote the probability that the i-th arrival visits j customers in the system upon his/her arrival by $\mathbb{P}\{N(t_i) =$

$j\}$, the expected waiting time of this arrival can be calculated as follows.

$\mathbb{E}W_i^m =$

$$\begin{cases} \sum_{j=1}^{i-1} \mathbb{P}\left\{N\left(t_i\right)=j\right\} \times \left(\frac{j}{\mu'}\right) & ;i \leq m \\ \sum_{j=1}^{i-1} \mathbb{P}\left\{N\left(t_i\right)=j\right\} \times \left(\frac{1}{\mu}+\frac{j-1}{\mu'}\right) & ;i=m+1 \\ \sum_{j=1}^{i-m-1} \mathbb{P}\left\{N\left(t_i\right)=j\right\} \times \left(\frac{j}{\mu'}\right) + \sum_{j=i-m}^{i-1} \mathbb{P}\left\{N\left(t_i\right)=j\right\} \times \left(\frac{1}{\mu}+\frac{j-1}{\mu'}\right) & ;i>m+1 \end{cases} \quad (2.2)$$

The visiting probabilities $\mathbb{P}\{N(t_i)=j\}$, $i=1,2,...,N$ and $j=1,2,...,i-1$, can be recursively computed by the following expressions.

$$\mathbb{P}\{N(t_i)=0\} = \sum_{k=1}^{i-1} \mathbb{P}\{N(t_{i-1})=k-1\} \times \mathbb{P}\{ND(t_{i-1})=k\} \quad (2.3)$$

and for $j=1,2,...,i-1$,

$$\mathbb{P}\{N(t_i)=j\} = \sum_{k=0}^{i-j-1} \mathbb{P}\{N(t_{i-1})=j+k-1\} \times \mathbb{P}\{ND(t_{i-1})=k\} \quad (2.4)$$

where $\mathbb{P}\{N(t_1)=0\}=1$ and $ND(t_i)$, $i=1,2,...,N-1$ denotes the number of customer departures in time period (t_i, t_{i+1}).

When the fast customer does not enter the system yet, we only have identical customers with exponential service durations with mean of $\frac{1}{\mu'}$ in the system, and therefore the departure process as long as the server is continuously busy is a Poisson process with rate μ'. In this case, when there are k customers in the system then the probability that all of them

will depart in a time period with the length of x is

$$1 - \sum_{q=0}^{k-1} \frac{(\mu' x)^q}{q!} e^{-\mu' x}$$

which is indeed the probability that x units of time is sufficient for k or more departures. Hence, from (2.3), we obtain

$$\mathbb{P}\{N(t_i) = 0\} = \sum_{k=1}^{i-1} \left(\mathbb{P}\left\{N\left(t_{i-1}\right) = k-1\right\} \times \left[1 - \sum_{q=0}^{k-1} \frac{(\mu' x)^q}{q!} e^{-\mu' x} \right] \right) \tag{2.5}$$

whenever $i \leq m$.

Also, in (2.4), where $i \leq m$, we can replace $\mathbb{P}\{ND(t_{i-1}) = k\}$ by

$$\frac{(\mu' x)^q}{q!} e^{-\mu' x},$$

which is the probability that there are exactly k departures in time period (t_{i-1}, t_i), to obtain the following equation.

$$\mathbb{P}\{N\left(t_i\right) = j\} = \sum_{k=0}^{i-j-1} \left(\mathbb{P}\{N\left(t_{i-1}\right) = j + k - 1\} \times \frac{(\mu' x)^k}{k!} e^{-\mu' x} \right) \tag{2.6}$$

where j is a positive integer less than i. Similar expressions could be found in the waiting time calculation method proposed for identical customers with exponential service times by Pegden and Rosenshine (1990).

For the customers scheduled after the fast customer, i.e. $i > m$, the calculation of $\mathbb{P}\{N(t_i) = j\}$, $j = 0, 1, ..., i-1$ is very complicated. We obtain more than ten different formulations for different situations that may happen in the time period (t_{i-1}, t_i). The detailed formulations are presented in Appendix B.2.

The complexity of the waiting time formulas makes it almost impossible to obtain any analytical result about the structure of the optimal sequence. It motivates us to apply *stochastic ordering* approach to tackle the problem.

We present some numerical results in the next section.

2.2.2 Numerical results

We have applied the calculation method proposed in the previous section to provide numerical results investigating whether SEPT/SV is really optimal. Table 2.1 shows the expected waiting time of three regular customers with exponential service rate of 1 and one fast customer with exponential service rate of 10 where the job allowance is $x = 1.5$.

Table 2.1: Expected waiting times, $S(1,3)/(SM, SM')/1, \mu = 10, \mu' = 1, x = 1.5$

	$\mathbb{E}W_1^m$	$\mathbb{E}W_2^m$	$\mathbb{E}W_3^m$	$\mathbb{E}W_4^m$	$\mathbb{E}W^m$	% gap
$m = 1$	0	0	0.2231	0.3476	0.5707	7.2
$m = 2$	0	0.2231	0.0553	0.254	**0.5324**	0
$m = 3$	0	0.2231	0.3476	0.1033	0.674	26.6
$m = 4$	0	0.2231	0.3476	0.4295	1.0003	87.9

As can be seen, surprisingly, it is optimal to schedule the fast customer in the second slot. Similar examples can be found for any $N > 3$. This

result implies that the SEPT/SV is not necessarily optimal for $N > 3$.

For $N = 4$, a sufficient condition which guarantees that SEPT/SV is not optimal is the following. Again, without loss of generality the service rate for the regular customers is assumed to be one and for the fast customer is $\mu > 1$.

Lemma 2.2. *For $S(1,3)/(SM, SM')/1$ with $\mu > 1$ and $\mu' = 1$, if $x > 0.80647$ and*

$$\mu > \frac{1}{\frac{x}{x+1} - e^{-x}} \tag{2.7}$$

then $\mathbb{E}W^1(x) > \mathbb{E}W^2(x)$.

Proof. Using the method presented in Appendix B.2, we have

$$\begin{aligned}
\mathbb{E}W^1(x) &= 2e^{-x} + (x+1)e^{-2x} + \frac{1}{\mu(\mu-1)^2}e^{-3\mu x} + \frac{1}{\mu(1-\mu)}e^{-2\mu x} \\
&+ \frac{1}{\mu-1}e^{-(\mu+1)x} + \frac{x+1}{\mu-1}e^{-(\mu+2)x} + \frac{1}{\mu}e^{-\mu x} - \frac{1}{(\mu-1)^2}e^{-(2\mu+1)x}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}W^2(x) &= 2e^{-x} + \frac{\mu}{\mu-1}e^{-2x} + \frac{\mu(x+1)}{\mu-1}e^{-3x} + \frac{1}{\mu(1-\mu)}e^{-2\mu x} \\
&- \frac{\mu}{(\mu-1)^2}e^{-(\mu+2)x} + \frac{1}{\mu}e^{-\mu x} + \frac{1}{(\mu-1)^2}e^{-(2\mu+1)x}
\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}W^1(x) - \mathbb{E}W^2(x) &= \frac{\mu x - x - 1}{\mu - 1}e^{-2x} + \frac{1}{\mu(\mu-1)^2}e^{-3\mu x} + \frac{\mu(x+1)}{1-\mu}e^{-3x} \\ &+ \frac{1}{\mu - 1}e^{-(\mu+1)x} + \frac{(\mu-1)x + 2\mu - 1}{(\mu-1)^2}e^{-(\mu+2)x} \\ &- \frac{2}{(\mu-1)^2}e^{-(2\mu+1)x}\end{aligned}$$

Since $\mu > 1$, then

$$\frac{1}{\mu(\mu-1)^2}e^{-3\mu x} + \frac{1}{\mu-1}e^{-(\mu+1)x} > 0 \tag{2.8}$$

Also, it follows from (2.7) that

$$\mu > \frac{(x+1)e^x}{xe^x - x - 1}$$

For $x > 0.80647$, we have $xe^x - x - 1 > 0$ and thus

$$\mu(xe^x - x - 1) - (x+1)e^x > 0$$

Dividing both sides by $(\mu - 1)e^{3x}$ gives

$$\frac{\mu x - x - 1}{\mu - 1}e^{-2x} + \frac{\mu(x+1)}{1-\mu}e^{-3x} > 0 \tag{2.9}$$

Moreover, for $x > 0.80647$, (2.7) implies that $\mu > \frac{x+1}{x}$. Then we have

$$\frac{(\mu-1)x+2\mu-1}{(\mu-1)^2}e^{-(\mu+2)x}-\frac{2}{(\mu-1)^2}e^{-(2\mu+1)x}>0 \qquad (2.10)$$

The desired inequality can be obtained from (2.8), (2.9) and (2.10). □

From the above lemma, we can draw some interesting insights. For $x > 0.80647$, the function $(\frac{x}{x+1}-e^{-x})^{-1}$ is decreasing and convex. Hence, when x and μ are relatively large, (2.7) is satisfied. Also, for a larger x, the condition (2.7) can be satisfied by a smaller μ. For example, if $x = 1$, then μ should be at least 7.57 to satisfy (2.7), i.e. the fast customer should be 7.57 times faster than a regular customer, while for $x = 2$, μ should be at least 1.89. Overall, we can conjecture that *when the job allowance x is large enough and the special customer is much faster than regular customers, then SEPT/SV is not optimal.*

Table 2.2 shows the total expected waiting time of 9 regular and 1 fast customers with service rates of 1 and 10 respectively for various x values.

An interesting phenomenon observed in this table is that the optimal position for the fast customer is always within the first half of the sequence. More exactly, it has been observed that when x is relatively small, as we would expect, SEPT and SV are optimal, and when x increases, the optimal position for the fast customer goes later but it becomes fixed after some x and never goes to the second half. A similar behaviour has been observed for a wide range of test problems. A similar behaviour also has been

Table 2.2: Total expected waiting times, $S(1,9)/(SM,SM')/1$, $\mu = 10$, $\mu' = 1$, various x values.

x	$\mathbb{E}W^1$	$\mathbb{E}W^2$	$\mathbb{E}W^3$	$\mathbb{E}W^4$	$\mathbb{E}W^5$	$\mathbb{E}W^6$	$\mathbb{E}W^7$	$\mathbb{E}W^8$	$\mathbb{E}W^9$	$\mathbb{E}W^{10}$
0	**36.9**	37.8	38.7	39.6	40.5	41.4	42.3	43.2	44.1	45
0.1	**32.76328**	33.38563	34.25423	35.14991	36.04926	36.94916	37.84915	38.74914	39.64914	40.54914
0.2	**29.09958**	29.23819	29.97428	30.83221	31.72054	32.6172	33.51624	34.41597	35.3159	36.21589
0.3	25.70796	**25.41655**	25.95087	26.71941	27.56871	28.44858	29.34056	30.23746	31.13635	32.03603
0.4	22.53396	**21.93663**	22.25454	22.8935	23.66876	24.50754	25.37744	26.26302	27.15659	28.05425
0.5	19.58328	**18.80518**	18.93017	19.42278	20.09595	20.86653	21.69261	22.55137	23.42982	24.32049
0.6	16.87891	16.02393	**15.99846**	16.34953	16.90652	17.58745	18.34741	19.15983	20.00826	20.88281
0.8	12.28261	11.47972	**11.29014**	11.42202	11.75719	12.23468	12.81879	13.48719	14.22623	15.03058
1	8.78016	8.15318	**7.93529**	7.94929	8.12337	8.4212	8.82284	9.31804	9.9048	10.59454
1.2	6.24234	5.80067	5.61894	**5.5888**	5.66837	5.83876	6.09252	6.43038	6.86272	7.42119
1.6	3.2238	3.03405	2.94851	**2.91932**	2.93046	2.97757	3.06333	3.19777	3.40358	3.73909
2	1.76333	1.68516	1.65286	**1.6407**	1.64098	1.65238	1.67786	1.72604	1.81794	2.01589
2.2	1.33377	1.28304	1.2638	1.25674	**1.25636**	1.2618	1.27531	1.3037	1.365	1.51839
3	0.48767	0.47767	0.47535	0.47474	**0.47465**	0.4749	0.47587	0.47921	0.49182	0.55095
4	0.15964	0.15814	0.15797	0.15795	**0.15795**	0.15795	0.15798	0.15822	0.16009	0.17987
8	0.00269	0.00269	0.00269	0.00269	**0.00269**	0.00269	0.00269	0.00269	0.00269	0.00303

The minimum cell in each row is highlighted. It is selected by MATLAB and here up to the fifth decimal digit is shown.

observed for deterministic appointment sequencing which is studied in Appendix C. We call this pattern First Half Rule (FHR) and will analytically obtain it in Theorem 3.1.

Table 2.3 indicates the expected waiting time of each arrival in various sequences for $S(1,9)/(SM,SM')/1, \mu = 10, \mu' = 1, x = 1.5$.

The results suggest that the expected waiting time of the n-th arrival decreases when the position of the fast customer moves from the first slot to the $(n-1)$-th slot. Moreover, it is maximized and constant as long as the fast customer is not scheduled behind him/her. Later in Corollary 3.1, this property will be analytically obtained. An important insight that can be drawn here is that *in order to minimize the waiting time of a customer, the fastest customer must be scheduled right before him/her.*

The next section explains why scheduling a slower customer first could decrease the total customer waiting time.

2.3 Why SEPT/SV may not be optimal?

As mentioned before, it seems strange to put a customer with a higher mean and variance of service duration first in the schedule to minimize the customers' waiting times. A reasonable explanation to this observation is as follows.

Suppose there are N identical regular customers scheduled at equally spaced appointment times with width of x and we want to exchange one

Table 2.3: Expected waiting times, $S(1,9)/(SM,SM')/1$, $\mu = 10$, $\mu' = 1$, $x = 1.5$

	$\mathbb{E}W_1^m$	$\mathbb{E}W_2^m$	$\mathbb{E}W_3^m$	$\mathbb{E}W_4^m$	$\mathbb{E}W_5^m$	$\mathbb{E}W_6^m$	$\mathbb{E}W_7^m$	$\mathbb{E}W_8^m$	$\mathbb{E}W_9^m$	$\mathbb{E}W_{10}^m$
$m = 1$	0	**0**	0.22313	0.3476	0.42953	0.48778	0.53112	0.56437	0.59043	0.6112
$m = 2$	0	0.22313	**0.05532**	0.25399	0.36791	0.44397	0.49852	0.53936	0.57083	0.59558
$m = 3$	0	0.22313	0.3476	**0.10332**	0.28386	0.38865	0.45919	0.5101	0.54838	0.57798
$m = 4$	0	0.22313	0.3476	0.42953	**0.14161**	0.30929	0.40696	0.47296	0.52075	0.55678
$m = 5$	0	0.22313	0.3476	0.42953	0.48778	**0.17199**	0.33036	0.42254	0.4849	0.53011
$m = 6$	0	0.22313	0.3476	0.42953	0.48778	0.53112	**0.19627**	0.34774	0.43566	0.49511
$m = 7$	0	0.22313	0.3476	0.42953	0.48778	0.53112	0.56437	**0.21587**	0.3621	0.44669
$m = 8$	0	0.22313	0.3476	0.42953	0.48778	0.53112	0.56437	0.59043	**0.23183**	0.37402
$m = 9$	0	0.22313	0.3476	0.42953	0.48778	0.53112	0.56437	0.59043	0.6112	**0.24494**
$m = 10$	0	0.22313	0.3476	0.42953	0.48778	0.53112	0.56437	0.59043	0.6112	0.62797

The minimum cell of each column is highlighted.

of them by a faster customer. If we exchange the kth customer, then the service time of the faster customer does not affect the waiting time of the customers who are scheduled before him (i.e. the first k customers) while it can reduce the waiting time of all customers scheduled after him through the waiting time of the $(k+1)$st customer, as the waiting time of the $(k+1)$st customer may affect the waiting time of the $(k+2)$nd customer which may affect the waiting time of the $(k+3)$th customer and so on. That is, if we exchange the last arrival, the service time of the faster customer does not affect the waiting time of all customers at all. From this perspective, we want the faster customer to be in the earlier slot as possible, because the time saving will affect the waiting times of more customers.

On the other hand, if we put the faster customer in the first slot, there would be another concern. Note that, the $(k+1)$st customer's waiting time depends on two factors: The service time of the kth customer and the waiting time of the kth customer. The effects of these two factors are intertwined. If the kth customer's waiting time is too short, the effect of the kth customer's service time on the $(k+1)$st customer's waiting time will reduce. For example, if x is 10 minutes, the service time of the kth customer is 8 minutes and if there is no wait for the kth customer, 1 minute's saving on the kth customer's service time will have no impact on the $(k+1)$st customer's waiting time. However, if the kth customer waits for 5 minutes, 1 minute's saving on his service time will reduce the $(k+1)$st customer's

waiting time by 1 minute. We call this second concern, the *voucher effect*. A rational customer, who has a \$100 voucher which cannot be partially used, will try to find an item which is around \$100, and not much cheaper. The potential saving can be provided by the faster customer can be viewed as a voucher which can help us to catch up when we are running behind the schedule after serving some customers. To make the best use of this potential saving, we need to possibly accumulate some delays.

In short, we hope the fast customer should be put in the early slot, because we hope the reduction of the customer's service time can affect more customer's waiting times. But on the other hand, we hope the fast customer can wait for sufficient time so that the reduction of his service time can generate an effect.

All these are due to the appointment intervals (job allowances). If there is no allowance (i.e., $x = 0$), every minute's saving will definitely reduce the waiting times of all the subsequent customers by one minute. Thus, there is no need to accumulate delay and the save, and so SEPT should be the best rule. On the other hand, even when the job allowance goes to infinity, it should be always impossible to put the faster customer at the end. Our numerical study shows that the faster customer should be scheduled latest at the middle slot, and thus motivates us to propose the First Half Rule (FHR) policy which is presented in the next section.

Chapter 3

Sequencing customers of two classes with stochastically ordered excess service times

In this section, we study a more general case where the appointment times are not necessarily equally spaced and the service durations are not necessarily exponentially distributed. We consider an $S(M, N-M)/(F, G)/1$ queueing system. As explained in §1.3, there are M fast and $N-M$ regular customers to be sequenced where the excess service time of a fast customer is stochastically smaller than the one of a regular customer in likelihood ratio order sense.

The queueing system investigated in the previous section is a special case of $S(M, N-M)/(F, G)/1$. In fact, under constant job allowance as-

sumption, the likelihood ratio order assumption of the excess service times is equivalent to the same assumption on the service times themselves. It is because the likelihood ratio order is preserved under adding or subtracting a constant (Shanthikumar and Yao, 1986). Therefore, any appointment system with fixed job allowance and two classes of customers from the same service time distribution family which has Monotone Likelihood Ratio Property (MLRP) can be considered as an $S(M, N-M)/(F, G)/1$ system. It turns out that most of the known families of distributions have monotone likelihood ratio in some statistics (Bartoszynski and Niewiadomska-Bugaj, 2008, p. 485). Exponential, Beta, Weibull, Normal with known variance, Uniform, Gamma, Poisson, Geometric, Binomial and Negative Binomial are some instances of MLRP distributions. It is worthwhile to note that scheduling patients in constant appointment intervals is a common practice in healthcare (Hall, 2012).

For the rest of the chapter, we first analytically establish FHR in §3.1 and then in §3.2 extend the application of FHR to the case with a late server. Finally, we extend the results to multiple fast customers in §3.3, and propose an effective FHR-based appointment sequencing heuristic algorithm in §3.4.

3.1 $S(1, N-1)/(F, G)/1$: The First Half Rule

Let Z denote the *excess* service time of the fast customer and Z_i ($i = 1, .., N-1$) denote the *excess* service time of the i-th regular customer in the appointment sequence. We assume that Z_i's are i.i.d. random variables which are also independent of Z. We also assume that the excess service time Z of the fast customer is stochastically smaller than the one of the regular customers in likelihood ratio order, i.e., $Z \leq_{lr} Z_i$ for all $i = 1, ..., N-1$.

Define *sequence* m as an arrival sequence in which the fast customer is scheduled in the m-th appointment slot. Let W_n^m be the waiting time of the n-th customer in sequence m. W_n^m can be represented in the following *Max-Flow* form (see (2.1)):

$$W_n^m = \begin{cases} 0 & ; \text{when } n = 1 \\ \max\{0, Z\} & ; \text{when } n = 2 \text{ and } m = 1 \\ \max\{0, Z_{n-1}, ..., \sum_{i=1}^{n-1} Z_i\} & ; \text{when } 1 < n < m+1 \\ \max\{0, Z, Z + Z_{n-2}, ..., Z + \sum_{i=1}^{n-2} Z_i\} & ; \text{when } n = m+1 > 2 \\ \max\{0, Z_{n-2}, ..., \sum_{i=m}^{n-2} Z_i, Z + \sum_{i=m}^{n-2} Z_i, ..., Z + \sum_{i=1}^{n-2} Z_i\} & ; \text{when } n > m+1 \end{cases} \quad (3.1)$$

The following lemma shows that the waiting time of the n-th arrival is stochastically minimized when the fast customer is scheduled at the $(n-1)$-

th slot.

Lemma 3.1. *The waiting time of the n-th customer has the following properties.*

(a) $W_n^m \geq_{st} W_n^{m+1}$ *for all* $m+1 < n$

(b) $W_n^1 \leq_{st} W_n^n$

(c) $W_n^n =_{st} W_n^{n+1} =_{st} ... =_{st} W_n^N$

(d) $W_n^m \leq_{st} W_{n+1}^m$ *for all* $m > n$

Proof. Comparing *sequence* m with $m+1$, as long as the fast customer is scheduled before the nth arrival (i.e. for $m < n-1$), based on the *Max-Flow* representation (2.1), we have

$$W_n^{m+1} = max\{0, Z_{n-2}, ..., \sum_{i=m+1}^{n-2} Z_i, Z+\sum_{i=m+1}^{n-2} Z_i, Z+\sum_{i=m}^{n-2} Z_i, Z+\sum_{i=m-1}^{n-2} Z_i, ..., Z+\sum_{i=1}^{n-2} Z_i\}$$

$$W_n^m = max\{0, Z_{n-2}, ..., \sum_{i=m+1}^{n-2} Z_i, \sum_{i=m}^{n-2} Z_i, Z+\sum_{i=m}^{n-2} Z_i, Z+\sum_{i=m-1}^{n-2} Z_i, ..., Z+\sum_{i=1}^{n-2} Z_i\}$$

According to Theorem A.5, since Z_i's are i.i.d and independent of Z and $Z \leq_{lr} Z_m$, then a set of partial sums for the sequence $\{0, Z_{n-2}, ..., Z_{m+1}, Z, Z_m, ..., Z_1\}$ is stochastically smaller than a set of partial sums for $\{0, Z_{n-2}, ..., Z_m, Z, Z_{m-1}, ..., Z_1\}$. That is

$$\begin{aligned}(Z_{n-2}, Z_{n-2} \quad + \quad & Z_{n-3}, ..., \sum_{i=m+1}^{n-2} Z_i, Z + \sum_{i=m+1}^{n-2} Z_i, ..., Z + \sum_{i=1}^{n-2} Z_i) \\ & \leq_{st} (Z_{n-2}, Z_{n-2} + Z_{n-3}, ..., \sum_{i=m}^{n-2} Z_i, Z + \sum_{i=m}^{n-2} Z_i, ..., Z + \sum_{i=1}^{n-2} Z_i).\end{aligned}$$

Moreover the function $\max\{\cdot\}$ is non-decreasing. Hence, for $m < n-1$, we obtain $W_n^{m+1} \leq_{st} W_n^m$ by Theorem A.1.

Parts (b), (c) and (d) hold even under the weaker assumption of $Z \leq_{st} Z_i$, $i = 1, \ldots, N-1$. For Part (b), we have

$$\begin{aligned}W_n^1 &= \max\{0, Z_{n-2}, Z_{n-2} + Z_{n-3}, ..., \sum_{i=1}^{n-2} Z_i, Z + \sum_{i=1}^{n-2} Z_i\} \\ W_n^n &= \max\{0, Z_{n-1}, Z_{n-1} + Z_{n-2}, ..., \sum_{i=2}^{n-1} Z_i, \quad \sum_{i=1}^{n-1} Z_i \ \}\end{aligned}$$

As Z_i's are identically and independently distributed, it can be easily obtained that $W_n^1 \leq_{st} W_n^n$ from Theorems A.3 and A.1.

The proof of Part (c) is trivial. It is intuitive that the waiting time of the n-th arrival is independent of the position of the fast customer as long as the fast customer is scheduled after him/her. In fact, for any $m \geq n$, the waiting time of the n-th arrival is constant and independent of m with the following expression.

$$W_n^m = max\{0, Z_{n-1}, Z_{n-1} + Z_{n-2}, ..., \sum_{i=2}^{n-1} Z_i, \quad \sum_{i=1}^{n-1} Z_i \ \}$$

Part (d) can be shown as follows.

$$W_1^m = 0 \leq_{st} \max\{0, Z_1\} = W_2^m$$

As Z_i's are i.i.d, we have $Z_{n-1} \leq_{st} Z_n$. Also, W_n^m is independent of Z_n. Therefore, by Theorem A.1 (b), if $W_{n-1}^m \leq_{st} W_n^m$ and $1 < n < m$, then $W_{n-1}^m + Z_{n-1} \leq_{st} W_n^m + Z_n$. It follows

$$W_n^m = \max\{0, W_{n-1}^m + Z_{n-1}\} \leq_{st} \max\{0, W_n^m + Z_n\} = W_{n+1}^m$$

By induction Part(d) is proved. Indeed, W_n^m is stochastically increasing and concave in n where $n < m$. An appointment-based queue with constant job allowance and homogeneous arrivals is investigated in Appendix D. It is shown in Lemma D.1 and Lemma D.3 that the waiting time of the n-th arrival in such a system is stochastically increasing and concave in n.

□

We can immediately obtain Corollary 3.1.

Corollary 3.1. *$\mathbb{E}W_n^m$ is decreasing in m for $m \leq n-1$, and constant for $m > n-1$. Specifically,*

$$\mathbb{E}W_n^{n-1} \leq \mathbb{E}W_n^{n-2} \leq ... \leq \mathbb{E}W_n^1 \leq \mathbb{E}W_n^n = \mathbb{E}W_n^{n+1} = ... = \mathbb{E}W_n^N$$

The following property of the customer waiting times plays a critical role in establishing the First Half Rule in Theorem 3.1.

Lemma 3.2. *For $n > m+1$,*

$$\mathbb{E}W_n^m - \mathbb{E}W_n^{m+1} \geq \mathbb{E}W_{n+1}^{m+1} - \mathbb{E}W_{n+1}^{m+2}$$

Proof. By the *Max-Flow* forms of the waiting times, for $n > m+2$

$$\begin{aligned}
W_n^m &= \max\{0, Z_{n-2}, ..., \sum_{i=m+1}^{n-2} Z_i, \quad \sum_{i=m}^{n-2} Z_i \quad , \; Z + \sum_{i=m}^{n-2} Z_i \quad , ..., Z + \sum_{i=1}^{n-2} Z_i\} \\
W_n^{m+1} &= \max\{0, Z_{n-2}, ..., \sum_{i=m+1}^{n-2} Z_i, Z + \sum_{i=m+1}^{n-2} Z_i, \; Z + \sum_{i=m}^{n-2} Z_i \quad , ..., Z + \sum_{i=1}^{n-2} Z_i\} \\
W_{n+1}^{m+1} &= \max\{0, Z_{n-1}, ..., \sum_{i=m+2}^{n-1} Z_i, \quad \sum_{i=m+1}^{n-1} Z_i \quad , Z + \sum_{i=m+1}^{n-1} Z_i, ..., Z + \sum_{i=1}^{n-1} Z_i\} \\
W_{n+1}^{m+2} &= \max\{0, Z_{n-1}, ..., \sum_{i=m+2}^{n-1} Z_i, Z + \sum_{i=m+2}^{n-1} Z_i, Z + \sum_{i=m+1}^{n-1} Z_i, ..., Z + \sum_{i=1}^{n-1} Z_i\}
\end{aligned}$$

We remove the term $Z + \sum_{i=1}^{n-1} Z_i$ from both W_{n+1}^{m+1} and W_{n+1}^{m+2} to represent $\widehat{W}_{n+1}^{m+1}$ and $\widehat{W}_{n+1}^{m+2}$.

$$\widehat{W}_{n+1}^{m+1} = max\{0, Z_{n-1}, ..., \sum_{i=m+2}^{n-1} Z_i, \quad \sum_{i=m+1}^{n-1} Z_i \quad , Z + \sum_{i=m+1}^{n-1} Z_i, ..., Z + \sum_{i=2}^{n-1} Z_i\}$$

$$\widehat{W}_{n+1}^{m+2} = max\{0, Z_{n-1}, ..., \sum_{i=m+2}^{n-1} Z_i, Z + \sum_{i=m+2}^{n-1} Z_i, Z + \sum_{i=m+1}^{n-1} Z_i, ..., Z + \sum_{i=2}^{n-1} Z_i\}$$

Replacing the sequence $Z_1, Z_2, ...Z_{n-2}$ by $Z_2, Z_3, ...Z_{n-1}$ in W_n^m and W_n^{m+1}, we reach exactly the same expressions as $\widehat{W}_{n+1}^{m+1}$ and $\widehat{W}_{n+1}^{m+2}$ re-

spectively. In addition, Z_i's are i.i.d. random variables. Thus,

$$W_n^m - W_n^{m+1} =_{st} \widehat{W}_{n+1}^{m+1} - \widehat{W}_{n+1}^{m+2}$$

and then

$$\mathbb{E}W_n^m - \mathbb{E}W_n^{m+1} = \mathbb{E}\widehat{W}_{n+1}^{m+1} - \mathbb{E}\widehat{W}_{n+1}^{m+2}$$

Now, in order to complete this proof we just need to show

$$\mathbb{E}\widehat{W}_{n+1}^{m+1} - \mathbb{E}\widehat{W}_{n+1}^{m+2} \geq \mathbb{E}W_{n+1}^{m+1} - \mathbb{E}W_{n+1}^{m+2} \tag{3.2}$$

which can be obtained from

$$W_{n+1}^{m+2} - \widehat{W}_{n+1}^{m+2} \geq_{st} W_{n+1}^{m+1} - \widehat{W}_{n+1}^{m+1} \tag{3.3}$$

We pull out $Z + \sum_{i=2}^{n-1} Z_i$ from $W_{n+1}^{m+1}, W_{n+1}^{m+2}, \widehat{W}_{n+1}^{m+1}$ and $\widehat{W}_{n+1}^{m+2}$ to obtain

$$W_{n+1}^{m+1} = Z + \sum_{i=2}^{n-1} Z_i - Y_{n+1}^{m+1}$$

$$W_{n+1}^{m+2} = Z + \sum_{i=2}^{n-1} Z_i - Y_{n+1}^{m+2}$$

$$\widehat{W}_{n+1}^{m+1} = Z + \sum_{i=2}^{n-1} Z_i - \widehat{Y}_{n+1}^{m+1}$$

$$\widehat{W}_{n+1}^{m+2} = Z + \sum_{i=2}^{n-1} Z_i - \widehat{Y}_{n+2}^{m+1}$$

where

$$Y_{n+1}^{m+1} = min\{Z+\sum_{i=2}^{n-1} Z_i, Z+\sum_{i=2}^{n-2} Z_i, ..., Z+\sum_{i=2}^{m+1} Z_i, Z+\sum_{i=2}^{m} Z_i, \sum_{i=2}^{m} Z_i, \sum_{i=2}^{m-1} Z_i, ..., Z_2, 0, -Z_1\}$$

$$Y_{n+1}^{m+2} = min\{Z+\sum_{i=2}^{n-1} Z_i, Z+\sum_{i=2}^{n-2} Z_i, ..., Z+\sum_{i=2}^{m+1} Z_i, \sum_{i=2}^{m+1} Z_i, \sum_{i=2}^{m} Z_i, \sum_{i=2}^{m-1} Z_i, ..., Z_2, 0, -Z_1\}$$

$$\widehat{Y}_{n+1}^{m+1} = min\{Z+\sum_{i=2}^{n-1} Z_i, Z+\sum_{i=2}^{n-2} Z_i, ..., Z+\sum_{i=2}^{m+1} Z_i, Z+\sum_{i=2}^{m} Z_i, \sum_{i=2}^{m} Z_i, \sum_{i=2}^{m-1} Z_i, ..., Z_2, 0\}$$

$$\widehat{Y}_{n+1}^{m+2} = min\{Z+\sum_{i=2}^{n-1} Z_i, Z+\sum_{i=2}^{n-2} Z_i, ..., Z+\sum_{i=2}^{m+1} Z_i, \sum_{i=2}^{m+1} Z_i, \sum_{i=2}^{m} Z_i, \sum_{i=2}^{m-1} Z_i, ..., Z_2, 0\}$$

Thus, (3.3) is equivalent to

$$\widehat{Y}_{n+1}^{m+2} - Y_{n+1}^{m+2} \geq_{st} \widehat{Y}_{n+1}^{m+1} - Y_{n+1}^{m+1} \tag{3.4}$$

According to Theorem A.5,

$$\left(Z_2, ..., \sum_{i=2}^{m+1} Z_i, Z + \sum_{i=2}^{m+1} Z_i, ..., Z + \sum_{i=2}^{n-1} Z_i\right) \geq_{st} \left(Z_2, ..., \sum_{i=2}^{m} Z_i, Z + \sum_{i=2}^{m} Z_i, ..., Z + \sum_{i=2}^{n-1} Z_i\right)$$

Furthermore, $\phi(x) = min\{0, x_1, x_2, ..., x_n\}$ is a non-decreasing function in $\mathbf{x} = (x_1, x_2, ..., x_n)$. Therefore, based on Theorem 1.A.3(a),

$$\widehat{Y}_{n+1}^{m+2} \geq_{st} \widehat{Y}_{n+1}^{m+1}$$

For any given realization of $Z_1 = \theta$, since $\widehat{Y}_{n+1}^{m+1}$ and $\widehat{Y}_{n+1}^{m+2}$ are independent of Z_1, we still have $\widehat{Y}_{n+1}^{m+2} \geq_{st} \widehat{Y}_{n+1}^{m+1}$. Moreover, $x - min\{x, -\theta\}$ is a

non-decreasing function of x. Therefore, given $Z_1 = \theta$,

$$\widehat{Y}_{n+1}^{m+2} - min\{\widehat{Y}_{n+1}^{m+2}, -\theta\} \geq_{st} \widehat{Y}_{n+1}^{m+1} - min\{\widehat{Y}_{n+1}^{m+1}, -\theta\}$$

which is equivalent to (3.4). Finally, according to Theorem 1.A.3(d), since (3.4) holds for any θ belonging to the support of Z_1, it holds in general and the proof is completed for $n > m + 2$. For $n = m + 2$, we can update W_n^{m+1}, W_{n+1}^{m+2}, $\widehat{W}_{n+1}^{m+2}$, Y_{n+1}^{m+2}, and $\widehat{Y}_{n+1}^{m+2}$ to the followings and develop a similar proof.

$$\begin{aligned}
W_n^{m+1} &= \max\{0, Z, Z + Z_{n-2}, ..., Z + \sum_{i=1}^{n-2} Z_i\} \\
W_{n+1}^{m+2} &= \max\{0, Z, Z + Z_{n-1}, ..., Z + \sum_{i=1}^{n-1} Z_i\} \\
\widehat{W}_{n+1}^{m+2} &= \max\{0, Z, Z + Z_{n-1}, ..., Z + \sum_{i=2}^{n-1} Z_i\} \\
Y_{n+1}^{m+2} &= min\{Z + \sum_{i=2}^{n-1} Z_i, \sum_{i=2}^{n-1} Z_i, \sum_{i=2}^{n-2} Z_i, ..., Z_2, 0, -Z_1\} \\
\widehat{Y}_{n+1}^{m+2} &= min\{Z + \sum_{i=2}^{n-1} Z_i, \sum_{i=2}^{n-1} Z_i, \sum_{i=2}^{n-2} Z_i, ..., Z_2, 0\}
\end{aligned}$$

□

We can quickly obtain the following corollary.

Corollary 3.2. *Given $n > m + 1$,*

$$\mathbb{E}W_n^m - \mathbb{E}W_n^{m+1} \geq \mathbb{E}W_{n+k}^{m+k} - \mathbb{E}W_{n+k}^{m+k+1}$$

for any non-negative integer $k \leq N - n$.

Denote the total waiting time of all customers in sequence m as W^m. The following theorem shows that, to minimize the total waiting time, the fast customer should be assigned to the first half of the sequence.

Theorem 3.1. First Half Rule (FHR): *The optimal slot for the fast customer is within the first half of the sequence* (*including* $\frac{N}{2}$ *for an even* N *and* $\frac{N+1}{2}$ *for an odd* N). *Specifically,*

$$\mathbb{E}W^{\lceil \frac{N}{2} \rceil} \leq \mathbb{E}W^{\lceil \frac{N}{2} \rceil + 1} \leq ... \leq \mathbb{E}W^{N}.$$

Proof. We want to show that for any $m \geq \lceil \frac{N}{2} \rceil$,

$$\mathbb{E}W^{m+1} - \mathbb{E}W^{m} \geq 0$$

By Corollary 3.1, we have $\mathbb{E}W_n^m = \mathbb{E}W_n^{m+1}$ given $n \leq m$. Therefore,

$$\mathbb{E}W^{m+1} - \mathbb{E}W^{m} = \sum_{n=m+1}^{N} \left(\mathbb{E}W_n^{m+1} - \mathbb{E}W_n^{m} \right)$$

Thus, to complete the proof we just need to show

$$\mathbb{E}W_{m+1}^{m+1} - \mathbb{E}W_{m+1}^{m} \geq \sum_{n=m+2}^{N} \left(\mathbb{E}W_n^{m} - \mathbb{E}W_n^{m+1} \right)$$

We divide the left hand side to two parts as follows.

$$\mathbb{E}W_{m+1}^{m+1} - \mathbb{E}W_{m+1}^{m} = (\mathbb{E}W_{m+1}^{m+1} - \mathbb{E}W_{m+1}^{1}) + (\mathbb{E}W_{m+1}^{1} - \mathbb{E}W_{m+1}^{m})$$

By Corollary 3.1, $\mathbb{E}W_{m+1}^{m+1} - \mathbb{E}W_{m+1}^{1} \geq 0$. Therefore, the following inequality can complete the proof.

$$\mathbb{E}W_{m+1}^{1} - \mathbb{E}W_{m+1}^{m} \geq \sum_{n=m+2}^{N} \left(\mathbb{E}W_{n}^{m} - \mathbb{E}W_{n}^{m+1}\right) \tag{3.5}$$

Again we divide the left hand side to smaller parts.

$$\mathbb{E}W_{m+1}^{1} - \mathbb{E}W_{m+1}^{m} = \sum_{k=1}^{m-1} (\mathbb{E}W_{m+1}^{k} - \mathbb{E}W_{m+1}^{k+1})$$

Since $m \geq \lceil \frac{N}{2} \rceil$, for each term $(\mathbb{E}W_{n}^{m} - \mathbb{E}W_{n}^{m+1})$ of $\sum_{n=m+2}^{N} (\mathbb{E}W_{n}^{m} - \mathbb{E}W_{n}^{m+1})$ there is a corresponding term $(\mathbb{E}W_{m+1}^{2m-n+1} - \mathbb{E}W_{m+1}^{2m-n+2})$ in $\sum_{k=1}^{m-1} (\mathbb{E}W_{m+1}^{k} - \mathbb{E}W_{m+1}^{k+1})$ which is larger according to Corollary 3.2. Therefore, for $m \geq \lceil \frac{N}{2} \rceil$,

$$\sum_{k=1}^{m-1} (\mathbb{E}W_{m+1}^{k} - \mathbb{E}W_{m+1}^{k+1}) \geq \sum_{n=m+2}^{N} \left(\mathbb{E}W_{n}^{m} - \mathbb{E}W_{n}^{m+1}\right)$$

which completes the proof. □

To make it more clear, we give an example. Figure 3.1 shows the expected customer waiting times of various sequences of an $S(1,9)/(SM, SM')/1$

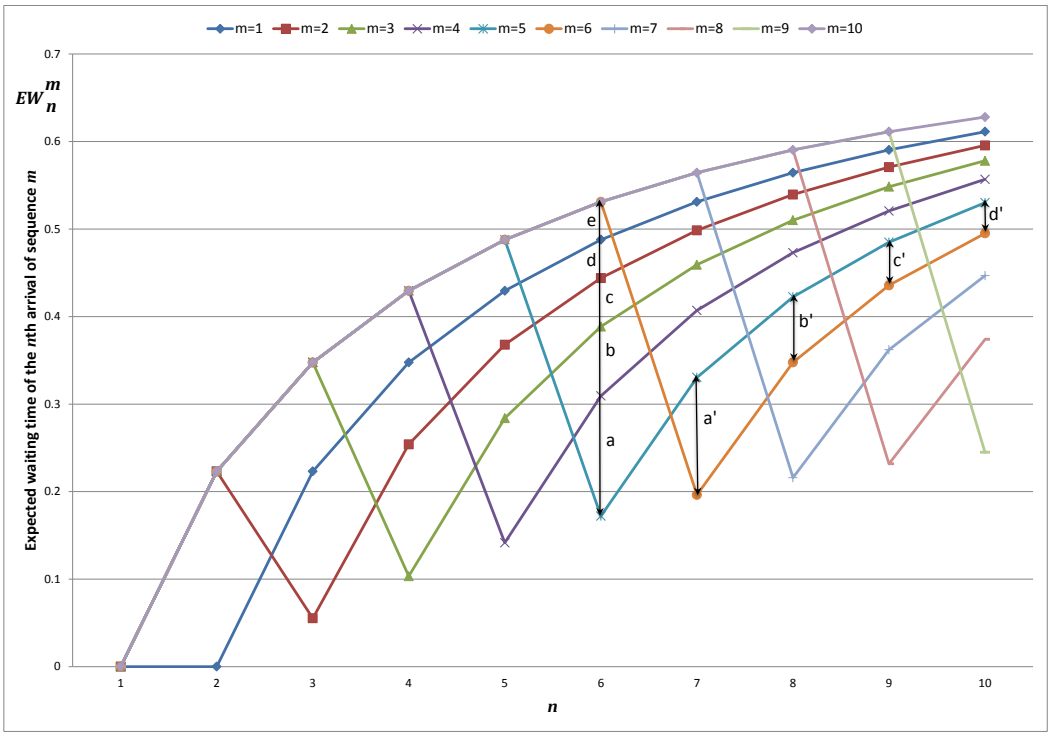

Figure 3.1: Comparison of sequences 5 and 6, $S(1,9)/(SM,SM')/1$,$\mu = 10, \mu' = 1, x = 1.5$

queue. Suppose we want to show

$$\mathbb{E}W^6 - \mathbb{E}W^5 \geq 0$$

According to Corollary 3.2, $a' \leq a$, $b' \leq b$, $c' \leq c$ and $d' \leq d$. Moreover, based on Corollary 3.1, $e \geq 0$. Therefore, we can conclude that

$$a + b + c + d + e \geq a' + b' + c' + d'$$

That is sequence 6 is not better than sequence 5 in terms of the total

customer waiting time.

In the next section, we consider the case in which the server could have some delay. For example, it could be related to the lateness of a doctor in a clinic or occupation of an operating room by the previous surgeon in a hospital. Another late server situation happens when a service provider (dentist, barber, consulter, etc.) stuck in traffic jam on the way of the service centre.

3.2 Late start of server in $S(1, N-1)/(F, G)/1$ queues

Let $W_n^m(\Theta)$ denote the waiting time of the n-th customer in *sequence* m given that the first customer has to wait for $\Theta \geq 0$ units of time to receive the service (i.e. the server is not available until Θ units of time after the first customer's arrival.) We assume that the server lateness Θ is random and independent of all the service times of the customers. Then

$W_n^m(\Theta) =$

$$\begin{cases} \Theta & ;n = 1 \\ \max\{0, Z + \Theta\} & ;n = 2 \text{ and } m = 1 \\ \max\{0, Z_{n-1}, ... \sum_{i=2}^{n-1} Z_i, \sum_{i=1}^{n-1} Z_i + \Theta\} & ;1 < n < m + 1 \\ \max\{0, Z, Z + Z_1 + \Theta\} & ;n = m + 1 = 3 \\ \max\{0, Z, Z + Z_{n-2}, ..., Z + \sum_{i=2}^{n-2} Z_i, Z + \sum_{i=1}^{n-2} Z_i + \Theta\} & ;n = m + 1 > 3 \\ \max\{0, Z_{n-2}, ..., \sum_{i=m}^{n-2} Z_i, Z + \sum_{i=m}^{n-2} Z_i, ..., Z + \sum_{i=1}^{n-2} Z_i + \Theta\} & ;n > m + 1 \end{cases} \tag{3.6}$$

Note that, given any realization θ of Θ, the waiting time $W_n^m(\theta)$ has similar properties to the ones in Lemma 3.1.

Lemma 3.3. *The waiting time of the n-th customer has the following properties:*

(a) $W_n^m(\Theta) \geq_{st} W_n^{m+1}(\Theta)$ *for all* $n > m + 1$

(b) $W_n^1(\Theta) \leq_{st} W_n^n(\Theta)$

(c) $W_n^n(\Theta) =_{st} W_n^{n+1}(\Theta) =_{st} ... =_{st} W_n^N(\Theta)$

(d) *Given* $n < m$, *if* $W_1^m(\Theta) \leq_{st} [\geq_{st}] W_2^m(\Theta)$, *then*

$$W_n^m(\Theta) \leq_{st} [\geq_{st}] W_{n+1}^m(\Theta).$$

(e) $W_n^m(\theta)$ *is stochastically increasing in* θ, *i.e.,* $W_n^m(\theta') \geq_{st} W_n^m(\theta)$, *whenever* $\theta' \geq \theta$.

Proof. For Part (a), if $m > 1$, then $W_n^m(\Theta)$ is the maximum of the set of partial sums of $\{0, Z_{n-2}, ..., Z_{m+1}, Z_m, Z, Z_{m-1}, ..., Z_1 + \Theta\}$ and $W_n^{m+1}(\theta)$ is the maximum of the set of partial sums of $\{0, Z_{n-2}, ..., Z_{m+1}, Z, Z_m, Z_{m-1}, ..., Z_1 + \Theta\}$. Since $Z \leq_{lr} Z_m$, $W_n^m(\Theta) \geq_{st} W_n^{m+1}(\Theta)$ by theorems A.5 and A.1.

For $m = 1$ and $n > 3$,

$$\begin{aligned} W_n^1(\Theta) &= \max\{0, Z_{n-2}, ..., \sum_{i=2}^{n-2} Z_i, \quad \sum_{i=1}^{n-2} Z_i \quad , Z + \sum_{i=1}^{n-2} Z_i + \Theta\} \\ W_n^2(\Theta) &= \max\{0, Z_{n-2}, ..., \sum_{i=2}^{n-2} Z_i, Z + \sum_{i=2}^{n-2} Z_i, Z + \sum_{i=1}^{n-2} Z_i + \Theta\} \end{aligned}$$

Thus,

$$\begin{aligned} W_n^1(\Theta) &= Z + \sum_{i=1}^{n-2} Z_i - \min\{A, -\Theta\} \\ W_n^2(\Theta) &= Z + \sum_{i=1}^{n-2} Z_i - \min\{B, -\Theta\} \end{aligned}$$

where A and B are the minimums of the sets of partial sums of $\{Z, Z_1, ..., Z_{n-2}\}$ and $\{Z_1, Z, Z_2, ..., Z_{n-2}\}$. Because $A \leq_{st} B$ and the function $-\min\{\cdot, -\Theta\}$ is non-increasing, we obtain $W_n^1(\Theta) \geq_{st} W_n^2(\Theta)$.

Finally, when $m = 1$ and $n = 3$,

$$W_3^1(\Theta) = \max\{0, Z_1, Z + Z_1 + \Theta\} = (Z + Z_1) - min\{Z, Z + Z_1, -\Theta\}$$

$$W_3^2(\Theta) = \max\{0, Z, Z + Z_1 + \Theta\} = (Z + Z_1) - min\{Z_1, Z + Z_1, -\Theta\}$$

Then, we can obtain $W_3^1(\Theta) \geq_{st} W_3^2(\Theta)$, since $(Z, Z + Z_1) \leq_{st} (Z_1, Z + Z_1)$ by Theorem A.5 and the function $-\min\{\cdot, -\Theta\}$ is non-increasing.

Parts (b) and (d) can be shown in a way similar to the proof of Lemma 3.1 (b) and (d) respectively. The proof of parts (c) and (e) are trivial. □

We can now extend Corollary 3.1 as follows.

Corollary 3.3. $\mathbb{E}W_n^m(\Theta)$ *is decreasing in* m *for* $m \leq n - 1$*, and constant for* $m > n - 1$*. Specifically,*

$$\mathbb{E}W_n^{n-1}(\Theta) \leq \mathbb{E}W_n^{n-2}(\Theta) \leq ... \leq \mathbb{E}W_n^1(\Theta) \leq \mathbb{E}W_n^n(\Theta) = \mathbb{E}W_n^{n+1}(\Theta) = ... = \mathbb{E}W_n^N(\Theta)$$

Given $\Theta = \theta \geq 0$, the following lemmas show that $\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta)$ is monotonic in θ and n. The first property helps in the generalization of the First Half Rule to the late server case.

Lemma 3.4. *Given* $n > m$, $\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta)$ *is non-increasing in* θ*.*

Proof. We first prove the lemma for $n > m + 1$. Let

$$A = \min\{Z_1, ..., \sum_{i=1}^{m-1} Z_i, Z + \sum_{i=1}^{m-1} Z_i, ..., Z + \sum_{i=1}^{n-2} Z_i\}$$

$$B = \min\{Z_1, ..., \sum_{i=1}^{m} Z_i, Z + \sum_{i=1}^{m} Z_i, ..., Z + \sum_{i=1}^{n-2} Z_i\}$$

Then,

$$W_n^m(\theta) = Z + \sum_{i=1}^{n-2} Z_i - Y_n^0(\theta)$$

$$W_n^{m+1}(\theta) = Z + \sum_{i=1}^{n-2} Z_i - Y_n^1(\theta)$$

where

$$Y_n^0(\theta) = min\{A, -\theta\}$$

$$Y_n^1(\theta) = min\{B, -\theta\}$$

By theorems A.5 and A.1, $A \leq_{st} B$. In addition, $\phi(x) = min\{x, -\theta'\} - min\{x, -\theta\}$ is non-increasing in x whenever $\theta \leq \theta'$. Therefore $\phi(A) \geq_{st} \phi(B)$, which leads to

$$Y_n^0(\theta') - Y_n^0(\theta) \geq_{st} Y_n^1(\theta') - Y_n^1(\theta)$$

It further implies that

$$W_n^m(\theta) - W_n^m(\theta') \geq_{st} W_n^{m+1}(\theta) - W_n^{m+1}(\theta') \tag{3.7}$$

Hence, for $\theta \leq \theta'$,

$$\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta) \geq \mathbb{E}W_n^m(\theta') - \mathbb{E}W_n^{m+1}(\theta')$$

That is $\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta)$ is non-increasing in θ for $n > m+1$. For $n = m+1$, we can establish a similar proof by the following amendments.

$$B = \min\{Z_1, Z_1 + Z_2, ..., \sum_{i=1}^{n-1} Z_i\}$$

$$W_n^{m+1}(\theta) = \sum_{i=1}^{n-2} Z_i - Y_n^1(\theta)$$

□

Lemma 3.5. *Given $\theta \geq 0$, $\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta)$ is non-increasing in n, for all $n > m+1$.*

Proof. We know that for $n > m+1$,

$$W_{n+1}^m(\theta) = max\{0, W_n^m(\theta) + Z_{n-1}\}$$

$$W_{n+1}^{m+1}(\theta) = max\{0, W_n^{m+1}(\theta) + Z_{n-1}\}$$

For any given realization of $Z_{n-1} = z_{n-1}$, $\phi(x) = x - max\{0, x + z_{n-1}\}$ is a non-decreasing function of x. From Lemma 3.3(a), we have $W_n^m(\theta) \geq_{st} W_n^{m+1}(\theta)$. Therefore, $\phi(W_n^m(\theta)) \geq_{st} \phi(W_n^{m+1}(\theta))$, which is equivalent to

$$W_n^m(\theta) - W_{n+1}^m(\theta) \geq_{st} W_n^{m+1}(\theta) - W_{n+1}^{m+1}(\theta)$$

According to Theorem A.1, we can ignore the condition $Z_{n-1} = z_{n-1}$. Hence,

$$\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta) \geq \mathbb{E}W_{n+1}^m(\theta) - \mathbb{E}W_{n+1}^{m+1}(\theta)$$

when $n > m + 1$, which completes the proof. □

The following lemma is slightly stronger than Lemma 3.2.

Lemma 3.6. *For $n > m + 1$,*

$$\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta) \geq \mathbb{E}W_{n+1}^{m+1}(\theta) - \mathbb{E}W_{n+1}^{m+2}(\theta)$$

Proof. We have,

$$[W_n^m(\theta) - W_n^{m+1}(\theta) | W_m^m(\theta) = \lambda] = W_{n-m+1}^1(\lambda) - W_{n-m+1}^2(\lambda)$$

$$[W_{n+1}^{m+1}(\theta) - W_{n+1}^{m+2}(\theta) | W_{m+1}^{m+1}(\theta) = \lambda'] = W_{n-m+1}^1(\lambda') - W_{n-m+1}^2(\lambda')$$

According to Lemma 3.4, $\mathbb{E}W_{n-m+1}^1(\lambda) - \mathbb{E}W_{n-m+1}^2(\lambda)$ is non-increasing in λ. In addition, having homogeneous customers in the system , as it is shown in Appendix D, the waiting time is stochastically increasing with respect to the arrival number. Thus, $W_m^m(\theta) \leq_{st} W_{m+1}^{m+1}(\theta)$. Therefore,

according to Theorem A.2,

$$\mathbb{E}[W_n^m(\theta) - W_n^{m+1}(\theta)|W_m^m(\theta)] \geq_{st} \mathbb{E}[W_{n+1}^{m+1}(\theta) - W_{n+1}^{m+2}(\theta)|W_{m+1}^{m+1}(\theta)]$$

It follows that,

$$\mathbb{E}[W_n^m(\theta) - W_n^{m+1}(\theta)] \geq \mathbb{E}[W_{n+1}^{m+1}(\theta) - W_{n+1}^{m+2}(\theta)]$$

□

We can quickly obtain the following corollary.

Corollary 3.4. *Given $n > m + 1$,*

$$\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta) \geq \mathbb{E}W_{n+k}^{m+k}(\theta) - \mathbb{E}W_{n+k}^{m+k+1}(\theta)$$

for any non-negative integer $k \leq N - n$.

Using this corollary and Lemma 3.3, it is not too difficult to obtain the following theorem which extends the application of FHR to the late server case.

Theorem 3.2. First Half Rule for Late Server (FHRL): *Given the server is late for $\Theta \geq 0$ units of time, the optimal slot for the fast customer is within the first half of the sequence (including $\frac{N}{2}$ for even N and $\frac{N+1}{2}$*

for odd N). Specifically,

$$\mathbb{E}W^{\lceil \frac{N}{2} \rceil}(\Theta) \leq \mathbb{E}W^{\lceil \frac{N}{2} \rceil+1}(\Theta) \leq ... \leq \mathbb{E}W^{N}(\Theta)$$

The proof is very similar to the proof of Theorem 3.1 and omitted.

Now, we can increase the number of fast customers to more than one which is investigated in the next section.

3.3 $S(M, N-M)/(F, G)/1$: Multiple fast customers

We can extend the case of single fast customer to the one with multiple fast customers. Suppose there are M fast customers whose excess service times are stochastically smaller than that of a regular customer in the likelihood ratio order. The server starts its service after Θ_0 units of time after the first customer arrives, where Θ_0 is a non-negative random variable independent of all the service times. Let $\tilde{m} = \{m_1, m_2, ..., m_M\}$, $1 \leq m_k < m_{k+1} \leq N$ for $k = 1, 2, ..., M-1$, and define *sequence* $\tilde{m}$ as an arrival sequence in which the k-th *fast* customer arrives in the m_k-th place ($k = 1, ..., M$.) Note that the waiting time $W_n^{\tilde{m}}(\Theta_0)$ of the customer arriving in the n-th place has a Markovian property. That is, the waiting time depends on the service history only through the previous customer's waiting and service times. As

a result, the waiting time of a customer who arrives in $(n+l)$-th place can be re-modelled as the waiting time of the l-th customer in another queue whose server starts late by $W_n^{\tilde{m}}(\Theta_0)$ units of time. Specifically speaking, let $m_0 = 0$ and $\bar{k} = \max\{k | m_k \leq n, k \in \mathbb{Z}^+\}$. If $\bar{k} < M$, we can define $\tilde{m}' = \{m'_1, ..., m'_{M-\bar{k}}\} = \{m_{\bar{k}+1} - n, ..., m_M - n\}$. We have

$$W_{n+l}^{\tilde{m}}(\Theta_0) =_{st} W_l^{\tilde{m}'}(\Theta_n) \quad ; \text{for } 1 \leq n \leq N-1 \text{ and } 1 \leq l \leq N-n,$$

where $\tilde{m}'$ is a sequence for an $S(M-\bar{k}, N-n-M+\bar{k})/(F,G)/1$ queue with a delayed start Θ_n, which follows the same distribution with $W_{n+1}^{\tilde{m}}(\Theta_0)$.

The previous sections identify some properties of the customer's waiting times when there is only one fast customer to be sequenced. For more than one fast customer, we start with an $S(1, N-1)/(F,G)/1$ queue and replace a regular customer with the second fast one.

Lemma 3.7. *Given an $S(1, N-1)/(F,G)/1$ queue (with a non-negative late start Θ_0 of the server) where the fast customer is assigned to the m_1-th place, and if the m_2-th place ($m_2 > m_1 + 1$) is assigned to another fast customer,*

$$\mathbb{E}W_n^{\tilde{m}_a}(\Theta_0) - \mathbb{E}W_n^{\tilde{m}'_a}(\Theta_0) \geq \mathbb{E}W_n^{\tilde{m}_b}(\Theta_0) - \mathbb{E}W_n^{\tilde{m}'_b}(\Theta_0), \text{ for } 1 \leq n \leq N$$

where $\tilde{m}_a = \{m_1\}, \tilde{m}'_a = \{m_1+1\}, \tilde{m}_b = \{m_1, m_2\}$ and $\tilde{m}'_b = \{m_1+1, m_2\}$

Proof. For $1 \leq n \leq m_2$, the lemma holds as $W_n^{\tilde{m}_a}(\Theta_0)$ and $W_n^{\tilde{m}'_a}(\Theta_0)$

have the same distribution with $W_n^{\tilde{m}_b}(\Theta_0)$ and $W_n^{\tilde{m}'_b}(\Theta_0)$ respectively. As $W_{m_2}^{\tilde{m}_a}(\Theta_0) \geq_{st} W_{m_2}^{\tilde{m}'_a}(\Theta_0)$ by Lemma 3.3(a), we can find two random variables $\hat{W}$ and $\hat{W}'$ whose distributions are identical respectively with $W_{m_2}^{\tilde{m}_a}(\Theta_0)$ and $W_{m_2}^{\tilde{m}'_a}(\Theta_0)$, independent of the n-th customer's service time for all $n \geq m_2$, and such that

$$P\{\hat{W} \geq \hat{W}'\} = 1$$

Define a multivariate function $h_j(w, z_1, ..., z_j)$ for $j \in \mathbb{N}$ as

$$h_j(w, z_1, ..., z_j) = \begin{cases} \max\{w + z_1, 0\} & ; \text{ if } j = 1 \\ \max\{h_{j-1}(w, z_1, ..., z_{j-1}) + z_j, 0\} & ; \text{ if } j > 1 \end{cases}$$

It is easy to see that, for example,

$$W_n^{\tilde{m}_a}(\Theta_0) = h_{n-m_2}(W_{m_2}^{\tilde{m}_a}(\Theta_0), Z_1, ..., Z_{n-m_2})$$

for all $n > m_2$.

We can further define

$$f_j(w, w', z_1, ..., z_j) \equiv h_j(w, z_1, ..., z_j) - h_j(w', z_1, ..., z_j)$$

It can be shown that, given $w \geq w'$, f_j is non-decreasing in z_i's for

$i = 1, ..., j$. We then have

$$f_j(\hat{W}, \hat{W}', Z_1, Z_2, ..., Z_j) \geq_{st} f_j(\hat{W}, \hat{W}', Z, Z_2, ..., Z_j)$$

due to the fact that Z_i's are stochastically larger than Z. As a result, given $n > m_2$, we can let $j = n - m_2$ and obtain

$$\begin{aligned}
\mathbb{E}W_n^{\tilde{m}_a}(\Theta_0) - \mathbb{E}W_n^{\tilde{m}'_a}(\Theta_0) &= \mathbb{E}(\hat{W}, Z_1, ..., Z_j)] - \mathbb{E}(\hat{W}', Z_1, ..., Z_j)] \\
&= \mathbb{E}(\hat{W}, \hat{W}', Z_1, Z_2, ..., Z_j)] \\
&\geq \mathbb{E}(\hat{W}, \hat{W}', Z, Z_2, ..., Z_j)] \\
&= \mathbb{E}W_n^{\tilde{m}_b}(\Theta_0) - \mathbb{E}W_n^{\tilde{m}'_b}(\Theta_0)
\end{aligned}$$

This completes the proof. □

Lemma 3.7 can be easily generalized to the case with more than two fast customers. By a similar proof, we have the following corollary:

Corollary 3.5. *Given a sequence $\tilde{m}_a = \{m_1, ..., m_{M_a}\}$ for an $S(M_a, N - M_a)/(F, G)/1$ queue and a sequence $\tilde{m}_b = \{m_1, ..., m_{M_a}, ..., m_{M_b}\}$ for an $S(M_b, N - M_b)/(F, G)/1$ queue, with a common non-negative random server delay Θ_0 for both queues, if $M_a \leq M_b$ and $m_2 > m_1 + 1$,*

$$\mathbb{E}W_n^{\tilde{m}_a}(\Theta_0) - \mathbb{E}W_n^{\tilde{m}'_a}(\Theta_0) \geq \mathbb{E}W_n^{\tilde{m}_b}(\Theta_0) - \mathbb{E}W_n^{\tilde{m}'_b}(\Theta_0), \text{ for } 1 \leq n \leq N$$

where $\tilde{m}'_a = \{m_1+1, m_2, ..., m_{M_a}\}$ *and* $\tilde{m}'_b = \{m_1+1, m_2, ..., m_{M_a}, ..., m_{M_b}\}$

Denote the total customers' waiting time of sequence $\tilde{m}$ as $W^{\tilde{m}}(\Theta_0)$. An extended version of the First Half Rule can be derived from the corollary.

Theorem 3.3. General First Half Rule (GFHR): *Given any arrival sequence* $\tilde{m}' = \{m'_1, m'_2, ..., m'_M\}$ *of an* $S(M, N-M)/(F,G)/1$ *queue with a non-negative random delay* Θ_0 *at the start of the server, if there exists a* k *such that* $1 \leq k \leq M-1$ *and*

$$m'_{k+1} > m'_k + \lceil \frac{N - m'_k}{2} \rceil,$$

we can always build another arrival sequence $\tilde{m} = \{m_1, m_2, ..., m_M\}$ *where* $m_i = m'_i$ *for* $i \neq k+1$ *and* $m_{k+1} = m'_{k+1} - 1 > m_k$. *We have*

$$\mathbb{E}W^{\tilde{m}}(\Theta_0) \leq \mathbb{E}W^{\tilde{m}'}(\Theta_0)$$

Proof. Let $\tilde{m}_a = \{m_{k+1} - m_k\}$ and $\tilde{m}'_a = \{m'_{k+1} - m'_k\}$ be arrival sequences of an $S(1, N - m'_k - 1)/(F,G)/1$ queue. Also let

$$\tilde{m}_b = \{m_{k+1} - m_k, m_{k+2} - m_k, ..., m_M - m_k\}$$

$$\tilde{m}'_b = \{m'_{k+1} - m'_k, m'_{k+2} - m'_k, ..., m'_M - m'_k\}$$

be arrival sequences of an $S(M-k, N-m'_k-M+k)/(F,G)/1$ queue. Both queues have a common start delay

$$\Theta_k =_{st} W^{\tilde{m}'}_{m'_k+1}(\Theta_0) =_{st} W^{\tilde{m}}_{m_k+1}(\Theta_0).$$

We then obtain

$$\begin{aligned}
\mathbb{E}W^{\tilde{m}}(\Theta_0) - \mathbb{E}W^{\tilde{m}'}(\Theta_0) &= \sum_{n=1}^{N} \mathbb{E}W_n^{\tilde{m}}(\Theta_0) - \sum_{n=1}^{N} \mathbb{E}W_n^{\tilde{m}'}(\Theta_0) \\
&= \sum_{n=m'_k+1}^{N} \mathbb{E}W_n^{\tilde{m}}(\Theta_0) - \sum_{n=m'_k+1}^{N} \mathbb{E}W_n^{\tilde{m}'}(\Theta_0) \\
&= \sum_{n=1}^{N-m'_k} \mathbb{E}W_n^{\tilde{m}_b}(\Theta_k) - \sum_{n=1}^{N-m'_k} \mathbb{E}W_n^{\tilde{m}'_b}(\Theta_k) \\
&\leq \sum_{n=1}^{N-m'_k} \mathbb{E}W_n^{\tilde{m}_a}(\Theta_k) - \sum_{n=1}^{N-m'_k} \mathbb{E}W_n^{\tilde{m}'_a}(\Theta_k) \\
&\leq 0.
\end{aligned}$$

which follows from Corollary 3.5 and Theorem 3.2 for the inequalities. Thus the First Half Rule holds. □

The proof of Theorem 3.3 also implies that, given *any* schedule for the first k slots, if there are still fast customers to be scheduled, then at least one of them should be scheduled in the first half of the remaining slots after slot k. In other words, at least one fast customer should be scheduled within the first half of the slots $k+1, k+2, ..., N$ (including $\frac{N-k}{2}$ for even $N-k$ and $\frac{N-k+1}{2}$ for odd $N-k$). The following corollary concludes the

section:

Corollary 3.6. *Let sequence* $\tilde{m}^* = \{m_1^*, m_2^*, ..., m_M^*\}$ *be the optimal arrival sequence of an* $S(M, N-M)/(F,G)/1$ *queue with a non-negative random delay* Θ_0 *at the start of the server. For* $1 \leq k \leq M-1$,

$$m_k^* < m_{k+1}^* \leq m_k^* + \lceil \frac{N - m_k^*}{2} \rceil$$

This corollary helps us to ignore more than half of the possible sequences which do not follow FHR, and perform an effective search to find the optimal sequence. For example, among 10 possible sequences of 2 fast and 3 regular customers, only 5 of them follow FHR and can be optimal. These 5 proper sequences are underlined as follows: FFRRR, FRFRR, FRRFR, FRRRF, RFFRR, RFRFR, RFRRF, RRFFR, RRFRF, RRRFF. The letter "F" stands for fast and "R" for regular customer customers in the sequence.

As another illustration, when there are 3 fast and 7 regular customers, there are 120 possible sequences, since $C(10,3) = 120$. Among these 120 sequences, only 50 of them are consistent with FHR which are shown in Table 3.1.

Table 3.1: FHR sequences, $S(3,7)/(F,G)/1$

m_1, m_2	Proper Sequences based on FHR
1,2	FFFRRRRRRR, FFRFRRRRRR, FFRRFRRRRR, FFRRRFRRRR
1,3	FRFFRRRRRR, FRFRFRRRRR, FRFRRFRRRR, FRFRRRFRRR
1,4	FRRFFRRRRR, FRRFRFRRRR, FRRFRRFRRR
1,5	FRRRFFRRRR, FRRRFRFRRR, FRRRFRRFRR
1,6	FRRRRFFRRR, FRRRRFRFRR
2,3	RFFFRRRRRR, RFFRFRRRRR, RFFRRFRRRR, RFFRRRFRRR
2,4	RFRFFRRRRR, RFRFRFRRRR, RFRFRRFRRR
2,5	RFRRFFRRRR, RFRRFRFRRR, RFRRFRRFRR
2,6	RFRRRFFRRR, RFRRRFRFRR
3,4	RRFFFRRRRR, RRFFRFRRRR, RRFFRRFRRR
3,5	RRFRFFRRRR, RRFRFRFRRR, RRFRFRRFRR
3,6	RRFRRFFRRR, RRFRRFRFRR
3,7	RRFRRRFFRR, RRFRRRFRFR
4,5	RRRFFFRRRR, RRRFFRFRRR, RRRFFRRFRR
4,6	RRRFRFFRRR, RRRFRFRFRR
4,7	RRRFRRFFRR, RRRFRRFRFR
5,6	RRRRFFFRRR, RRRRFFRFRR
5,7	RRRRFRFFRR, RRRRFRFRFR
5,8	RRRRFRRFFR

In the next section, we propose a simple and effective FHR-based appointment sequencing heuristic algorithm to minimize the total expected customer waiting time.

3.4 An Effective FHR-based Appointment Sequencing Heuristic Algorithm

Although the optimal sequence cannot be completely characterized by FHR, we can propose effective sequencing heuristics based on this rule. To develop a good heuristic, we first need to investigate the structure of the optimal sequence through numerical experiments.

Table 3.2 shows the optimal sequences for an $S(3,7)/(SM,SM')/1$ queue, for various values of the job allowance x, and the fast customers' service rate μ. The optimal sequence is found by complete enumeration.

Table 3.2: Optimal sequences, $S(3,7)/(SM,SM')/1$, $\mu' = 1$, various x and μ values.

$\mu = 1.5$		$\mu = 5$		$\mu = 10$	
Job allowance	$\tilde{m}^*$	Job allowance	$\tilde{m}^*$	Job allowance	$\tilde{m}^*$
$0.00 \leq x < 0.79$	1,2,3	$0.00 \leq x < 0.40$	1,2,3	$0.00 \leq x < 0.30$	1,2,3
$0.79 \leq x < 1.24$	2,3,4	$0.40 \leq x < 0.60$	2,3,4	$0.30 \leq x < 0.49$	2,3,4
$1.24 \leq x < 1.36$	2,3,5	$0.60 \leq x < 0.81$	2,3,5	$0.49 \leq x < 0.70$	2,3,5
$1.36 \leq x < 1.59$	2,4,5	$0.81 \leq x < 0.89$	2,4,5	$0.70 \leq x < 0.79$	2,4,5
$1.59 \leq x < 2.56$	2,4,6	$0.89 \leq x < 1.72$	2,4,6	$0.79 \leq x < 1.60$	2,4,6
$2.56 < x$	3,5,7	$1.72 < x$	3,5,7	$1.60 \leq x$	3,5,7
$x = 0.5$		$x = 1$		$x = 1.5$	
Fast service rate	$\tilde{m}^*$	Fast service rate	$\tilde{m}^*$	Fast service rate	$\tilde{m}^*$
$1.00 < \mu < 3.12$	1,2,3	$1.00 < \mu < 1.08$	1,2,3	$1.00 < \mu < 1.01$	1,2,3
$3.12 \leq \mu < 9.12$	2,3,4	$1.08 \leq \mu < 1.93$	2,3,4	$1.01 \leq \mu < 1.26$	2,3,4
$9.12 \leq \mu$	2,3,5	$1.93 \leq \mu < 2.66$	2,3,5	$1.26 \leq \mu < 1.31$	2,3,5
		$2.66 \leq \mu < 3.42$	2,4,5	$1.31 \leq \mu < 1.61$	2,4,5
		$3.42 \leq \mu$	2,4,6	$1.61 \leq \mu < 63.3$	2,4,6
				$63.33 \leq \mu$	3,5,7

As can be seen, when x and μ are relatively small, the optimal sequence follows SEPT/SV. In addition, as either x or μ increases, the fast customers move toward the middle position in the optimal sequence. Furthermore, the movement of the fast customers occurs iteratively from the last fast customer to the first one. A similar behaviour is observed for a wide range of sequencing problems with exponential service times.

Inspired by this observation and based on FHR, we propose the following heuristic algorithm which shows an impressive performance in our numerical experiments.

Step 1: For $k = 1$ to M
set $m_k = m'_k = k$
Step 2: While ($m_M < m_{M-1} + \lceil \frac{N-m_{M-1}}{2} \rceil$ and the total waiting time of sequence $\{m_1, ..., m_{M-1}, m_M + 1\}$ is less than that of sequence $\{m_1, ..., m_{M-1}, m_M\}$)
set $m_M = m_M + 1$
Step 3: For $k = M - 1$ to 2
while ($m_k < m_{k-1} + \lceil \frac{N-m_{k-1}}{2} \rceil$, $m_k < m_{k+1} - 1$, and the total waiting time of sequence $\{m_1, ..., m_k + 1, ..., m_M\}$ is less than that of sequence $\{m_1, ..., m_k, ..., m_M\}$)
set $m_k = m_k + 1$
Step 4: While ($m_1 < \lceil \frac{N}{2} \rceil$, $m_1 < m_2 - 1$, and the total waiting time of sequence $\{m_1 + 1, ..., m_M\}$ is less than that of sequence $\{m_1, ..., m_M\}$)
set $m_1 = m_1 + 1$
Step 5 If there exists $k \in \{1, ..., M\}$ such that $m'_k \neq m_k$, then
for $k = 1$ to M
set $m'_k = m_k$
go to Step 2

To investigate the performance of the proposed heuristic, four $S(M, N-M)/(SM, SM')/1$ appointment-based queues with $N = 10$ and $M = 2, 4, 6, 8$ are considered. For each queueing system, one thousand test problems are randomly generated, where the service rate for the regular customers is assumed to be one, the service rate for the fast customers follows a continuous uniform $(1, 20)$ distribution and the job allowance x is generated by a continuous uniform $(0, 2)$ distribution. For each test problem, the optimal sequence is found by complete enumeration and compared with the heuristic sequence. Table 3.3 shows the numerical results. The second column shows the percentage of the optimal sequences found by the proposed heuristic. The third column shows the average percentage of the difference between the total waiting time of the heuristic sequence and that of the optimal sequence, and the forth column presents this difference in the

worst case. The last column shows the computation time of the heuristic algorithm compare to the computation time of the complete enumeration method.

Table 3.3: FHR-based heuristic performance,
$S(M, N-M)/(SM, SM')/1$, $N = 10$, $\mu \sim U(1, 20)$, $\mu' = 1$, $x \sim U(0, 2)$

Num. of fast cus.	OS found %	Av. gap %	Wst gap %	CT_h/CT_E %
$M = 2$	93.5	0.23	12.08	10.48
$M = 4$	86	0.18	8.87	3.21
$M = 6$	99.6	7.04E-5	3.19	2.89
$M = 8$	100	0	0	2.22

As can be seen, the proposed heuristic can find the optimal sequence in most cases by much less computational effort compare to the complete enumeration. Moreover, the total waiting time of the heuristic solution on average is less that 1% higher than that of the optimal sequence and in the worst case the performance of the heuristic is still very good.

We have also compared the performance of our FHR-based heuristic with SEPT/SV. To this end, two thousands test problems (100 tests for each combination of N and M) are generated using the following parameters.

N	10, 20, 30, 40, 50
M	20% of N, 40% of N, 60% of N, 80% of N
μ	$U(1, 20)$
μ'	1
x	$U(0, 2)$

Table 3.4 shows the numerical results for the comparison of the proposed heuristic with SEPT/SV.

Table 3.4: FHR-based heuristic algorithm vs SEPT/SV

N	M	TW-S	TW-H	A-IM %	B-IM %	TNS	TNH	$\frac{TNH}{TNS}$ %
10	2	9.94	9.31	10.81	18.71	4.50E+01	4.67	2.21E+01
	4	5.70	5.21	15.12	25.93	2.10E+02	6.88	2.71E+00
	6	2.99	2.87	9.4	19.13	2.10E+02	6.32	1.42E+00
	8	1.20	1.20	0	0	4.50E+01	1.00	2.67E+00
20	4	34.97	31.33	17.14	28.31	4.85E+03	19.10	7.22E-01
	8	25.11	22.13	23.33	40.79	1.26E+05	26.19	1.99E-02
	12	15.47	13.94	21.52	40.24	1.26E+05	24.57	1.23E-02
	16	10.73	10.62	8.87	19.3	4.85E+03	14.76	2.21E-01
30	6	83.24	73.66	20.21	35.16	5.94E+05	42.08	1.40E-02
	12	58.00	47.46	30.09	50.18	8.65E+07	63.18	6.71E-05
	18	25.81	21.45	33.4	52.03	8.65E+07	68.59	2.98E-05
	24	17.49	16.83	19.37	33.72	5.94E+05	40.66	2.94E-03
40	8	110.21	96.15	21.02	39.96	7.69E+07	86.58	1.43E-04
	16	99.00	81.72	31.36	55.59	6.29E+10	118.82	1.58E-07
	24	45.68	36.14	36.04	58.9	6.29E+10	125.92	7.27E-08
	32	17.82	16.07	26.34	43.79	7.69E+07	79.73	2.32E-05
50	10	192.45	166.34	23.65	44.22	1.03E+10	132.42	1.87E-06
	20	147.86	114.90	36.7	60.62	4.71E+13	192.51	3.14E-10
	30	81.98	66.96	38.01	64.55	4.71E+13	196.45	1.74E-10
	40	23.15	19.57	32.06	51.45	1.03E+10	127.43	2.25E-07

TW-S: Average total waiting time provided by SEPT/SV
TW-H: Average total waiting time provided by heuristic
A-IM: Average improvement provided by heuristic compare to SEPT/SV
B-IM: Best improvement provided by heuristic compare to SEPT/SV
TNS: Total number of possible sequences
TNH: Total number of sequences explored by heuristic

As the problem size increases, the performance of the proposed heuristic becomes more impressive. It can be seen that the heuristic could find a sequence with about 60% lower waiting time compare to SEPT/SV where $N = 50$ and $M = 20$ or 30. The proposed heuristic is very effective,

since it can find such good sequences by exploring a very tiny piece of the feasible area. It is very promising that this simple heuristic can improve the waiting time of SEPT/SV by more than 21 % where there are 40 or 50 customers to be sequenced. For a given total number of customers, the improvement provided by the heuristic is more significant when around half of the customers are fast.

Although, the proposed heuristic is applied only to the case with exponential service times in this section, it can be used for general service time distribution as well. In case of general service time distribution, we can apply the calculation method proposed in Millhiser and Valenti (2012) to numerically compute the total expected customers' waiting time for each sequence.

In the next chapter, we address an important practical concern in the appointment scheduling problem, the no-show phenomenon.

Chapter 4

Appointment sequencing with no-shows

An important aspect of customer behavior that influences the overall efficiency of an appointment system is the phenomenon of no-shows (Hassin and Mendel, 2008). A no-show happens when a customer either does not show up for service or cancels his/her appointment too late so that the service provider is not able to replace him/her by another customer.

We consider the case where the customers have different probabilities to show up at their appointment times. Suppose that, among the N customers in the appointment schedule, there are M special customers whose show up probability is p_s. On the other hand, each of the remaining $N-M$ regular customers has a show up probability $p_r > p_s$. If a customer in the n-th place of an appointment sequence shows up, his/her service time S_n,

$n = 1, ..., N$, is independently and identically distributed with the distribution function F, regardless of the customer class. As the service time distributions are identical, we also let the job allowance x to be a constant for all the customers. The objective is to minimize the total expected waiting time of the customers *who actually show up*. Let $\mathbf{P}$ denote the sequencing problem of no-shows.

A practical situation motivates us to explore this problem is the common situation in healthcare clinics when there are *new* and *repeat* patients with the same service time distributions but different show up probabilities in a clinic.

4.1 $S(1, N-1)/(F, F)/1$ queue with no-shows

We start from the simplest case where there is only one special customer. We assume that the server is available at the first appointment time. That is, the server delay Θ is negligible. Following the notations in Section 3.1, let *sequence* m be an arrival sequence in which the *special* customer is assigned to the m-th appointment slot, $m = 1, ..., N$. Let W_n^m and W^m each denote the waiting time of the n-th arrival and the total waiting time of sequence m respectively. That is $W^m = \sum_{n=1}^{N} W_n^m$.

Let $\{I_1^m, ..., I_N^m\}$ be a set of mutually independent showing up indicator

random variables for sequence m where

$$\begin{aligned} \mathbb{P}(I_n^m = 1) &= \begin{cases} p_s & ; \text{when } n = m \\ p_r & ; \text{otherwise} \end{cases} \\ \mathbb{P}(I_n^m = 0) &= 1 - \mathbb{P}(I_n^m = 1) \end{aligned}$$

The waiting time W_n^m is considered to be zero if the n-th arrival does not show up for service. Otherwise, it can be calculated using the following *Max-Flow* expression.

$$W_n^m = \max\{0, Z_{n-1}^m, Z_{n-1}^m + Z_{n-2}^m, ..., \sum_{i=1}^{n-1} Z_i^m \ \} \tag{4.1}$$

where $Z_i^m = I_i^m S_i - x$ for $i = 1, ..., n$.

The following theorem extends the application of FHR to the no-show problem **P** where there is only one special customer.

Theorem 4.1. ***First Half Rule for No-Shows (FHRNS):*** *In* $S(1, N-1)/(F, F)/1$ *with a lower show up probability for the special customer, the special customer should be scheduled in the first half of the sequence (including* $\frac{N}{2}$ *for even* N *and* $\frac{N+1}{2}$ *for odd* N*). Specifically,*

$$\mathbb{E}W^{\lceil \frac{N}{2} \rceil} \leq \mathbb{E}W^{\lceil \frac{N}{2} \rceil + 1} \leq ... \leq \mathbb{E}W^N.$$

Proof. Consider an analogous problem $\mathbf{P'}$ with equivalent service time distributions instead of no-shows. That is, we assume all the customers would

definitely show up, but their service times follow a mixture of the distribution of a degenerate random variable 0 and the distribution F. The service time of the n-th arrival of sequence m in problem $\mathbf{P'}$ can be represented as $S_n^m = I_n^m S_n$.

For problem $\mathbf{P'}$, denote the waiting time of the n-th arrival of sequence m as $W_n^{'m}$ and the total waiting time of this sequence as $W^{'m}$.

Considering the same arrival sequence for $\mathbf{P}$ and $\mathbf{P'}$, the expected waiting time of the n-th arrival in problem $\mathbf{P}$ is equal to his/her expected waiting time in $\mathbf{P'}$ if s/he shows up, otherwise it is zero. Thus we have

$$\mathbb{E}W_n^m = \begin{cases} p_s \;\; \mathbb{E}W_n^{'m} & ; \text{when } n = m \\ p_r \;\; \mathbb{E}W_n^{'m} & ; \text{otherwise} \end{cases} \tag{4.2}$$

Let $Z_n'^m$ be the excess service time of the n-th arrival of sequence m for problem $\mathbf{P'}$, i.e. $Z_n'^m = S_n^m - x$. It is not too difficult to show that for any m, by Theorem A.4, $S_m^m \leq_{lr} S_n^m$ where $n \neq m$. Moreover, for two density functions f and g, if $\frac{f(t)}{g(t)}$ is non-decreasing in t, then $\frac{f(t+c)}{g(t+c)}$ is non-decreasing in t. Thus, following the definition of the likelihood ratio order presented in §1.3, for two random variables X and Y, if $X \leq_{lr} Y$, then $X - c \leq_{lr} Y - c$, where c is a constant. Therefore, it follows from $S_m^m \leq_{lr} S_n^m$ that $Z_m'^m \leq_{lr} Z_n'^m$, where $n \neq m$. It implies that the FHR is valid for problem $\mathbf{P'}$ in minimizing the total waiting time. Hence, we have

$$\mathbb{E}W'^{1}_{m+1} - \mathbb{E}W'^{m}_{m+1} + \sum_{n=m+2}^{N} (\mathbb{E}W'^{m+1}_{n} - \mathbb{E}W'^{m}_{n}) \geq 0 \tag{4.3}$$

where $m \geq \lceil \frac{N}{2} \rceil$ (see the proof of Theorem 3.1).

To prove FHR for the original problem $\mathbf{P}$, we need to show that for $m \geq \lceil \frac{N}{2} \rceil$,

$$\mathbb{E}W^{m+1} - \mathbb{E}W^{m} \geq 0.$$

Since $\mathbb{E}W^{m}_{n} = \mathbb{E}W^{m+1}_{n}$ for $n = 1, 2, ..., m-1$, it is equivalent to

$$(\mathbb{E}W^{m+1}_{m} - \mathbb{E}W^{m}_{m}) + (\mathbb{E}W^{m+1}_{m+1} - \mathbb{E}W^{m}_{m+1}) + \sum_{n=m+2}^{N} (\mathbb{E}W^{m+1}_{n} - \mathbb{E}W^{m}_{n}) \geq 0 \tag{4.4}$$

We know that $\mathbb{E}W^{m}_{n} = p_r \mathbb{E}W'^{m}_{n}$ for $n \neq m$ and $\mathbb{E}W^{m}_{m} = p_s \mathbb{E}W'^{m}_{m}$. Thus, (4.4) is equivalent to

$$(p_r \mathbb{E}W'^{m+1}_{m} - p_s \mathbb{E}W'^{m}_{m}) + (p_s \mathbb{E}W'^{m+1}_{m+1} - p_r \mathbb{E}W'^{m}_{m+1}) + p_r \sum_{n=m+2}^{N} (\mathbb{E}W'^{m+1}_{n} - \mathbb{E}W'^{m}_{n}) \geq 0$$

and since $\mathbb{E}W'^{m}_{m} = \mathbb{E}W'^{m+1}_{m}$, it is equivalent to

$$(1 - \frac{p_s}{p_r})\mathbb{E}W'^{m}_{m} + \frac{p_s}{p_r}\mathbb{E}W'^{m+1}_{m+1} - \mathbb{E}W'^{1}_{m+1} + \left(\mathbb{E}W'^{1}_{m+1} - \mathbb{E}W'^{m}_{m+1} + \sum_{n=m+2}^{N} (\mathbb{E}W'^{m+1}_{n} - \mathbb{E}W'^{m}_{n}) \right) \geq 0 \tag{4.5}$$

According to (4.3), the last term of the left hand side is non-negative. Thus, we just need to show

$$\Delta = (1 - \frac{p_s}{p_r})\mathbb{E}W_m^{'m} + \frac{p_s}{p_r}\mathbb{E}W_{m+1}^{'m+1} - \mathbb{E}W_{m+1}^{'1} \geq 0 \tag{4.6}$$

The waiting time of the n-th arrival in problem $\mathbf{P'}$ can be represented as follows.

$$W_n'^m = \max\{0, Z_{n-1}'^m, Z_{n-1}'^m + Z_{n-2}'^m, ..., \sum_{i=1}^{n-1} Z_i'^m \ \ \}$$

For any $m \geq 2$, let

$$A = \max\{0, Z_m'^1, Z_m'^1 + Z_{m-1}'^1, ..., \sum_{i=2}^{m} Z_i'^1\}$$

$$B = \max\{0, Z_m'^1, Z_m'^1 + Z_{m-1}'^1, ..., \sum_{i=2}^{m} Z_i'^1, S_1 - x + \sum_{i=2}^{m} Z_i'^1\}.$$

Conditioning on whether the first arrival would show up or not, we have

$$\mathbb{E}W_{m+1}^{'1} = p_s\mathbb{E}B + (1 - p_s)\mathbb{E}A$$

$$\mathbb{E}W_{m+1}^{'m+1} = p_r\mathbb{E}B + (1 - p_r)\mathbb{E}A.$$

Moreover, since $Z_i'^m$ are i.i.d where $i \neq m$, $\mathbb{E}W_m^{'m} = \mathbb{E}A$. Therefore,

$$\Delta = (1 - \frac{p_s}{p_r})\mathbb{E}A + \frac{p_s}{p_r}(p_r\mathbb{E}B + (1 - p_r)\mathbb{E}A) - (p_s\mathbb{E}B + (1 - p_s)\mathbb{E}A) = 0$$

which completes the proof. □

In the next section, we incorporate the server unpunctuality.

4.2 Late start of server in $S(1, N-1)/(F, F)/1$ queue with no-shows

Now we extend the result considering a possibly late server. We upgrade problem **P** to consider the server lateness $\Theta \geq 0$ and revise the waiting time notations to $W_n^m(\Theta)$ and $W^m(\Theta)$. The following theorem shows that FHR works for the no-show case even with unpunctual server.

Theorem 4.2. First Half Rule for No-Shows with Late Server (FHRNSL): *Given the server is late for $\Theta \geq 0$ units of time, the optimal slot for the fast customer for problem* **P** *is within the first half of the sequence (including $\frac{N}{2}$ for even N and $\frac{N+1}{2}$ for odd N). Specifically,*

$$\mathbb{E}W^{\lceil \frac{N}{2} \rceil}(\Theta) \leq \mathbb{E}W^{\lceil \frac{N}{2} \rceil + 1}(\Theta) \leq ... \leq \mathbb{E}W^{N}(\Theta).$$

Proof. We show that for any $m \geq \lceil \frac{N}{2} \rceil$ and $\Theta = \theta \geq 0$,

$$\mathbb{E}W^{m}(\theta) - \mathbb{E}W^{m+1}(\theta) \leq 0$$

We recall problem $\mathbf{P}'$ as defined in the proof of Theorem 4.1 and upgrade it to consider the server lateness. We use notations $W_n'^m(\Theta)$ and $W'^m(\Theta)$ to represent the waiting time of the nth arrival and the total waiting time

of sequence m for upgraded $\mathbf{P}'$. Then, Equation (4.2) would change to

$$\mathbb{E}W_n^m(\Theta) = \begin{cases} p_s \;\; \mathbb{E}W_n'^m(\Theta) & ; \text{when } n = m \\ p_r \;\; \mathbb{E}W_n'^m(\Theta) & ; \text{otherwise} \end{cases} \tag{4.7}$$

We know that $\mathbb{E}W_n^m(\theta) = \mathbb{E}W_n^{m+1}(\theta)$ given $n < m$. As a result, we obtain

$$\begin{aligned} \mathbb{E}W^m(\theta) - \mathbb{E}W^{m+1}(\theta) &= \sum_{n=1}^{N} \mathbb{E}W_n^m(\theta) - \sum_{n=1}^{N} \mathbb{E}W_n^{m+1}(\theta) \\ &= \sum_{n=m}^{N} [\mathbb{E}W_n^m(\theta) - \mathbb{E}W_n^{m+1}(\theta)] \\ &= (p_r - p_s)[\mathbb{E}W_{m+1}'^{m+1}(\theta) - \mathbb{E}W_m'^m(\theta)] \\ &\quad + p_r \sum_{n=m+1}^{N} [\mathbb{E}W_n'^m(\theta) - \mathbb{E}W_n'^{m+1}(\theta)] \\ &\leq (p_r - p_s)[\mathbb{E}W_{m+1}'^{m+1}(0) - \mathbb{E}W_m'^m(0)] \\ &\quad + p_r \sum_{n=m+1}^{N} [\mathbb{E}W_n'^m(0) - \mathbb{E}W_n'^{m+1}(0)] \\ &= \sum_{n=m}^{N} [\mathbb{E}W_n^m(0) - \mathbb{E}W_n^{m+1}(0)] \\ &\leq 0. \end{aligned}$$

As shown in the proof of Theorem 4.1, in problem $\mathbf{P}'$ the excess service time of the special customer is smaller than that of a regular customer in likelihood ratio order. Therefore according to Lemma 3.4, $\mathbb{E}W_n'^m(\theta) - \mathbb{E}W_n'^{m+1}(\theta)$ is non-increasing in θ for $n > m$. Moreover, since $W_{m+1}'^{m+1}(\theta) = [W_m'^m(\theta) + Z_m'^m]^+$ and $W_m'^m(\theta)$ is stochastically increasing in θ, using Theorem A.2, it is not too difficult to show that $W_{m+1}'^{m+1}(\theta) - W_m'^m(\theta)$ is stochastically

decreasing in θ. Therefore, $\mathbb{E}W'^{m+1}_{m+1}(\theta) - \mathbb{E}W'^{m}_{m}(\theta)$ is also non-increasing in θ. The first inequality comes from these two properties and the last inequality follows from the application of FHRNS for problem $\mathbf{P}$.

As the inequality holds for any $\Theta = \theta \geq 0$, the proof is completed. $\square$

4.3 $S(M, N-M)/(F, F)/1$: More than one fast customer with no-shows

We can extend the no-show case of single fast customer with unpunctual server to the one with multiple fast customers. Following the notations in Section 3.3, let *sequence* $\tilde{m} = \{m_1, ..., m_M\}$ be an arrival sequence in which the k-th *special* customer is assigned to the m_k-th appointment slot, $k = 1, ..., M$. Denote the total customer waiting time of sequence $\tilde{m}$ as $W^{\mathbf{P}[\tilde{m}]}(\Theta)$, while the waiting time of the n-th arrival of this sequence is $W_n^{\mathbf{P}[\tilde{m}]}(\Theta)$.

We consider an analogous problem $\mathbf{Q}$ with equivalent service time distributions instead of no-shows. Let $W_n^{\mathbf{Q}[\tilde{m}]}(\Theta)$ and $W^{\mathbf{Q}[\tilde{m}]}(\Theta)$ each denote the waiting time of the n-th arrival and the total waiting time of sequence $\tilde{m}$ for problem $\mathbf{Q}$ respectively. We then have

$$\mathbb{E}W_n^{\mathbf{P}[\tilde{m}]}(\Theta) = \begin{cases} p_s \;\; \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}]}(\Theta) & ; \text{when } n \in \tilde{m} \\ p_r \;\; \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}]}(\Theta) & ; \text{otherwise} \end{cases} \tag{4.8}$$

We extend Corollary 3.5 to the following lemma.

Lemma 4.1. *Considering the no-show case, given a sequence* $\tilde{m}_a = \{m_1, ..., m_{M_a}\}$ *for an* $S(M_a, N-M_a)/(F,F)/1$ *queue and a sequence* $\tilde{m}_b = \{m_1, ..., m_{M_a}, ..., m_{M_b}\}$ *for an* $S(M_b, N-M_b)/(F,F)/1$ *queue, with a common non-negative random server delay* Θ_0 *for both queues, for any given* n*, if* $M_a \leq M_b$ *and* $m_2 > m_1 + 1$*,*

$$\mathbb{E}W_n^{\boldsymbol{P}[\tilde{m}_a]}(\Theta_0) - \mathbb{E}W_n^{\boldsymbol{P}[\tilde{m}'_a]}(\Theta_0) \geq \mathbb{E}W_n^{\boldsymbol{P}[\tilde{m}_b]}(\Theta_0) - \mathbb{E}W_n^{\boldsymbol{P}[\tilde{m}'_b]}(\Theta_0)$$

where $\tilde{m}'_a = \{m_1+1, m_2, ..., m_{M_a}\}$ *and* $\tilde{m}'_b = \{m_1+1, m_2, ..., m_{M_a}, ..., m_{M_b}\}$

Proof. According to Corollary 3.5, given $1 \leq n \leq N$,

$$\mathbb{E}W_n^{\mathbf{Q}[\tilde{m}_a]}(\Theta_0) - \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}'_a]}(\Theta_0) \geq \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}_b]}(\Theta_0) - \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}'_b]}(\Theta_0). \quad (4.9)$$

Then we have,

for $n \in \tilde{m}_a, n \notin \{m_1, m_1+1\}$,

$$p_s\left(\mathbb{E}W_n^{\mathbf{Q}[\tilde{m}_a]}(\Theta_0) - \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}'_a]}(\Theta_0)\right) \geq p_s\left(\mathbb{E}W_n^{\mathbf{Q}[\tilde{m}_b]}(\Theta_0) - \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}'_b]}(\Theta_0)\right),$$

for $n \notin \tilde{m}_a, n \in \tilde{m}_b, n \notin \{m_1, m_1+1\}$,

$$p_r\left(\mathbb{E}W_n^{\mathbf{Q}[\tilde{m}_a]}(\Theta_0) - \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}'_a]}(\Theta_0)\right) \geq p_s\left(\mathbb{E}W_n^{\mathbf{Q}[\tilde{m}_b]}(\Theta_0) - \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}'_b]}(\Theta_0)\right),$$

for $n \notin \tilde{m}_a, n \notin \tilde{m}_b, n \notin \{m_1, m_1+1\}$,

$$p_r\left(\mathbb{E}W_n^{\mathbf{Q}[\tilde{m}_a]}(\Theta_0) - \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}'_a]}(\Theta_0)\right) \geq p_r\left(\mathbb{E}W_n^{\mathbf{Q}[\tilde{m}_b]}(\Theta_0) - \mathbb{E}W_n^{\mathbf{Q}[\tilde{m}'_b]}(\Theta_0)\right).$$

Therefore, by Equation(4.8), the lemma holds for $n \notin \{m_1, m_1+1\}$.

In addition, for $n \in \{m_1, m_1+1\}$,

$$\mathbb{E}W_n^{\mathbf{P}[\tilde{m}_a]}(\Theta_0) = \mathbb{E}W_n^{\mathbf{P}[\tilde{m}_b]}(\Theta_0)$$

and

$$\mathbb{E}W_n^{\mathbf{P}[\tilde{m}'_a]}(\Theta_0) = \mathbb{E}W_n^{\mathbf{P}[\tilde{m}'_b]}(\Theta_0)$$

which completes the proof. □

Now, using Lemma 4.1 and Theorem 4.2, we can develop a similar proof as in Theorem 3.3 and obtain the following important result.

Theorem 4.3. General First Half Rule for No-Shows (GFHRNS): *Considering the no-show case, given any arrival sequence* $\tilde{m}' = \{m'_1, m'_2, ..., m'_M\}$ *of an* $S(M, N-M)/(F,F)/1$ *queue with a non-negative random delay* Θ_0 *at the start of the server, if there exists a* k *such that* $1 \leq k \leq M-1$ *and*

$$m'_{k+1} > m'_k + \lceil \frac{N - m'_k}{2} \rceil,$$

we can always build another arrival sequence $\tilde{m} = \{m_1, m_2, ..., m_M\}$ *where*

$m_i = m'_i$ for $i \neq k+1$ and $m_{k+1} = m'_{k+1} - 1 > m_k$. We have

$$\mathbb{E}W^{\boldsymbol{P}[\tilde{m}]}(\Theta_0) \leq \mathbb{E}W^{\boldsymbol{P}[\tilde{m}']}(\Theta_0)$$

Now we can partially obtain the optimal sequence for problem **P** by the following corollary.

Corollary 4.1. *Let sequence $\tilde{m}^* = \{m_1^*, m_2^*, ..., m_M^*\}$ be the optimal arrival sequence of an $S(M, N-M)/(F,F)/1$ queue with a non-negative random server delay Θ_0 where the fast customers are distinguished by their lower show up probabilities. For $1 \leq k \leq M-1$,*

$$m_k^* < m_{k+1}^* \leq m_k^* + \lceil \frac{N - m_k^*}{2} \rceil$$

We conclude this section with two significant conclusions of our no-show results. Firstly, the results could be applied for any service time distribution. In other words, there is no special stochastic ordering assumption required to apply GFHR for no-shows. Secondly, the results could be applied to schedule the break times. The following remark represents this application.

Remark 4.1. ***Break Time Scheduling:*** *When $p_s = 0$, the time allocated to the special customers can be considered as breaks for the server. Thus, GFHRNS can be used to schedule the breaks to minimize the total customer waiting time in an equally spaced appointment system with homogeneous*

arrivals.

Chapter 5

Conclusions and future works

In this thesis, we have studied an appointment-based queue with two classes of customers whose excess service times are stochastically ordered. The operational target is to find the optimal sequence of arrivals to minimize the total waiting time of customers. We first show that the optimal sequence does not follow the Shortest Expected Processing Time first (SEPT) and the Smallest Variance first (SV) rules in general. A new concept, the *voucher effect*, has been introduced to explain this counter-intuitive observation. He have shown that the optimal sequence structure is influenced by the interaction between two effects in sequencing heterogeneous customers or services: The snowball effect which drives the stochastically faster customers toward the beginning of the optimal sequence, and the voucher effect which pushes the faster customers toward to the end of the sequence.

We have then identified and proved an important property, the First

Half Rule (FHR), of the optimal sequence. FHR implies that each fast customer should be scheduled in a position that is in the first half of the positions after the previous customer in the schedule. While the application of FHR is not limited to appointment systems with constant intervals, it could be applied to any appointment system with constant intervals and any service time distributions with Monotone Likelihood Ratio Property (MLRP). The main tool for providing this result is stochastic order arguments. Based on the same framework, we have considered the appointment sequencing problem in the presence of late server starting.

We have also extended our analysis to a system where the two customer classes are different in their no-show probabilities instead of the excess service times. For the optimal sequence to minimize the total waiting time of the customers who show up, the first half rule still applies. This results is shown to be applicable in scheduling breaks in the appointment systems with constant intervals and homogeneous customers.

Moreover, an effective FHR-based appointment sequencing heuristic algorithm is developed in this thesis. This heuristic could find the optimal arrival sequence in most cases over a wide range of test problems. By exploring a small piece of the feasible area, the proposed heuristic is able to improve the total wanting time by more than 60% in comparison with SEPT and SV rules.

For future work, we are interested

- to investigate the applicability of FHR where the job allowances are optimized for each sequence.
- to address more practical situations where FHR could be applied. For example, the case in which the job allowance for each customer is proportional to his/her mean service time.
- to develop a an appointment sequencing heuristic which is guaranteed to find the optimal sequence.
- to find a condition that guarantees SEPT or SV is optimal.
- to address customer unpunctuality and investigate the applicability of FHR when the fast customer is the customer who is more likely to arrive earlier than his/her appointment time.
- to develop an easy to implement sequencing heuristic based on FHR which can be used in hospitals.
- to consider an appointment system with heterogeneous customers and investigate if we can categorize the customers in two groups and apply FHR.
- to analytically model the impact of the increase of the job allowance on the structure of the optimal sequence. It can make a bridge between machine scheduling and appointment scheduling.

Bibliography

Arbitman, D. B. (1986). A primer on patient classification systems and their relevance to ambulatory care. *The Journal of Ambulatory Care Management*, 9(1):58–81.

Bailey, N. (1952). A study of queues and appointment systems in hospital outpatient departments with fast reference to waiting times. *Journal of Royal Statistical Society*, 14:185–199.

Bartoszynski, R. and Niewiadomska-Bugaj, M. (2008). *Probability and statistical inference*. John Wiley & Sons, 2nd edition.

Begen, M. A. and Queyranne, M. (2011). Appointment scheduling with discrete random durations. *Mathematics of Operations Research*, 36(2):240–257.

Boland, P. J., Proschan, F., and Tong, Y. (1992). A stochastic ordering of partial sums of independent random variables and of some random processes. *Journal of Applied Probability*, pages 645–654.

Bosch, P. M. V. and Dietz, D. C. (2000). Minimizing expected waiting in a medical appointment system. *IIE Transactions*, 32(9):841–848.

Bosch, P. M. V. and Dietz, D. C. (2001). Scheduling and sequencing arrivals to an appointment system. *Journal of Service Research*, 4(1):15–25.

Cayirli, T. and Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549.

Cayirli, T., Veral, E., and Rosen, H. (2006). Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9(1):47–58.

Cayirli, T., Veral, E., and Rosen, H. (2008). Assessment of patient classification in appointment system design. *Production and Operations Management*, 17(3):338–353.

Choi, S. and Wilhelm, W. E. (2012). An analysis of sequencing surgeries with durations that follow the lognormal, gamma, or normal distribution. *IIE Transactions on Healthcare Systems Engineering*, 2(2):156–171.

Denton, B. and Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35(11):1003–1016.

Denton, B., Viapiano, J., and Vogl, A. (2007). Optimization of surgery se-

quencing and scheduling decisions under uncertainty. *Health Care Management Science*, 10(1):13–24.

Dexter, F. and Ledolter, J. (2005). Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. *Anesthesiology*, 103(6):1259–1167.

Erdogan, S. A., Denton, B. T., Cochran, J., Cox, L., Keskinocak, P., Kharoufeh, J., and Smith, J. (2010). Surgery planning and scheduling. *Wiley Encyclopedia of Operations Research and Management Science*.

Ge, D., Wan, G., Wang, Z., and Zhang, J. (2014). A note on appointment scheduling with piecewise linear cost functions. *Mathematics of Operations Research*, 39(4):1244–1251.

Gul, S., Denton, B. T., Fowler, J. W., and Huschka, T. (2011). Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production and Operations management*, 20(3):406–417.

Gupta, D. (2007). Surgical suites' operations management. *Production and Operations Management*, 16(6):689–700.

Gupta, D. and Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9):800–819.

Hall, R. W. (2012). *Handbook of Healthcare System Scheduling*. Springer.

Hassin, R. and Mendel, S. (2008). Scheduling arrivals to queues: A single-server model with no-shows. *Management Science*, 54(3):565–572.

Jebali, A., Hadj Alouane, A. B., and Ladet, P. (2006). Operating rooms scheduling. *International Journal of Production Economics*, 99(1):52–62.

Kaandorp, G. C. and Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science*, 10(3):217–229.

Kendall, D. G. et al. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. *The Annals of Mathematical Statistics*, 24(3):338–354.

Klassen, K. J. and Rohleder, T. R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of Operations Management*, 14(2):83–101.

Kong, Q., Lee, C.-Y., Teo, C.-P., and Zheng, Z. (2013). Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research*, 61(3):711–726.

Kong, Q., Lee, C.-Y., Teo, C.-P., and Zheng, Z. (2014). Appointment sequencing: moving beyond the smallest-variance-first rule. In *20th Conference of the International Federation of Operational Research Societies*.

Lebowitz, P. (2003). Schedule the short procedure first to improve or efficiency. *AORN Journal*, 78(4):651–659.

Lehaney, B., Clarke, S., and Paul, R. (1999). A case of an intervention in an outpatients department. *Journal of the Operational Research Society*, 50(9):877–891.

Mancilla, C. and Storer, R. (2012). A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions*, 44(8):655–670.

Marcon, E. and Dexter, F. (2006). Impact of surgical sequencing on post anesthesia care unit staffing. *Health Care Management Science*, 9(1):87–98.

Meza, J. P. (1998). Patient waiting times in a physician's office. *The American Journal of Managed Care*, 4(5):703–712.

Millhiser, W. P. and Valenti, B. C. (2012). Delay distributions in appointment systems with generally and non-identically distributed service times and no-shows. Available at http://dx.doi.org/10.2139/ssrn.2045074.

Pegden, C. D. and Rosenshine, M. (1990). Scheduling arrivals to queues. *Computers & Operations Research*, 17(4):343–348.

Pinedo, M. (2009). *Planning and scheduling in manufacturing and services*. Springer, 2nd edition.

Pinedo, M. (2012). *Scheduling: theory, algorithms, and systems*. Springer, 4th edition.

Rohleder, T. R. and Klassen, K. J. (2000). Using client-variance information to improve dynamic appointment scheduling performance. *Omega*, 28(3):293–302.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic orders*. Springer.

Shanthikumar, J. G. and Yao, D. D. (1986). The preservation of likelihood ratio ordering under convolution. *Stochastic Processes and Their Applications*, 23(2):259–267.

Wang, P. P. (1993). Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics (NRL)*, 40(3):345–360.

Wang, P. P. (1999). Sequencing and scheduling n customers for a stochastic server. *European Journal of Operational Research*, 119(3):729–738.

Weiss, E. N. (1990). Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions*, 22(2):143–150.

Welch, J. and Bailey, N. J. (1952). Appointment systems in hospital outpatient departments. *The Lancet*, 259(6718):1105–1108.

Appendix A

Some useful properties of stochastic orders

Many useful properties of stochastic ordering can be found in Shaked and Shanthikumar (2007). We recall some of them as follows to make our proofs self-contained.

Theorem A.1. *(Recall theorems 1.A.3 of Shaked and Shanthikumar 2007)*
(a) *For random variables X and Y, if $X \leq_{st} Y$ and g is any non-decreasing [non-increasing] function, then $g(X) \leq_{st} [\geq_{st}] g(Y)$. The same property holds for random vectors $\mathbf{X}$ and $\mathbf{Y}$.*
(b) *Let $X_1, X_2, ..., X_m$ be a set of independent random variables and let $Y_1, Y_2, ..., Y_m$ be another set of independent random variables. If $X_i \leq_{st} Y_i$,*

for $i = 1, 2, ..., m$, then, for any increasing function $\phi : \mathbb{R}^m \to \mathbb{R}$, one has

$$\phi(X_1, X_2, ..., X_m) \leq_{st} \phi(Y_1, Y_2, ..., Y_m)$$

In particular,

$$\sum_{j=1}^{m} X_j \leq_{st} \sum_{j=1}^{m} Y_j$$

(c) *Let X, Y and Θ be random variables such that $[X|\Theta = \theta] \leq_{st} [Y|\Theta = \theta]$ for all θ in the support of Θ. Then $X \leq_{st} Y$.*

Theorem A.2. *(Recall Theorem 1.A.6 of Shaked and Shanthikumar 2007) Consider a family of distribution functions $\{G_\theta, \theta \in \chi\}$. Let Θ_1 and Θ_2 be two random variables with supports in χ and distribution functions F_1 and F_2, respectively. Let Y_1 and Y_2 be two random variables such that $Y_i =_{st} X(\Theta_i), i = 1, 2$. That is the distribution function of Y_i is given by*

$$H_i(y) = \int_\chi G_\Theta(y) dF_i(\theta), \quad y \in \mathbb{R}, i = 1, 2.$$

If $X(\theta) \leq_{st} X(\theta')$, whenever $\theta \leq \theta'$, and if $\Theta_1 \leq_{st} \Theta_2$, then $Y_1 \leq_{st} Y_2$.

Theorem A.3. *(Recall Theorem 6.B.3 of Shaked and Shanthikumar 2007)*

Let $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ *and* $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ *be two n-dimensional random vectors. If*

$$X_1 \leq_{st} Y_1$$

and for $i = 2, 3, ..., n$,

$$[X_i | X_1 = x_1, ..., X_{i-1} = x_{i-1}] \leq_{st} [Y_i | Y_1 = y_1, ..., Y_{i-1} = y_{i-1}]$$

whenever $x_j \leq y_j, j = 1, 2, ..., i-1$, *then* $\boldsymbol{X} \leq_{st} \boldsymbol{Y}$.

Theorem A.4. *(Recall Theorem 1.C.30 of Shaked and Shanthikumar 2007) Let* X *and* Y *be two random variables with distribution functions* F *and* G, *respectively. Let* W *be a random variable with the distribution function* $pF + (1-p)G$ *for some* $p \in (0, 1)$. *If* $X \leq_{lr} Y$, *then* $X \leq_{lr} W \leq_{lr} Y$.

We also recall a theorem from Boland et al. (1992) as follows.

Theorem A.5. *(Recall Theorem 1 of Boland et al. 1992) Let* $\{X_1, ..., X_n\}$ *be a sequence of independent random variables. If*

$$X_1 \leq_{lr} ... \leq_{lr} X_n$$

then

(a) *for any permutation* $\pi = (\pi_1, ..., \pi_n)$ *of* $(1, 2, ..., n)$

$$(X_1, X_1+X_2, ..., \sum_{i=1}^{n} X_i) \leq_{st} (X_{\pi_1}, X_{\pi_1}+X_{\pi_2}, ..., \sum_{i=1}^{n} X_{\pi_i}) \leq_{st} (X_n, X_n+X_{n-1}, ..., \sum_{i=1}^{n} X_i)$$

(b) *for a given permutation* $\pi = (\pi_1, ..., \pi_r, ..., \pi_s, ..., \pi_n)$ *of* $(1, 2, ..., n)$ *where* $\pi_r < \pi_s$*, if one swaps* π_r *and* π_s *to build a new permutation* $\pi' = (\pi'_1, ..., \pi'_n) = (\pi_1, ..., \pi_s, ..., \pi_r, ..., \pi_n)$*, then*

$$(X_{\pi_1}, X_{\pi_1} + X_{\pi_2}, ..., \sum_{i=1}^{n} X_{\pi_i}) \leq_{st} (X_{\pi'_1}, X_{\pi'_1} + X_{\pi'_2}, ..., \sum_{i=1}^{n} X_{\pi'_i})$$

Appendix B

Expected waiting time calculation

B.1 $S(1,2)/(SM, SM')/1$

Consider an appointment system with fixed job allowance x and three customers with exponential service durations where the service rate for the nth arrival is $\mu_n, n = 1, 2, 3$. Let $p_n^j(x)$, $n = 1, 2, 3$ and $j = 1, ..., n-1$, denote the probability that the nth arrival visits j customers in the system upon his/her arrival. The expected waiting time of the nth arrival, denoted by $\mathbb{E}W_n(x)$ can be calculated as follows.

$$\mathbb{E}W_1(x) = 0$$

$$\mathbb{E}W_2(x) = \frac{1}{\mu_1} p_2^1(x)$$

$$\mathbb{E}W_3(x) = \left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right) p_3^2(x) + \frac{1}{\mu_2} p_3^1(x)$$

where

$$p_2^1(x) = e^{-\mu_1 x}$$

$$p_3^2(x) = e^{-2\mu_1 x}$$

$$p_3^1(x) = \frac{\mu_2}{\mu_1 - \mu_2} e^{-(\mu_1+\mu_2)x} - \frac{\mu_1}{\mu_1 - \mu_2} e^{-2\mu_1 x} + e^{-\mu_2 x}$$

B.2 $S(1, N-1)/(SM, SM')/1$, $N > 3$

Based on Equation (2.3), for $l = 1, 2, ..N - m$, $\Pr\{N(t_{m+l}) = 0\}$ can be calculated as follows.

For $m = 1$ and $l = 1$,

$$\Pr\left\{N\left(t_{m+l}\right) = 0\right\} = \Pr\left\{N\left(t_2\right) = 0\right\} = 1 - e^{-\mu x}$$

For $m > 1$ and $l = 1$,

$$\begin{aligned}\Pr\{N(t_{m+l}) = 0\} &= \Pr\left\{N\left(t_m\right) = 0\right\} \times \left(1 - e^{-\mu x}\right) \\ &\quad + \sum_{k=1}^{m-1} \left(\Pr\left\{N\left(t_m\right) = k\right\} \times \int_0^x \frac{\mu'^k t^{k-1}}{(k-1)!} e^{-\mu' t} \left(1 - e^{-\mu(x-t)}\right) dt\right)\end{aligned}$$

For $m = 1$ and $l > 1$,

$$\Pr\{N(t_{m+l}) = 0\} = \sum_{k=1}^{l-1} \left(Pr\left\{N\left(t_l\right) = k-1\right\} \times \left[1 - \sum_{q=0}^{k-1} \frac{(\mu' x)^q}{q!} e^{-\mu' x} \right] \right)$$

$$+ Pr\left\{N\left(t_l\right) = l-1\right\} \times \left[1 - \left(e^{-\mu x} + \int_0^x \mu e^{-\mu t} \times \sum_{q=0}^{l-2} \frac{(\mu'(x-t))^q}{q!} e^{-\mu'(x-t)} \, dt \right) \right]$$

For $m > 1$ and $l > 1$,

$$\Pr\left\{N\left(t_{m+l}\right) = 0\right\} = \sum_{k=1}^{l-1} \left(Pr\left\{N\left(t_{m+l-1}\right) = k-1\right\} \times \left[1 - \sum_{q=0}^{k-1} \frac{(\mu' x)^q}{q!} e^{-\mu' x} \right] \right)$$

$$+ Pr\left\{N\left(t_{m+l-1}\right) = l-1\right\} \times \left[1 - \left(e^{-\mu x} + \int_0^x \mu e^{-\mu t} \times \sum_{q=0}^{l-2} \frac{(\mu'\left(x-t\right))^q}{q!} e^{-\mu'(x-t)} \, dt \right) \right]$$

$$+ \sum_{k=l+1}^{m+l-1} \left(\begin{array}{l} Pr\left\{N\left(t_{m+l-1}\right) = k-1\right\} \\ \times \left[1 - \left(\begin{array}{l} \int_x^\infty \frac{\mu'^{k-l} t^{k-l-1}}{(k-l-1)!} e^{-\mu' t} dt + \int_0^x \frac{\mu'^{k-l} t^{k-l-1}}{(k-l-1)!} e^{-\mu' t} \times e^{-\mu(x-t)} dt \\ + \int_0^x \int_t^x \frac{\mu'^{k-l} t^{k-l-1}}{(k-l-1)!} e^{-\mu' t} \times \mu e^{-\mu\left(t'-t\right)} \\ \times \sum_{q=0}^{l-2} \frac{\left(\mu'\left(x-t'\right)\right)^q}{q!} e^{-\mu'\left(x-t'\right)} \, dt' dt \end{array} \right) \right] \end{array} \right)$$

Based on Equation (2.4), for a given m and $l = 1, 2, ..N - m$, where j is a positive integer less than $m + l - 1$, $\Pr\{N(t_{m+l}) = j\}$ can be calculated as follows.

For $m = 1$ and $l < j$,

$$\Pr\{N(t_{m+l}) = j\} = \sum_{k=0}^{l-j-1} \left(\Pr\left\{N\left(t_l\right) = j + k - 1\right\} \times \frac{\left(\mu' x\right)^k}{k!} e^{-\mu' x} \right)$$

$$+ \Pr\left\{N\left(t_l\right) = l - 1\right\} \times \int_0^x \mu e^{-\mu t} \frac{\left(\mu'(x-t)\right)^{l-j-1}}{(l-j-1)!} e^{-\mu'(x-t)} dt$$

For $m > 1$ and $l = j$,

$$\Pr\{N(t_{m+l}) = j\} = \Pr\left\{N\left(t_{1+l}\right) = l\right\} = \Pr\left\{N\left(t_l\right) = l - 1\right\} \times e^{-\mu x} = e^{-\mu(lx)}$$

For $m > 1$ and $l < j < m + l$,

$$\Pr\left\{N\left(t_{m+l}\right) = j\right\} = \sum_{k=0}^{m+l-j-1} \left(\Pr\left\{N\left(t_{m+l-1}\right) = j + k - 1\right\} \times \frac{\left(\mu' x\right)^k}{k!} e^{-\mu' x} \right)$$

For $m > 1$ and $l = j$,

$$\Pr\{N(t_{m+l}) = j\} = \Pr\{N(t_{m+l-1}) = l-1\} \times e^{-\mu x}$$

$$+ \sum_{k=1}^{m-1} \left(\Pr\{N(t_{m+l-1}) = l+k-1\} \times \int_0^x \frac{t^{k-1}\mu'^k}{(k-1)!} e^{-\mu' t} e^{-\mu(x-t)} dt \right)$$

For $m > 1$ and $l > j$,

$$\Pr\{N(t_{m+l}) = j\} = \sum_{k=0}^{l-j-1} \left(\Pr\{N(t_{m+l-1}) = j+k-1\} \times \frac{(\mu' x)^k}{k!} e^{-\mu' x} \right)$$

$$+ \Pr\{N(t_{m+l-1}) = l-1\} \times \int_0^x \mu e^{-\mu t} \frac{(\mu'(x-t))^{l-j-1}}{(l-j-1)!} e^{-\mu'(x-t)} dt$$

$$+ \sum_{k=l-j+1}^{m+l-j-1} \left(\begin{array}{l} \Pr\{N(t_{m+l-1}) = j+k-1\} \\ \times \int_0^x \int_t^x \frac{t^{j+k-l-1}\mu'^{j+k-l}}{(j+k-l-1)!} e^{-\mu' t} \times \mu e^{-\mu\left(t'-t\right)} \frac{\left(\mu'\left(x-t'\right)\right)^{l-j-1}}{(l-j-1)!} e^{-\mu'\left(x-t'\right)} dt' dt \end{array} \right)$$

Appendix C

$S(M, N)/(D, D')/1$

As mentioned in §1, the appointment sequencing problem is still NP-hard even under deterministic model where the time interval allocated to each customer is set to a constant, and the service duration for each customer is known in advance (Kong et al., 2014).

In this section, we consider a deterministic appointment sequencing problem with the following characteristics. There are M special customers with deterministic service time of s and job allowance of x_s, and N regular customers with deterministic service time of r and job allowance of x_r. Sequence $m = \{m_1, m_2, ..., m_M\}$ is the arrival sequence in which the special customers are scheduled in positions $m_1, m_2, ..., m_M$. Let w_i^m denote the waiting time of the ith arrival of sequence m, and w^m the total waiting time of sequence m. m^* denotes the optimal sequence to minimize the total customers' waiting time. The excess service times of the special and

regular customers are denoted by z_s and z_r respectively, and it is assumed that $z_s < z_r$. Note that in other sections, N denotes to the total number of customers, but here it denotes to the number of regular customers.

Lemma C.1. *Assuming $z_s < z_r$,*

(a) if $z_r \leq 0$, then the total waiting time for all possible sequences is zero.

(b) if $z_s \geq 0$, then the special customers should be scheduled first, i.e. $m^* = 1, 2, ..., M$.

Proof. Where $z_r \leq 0$, the service time of no customer exceeds his allocated allowance time and therefore no one has to wait. If $z_s \geq 0$, then the server has to continuously serve all the customers and is never idle during the session. Thus,

$$w_i^m = k_i^m z_s + (i - 1 - k_i^m) z_r$$

where k_i^m is the number of special customers scheduled before the ith arrival in sequence m. For any given sequence, if we can find a regular customer followed by a special customer, we can reduce the total waiting time by $z_r - z_s$ by swapping these two customers. We repeat this procedure to reach the optimal sequence $m^* = 1, 2, ..., M$. □

Now we investigate the structure of the optimal sequence with one special customer, where $z_s < 0 < z_r$. Without loss of generality, we can assume that $z_r = 1$. Let $\alpha = -z_s$, $0 < \alpha$, which shows the potential saving can be provided by a special customer. For $m = 1, 2, ..., N - 1$,

$$w_i^m = \begin{cases} i-1 & ;when\ i \le m \\ i-m-1+[(m-1)-\alpha]^+ & ;when\ i > m \end{cases}$$

$$w^m = \sum_{i=0}^{m-1} i + \sum_{j=0}^{N-m} \left(j + [(m-1)-\alpha]^+\right)$$

Hence, for $m = 1, 2, ..., N-1$,

$$\Delta w^m = w^{m+1} - w^m = \begin{cases} \alpha + 1 & ;when\ \alpha \le m-1 \\ 2m - N + (N-m)(m-\alpha) & ;when\ m-1 < \alpha < m \\ 2m - N & ;when\ \alpha \ge m \end{cases}$$

It is not difficult to show that for any $m \ge \lceil \frac{N+M}{2} \rceil$, $\Delta w^m \ge 0$. Therefore we have the the following lemma which validates FHR under deterministic service time assumption.

Lemma C.2. *The special customer in $S(1, N+1)/(D, D')/1$ must be scheduled in the first half of the sequence.*

Also, considering equally spaced appointment times with length of x, i.e. $x_s = x_r = x$, it is not too difficult to show the optimal slot for the special customer moves from the first slot toward the middle of the sequence as x increases.

In summary, the deterministic results are consistent with our observations for $S(1, N-1)/(SM, SM')/1$ in Section 2.2.2.

Appendix D

Homogeneous customers arriving at equally spaced appointment times

In this section, we show that in an appointment system with constant job allowance and homogeneous customers, the expected waiting time of the n-th arrival is increasing and concave in n.

Consider an appointment system with homogeneous customers with i.i.d service time S with distribution function F.

Lemma D.1. *W_n is stochastically increasing in n.*

Proof. We first show that $[W_{n+1}|W_n = \theta]$ is stochastically increasing in θ.

$$W_{n+1} = [W_n + S - x]^+ \tag{D.1}$$

Thus

$$[W_{n+1}|W_n = \theta] = [\theta + S - x]^+$$

and

$$Pr\{[\theta + S - x]^+ > t\} = \begin{cases} \overline{F}(t + x - \theta) & t \geq 0 \\ 1 & otherwise \end{cases}$$

where F is the distribution function of S, $F(s) = 0$ for $s < 0$, and $\overline{F}$ is $1 - F$. Since $\overline{F}$ is a non-increasing function, for all $t \in (-\infty, +\infty)$

$$[W_{n+1}|W_n = \theta] \leq_{st} [W_{n+1}|W_n = \theta']$$

whenever $\theta < \theta'$. Hence, $[W_{n+1}|W_n = \theta]$ is stochastically increasing in θ. Also $W_1 \leq_{st} W_2$, since $W_1 = 0$, and W_2 is a non-negative random variable. Therefore according to Theorem A.2, $W_2 \leq_{st} W_3$. Similarly, from $W_2 \leq_{st} W_3$, it can be proved that $W_3 \leq_{st} W_4$, and so on. Theorem A.2 can thus be used recursively to show $W_n \leq_{st} W_{n+1}$.

□

Lemma D.2. *$[W_{n+1} - W_n|W_n = \theta]$ is stochastically decreasing in θ.*

Proof. We show that $[W_{n+1} - W_n|W_n = \theta'] \leq_{st} [W_{n+1} - W_n|W_n = \theta]$ whenever $0 \leq \theta < \theta'$. According to (D.1), we have

$$[W_{n+1} - W_n|W_n = \theta] = [\theta + S - x]^+ - \theta.$$

Let G_θ ($G_{\theta'}$) be the distribution function of $[W_{n+1}-W_n|W_n=\theta(W_n=\theta')]$.

Then

$$G_\theta(t)=\begin{cases} F(t+x) & (\theta\geq x\ \&\ t\geq -x)\ or\ (\theta<x\ \&\ t\geq-\theta) \\ 0 & otherwise \end{cases}$$

We have a similar expression for $G_{\theta'}$, $\theta'>\theta$. Therefore, as can be seen in Figure C.1,

$$G_{\theta'}(t)-G_\theta(t)=\begin{cases} F(t+x) & (\theta<x<\theta'\&-x<t<-\theta)or(\theta'<x\&-\theta'\leq t<\theta) \\ 0 & otherwise \end{cases}$$

Hence, $G_{\theta'}(t)\geq G_\theta(t)$ for all $t\in(-\infty,+\infty)$ whenever $\theta<\theta'$. That is

$$[W_{n+1}-W_n|W_n=\theta]\geq_{st}[W_{n+1}-W_n|W_n=\theta']$$

which means that $[W_{n+1}-W_n|W_n=\theta]$ is stochastically decreasing in θ. □

Lemma D.3. *$W_{n+1}-W_n$ is stochastically decreasing in n.*

Proof. According to Lemma D.1, $W_n\leq_{st}W_{n+1}$ and it follows from Lemma D.2 that

$$[W_n-W_{n+1}|W_n=\theta]\leq_{st}[W_{n+1}-W_{n+2}|W_{n+1}=\theta']$$

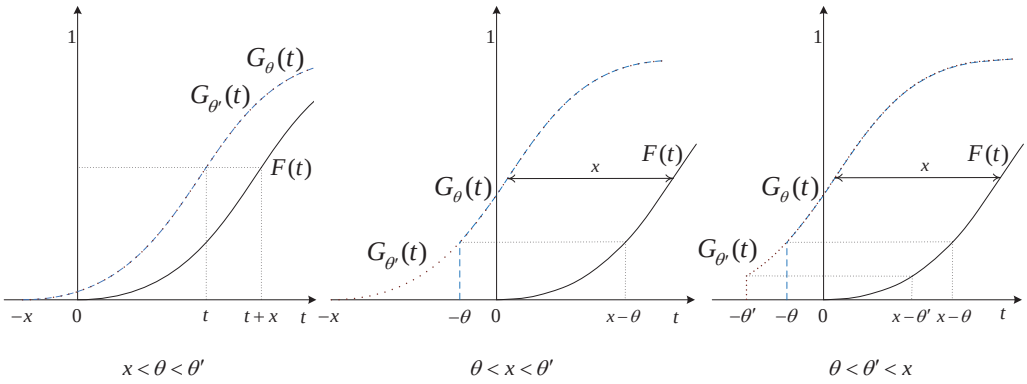

Figure D.1: $[W_{n+1} - W_n | W_n = \theta]$ and $[W_{n+1} - W_n | W_n = \theta']$ distributions, where $0 \leq \theta < \theta'$

whenever $\theta \leq \theta'$. Let $Y_n = W_n - W_{n+1}$. Then based on Theorem A.2

$$Y_n \leq_{st} Y_{n+1}$$

Thus, $W_{n+1} - W_n \geq_{st} W_{n+2} - W_{n+1}$ (i.e. $W_{n+1} - W_n$ is stochastically decreasing in n). □

Lemma D.4. $\mathbb{E}[W_n]$ *is increasing and concave in* n.

Proof. According to Lemma D.1, $\mathbb{E}[W_n]$ is increasing in n. Also, from Lemma D.3, we have

$$\mathbb{E}[W_{n+1}] - \mathbb{E}[W_n] \geq \mathbb{E}[W_{n+2}] - \mathbb{E}[W_{n+1}].$$

Hence $\mathbb{E}[W_n]$ is increasing and concave with respect to n. □