# Constant Job-Allowance Policies for Appointment Scheduling: Performance Bounds and Numerical Analysis

Shenghai Zhou

School of Business, Central South University, Changsha, China 410083, zshsjtu2014@sjtu.edu.cn

Yichuan Ding

Desautels Faculty of Management, McGill University, Montreal, Quebec H3A 1G5, Canada, daniel.ding@mcgill.ca

#### Woonghee Tim Huh

Sauder School of Business, University of British Columbia, Vancouver, Canada V6T 1Z2, tim.huh@sauder.ubc.ca

#### Guohua Wan

Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai, China 200030, ghwan@sjtu.edu.cn

We consider the classical appointment scheduling problem that does not have a session length constraint. The decision is to determine job allowances to minimize total expected weighted costs of waiting times of jobs and idle times of the service provider. In particular, we study a simple yet effective scheduling policy - the so-called "constant" policy, which allocates a constant job allowance for each appointment. Prior studies on appointment scheduling suggest a "dome"-shape structure for the optimal job allowance over the planning horizon. This implies that job allowance does not vary significantly in the middle of the schedule sequence, but varies at the beginning as well as the end of the optimal schedule. We show that an even simple scheduling policy - the constant policy, is asymptotically optimal thus provide theoretical justification for such a widely used policy. Using a dynamic programming formulation, we express each job's waiting time as the maximum of a random walk, which allows us to bound the performance gap between the constant policy and the optimal schedule based on classical results on D/G/1 queues. We derive an explicit upper bound for the performance gap when either of the following conditions holds: (1) the server idling cost is relatively small compared to the job waiting cost; (2) the number of appointments is sufficiently large. We also extend this result to a more general setting with multiple service types. Numerical experiments show that the constant policy is near optimal even when the number of appointments is small or when the server idling cost is moderately large, which complements our theoretical results. Our result provides a justification and strong support for the constant policy under certain mild conditions. Moreover, with minor modifications, the constant policy can be adapted to more general scenarios with patient no-shows, or with heterogeneous appointment types.

Key words: appointment scheduling, constant policy, asymptotic optimality, D/G/1 queue History: Received: October 2019; Accepted: December 2020 by Zuo-Jun Max Shen, after 2 revisions

## 1. Introduction

Appointment scheduling is commonly used in services such as personal consulting, car repairing, and healthcare (e.g., CT scan, magnetic resonance imaging (MRI) and outpatient service). As the first step in making appointments, one needs to determine the start time for each appointment slot, or equivalently the job allowance for each patient or customer. This problem has been referred to as appointment scheduling in the literature (see, for example, Denton and Gupta 2003, Hassin and Mendel 2008). The objective is to maximize the servers' utilization while minimizing the patients' wait times, assuming that all the scheduled patients arrive at the servers punctually. (In this paper, we use the terms patients and customers interchangeably.) Due to the random service times, sometimes the previous patient spent less time than the scheduled job allowance, leaving the server idled; at other times, the previous patient cannot finish on time, resulting in delay (waiting) of the next appointment. In either case, it results in efficiency loss to the system. In some applications, the completion time of the last appointment is subject to a "session length" constraint. An overtime cost would be incurred if the completion time has exceeded the session length. In our paper, however, we do consider the idling cost and the waiting cost, but not the overtime cost. The appointment scheduling problem with session length constraint and overtime cost would require different proof framework which is out of the scope of this paper.

In general, solving the appointment scheduling problem to optimality is intractable (Robinson and Chen 2003). Even for a high quality solution with SAA (Sample Average Approximation), the number of samples is a polynomial in the number of appointments and the accuracy level (Begen et al. 2012). An alternative approach to this problem is to study the properties of the optimal schedules and derive insights for practical use. One of the most important findings from numerical study is that, when the service times are independently and identically distributed (i.i.d.), the optimal schedule exhibits a "dome" shape (Wang 1993, 1997, Denton and Gupta 2003, Kaandorp and Koole 2007, Hassin and Mendel 2008, Klassen and Yoogalingam 2009). A "dome" shape means that the job allowance increases sharply for the first few appointments, then climbs slowly before the peak, decreases slowly after the peak, and finally, descends quickly for the last few appointments. Figure 1 plots the optimal schedule that minimizes the sum of idling cost and waiting cost. The height of the curve corresponds to the job allowance and the curve exhibits a "dome" shape. To provide some intuition behind the "dome" shape, the optimal schedule will assign relatively smaller job allowances to the first few appointments because, at the beginning of the day, there are no backlogs and the risk of idleness is high. The job allowances then increase because the risk of idling decreases with the number of cumulative arrivals. During the middle of the day, the queue dynamics is stationary, and it is reasonable to expect that the job allowance remains close to a constant. Near the the end of the day, backlogs do not result in the same level of cascading impact into the future, and consequently, the job allowances decrease.



Figure 1 Dome Shape of Optimal Schedule (service time follows a normal distribution with mean 20 and variance 16, with unit idle time cost 3 and unit waiting time cost 1)

Knowing that the optimal schedule is "dome"-shaped, however, does not directly lead to an implementable schedule, because the job allowance for each appointment is still unknown. However, one may observe that the differences of the job allowances become smaller in the middle of the scheduling period. In fact, it is reported that the "dome"shaped curve may become more flattened in the middle when the number of appointments increases (Klassen and Yoogalingam 2009), or when the unit waiting cost is relatively larger compared to the unit idle cost (Hassin and Mendel 2008). Inspired by this observation, Klassen and Yoogalingam (2009) propose a simple policy, i.e., a "plateau-dome" policy, where the job allowances in the middle of a day are constrained to all be equal. They show that the plateau-dome policy performs robustly in various parametric settings.

This motivates us to consider an even simpler policy, i.e., the *constant job-allowance* policy, which simply allocates the same job allowance to every appointment. Here, a natural question may be how well the constant policy performs in comparison to more sophisticated policies, particularly, to the optimal schedule. This paper answers this question by deriving

a closed-form upper bound for the optimality gap of the constant policy when either of the following conditions holds: (1) the server idling cost is relatively small compared to the job waiting cost; (2) there are sufficiently many appointments to be scheduled in a continuous session. This upper bound also leads to the asymptotical optimality of the constant policy under the above conditions. To the best of our knowledge, this is the first theoretical performance bound derived for this classic problem. In fact, the existing studies on the properties of the optimal schedule, e.g., the dome-shape characterization, are all based on numerical analyses.

The constant policy is used widely in practice. For example, it has been used in hospitals in Ontario which provide 24-hour MRI service to patients who have booked appointments in advance.<sup>1</sup> It has also been used in many outpatient care clinics, such as the British Columbia Children's Hospital. The reasons for the constant policy being so commonly used include that this policy is easy to implement, and is also fair by assigning an equal service time slot to each patient. Our research work provides both the theoretical analysis and numerical examination for this important policy. In particular, our theory implies that the constant policy achieves near optimality either when the patient waiting cost is relatively higher than the server idling cost, or when a large number (e.g., typically more than 400) of appointments have been served during a consecutive period without service interruption. The first condition applies to many service industries which target at highvalued customers, for example, consulting service for making investments, or exclusive services for VIP customers. The second condition is suitable in a few practical appointment systems, such as the aforementioned 24-hour MRI or CT scan. It has been reported that a few hospitals have done more than 400 MRIs on average in a time window of 7-14 days until service interruption (e.g., maintenance of the MRI machine), and most of those MIRs are booked in well advance. Therefore, our analysis provides a theoretical guarantee for the constant policy when being used in those scenarios. Nevertheless, neither condition holds in a typical outpatient care setting, suggesting that there could be potential improvement by considering more sophisticated scheduling policies instead of a constant policy.

Our theoretical framework of analyzing the constant policy can be adapted to variants of model assumptions. Specifically, the same approach can be used to derive the theoretical bound for the constant policy with i.i.d. patient no-shows, given the sequence of the

 $<sup>^{1}</sup>$  According to the historical data, except for less than 2% urgent cases, the majority of the patients need to book their appointments one week before their visit date.

services. We also extend this framework to the case when the service time distributions differ in varying time blocks of a day. We propose a piecewise-constant policy in which patients in the same time block (thus with the same service time distributions) are assigned with the same job allowance. We derive theoretical upper bounds for the optimality gap of this piecewise-constant policy and show that asymptotic optimality can be achieved by a certain constant policy.

Our main contribution is three-fold as follows.

• First, under the assumption of i.i.d. service durations, we provide an explicit upper bound for the optimality gap of the constant policy under either of the aforementioned conditions. To the best of our knowledge, this is the first proven result to be within  $1 + \epsilon$ of the optimal schedule, from theoretical aspect. This upper bound implies asymptotic optimality of the constant policy, providing a theoretical basis for the implementation of such a simple policy in practice when either condition holds. Our numerical study also shows that the constant policy performs well for a small number of appointments (e.g., n = 16). Our theoretical and numerical results thus characterize appointment systems in which a constant policy achieves near optimal performance.

• Second, the theoretical bound we derive in Section 4 for the constant policy validates several findings regarding the optimal appointment schedule previously reported in the literature. For example, the upper bound for the optimality gap becomes smaller when the unit waiting cost is larger compared to the unit idling cost. This is consistent with the numerical results reported by Hassin and Mendel (2008) that the "dome" tends to be more like a constant when the ratio between waiting cost and idling cost is larger. The upper bound is also closer to zero when the number of appointments is larger. This is consistent with the numerical findings reported by the literature on optimal schedule patterns (Klassen and Yoogalingam 2009).

• Third, this is the first attempt to adapt the proof framework, which was initially developed by Goldberg et al. (2016) and Xin and Goldberg (2016) for inventory models, to the appointment scheduling setting. Such an adaption, as we will point out later in the technical analysis part, is not trivial. With this connection being made, it opens the possibility for analyzing more general settings and making more interesting findings. In fact, we have made two extensions, an appointment system with patient no-shows, or with heterogeneous service time distributions. This demonstrates the robustness of our

theoretical characterization of the constant policy. The analysis of the piecewise constant policy takes the first step in studying the optimal scheduling policy when service time distributions are not i.i.d.

The remainder of this paper is organized as follows. Section 2 is the literature review. Section 3 presents the description of the problem and the model setting. Section 4 presents the main theoretical result – an upper bound for the optimality gap of the constant policy and its asymptotical optimality, as well as a formal proof for this result. We also extend the asymptotic optimality result to incorporating patient no-shows in Section 4. Section 5 extends the result for i.i.d. service durations to the case of piecewise i.i.d. service durations, and show in this case that a piece-wise constant scheduling policy is asymptotically optimal. Section 6 presents a numerical study, where the optimal policy is approximately computed using the sample average approximation (SAA) method and compared to the optimal constant policy. Section 7 concludes the paper and discusses future research.

### 2. Literature Review

The literature on appointment scheduling has been growing rapidly in recent years, especially in the healthcare services. Most of the literature models the appointment scheduling problem by focusing on making two types of decisions: (1) determining the number of patients scheduled in time blocks, or (2) the start time for each appointment. In the models for the first type of decision (Kaandorp and Koole 2007, LaGanga and Lawrence 2012), it is often assumed that a working day is divided into multiple time blocks and each time block can accommodate multiple appointment slots. The actual consultation time of each appointment is often assumed to have deterministic length. Some literature (Robinson and Chen 2010, Zacharias and Pinedo 2014, 2017) also considers patient no-shows in the models.

In models for the second type of decision, some of the literature considers the sequence of appointments as well as the job allowance for each appointment as the decisions (Mak et al. 2014b, Chen and Robinson 2014, Mancilla and Storer 2012). However, most of the existing research assumes that the sequence of the appointments are given and only the job allowance for each appointment is to be determined (Kong et al. 2013, Mak et al. 2014a). These types of models have been studied by a rich set of literature for different service time distributions, e.g., uniform distribution (Ho and Lau 1992, Denton and Gupta 2003), normal distribution (Denton and Gupta 2003), log-normal distribution (Cayirli et al. 2008, Chen and Robinson 2014), exponential distribution (Ho and Lau 1992, Kaandorp and Koole 2007, Hassin and Mendel 2008, Zeng et al. 2010), and gamma distribution (Bailey 1952, Soriano 1966, Denton and Gupta 2003). In contrast, our analysis does not impose any assumption on the service time distribution.

A common optimization objective in the appointment scheduling literature is to minimize the weighted expected value of the waiting cost and the idling cost, given the corresponding cost rates (Robinson and Chen 2010, Yang et al. 1998, Weiss 1990, Lau and Lau 2000, Mak et al. 2014b, Kuiper et al. 2015). Problems with such an objective is referred to as the WIT (waiting and idle time) model (Klassen and Yoogalingam 2009). Examples that suit for the WIT model include 24 hour CT scan, 24 hour MRI, or other applications in which the fixed session length does not impose a hard constraint; see Weiss (1990), Lau and Lau (2000), Robinson and Chen (2003), Kuiper et al. (2015) for more examples. Some studies aim to minimize the day (session) end time (i.e., the completion time of the last appointment) (Klassen and Rohleder 2004, Hassin and Mendel 2008), referred to as the WITDET (waiting, idle, and day end time) model. Since the expected session end time is the sum of the expected idling time and expected service time, with the latter being a constant, minimizing the expected session end time is equivalent to minimizing the total idling time. Therefore, the WITDET model is equivalent to the WIT model. One will get a different model called WITOT (waiting, idle, and overtime) by assuming that the session has a fixed length and including the overtime cost (Denton and Gupta 2003, Kaandorp and Koole 2007). By including the overtime cost, the WITOT model becomes very different from the WIT, and our method can no longer apply to. However, the numerical results show that the constant policy performs near optimally even for the WITOT model; see Appendix C.2.

When there is uncertainty surrounding the parameters, the appointment scheduling problem is usually formulated as a stochastic programming problem. However, there are no efficient methods to solve stochastic programming problems in general. The most popular method, the sample average approximation (SAA) method (Begen et al. 2012), is computationally expensive and is unable to solve the model to optimality if the problem size is large. Being aware of the computational challenge in obtaining the optimal schedule, researchers turn to simple yet effective scheduling policies. For example, Bailey (1952) propose to schedule two customers at the beginning of each time block and the third one at the 2/3 mark of the time block. Researchers also investigate other scheduling policies/rules, e.g., the Yang's rule (Yang et al. 1998) and multiple block rule (Soriano 1966). For a summary of the schedule policies, we refer the reader to a survey paper by Cayirli and Veral (2003). Although these policies perform quite well in numerical experiments, and some of them also prevail in practice, there is no theoretical analysis, in particular with regards to the performance bounds for the rules, reported in the literature.

While it is challenging to compute the optimal schedule, it is desirable to identify the structural properties of the optimal schedule. In the presence of patient no-shows, Robinson and Chen (2010) prove that the optimal schedule exhibits a "no hole" structure when the scheduler's decision is the number of appointments scheduled in each time block. When the decision variables are the job allowances, Wang (1993) show that that the optimal job allowance exhibits a "dome"-shape when the service times are exponentially distributed. In recent decades, this "dome"-shape of the optimal schedule has been illustrated for i.i.d. service durations in numerical experiments (Denton and Gupta 2003, Kaandorp and Koole 2007, Hassin and Mendel 2008, Klassen and Yoogalingam 2009). However, there is scant theoretical evidence to support these findings. In this work, the constant schedule policy with i.i.d. service durations is analyzed and the asymptotic optimality is then proved.

For the "dome"-shape scheduling policy, it is reported that the shape of the "dome" is more flattened as the ratio of the idle time penalty parameter to the waiting time cost parameter decreases for the i.i.d. uniform service durations (Denton and Gupta 2003). This flattening of the "dome"-shape is also illustrated with the i.i.d. exponential service durations (Hassin and Mendel 2008). The underlying intuition of the impact of the unit idle time cost parameter is that when this parameter is high, the total cost related to idling is high, providing an incentive to reduce total idle time by scheduling the first few customers to arrive very closely or even together. Furthermore, at the end of the planning horizon, the last few patients are scheduled to arrive closer to each other to avoid the server from idling because only a few patients arrive. Our theoretical results also explain the impact of the relative costs associated with server idling and customer waiting.

9

The constant job allowance policy has been studied in a few previous works. Jansson (1966) considers an appointment booking system with constant job allowances and exponentially distributed service durations. He derives the constant job allowance which minimizes the expected total cost in such an appointment booking system (which is essentially a D/M/1 queue). Mercer (1960) and Sabria and Daganzo (1989) study the steady-state behavior of an appointment system in the presence of customer unpunctuality. Charnetski (1984) uses service times of over 2,000 real surgical procedures to evaluate the performance of a constant job allowance policy in an operating room. Wang (1997) derives the optimal constant job allowance when the service durations can be approximated by a phase-type distribution. However, those works have focused either on searching for the optimal job allowance or analyzing the steady-state queue, while our paper aims to bound the gap between the optimal constant job-allowance schedule and the optimal schedule. Chen et al. (2016) study a new appointment system in which customers are given the earliest possible appointment times under the service level constraints, and derive some asymptotic properties of this system when the number of arrivals approaches infinity.

The performance of an appointment scheduling policy depends on the probability of no-shows (Gupta and Denton 2008, Cayirli et al. 2006, 2008). It is reported that the "dome"-shape is more pronounced as the show-up probability decreases (Kaandorp and Koole 2007, Hassin and Mendel 2008). This means that high show-up probability may result in a more flattened optimal scheduling pattern, and our theoretical analysis supports this claim. During the past decade, patients' no-shows have been studied extensively in the literature such as Muthuraman and Lawley (2008), Robinson and Chen (2010), Cayirli et al. (2012), Luo et al. (2012), Zacharias and Pinedo (2014), Zacharias and Pinedo (2017), and Kong et al. (2020). A well-known strategy to offset the impact of no-shows is overbooking, i.e., booking more customers in a certain time slot than the service capacity. For example, Muthuraman and Lawley (2008) develop a stochastic overbooking policy to compensate for patients with no-shows in an outpatient clinic. LaGanga and Lawrence (2012) derive analytical bounds for the optimal number of patients to be scheduled in each time block with customers' no-shows. In addition to overbooking, other approaches can be used to mitigate the detrimental effects of patient no-shows; for example, Cayirli et al. (2012) propose a universal appointment rule to reduce the disruptive impact of no-shows. However, the theoretical analysis for appointment scheduling with patient no-shows is not considered and is usually quite challenging.

### 3. Problem Formulation

Consider a generic clinic session with a single doctor and n + 1 appointments to be scheduled in a finite time horizon (e.g., a day). The service times of those appointments are independent and identically distributed (i.i.d.) random variables, denoted by  $\mathbf{A} = (A_1, A_2, \dots, A_n)$ . (Throughout this paper, we distinguish between a random variable and its realization (or constant) using capital letter and lower case letter, respectively. We use bold-faced letters to denote vectors.) We use A to denote the random variable having the same distribution as each  $A_t$ . While the literature has characterized the optimal policies when the random service duration A follows certain distributions (Kaandorp and Koole 2007, Hassin and Mendel 2008, Zeng et al. 2010), our main results (i.e., upper bounds (13) and (29)) allow the service times to follow any continuous distribution.

We assume that the n+1 appointments, indexed by  $t = 1, 2, \dots, n+1$ , have already been sequenced. The clinic manager, being aware of the fixed sequence as well as the distribution of the i.i.d. service times, needs to determine the job allowance of each appointment, that is, the time separation between the start times of two consecutive appointments. We use  $s_t$ to denote the job allowance scheduled for the  $t^{th}$  appointment, and use  $s = (s_1, s_2, \cdots, s_n)$ to denote the vector of all job allowances. We do not include the job allowance of the last job  $s_{n+1}$  as a decision variable because that will not affect the objective value. The start time of the first appointment is 0, and thus, given s, the appointment start times are scheduled as  $\{0, s_1, s_1 + s_2, \cdots, \sum_{t=1}^n s_t\}$ , which can also be interpreted as the scheduled arrival times of patients. According to the usual practice, the scheduled arrival times have to be determined and released to patients before the beginning of the horizon (time 0) and no further adjustment can be made after the patients have arrived. We assume that all patients will show up punctually at their scheduled times, but will later relax this assumption and consider patient no-shows in Section 4.3. Finally, we assume that the job allowance for each appointment is subject to an upper bound constraint, i.e., there exists d such that  $s_t \leq d$  for  $t = 1, \ldots, n$ .

Due to the uncertainty of the service durations, a doctor may not finish serving each patient on time. If the doctor finishes serving the  $(t-1)^{th}$  patient prior to the scheduled start time of the  $t^{th}$  appointment,  $\sum_{k=1}^{t-1} s_k$ , then the doctor will stay idle until that time; otherwise, if the  $(t-1)^{th}$  patient has not completed service by the scheduled time  $\sum_{k=1}^{t-1} s_k$ , then the next patient has to wait until the service completes. For  $t = 1, \ldots, n+1$ , let  $W_t$ 

denote the waiting time of  $t^{th}$  patient and  $I_t$  denote the doctor's idle time before seeing the  $t^{th}$  patient. Neither idling nor waiting cost would be incurred after the doctor serves the last appointment. We assume that the doctor is always available at time 0, so that the first patient can see the doctor immediately, i.e.,  $W_1 = 0$ . A graphical illustration of the variables introduced above is provided in Figure 2.



Figure 2 Variables in Our Model

Note that the scheduled start time or the appointment for the  $t^{th}$  patient is  $\sum_{k=1}^{t-1} s_k$ , and the actual start time of this patient is  $\sum_{k=1}^{t-1} s_k + W_t$ , allowing for any possible waiting. Thus, the ending time for the  $t^{th}$  patient is  $\sum_{k=1}^{t-1} s_k + W_t + A_t$ . Now, since the scheduled start time for the  $(t+1)^{th}$  patient is  $\sum_{k=1}^{t} s_k$ , the actual start time of this patient is the larger of  $\sum_{k=1}^{t-1} s_k + W_t + A_t$  and  $\sum_{k=1}^{t} s_k$ , and it follows that the waiting times and idle times for the  $(t+1)^{th}$  patient can be recursively expressed as follows: for each  $t = 1, 2, \dots, n$ ,

$$W_{t+1} = [W_t + A_t - s_t]^+$$
 and (1)

$$I_{t+1} = [W_t + A_t - s_t]^- , \qquad (2)$$

where  $[x]^+ \max\{x, 0\}$  and  $[x]^- = \max\{-x, 0\}$  denote the positive and negative part of x, respectively.

Following the literature (Weiss 1990, Lau and Lau 2000, Mak et al. 2014b, Kuiper et al. 2015), we assume that the scheduler aims at minimizing the expected total cost, which consists of expected waiting cost and idle cost and has the following expression,

$$\nu(s) = E\left[\sum_{t=1}^{n} \left\{ c_W \cdot W_{t+1}(s) + c_I \cdot I_{t+1}(s) \right\} \right].$$
(3)

The parameters  $c_W$  and  $c_I$  in the above equation stand for the unit waiting time cost and idle time cost, respectively. We use the notations  $W_{t+1}(s)$  and  $I_{t+1}(s)$  to highlight their dependence on the job allowance vector s.

It is usually difficult to solve the above problem in large sizes (e.g., n > 50) to optimality in real time. Heuristic methods have been proposed in the literature to obtain near-optimal and easy-to-implement policies. In this paper, we provide a theoretical justification of the constant policy by proving its asymptotic optimality for the above problem under certain conditions. We first prove the result for i.i.d. service durations. Later, we prove an analog of the result when the service durations are allowed to have different distributions for patients scheduled in different time blocks.

# 4. The Constant Policy

Throughout this section, we assume that the service durations  $\{A_t\}$ , t = 1, 2, ..., n, are i.i.d. with the common probability density function f and cumulative distribution function F. Let  $\mu$  and  $\sigma$  represent the mean and standard deviation of  $A_t$ , respectively. In this section, we show under certain conditions that there exists a constant policy which is asymptotically optimal when the total number of appointments n grows large.

We first consider the case with n = 1. In this case, there are only two appointments, and this appointment scheduling problem is equivalent to a Newsvendor problem (Weiss 1990). The first assignment is at time 0, and the second appointment is at time  $s_1$ , which is the only decision variable here. The objective function given in (3) can be written as:

$$\nu(s_1) = E[c_W \cdot W_2(s_1) + c_I \cdot I_2(s_1)] = c_W \cdot E[A_1 - s_1]^+ + c_I \cdot E[A_1 - s_1]^-$$

where the last equality holds since (1) and (2) imply  $W_2 = [A_1 - s_1]^+$  and  $I_2 = [A_1 - s_1]^-$ . It is easy to verify that this function is a Newsvendor cost function, which is convex in  $s_1$ . Let  $\underline{s}$  and  $\underline{g}$  denote the minimizer and the minimum value of  $\nu(s_1)$  given above, i.e.,

$$\underline{s} = \arg\min_{s \in \Re^+} c_W \cdot E[A-s]^+ + c_I \cdot E[A-s]^- \quad \text{and} \quad (4)$$

$$g = \min_{s \in \Re^+} c_W \cdot E[A - s]^+ + c_I \cdot E[A - s]^- .$$
(5)

Note that the above minimization problem faces uncertainty arising only from the service duration A. In the problem with an arbitrary number of appointments, the job allowance for the  $t^{th}$  patient needs to consider not only the service duration  $A_{t+1}$  for the next patient but also possibly delay in starting the service for the  $t^{th}$  patient. With this additional uncertainty, it can be shown that  $\underline{s}$  is a lower bound on the job allowance for the  $t^{th}$  patient under the optimal policy, denoted by  $s_t^*$ . Similarly, g is a lower bound on the cost associated with the  $t^{th}$  patient given by  $C_t = c_W \cdot W_{t+1} + c_I \cdot I_{t+1}$ . Let  $C_t^*$  denote this cost under the optimal policy. These findings are summarized in the next lemma. The proof of the lemma is attached in Appendix A.1.

LEMMA 1. For  $t \in \{1, \ldots, n\}$ ,  $s_t^* \geq \underline{s}$  and  $E[C_t^*] \geq g$ .

### 4.1. Statement of Theorem 1 and Discussion

To formally state the main result of this section, Theorem 1, we need to introduce several notations. Recall that  $c_W$  and  $c_I$  are the per-unit cost parameters associated with the patient's waiting and the server's idling. Also, F is the cumulative distribution function of the service time A. Furthermore, d is an upper bound imposed on the choice of each  $s_t$ , i.e.,  $s_t \leq d$  for t = 1, ..., n. Let  $\zeta$  denote the skewness of the random service time A and define  $\tau$  to be a function of  $\zeta$  and two cost parameters  $c_W$  and  $c_I$ :

$$\tau = \left[ \left\{ 8 \left( 3\zeta + \sqrt{\frac{2c_I}{c_W}} + \sqrt{\left(\frac{c_I}{2c_W}\right)^3} \right) \right\}^2 \right] \quad \text{where} \quad \zeta = \frac{E[|A - \mu|^3]}{\sigma^3} . \quad (6)$$

From the definition of  $\tau$ , larger skewness of the service duration distribution or a large ratio of  $c_I/c_W$  would lead to a large value of  $\tau$ . For fixed r, we let random variable  $\eta_{\infty}^r$ denote the hitting time of the maximum of a random walk. In the random walk that we consider, the increment quantity in each epoch i is  $A_i - r$ . Define, for each  $t \in \{0, 1, 2, \ldots\}$ , the following random variable:

$$\eta_t^r = \min\left\{j^* \in \{0, 1, \dots, t\} \mid \sum_{i=1}^{j^*} (A_i - r) = \max_{j \in \{0, 1, \dots, t\}} \sum_{i=1}^{j} (A_i - r)\right\} .$$
(7)

Without loss of generality, we define  $\sum_{i=l}^{k} \cdot = 0$  for  $0 \le k < l$  in this paper. Since the summation in the above equation represents a random walk, an explicit, but still complicated, expression for the probability distribution of  $\eta_t^r$  is given in Theorem 1 of Andersen (1955).

Then, for each  $n \ge 1$ , define

$$\epsilon_n(r) := \frac{c_W r E[(\eta_n^r)^2 - \eta_n^r]}{ng} + \frac{c_I \sigma^2}{2ng(r - \mu)} .$$
(8)

 $\epsilon_n(r)$  will be used to provide a bound on the performance of the constant policy. In its expression above, the second term comes from the bound for the expected waiting of a steady D/G/1 queue. The quantity  $E[(\eta_n^r)^2 - \eta_n^r]$  in the first term can be numerically computed and be upper bounded uniformly in n; see the next lemma. Both the numerical algorithm and the derivation of the upper bound are provided in Appendix A.2.

LEMMA 2. When  $r > \mu$ ,  $E[(\eta_n^r)^2 - \eta_n^r]$  has a uniform upper bound

$$E[(\eta_n^r)^2 - \eta_n^r] \le \frac{2\varphi_r^2}{(1 - \varphi_r)^2},\tag{9}$$

where the scalar  $\varphi_r$  is defined as follows,

$$\varphi_r = \inf_{\psi \ge 0} \phi_r(\psi) \quad and \quad \psi_r = \arg \inf_{\psi \ge 0} \phi_r(\psi), \qquad where \qquad \phi_r(\psi) = E[\exp(\psi(A-r))].$$
(10)

To elaborate on the upper bound (9),  $\phi_r(\psi)$  can be regarded as the moment generating function of random variable A - r. The function attains its minimal value  $\varphi_r$  when  $\psi = \psi_r$ . The proof of Lemma 2 is provided in Appendix A.2.

A constant policy r refers to a schedule in which all job allowances are set to a constant value. We use  $\nu(r)$  and  $\nu^*$  to denote the expected total cost under the constant policy r and the optimal (possibly non-constant) policy, respectively. Let  $r^*$  denote the job allowance used by the optimal constant policy, so  $r^* = \arg\min\{\nu(r)|r \ge 0\}$ . We use a vector  $s^*$  to represent the sequence of job allowances under the optimal policy, where  $s_t^*$  denotes the job allowance of the  $t^{th}$  appointment. We use  $\overline{s}^*$  to denote the average job allowance under the optimal policy, that is,

$$\overline{s}^* = \frac{1}{n} \sum_{t=1}^n s_t^*.$$
 (11)

THEOREM 1. Suppose at least one of the following conditions hold:

- Condition 1:  $F(\mu) < c_W / (c_W + c_I);$
- Condition 2:  $n \ge \max\{2\tau, 6\tau^{\frac{3}{2}}\sigma^{-1}(d-\underline{s})\}.$

Then we have  $\overline{s}^* > \mu$ , and

$$\nu^* \leq \nu(r^*) \leq \nu(\overline{s}^*) \leq (1 + \epsilon_n(\overline{s}^*))\nu^* .$$
(12)

Furthermore,

$$\epsilon_n(\overline{s}^*) \leq \frac{2c_W \overline{s}^* \varphi_{\overline{s}^*}^2}{ng(1-\varphi_{\overline{s}^*})^2} + \frac{c_I \sigma^2}{2ng(\overline{s}^*-\mu)} \to 0 \qquad as \ n \to \infty.$$
(13)

While the formal proof of this theorem is provided in Section 4.2, we make several comments regarding the statement of the theorem.

REMARK 1. The relationship in (12) can be written as  $0 \le \nu(\bar{s}^*)/\nu^* - 1 \le \epsilon_n(\bar{s}^*)$ . We refer to the ratio  $\nu(\bar{s}^*)/\nu^* - 1$  as the *relative optimality gap* for the constant policy r. Theorem 1 provides two upper bounds for the relative optimality gap: one is the expression of  $\epsilon_n(\bar{s}^*)$  on the right-hand-side of (8) and the other in (13). The first upper bound can be computed only when  $A_t$  follows a given continuous distributions, for example, a normal distribution. The second upper bound is weaker than the first one, but it has better analytic properties, providing qualitative insights. For example, the second upper bound monotonically decreases with n, suggesting that the constant policy has a better performance guarantee when the time horizon is longer. The limiting property in (13) shows that a simple constant policy  $\overline{s}^*$  is asymptotically optimal.

REMARK 2. For general constant policy r, the optimality gap, which refers to the difference  $\nu(r) - \nu(s^*)$ , cannot be computed explicitly without knowing the optimal schedule  $s^*$ . However, by selecting  $r = \bar{s}^*$ , the relative optimality gap can be bounded by the right-handside of Equation (13), which converges to 0 as  $n \to \infty$ . In fact, since the optimal constant policy  $r^*$  always outperforms the constant policy r, we expect the gap  $\nu(r^*) - \nu(s^*)$  can be even smaller. Our numerical experiments in Appendix C.3 show that the values of  $r^*$ and  $\bar{s}^*$  stay very close, which suggests that we have not relaxed much by analyzing the constant policy  $\bar{s}^*$  instead of  $r^*$  in our analysis. For practical applications, however, we always use  $r^*$  instead of  $\bar{s}^*$ , because  $r^*$  can be efficiently computed by a one-dimensional search, whereas computation of  $\bar{s}^*$  requires the knowledge of optimal schedule  $s^*$  and is much more involved.

We comment on Conditions 1 and 2 used in Theorem 1. Each of them provides a sufficient condition for the value of  $\overline{s}^* = (s_1^* + \cdots + s_n^*)/n$  given in (11) to satisfy  $\overline{s}^* > \mu$ . This ensures that  $\overline{s}^*$  is sufficiently high so that the random walk used in (14) has a negative drift, implying that  $\eta_{\infty}^r$  has a finite expectation.

### LEMMA 3. If at least one of the two conditions in Theorem 1 holds, then $\overline{s}^* > \mu$ .

While a rigorous proof for the above lemma is provided in Appendix A.3, we provide below some key ideas in the proof. The main part of Condition 1,  $F(\mu) < c_W/(c_W + c_I)$ , ensures that  $\mu < \underline{s}$  since  $\underline{s}$  defined in (4) is a newsvendor solution satisfying  $F(\underline{s}) = c_W/(c_W + c_I)$ . Since Lemma 1 implies  $\overline{s}^* = (s_1^* + \cdots + s_n^*)/n \ge \underline{s}$ , it follows that  $\overline{s}^* > \mu$ . An implication of Condition 1 is that when the ratio  $c_W/c_I$  is sufficiently large,  $c_W/(c_W + c_I)$  is close to 1, satisfying Condition 1. Then, for any value of  $n \ge 1$ , the bound in (12) is valid and the asymptotic optimality holds. This is consistent with the numerical studies in Denton and Gupta (2003) and Hassin and Mendel (2008), which show that the "dome"-shape becomes more flattened when the ratio  $c_W/c_I$  is larger. When Condition 1 fails, the unit waiting cost  $c_W$  is relatively small compared to the doctor idling cost parameter  $c_I$ . In this case, the scheduler would rather keep the patients waiting than making the doctor idle, and may choose the job allowances to be small. This reasoning is valid when the total number of appointments, n, is small. However, when n is large and approaches infinity, the job allowances under the optimal policy  $s^* = (s_1^*, s_2^*, \dots, s_n^*)$  should be large enough so that their average,  $\overline{s}^*$ , should be at least  $\mu$ ; otherwise, the policy would result in very high waiting time by queuing theory, failing to be optimal. Condition 2 spells the a sufficient condition on how large n should be.

We now elaborate on  $\eta_t^r$ , which is used to calculate  $\epsilon_n(r)$  for any  $r > \mu$ . We explore its connection to the waiting time. Let  $W_{t+1}^r$  denote the waiting time of the  $(t+1)^{th}$  patient under this constant policy r. Using the Lindley's recursion and the recursive definition of  $W_t$  given in (1), we have

$$W_{t+1}^r = [W_t + A_t - r]^+ = \max\{0, W_t + A_t - r\}$$
  
=  $\max\{0, A_t - r, W_{t-1} + A_{t-1} - r\} = \cdots = \max_{j \in \{0, 1, \dots, t\}} \left\{ \sum_{i=t+1-j}^t (A_i - r) \right\}.$ 

From the i.i.d. assumption of  $A_i$ 's, the rightmost expression above has the same distribution as  $\max_{j \in \{0,1,\dots,t\}} \left\{ \sum_{i=1}^{j} (A_i - r) \right\}$ , which in turn has the same distribution as  $W_{t+1}^r$ , i.e.,

$$W_{t+1}^r \stackrel{d}{=} \max_{j \in \{0,1,\dots,t\}} \left\{ \sum_{i=1}^j (A_i - r) \right\} .$$
(14)

We have defined  $\eta_t^r$  in (7) to be a random variable representing the index corresponding to the maximum of the random walk used in the right-side of (14), where ties are broken in favor of the smaller index.

Finally, we elaborate on why our analysis is restricted to constant policies with  $r > \mu$ . Note that  $\epsilon_n(r)$  is an upper bound on the relative optimality gap. According to classic random walk theory (see e.g., Karlin 2014),  $\eta_n^r$ , used in the definition of  $\epsilon_n(r)$  given in (8), has finite first and second moments if and only if  $r > \mu$ . Thus,  $E[(\eta_n^r)^2 - \eta_n^r] < +\infty$  is finite if and only if  $r > \mu$ . Therefore,  $r > \mu$  is a sufficient and necessary condition for the upper bound to converge to 0.

### 4.2. Proof of Theorem 1

Now we prove Theorem 1. We first derive an alternative expression for  $\nu(r) - \nu^*$  for general  $r > \mu$ , and provide an upper bound for this difference. We then show that this upper bound has a simpler expression when we choose  $r = \overline{s}^*$ . Finally, by lower bounding  $\nu^*$ , we derive an upper bound for the relative optimality gap  $(\nu(\overline{s}^*) - \nu^*)/\nu^*$ .

The following lemma derives an alternative expression for  $\nu(s)$ , originally defined in (3). A proof for this lemma is in Appendix A.4.

LEMMA 4. For any given schedule s,

$$\nu(s) = E\left[\sum_{t=1}^{n} c_{W}W_{t+1}\right] + c_{I}\sum_{t=1}^{n} s_{t} + c_{I}E[W_{n+1}] - c_{I}n\mu .$$

According to Lemma 4, for any constant policy r, we have

$$\nu(r) - \nu(s^*) = c_W \sum_{t=1}^n E\left[W_{t+1}^r - W_{t+1}^*\right] + c_I \left(nr - \sum_{t=1}^n s_t^*\right) + c_I E\left[W_{n+1}^r - W_{n+1}^*\right], \quad (15)$$

where  $W_t^r$  and  $W_t^*$  represent the waiting time under a constant policy r and the optimal policy, respectively. Next we upper bound each of the three terms on the right side of the above equation.

To upper bound the first term, consider an alternative expression for  $E[W_{t+1}^r]$  using (14), and the definition of  $\eta_t^r$  given in (7):

$$E\left[W_{t+1}^{r}\right] = E\left[\max_{j\in\{0,\cdots,t\}}\left\{\sum_{i=1}^{j}(A_{i}-r)\right\}\right] = E\left[\left(\sum_{i=1}^{\eta_{t}^{r}}A_{i}\right) - \eta_{t}^{r}r\right]$$

Now, we obtain the following lower bound for  $W_t^*$ , we use an argument similar to (14):

$$E\left[W_{t+1}^*\right] = E\left[\max_{j=0,1,\cdots,t} \left\{\sum_{i=t+1-j}^t (A_i - s_i^*)\right\}\right]$$
$$\geq E\left[\sum_{i=t+1-\eta_t^r}^t (A_i - s_i^*)\right] = E\left[\left(\sum_{i=1}^{\eta_t^r} A_i\right) - \left(\sum_{i=t+1-\eta_t^r}^t s_i^*\right)\right],$$

where the inequality holds since we have used  $\eta_t^r$  instead of the true maximizer, and the second equality holds since  $A_i$ 's are i.i.d.. Then,

$$E\left[W_{t+1}^r - W_{t+1}^*\right] \leq E\left[\left(\sum_{i=t+1-\eta_t^r}^t s_i^*\right) - r\eta_t^r\right].$$
(16)

Thus, we have

$$\sum_{t=1}^{n} E\left[W_{t+1}^{r} - W_{t+1}^{*}\right] \leq \sum_{t=1}^{n} E\left[\left(\sum_{i=t+1-\eta_{t}^{r}}^{t} s_{i}^{*}\right) - r\eta_{t}^{r}\right] \leq r E\left[(\eta_{n}^{r})^{2} - \eta_{n}^{r}\right], \quad (17)$$

where the second inequality holds provided that the following Lemma 5 holds. The proof of the lemma below is in Appendix A.5.

Lemma 5.

$$\sum_{t=1}^{n} E\left[\sum_{i=t+1-\eta_{t}^{r}}^{t} s_{i}^{*}\right] - \sum_{t=1}^{n} r E\left[\eta_{t}^{r}\right] \leq r E\left[(\eta_{n}^{r})^{2} - \eta_{n}^{r}\right] .$$

Hence, with (17), the first term in (15) can be bound as follows

$$c_W \sum_{t=1}^n E\left[W_{t+1}^r - W_{t+1}^*\right] \leq c_W r E[(\eta_n^r)^2 - \eta_n^r] .$$
(18)

To upper bound the third term in (15), which is  $c_I E[W_{n+1}^r - W_{n+1}^*]$  when  $r > \mu$ , it suffices to upper bound  $E[W_{n+1}^r]$ . Note from the recursive definition given in (1) that  $W_t^r$  can be regarded as the waiting time of  $t^{th}$  customer in a D/G/1 queue with constant interarrival time r and random service time A, with an initial state  $W_1^r = 0$ . By the property of D/G/1 queue, the expected value  $E[W_t^r]$  monotonically increases from 0 to  $E[W_{\infty}^r]$  (see, for example, Theorem 4 of Kingman (1962)). Thus, we have

$$E[W_t^r] \le E[W_\infty^r] \qquad \text{for all } t \ge 1.$$
(19)

We now state a lemma that provides an analytic upper bound for  $E[W_{\infty}^r]$ . Let  $\rho = \mu/r$ . Also, define

$$\overline{\Upsilon}_1 = \frac{\sigma^2}{2r(1-\rho)}$$
 and  $\overline{\Upsilon}_2 = \frac{4}{e^2(\psi_r)^2}\log\frac{1}{1-\varphi_r} + \left[\frac{\sigma^2}{2r(1-\rho)}\right]^2$ . (20)

LEMMA 6. Consider a D/G/1 queue with a deterministic inter-arrival time r and random service time A, with mean  $\mu$  and standard deviation  $\sigma$ . If  $\rho = \mu/r < 1$ , then  $\psi_r > 0$ holds, and the steady-state waiting distribution  $W_{\infty}$  has the following upper bounds on the first and second moments:

$$E[W_{\infty}] \leq \overline{\Upsilon}_1 \quad and \quad E[W_{\infty}^2] \leq \overline{\Upsilon}_2$$

The first part of the lemma,  $\psi_r > 0$ , is due to Kingman (1962), and the bound for the first moment appears in Marshall (1968). The bound for the second moment comes from Theorem 6 of Kingman (1962).

An implication of the above lemma is an upper bound on the third term in (15). Since  $r > \mu$  holds by Lemma 3, we can apply Lemma 6 to obtain

$$c_I E[W_{n+1}^r - W_{n+1}^*] \le c_I E[W_{n+1}^r] \le c_I \overline{\Upsilon}_1 = \frac{c_I \sigma^2}{2r(1-\rho)}.$$
 (21)

We have thus derived upper bounds in (18) and (21) for the first and the third terms at the right-hand-side of (15). Furthermore, the second term at the right-hand-side of (15) equals to zero when we choose  $r = \overline{s}^*$ . Therefore, we have

$$\nu(\bar{s}^*) - \nu(s^*) \leq c_W \bar{s}^* E[(\eta_n^{\bar{s}^*})^2 - \eta_n^{\bar{s}^*}] + 0 + \frac{c_I \sigma^2}{2\bar{s}^*(1-\rho)}$$

Since Lemma 1 states  $E[C_t^*] \ge g$  for each t, it follows that  $\nu^* = \sum_{t=1}^n E[C_t^*] \ge ng$ . Thus,

$$\frac{\nu(\overline{s}^*) - \nu^*}{\nu^*} \le \frac{c_W \overline{s}^* E[(\eta_n^{\overline{s}^*})^2 - \eta_n^{\overline{s}^*}]}{ng} + \frac{1}{ng} \cdot \frac{c_I \sigma^2}{2\overline{s}^* (1 - \rho)} = \epsilon_n(\overline{s}^*)$$

where the last equality follows from the definition of  $\epsilon_n(\cdot)$  in (8). This establishes (12), the first result in the statement of Theorem 1.

Finally, by substituting the upper bound for  $E[(\eta_n^r)^2 - \eta_n^r]$  (i.e., (9)) to the expression of  $\epsilon_n(r)$  (i.e., (8)), we obtain

$$\epsilon_n(r) = \frac{c_W r E[(\eta_n^r)^2 - \eta_n^r]}{ng} + \frac{c_I \sigma^2}{2ng(r-\mu)} \le \frac{2c_W r \varphi_r^2}{ng(1-\varphi_r)^2} + \frac{c_I \sigma^2}{2ng(r-\mu)}$$

which is the inequality in (13) in the statement of Theorem 1. It is easy to show that the rightmost expression above converges to 0 as  $n \to \infty$ . This completes the proof of the theorem.

We make a few comments on the proof of Theorem 1. The structure of our proof follows the same framework as Goldberg et al. (2016) and Xin and Goldberg (2016). These two papers have studied a constant order quantity policy for a lost sales inventory control problem. However, an adaptation of their idea to the appointment scheduling setting requires additional work for two reasons. First, the recursive formulae of the waiting time and a corresponding quantity in inventory papers are different:  $W_{t+1} = [W_t + s_t - A_t]^+$  in the lost sales model and  $W_{t+1} = [W_t + A_t - s_t]^+$  in the appointment scheduling model. Despite apparent similarity, the two formulae lead to a different analysis. For example, Lemma 5 in our proof has used a different method than the counterpart of Theorem 2 in Goldberg et al. (2016). The random walk used in (7) of our analysis starts with an initially empty state whereas the random walk used in Goldberg et al. (2016) starts from a steady state; in analyzing  $\eta_t^r$ , we cannot use a key result identified in Lemma 6 of Goldberg et al. (2016), and have to derive a different relationship (see, for example,  $\eta_t^r$  used in (A.16) in the appendix). Second, the inventory model in the literature and the appointment scheduling model studied here use different asymptotic regimes to establish asymptotic optimality of the constant order quantity or job allowance policy. In the inventory models by Goldberg et al. (2016) and Xin and Goldberg (2016), there are two time durations, the first one being the fixed lead time L that determines the state of the dynamic programming and the second one being the rolling horizon T of decision epochs. The asymptotic optimality in Goldberg et al. (2016) requires that  $L \to \infty$  and  $T/L \to \infty$  as the relative performance gap becomes close to 0. In Xin and Goldberg (2016), they use the asymptotic regime where  $T \to \infty$  while L is fixed. Thus, in both of these papers, it is assumed that  $T/L \to \infty$ holds, implying that the planning horizon contains infinitely many information updates. In contrast, in our paper, we have no information updates, which is essentially equivalent to L = T, which is not covered by Goldberg et al. (2016) and Xin and Goldberg (2016).

It would be challenging to further strengthen the upper bound  $\epsilon_n(r)$ , as it is difficult to further tighten the upper bound in Lemma 5.

### 4.3. When Patients Have No-Shows

An underlying assumption in Theorem 1 is that all the patients would show up. We now discuss briefly how Theorem 1 can be extended to cover the case with the possibility of patient no-shows. To model no-shows, we introduce a random binary variable  $Z_t \in \{0, 1\}$  to indicate whether the  $t^{th}$  patient in the sequence shows up for her appointment  $(Z_t = 1)$  or not  $(Z_t = 0)$ . We assume  $Z_t$ 's are i.i.d. and have an average show-up probability p, i.e.,  $E[Z_t] = p$ .

The key idea in our approach is that we treat no-show patients as regular patients who show up, but consider the service times of these patients to be 0. More precisely, let  $(\check{A}_1, \check{A}_2, \dots, \check{A}_n)$  denote the service duration vector in the presence of no-shows. Then, for each  $t \in \{1, 2, \dots, n\}$ , we have  $\check{A}_t = Z_t A_t$ , where  $(A_1, \dots, A_n)$  is the service duration vector without no-shows studied previously. On this basis, the waiting times and idling times account for no-shows, which we now denote by  $\{\check{W}_t\}$  and  $\{\check{I}_t\}$ , respectively, can be adapted from (1) and (2) as follows: for  $t \in \{1, ..., n\}$ ,

$$\check{W}_{t+1} = [\check{W}_t + \check{A}_t - s_t]^+$$
 and  
 $\check{I}_{t+1} = [\check{W}_t + \check{A}_t - s_t]^-$ ,

where  $\check{W}_1 = 0$ .

When we count the waiting cost, we only need to count the waiting time of those who actually show up. Thus, the total cost for a given schedule s can be formulated as, similar to (3),

$$\nu(\boldsymbol{s}) = E\left[\sum_{t=1}^{n} \left\{ c_W \cdot Z_{t+1} \cdot \check{W}_{t+1}(\boldsymbol{s}) + c_I \cdot \check{I}_{t+1}(\boldsymbol{s}) \right\} \right].$$

Since  $Z_{t+1}$  is independent of  $(Z_1, Z_2, \dots, Z_t)$ , it must be independent of  $W_{t+1}$ . Thus,  $E[Z_{t+1}\check{W}_{t+1}] = pE[\check{W}_{t+1}]$ . Note that the idle time before the  $(t+1)^{th}$  appointment,  $I_{t+1}$ , does not depend on whether the  $(t+1)^{th}$  patient would show up or not. Therefore, omitting the explicit dependence on s, we have

$$\nu(\mathbf{s}) = E\left[\sum_{t=1}^{n} \left\{ (c_W p) \cdot \check{W}_{t+1} + c_I \cdot \check{I}_{t+1} \right\} \right]$$
  
=  $E\left[\sum_{t=1}^{n} \left\{ \check{c}_W \cdot [\check{W}_t + \check{A}_t - s_t]^+ + c_I \cdot [\check{W}_t + \check{A}_t - s_t]^- \right\} \right]$ 

where  $\check{c}_W = c_W p$ . Thus, with the modified service durations  $\check{A}_t$  and modified unit waiting cost  $\check{c}_W$ , we can apply Theorem 1 and derive analogous results.

# 5. Multiple Types of Appointments: The Piecewise Constant Policy

In this section, we relax the assumption of homogeneous appointments by considering a scenario where appointments are classified into  $M \ge 2$  types. The classification can be based on factors such as a patient's physiological attributes, time preference (e.g., morning or afternoon), first visit or revisit, etc. Appointments of the same type follow the identical service duration distribution, which is known by the scheduler as a priori. We let

$$n^m = q^m n$$

denote the number of appointments of type m, where n denotes the total number of patients of all types, and  $q^m$  denotes the proportion of the type-m appointments. Let  $F^m(\cdot)$  be the cumulative distribution function of the service duration of the type-*m* appointments, and we denote its mean and standard deviation by  $\mu^m$ , and  $\sigma^m$ , respectively. Appointments of the same type have i.i.d. service durations.

We consider a scenario in which appointments in the same day are sequenced in the following way: type-1 appointments are scheduled to the first  $n^1$  slots, say, the first block; type-2 appointments are scheduled to slots between  $n^1 + 1$  till  $n^2$ , say, the second block, etc. Such a sequencing strategy is practical because the classification of appointment types have taken into account a patient's time preferences. For example, patients who prefer morning slots can be scheduled in the morning. The scheduler can further classify the morning slots into multiple types and schedule them into different blocks. Patients usually have some flexibility with their appointment times, as long as their preference to morning or afternoon slots has been respected.

The policy that we consider in this section, called the *piecewise constant policy*, schedules the same job allowance for all appointments of the same type. We show that, under the above assumptions, this policy is asymptotically optimal. To facilitate the discussion, we use the superscript m and subscript t to index the blocks and the position in that block, respectively; for example,  $A_t^m$  stands for the random service duration of the  $t^{th}$  patient in the  $m^{th}$  block. Let  $s_t^{m*}$  denote the optimal job allowance of the  $t^{th}$  patient in the  $m^{th}$  block under the optimal schedule. Let

$$\bar{s}^{m*} = \frac{1}{n^m} \sum_{t=1}^{n^m} s_t^{m*}$$
(22)

denote the average job allowance of type-*m* appointments under the optimal schedule. We consider a piecewise constant policy  $\overline{s}^* = (\overline{s}^{1*}, \overline{s}^{2*}, \dots, \overline{s}^{M*})$ , where  $\overline{s}^*$  is bolded to signify that it is a vector representing a piecewise constant policy. Consistent with the previous section, we also use  $\nu(\overline{s}^*)$  and  $\nu^*$  to denote the expected total cost under the piecewise constant policy  $\overline{s}^*$  and the optimal schedule, respectively.

#### 5.1. Statement of Theorem 2 and Discussion

We present our result in Theorem 2 below, which provides an approximation ratio for the piecewise constant policy. Most of the notations used in Theorem 2 follow the same definition and interpretation that we introduced in Section 4 for Theorem 1, by adapting to a system which consists of only one type of patients. For  $m \in \{1, \ldots, M\}$ , we define  $\underline{s}^m$  and  $g^m$  in the exact same way as  $\underline{s}$  and g given in (4) and (5), i.e., the minimizer and minimum value of a Newsvendor function:

$$\underline{s}^{m} = \arg\min_{s\in\Re^{+}} c_{W} \cdot E[A^{m} - s]^{+} + c_{I} \cdot E[A^{m} - s]^{-} \quad \text{and} \\ g^{m} = \min_{s\in\Re^{+}} c_{W} \cdot E[A^{m} - s]^{+} + c_{I} \cdot E[A^{m} - s]^{-} .$$
(23)

where  $A^m$  denotes the i.i.d. service duration for type m. Similar to (6), define  $\tau^m$  to be the skewness of  $A^m$  and define  $\zeta^m$  as a function of  $\tau^m$ :

$$\tau^{m} = \left[ \left\{ 8 \left( 3(\zeta^{m}) + \sqrt{\frac{2c_{I}}{c_{W}}} + \sqrt{\left(\frac{c_{I}}{2c_{W}}\right)^{3}} \right) \right\}^{2} \right] \quad \text{where} \quad \zeta^{m} = \frac{E[|A^{m} - \mu^{m}|^{3}]}{(\sigma^{m})^{3}} \quad (\zeta^{m} - \mu^{m})^{3} \quad ($$

Let  $\rho^m = \mu^m / \overline{s}^{m*}$ . Similar to (10), define for  $\overline{s}^{m*} > \mu^m$ ,

$$\varphi_{\overline{s}^{m*}}^m = \inf_{\psi \ge 0} \phi_{\overline{s}^{m*}}^m(\psi) \quad \text{and} \quad \psi_{\overline{s}^{m*}}^m = \arg \inf_{\psi \ge 0} \phi_{\overline{s}^{m*}}^m(\psi) \quad \text{where} \quad \phi_{\overline{s}^{m*}}^m(\psi) = E[\exp(\psi(A^m - \overline{s}^{m*}))]$$

$$\tag{24}$$

Similarly, we can define  $\phi_{\bar{r}^m}^m(\psi)$ ,  $\psi_{\bar{r}^m}^m$ ,  $\varphi_{\bar{r}^m}^m$  for a different constant policy  $\bar{r}^m$ , with

$$\bar{r}^m = \bar{s}^{m*} - \frac{1}{\bar{K}^m}, \quad \text{where} \quad \bar{K}^m = \left\lceil \frac{2}{\bar{s}^{m*} - \mu^m} \right\rceil .$$
 (25)

Thus,  $\bar{r}^m > \mu^m$ .

We also define  $\eta_t^{m,\overline{s}^{m*}}$  in the same way as  $\eta_t^r$  in (7) except that  $A_i$  and r are replaced by  $A_i^m$  and  $\overline{s}_i^{m*}$ , i.e.,  $\eta_t^{m,\overline{s}^{m*}}$  is the first index when the random walk  $\sum_{k=1}^t (A_k^m - \overline{s}^{m*})$  achieves its maximum value:

$$\eta_t^{m,\overline{s}^{m*}} = \min\left\{ j^* \in \{0, 1, \dots, t\} \mid \sum_{i=1}^{j^*} (A_i^m - \overline{s}^{m*}) = \max_{j \in \{0, 1, \dots, t\}} \sum_{i=1}^j (A_i^m - \overline{s}^{m*}) \right\} .$$

Furthermore, let  $\eta_{\infty}^{m,\overline{s}^{m*}}$  be the limiting distribution of  $\eta_t^{m,\overline{s}^{m*}}$  as  $t \to \infty$ . We let, similar to (20),

$$\overline{\Upsilon}_{1}^{m} = \frac{(\sigma^{m})^{2}}{2\overline{s}^{m*}(1-\rho^{m})} \quad \text{and} \quad \overline{\Upsilon}_{2}^{m} = \frac{4}{e^{2}(\psi_{\overline{s}^{m*}}^{m})^{2}}\log\frac{1}{1-\varphi_{\overline{s}^{m*}}^{m}} + \left[\frac{(\sigma^{m})^{2}}{2\overline{s}^{m*}(1-\rho^{m})}\right]^{2} ,$$

and define

$$\Omega^m = (m-1)\bar{K}^m \sum_{k=1}^{m-1} \overline{\Upsilon}_2^k + \frac{\tilde{\varphi}}{1-\tilde{\varphi}} \sum_{k=1}^{m-1} \overline{\Upsilon}_1^k$$
(26)

where

$$\tilde{\varphi} = \max_{m \in \{1,2,\dots,M\}} \varphi^m_{\bar{r}^m} .$$
(27)

Finally, similar to the definition of  $\epsilon_n(\overline{s}^*)$  in (8), we define

$$\epsilon_n(\overline{\boldsymbol{s}}^*) = \frac{c_W \sum_{m=1}^M \left\{ \overline{\boldsymbol{s}}^{m*} E\left[ \left( \eta_n^{m, \overline{\boldsymbol{s}}^{m*}} \right)^2 - \eta_n^{m, \overline{\boldsymbol{s}}^{m*}} \right] + \Omega^m \right\} + c_I \sum_{m=1}^M \overline{\Upsilon}_1^m}{\sum_{m=1}^M n^m g^m} .$$
(28)

THEOREM 2. Suppose at least one of the following conditions hold for each  $m \in \{1, ..., M\}$ :

- Condition 1':  $F^m(\mu^m) < c_W/(c_W + c_I);$
- Condition 2':  $n^m \ge \max\{2(\tau^m), 6(\tau^m)^{\frac{3}{2}}(\sigma^m)^{-1}(d-\underline{s}^m)\}.$

Let  $n^m \ge 1$  hold for each  $m \in \{1, \ldots, M\}$ . Then, we have  $\overline{s}^{m*} > \mu^m$  for each  $m \in \{1, \ldots, M\}$ , and

$$\nu^* \leq \nu(\overline{\boldsymbol{s}}^*) \leq (1 + \epsilon_n(\overline{\boldsymbol{s}}^*))\nu^*$$

Furthermore,

$$\epsilon_n(\overline{\boldsymbol{s}}^*) \leq \frac{c_W \sum_{m=1}^M \left\{ \frac{2\overline{\boldsymbol{s}}^{m*} \left(\varphi_{\overline{\boldsymbol{s}}^{m*}}^m\right)^2}{(1-\varphi_{\overline{\boldsymbol{s}}^{m*}}^m)^2} + \Omega^m \right\} + c_I \sum_{m=1}^M \overline{\Upsilon}_1^m}{\sum_{m=1}^M n^m g^m} \to 0 \qquad \text{as } n \to \infty.$$
(29)

Similar to Theorem 1, Theorem 2 provides two upper bounds for the relative optimality gap  $\nu(\bar{s}^*)/\nu^* - 1$ . The first upper bound, which is  $\epsilon_n(\cdot)$  itself given in (28), can be computed only when the inter-arrival times  $\{A_t^m\}$  follow certain distributions, e.g., a normal distribution. The second bound, given in (29), is an upper bound on  $\epsilon_n(\cdot)$ , and it converges to zero when  $n \to \infty$ , proving the asymptotic optimality of the piecewise constant policy  $\bar{s}^*$ .

Condition 1' and 2' are simply adapting Condition 1 and 2 to a system with only one type of appointment. By Lemma 3, Condition 1' and 2' provide a sufficient condition for  $\bar{s}^{m*} > \mu^m$  for each m, which gives a sufficient and necessary condition for the first and second moment of  $\eta_{\infty}^{m,\bar{s}^{m*}}$  to be finite. To achieve this, we consider the waiting-time distributions  $\{W_t^{m,\bar{s}^{m*}}\}$ , where t = 1, 2, ..., in a D/G/1 queue with the constant inter-arrival time  $\bar{s}^{m*}$ and random service times  $\{A_t^m\}$ , where the system is initially empty. In our analysis, we use the following bounds, given in Lemma 6 of Section 4, on the first two moments of the steady-state waiting-time distribution  $W_{\infty}^{m,\bar{s}^{m*}}$ :

$$E\left[W_{\infty}^{m,\overline{s}^{m*}}\right] \leq \overline{\Upsilon}_{1}^{m} \quad \text{and} \quad E\left[\left(W_{\infty}^{m,\overline{s}^{m*}}\right)^{2}\right] \leq \overline{\Upsilon}_{2}^{m}.$$
 (30)

Earlier in Section 4, recall from (19) that we used the steady-state distribution to bound the distribution of waiting time for any t. Such a bound is valid if the system is initially empty. Thus, in applying this result to our setting, such a bound is valid for the first block of patients, i.e., for m = 1. In fact, if there is only one block, i.e., M = 1, then the definition of  $\epsilon_n(\cdot)$  in (28) is equivalent to the corresponding definition in (8) of Section 4, and the bound in (29) simplifies to (13).

However, for subsequent blocks  $(m \ge 2)$ , some patients from the previous subsequence need to be seen before the patients from this block can be served. In other words, the waiting time of the first patient in the current subsequence is not necessarily 0. As a result, we can neither consider the M blocks as M separate systems, nor apply the upper bound of Theorem 1 to each block. Instead, we need to derive an upper bound for a new system with nonzero initial stocks which are unfinished appointments from the previous block. A careful analysis of this treatment is required in the proof that appears in the appendix, accounting for the extra  $\Omega^m$  term included in the upper bounds in (28) and (29) above.

We now investigate the  $\Omega^m$  expressions to gain insight into how the order of the blocks affects the overall performance. Recall from the definition of each  $\Omega^m$  in (26) that  $\Omega^m$ contains  $\sum_{k=1}^{m-1} \overline{\Upsilon}_1^k$  and  $\sum_{k=1}^{m-1} \overline{\Upsilon}_2^k$ . The summation  $\sum_{m=1}^M \Omega^m$ , appearing in (28) and (29), contains terms involving  $(M-k)\overline{\Upsilon}_1^k$  and  $\sum_{m=k+1}^M (m-1)\overline{K}^m\overline{\Upsilon}_2^k$ . In other words, a block scheduled earlier in the schedule has a smaller k and thus a larger weight, (M-k) or  $\sum_{m=k+1}^M (m-1)\overline{K}^m$ , contributing towards the computation of the upper bounds. From the property of the random walk used in the definition of  $\overline{\Upsilon}_1^k$  and  $\overline{\Upsilon}_2^k$ , it can be shown that these quantities increase with the standard deviation of the service time distribution  $A^k$ . Thus, it follows that the summation  $\sum_{k=1}^M \Omega^k$  included in the upper bounds can be made smaller by arranging appointment blocks with lower service time standard deviation before those with higher service time standard deviation. This observation is consistent with the computational result of Denton et al. (2007), in which they show that sequencing appointments in the decreasing order of their variances performs well, better than the other two heuristic sequencing algorithms studied in their paper.

### 5.2. Proof of Theorem 2

The proof of Theorem 2 follows the same high-level structure as the proof of Theorem 1 consisting of these three steps: we first derive an alternative expression for  $\nu(\bar{s}^*) - \nu^*$ , then find an upper bound for that difference by choosing a particular piecewise constant policy  $\bar{s}^*$ , and finally derive a lower bound for  $\nu^*$ , leading to an upper bound for the relative optimality gap  $\nu(\bar{s}^*)/\nu^* - 1$ . However, there is a notable difference between the

homogeneous service model of Section 4 and the multiple type of appointments considered in this section – the queue may not be empty at the beginning of each block, except for the first one. This requires a different approach in finding an upper bound on the difference  $\nu(\bar{s}^*) - \nu^*$ .

Throughout the proof, we use  $\mathbf{s}^* = (s_t^{m*})$  to denote the optimal schedule, where  $s_t^{m*}$  denotes the job allowance for the  $t^{th}$  patient in the  $m^{th}$  block, and let  $\overline{W}_t^{m*}$  represent the waiting time for the  $t^{th}$  patient in the  $m^{th}$  block in the optimal schedule. We use  $\overline{W}_t^m$  to denote the waiting time for the  $t^{th}$  patient in the  $m^{th}$  block under the piecewise constant policy  $\overline{\mathbf{s}}^* = (\overline{\mathbf{s}}^{1*}, \ldots, \overline{\mathbf{s}}^{M*})$ , where each  $\overline{\mathbf{s}}^{m*}$  is defined according to (22). In comparison, for fixed m, we use the notation  $W_t^{m,\overline{\mathbf{s}}^{m*}}$  to denote the waiting time for the  $t^{th}$  patient if the inter-arrival times are deterministically  $\overline{\mathbf{s}}^{m*}$ , and the i.i.d. service times are given by the random variables  $(A_t^m)$ , where  $t = 1, 2, \ldots$ , provided that the buffer is empty initially. Thus,  $W_t^{m,\overline{\mathbf{s}}^{m*}}$  is the initial state: the system with  $\overline{W}_t^m$  may have an initial buffer resulting from the previous block, but the system with  $W_t^{m,\overline{\mathbf{s}}^{m*}}$  is initially empty.

For the first step of finding an upper bound for  $\nu(\bar{s}^*) - \nu^*$ , we derive an alternative expression for this quantity, similar to (15).

Lemma 7.

$$\nu(\overline{\boldsymbol{s}}^*) - \nu(\boldsymbol{s}^*) = c_W \sum_{m=1}^M \sum_{t=1}^{n^m} E\left[\overline{W}_{t+1}^m - \overline{W}_{t+1}^{m*}\right] + c_I E\left[\overline{W}_{n^M+1}^M - \overline{W}_{n^M+1}^*\right].$$
(31)

A formal proof of Lemma 7 is attached in Appendix B.1. We develop an upper bound for the right-side of (31). From the choice of  $\overline{s}^{m*} = (s_1^{m*} + s_2^{m*} + \cdots + s_{n^m}^{m*})/n^m$  given in (22),  $\overline{s}^{m*}$  corresponds to the average job allowance of type-*m* appointments in the optimal schedule. Thus, we find each of the two terms on the right side of (31).

In the proof of Theorem 1, we exploit the property of the D/G/1 queue and derived an upper bound for the waiting-time difference between the constant policy r and the optimal policy for a single type of patients, which is given in (16). To bound the first term in (31), it may be tempting to apply the same method to each block m to obtain an upper bound for  $\sum_{t=1}^{n^m} E\left[\overline{W}_{t+1}^m - \overline{W}_{t+1}^{m*}\right]$ . Unfortunately, this approach does not work out easily due to a key difference between the models in Section 4 and Section 5 – that in the  $m^{th}$ block,  $\overline{W}_1^m$ , may not be zero as all the patients from the previous block may not have been served by the allocated time. Therefore, it calls for some additional work to upper bound  $\sum_{t=1}^{n^m} E\left[\overline{W}_{t+1}^m - \overline{W}_{t+1}^{m*}\right]$  when  $\overline{W}_1^m$  is possibly strictly positive.

To this end, we consider an alternate queuing system. A key idea in this alternate system is that it is a priority queue where the patients from a later block (i.e., larger m) is given priority over the patients from an earlier block (i.e., smaller m). Patients within the same block are served in the order of scheduled arrival times (i.e., smaller t index). Thus, the first patient in any block is served starting from the appointment time without any waiting delay. Furthermore, the alternative system allows preemption. Figure 3 illustrates both the original system (under "System 1") and the alternative system (under "System 2").



Figure 3 An Alternative System that Holds the Jobs from the Previous Subsequence tills the Server Is Idle

It is easy to verify that the server is busy in the original system if and only if the server is busy in the alternative system. Since  $\overline{W}_1^m$  is the amount of waiting time for the first patient in block m in the original system, it also corresponds to the amount of time needed to serve all backlogged patients at the appointment time of the first patient in block min either the original system or the alternative system. Recall that the job allowance for each patient in block m is  $\overline{s}^{m*}$  under the piecewise constant policy. If w is the amount of server time required to clear all backlog present at the first appointment of block m, let  $\Delta^m(w)$  denote the random variable representing that index of the earliest type-m patient experiencing no waiting if there were infinitely many type-m patients, i.e.,  $\Delta^m(\omega)$  is the smallest integer such that w plus the total of t-1 service durations is at most the total job allowance of t-1 patients:

$$\Delta^{m}(w) = \min\left\{t \in \{1, \ldots\} \mid w + (A_{1}^{m} + \cdots + A_{t-1}^{m}) \le (t-1) \cdot \overline{s}^{m*}\right\} .$$
(32)

Recall that  $W_t^{m,\bar{s}^{m*}}$  is the waiting time of the  $t^{th}$  patient in block m if there is no initial backlog at the beginning of block m. Based on the definition of the  $\Delta^m$  function above, it is

easy to see that, if  $t \ge \Delta^m(\overline{W}_1^m)$ , the starting time of the service for the  $t^{th}$  patient in block m is the same regardless of whether the initial backlog is empty or it is  $\overline{W}_1^m$  amount of time; consequently, the waiting time for this patient is unaffected under these two cases. Thus,  $\overline{W}_t^m = W_t^{m,\overline{s}^{m*}}$  provided that  $t \ge \Delta^m(\overline{W}_1^m)$ . Otherwise, the waiting time of this customer under the original system with the initial backlog of  $\overline{W}_1^m$  may be larger than the waiting time of this customer had the backlog been empty at the beginning of this block; however, it can be argued that this difference does not exceed  $\overline{W}_1^m$ , i.e.,  $W_t^{m,\overline{s}^{m*}} \le \overline{W}_t^m \le W_t^{m,\overline{s}^{m*}} + \overline{W}_1^m$ . (To see this, note that the original system is busy whenever the system corresponding to  $W_t^{m,\overline{s}^{m*}}$ , i.e., starting with an empty buffer, is busy.) Thus,

$$\overline{W}_{t}^{m} \begin{cases} \leq W_{t}^{m,\overline{s}^{m*}} + \overline{W}_{1}^{m} & \text{if } t < \Delta^{m}(\overline{W}_{1}^{m}) \\ = W_{t}^{m,\overline{s}^{m*}} & \text{if } t \ge \Delta^{m}(\overline{W}_{1}^{m}). \end{cases}$$
(33)

To find an upper bound on the first term of (31), the property given in (33) will be used, along with the following lemma on the expected value of  $\Delta^m(w)$ . The parameters  $\bar{K}^m$  and  $\tilde{\varphi}$  were defined in (25) and (27). The proof of this lemma appears in Appendix B.2.

LEMMA 8. For any w > 0,

$$E[\Delta^m(w)|w] \leq \bar{K}^m w + \frac{\tilde{\varphi}}{1-\tilde{\varphi}}$$

From (33),

$$\sum_{t=1}^{n^m} E\left[\overline{W}_{t+1}^m - \overline{W}_{t+1}^{m*}\right] \leq \sum_{t=1}^{n^m} E\left[W_{t+1}^{m,\overline{s}^{m*}} - \overline{W}_{t+1}^{m*}\right] + E\left[\sum_{t=1}^{n^m} \overline{W}_1^m \cdot \mathbf{1}[t < \Delta^m(\overline{W}_1^m)]\right] (34)$$

Since  $\overline{W}_{t+1}^{m*}$  denoting the waiting time under the optimal policy would have been the same or smaller if the initial buffer was empty, we can use the results from Section 4 to bound the first term on the right above using (17):

$$\sum_{t=1}^{n^m} E\left[W_{t+1}^{m,\overline{s}^{m*}} - \overline{W}_{t+1}^{m*}\right] \leq \overline{s}^{m*} E\left[(\eta_n^{m,\overline{s}^{m*}})^2 - \eta_n^{m,\overline{s}^{m*}}\right] .$$
(35)

To bound the second term on the right side of (34), we use Lemma 8 to obtain

$$E\left[\sum_{t=1}^{n^m} \overline{W}_1^m \cdot \mathbf{1}[t < \Delta^m(\overline{W}_1^m)]\right] \le E\left[\Delta^m(\overline{W}_1^m)\overline{W}_1^m\right] \le \bar{K}^m E\left[\left(\overline{W}_1^m\right)^2\right] + \frac{\tilde{\varphi}}{1-\tilde{\varphi}}E\left[\overline{W}_1^m\right]_{36}$$

We proceed to find upper bounds for  $E[(\overline{W}_1^m)^2]$  and  $E[\overline{W}_1^m]$ . From (33),

$$\overline{W}_{1}^{m} = \overline{W}_{n^{m-1}+1}^{m-1} \leq W_{n^{m-1}+1}^{m-1,\overline{s}^{m-1,*}} + \overline{W}_{1}^{m-1}$$

By repeatedly applying the above inequality, we obtain

$$\overline{W}_{1}^{m} \leq W_{n^{m-1}+1}^{m-1,*} + \overline{W}_{1}^{m-1} \leq \cdots \leq \sum_{k=1}^{m-1} W_{n^{k}+1}^{k,\overline{s}^{k*}}$$
(37)

As we argued in the proof of Theorem 1,  $E[W_t^{k,\overline{s}^{k*}}]$  monotonically increases from zero to  $E[W_{\infty}^{k,\overline{s}^{k*}}]$  as  $t \to \infty$ . Furthermore, we have  $E[W_{\infty}^{k,\overline{s}^{k*}}] \leq \widetilde{\Upsilon}_1^k$  by (30). Thus, we have a bound for  $E[\overline{W}_1^m]$ :

$$E\left[\overline{W}_{1}^{m}\right] \leq \sum_{k=1}^{m-1} E\left[W_{n^{k}+1}^{k,\overline{s}^{k*}}\right] \leq \sum_{k=1}^{m-1} \overline{\Upsilon}_{1}^{k} .$$

$$(38)$$

Also using (37) and a similar argument, we can obtain a bound for  $E[(\overline{W}_1^m)^2]$ :

$$E\left[\left(\overline{W}_{1}^{m}\right)^{2}\right] \leq E\left[\left(\sum_{k=1}^{m-1} W_{\infty}^{k,\overline{s}^{k*}}\right)^{2}\right] \leq (m-1)\sum_{k=1}^{m-1} E\left[\left(W_{\infty}^{k,\overline{s}^{k*}}\right)^{2}\right] \leq (m-1)\sum_{k=1}^{m-1} \overline{\Upsilon}_{2}^{k},$$
(39)

where the second inequality follows from Cauchy's inequality, i.e.,  $(a_1b_1 + \cdots + a_{m'}b_{m'})^2 \leq (a_1^2 + \cdots + a_{m'}^2)(b_1^2 + \cdots + b_{m'}^2)$ , and the last inequality follows from (30).

Putting together, we have

$$\begin{split} \sum_{t=1}^{n^m} E\left[\overline{W}_{t+1}^m - \overline{W}_{t+1}^{m*}\right] &\leq \sum_{t=1}^{n^m} E\left[W_{t+1}^{m,\bar{s}^{m*}} - \overline{W}_{t+1}^{m*}\right] + E\left[\sum_{t=1}^{n^m} \overline{W}_1^m \cdot \mathbf{1}[t < \Delta^m(\overline{W}_1^m)]\right] \\ &\leq \overline{s}^{m*} E\left[(\eta_n^{m,\bar{s}^{m*}})^2 - \eta_n^{m,\bar{s}^{m*}}\right] + \bar{K}^m E\left[\left(\overline{W}_1^m\right)^2\right] + \frac{\tilde{\varphi}}{1 - \tilde{\varphi}} E\left[\overline{W}_1^m\right] \\ &\leq \overline{s}^{m*} E\left[(\eta_n^{m,\bar{s}^{m*}})^2 - \eta_n^{m,\bar{s}^{m*}}\right] + (m-1)\bar{K}^m \sum_{k=1}^{m-1} \overline{\Upsilon}_2^k + \frac{\tilde{\varphi}}{1 - \tilde{\varphi}} \sum_{k=1}^{m-1} \overline{\Upsilon}_1^k \\ &= \overline{s}^{m*} E\left[(\eta_n^{m,\bar{s}^{m*}})^2 - \eta_n^{m,\bar{s}^{m*}}\right] + \Omega^m \;. \end{split}$$

Above, the first inequality follows from (34), the second inequality follows from (35) and (36), the third inequality follows from (38) and (39), and the final equality follows from the definition of  $\Omega^m$  in (26). Thus, the first term in (31) is bounded above as follows:

$$c_{W} \sum_{m=1}^{M} \sum_{t=1}^{n^{m}} E\left[\overline{W}_{t+1}^{m} - \overline{W}_{t+1}^{m*}\right] \leq c_{W} \sum_{m=1}^{M} \left\{\overline{s}^{m*} E\left[(\eta_{n}^{m,\overline{s}^{m*}})^{2} - \eta_{n}^{m,\overline{s}^{m*}}\right] + \Omega^{m}\right\}.$$

Finally, the second term in (31) satisfies

$$c_I E\left[\overline{W}_{n^M+1}^M - \overline{W}_{n^M+1}^*\right] \leq c_I E\left[\overline{W}_{n^M+1}^M\right] = c_I E\left[\overline{W}_{1}^{M+1}\right] \leq c_I \sum_{m=1}^M \overline{\Upsilon}_{1}^m,$$

where the last inequality follows from (38). Therefore, from Lemma 7, we conclude

$$\nu(\overline{s}^{*}) - \nu^{*} = \nu(\overline{s}^{*}) - \nu(s^{*}) \leq c_{W} \sum_{m=1}^{M} \left\{ \overline{s}^{m*} E\left[ (\eta_{n}^{m, \overline{s}^{m*}})^{2} - \eta_{n}^{m, \overline{s}^{m*}} \right] + \Omega^{m} \right\} + c_{I} \sum_{m=1}^{M} \widetilde{\Upsilon}_{1}^{m}$$

This establishes an upper bound for  $\nu(\overline{s}^*) - \nu^*$ .

Now, since the expected total cost associated with each appointment in each block m is bounded below by  $g^m$  defined in (23) from Lemma 1, it follows

$$\nu^* \geq \sum_{i=1}^m n^m g^m$$

Therefore,

$$\frac{\nu(\overline{\boldsymbol{s}}^{*}) - \nu^{*}}{\nu^{*}} \leq \frac{c_{W} \sum_{m=1}^{M} \left\{ \overline{\boldsymbol{s}}^{m*} E\left[ \left( \eta_{n}^{m, \overline{\boldsymbol{s}}^{m*}} \right)^{2} - \eta_{n}^{m, \overline{\boldsymbol{s}}^{m*}} \right] + \Omega^{m} \right\} + c_{I} \sum_{m=1}^{M} \overline{\Upsilon}_{1}^{m}}{\sum_{m=1}^{M} n^{m} g^{m}} = \epsilon_{n}(\overline{\boldsymbol{s}}^{*}) ,$$

$$(40)$$

where the equality follows from the definition of  $\epsilon_n(\cdot)$  in (28). Furthermore, since (A.6) implies

$$\overline{s}^{m*}E\left[(\eta_n^{m,\overline{s}^{m*}})^2 - \eta_n^{m,\overline{s}^{m*}}\right] \leq \frac{2 \cdot \overline{s}^{m*} \cdot (\varphi_{\overline{s}^{m*}}^m)^2}{(1 - \varphi_{\overline{s}^{m*}}^m)^2} ,$$

we obtain the inequality in (29). The convergence result in (29) holds since the numerator in the bound is independent of n and the denominator increases to become arbitrarily large as  $n \to \infty$ . This completes the proof of Theorem 2.

# 6. Computational Results: Performance of the Constant and Piecewise Constant Policies

In this section, we report on our computational investigation of the performance of the constant policy and the piecewise constant policy. We study how much the restriction imposed by these simple policies increases the overall cost compared to a more general policy.

We use the sample average approximation approach (Kleywegt et al. 2002) in our computation. Specifically, we randomly generate K i.i.d. scenarios with a given distribution of service time and then solve the resulting linear program that optimizes the schedule for these sample service times. If the job allowance is allowed to be different for each job, the resulting solution, denoted by  $s_{SAA}$ , is a proxy for the optimal schedule. If the linear program constrains the job allowance to be the same, the resulting solution, denoted by  $r_{SAA}$ , is the best constant policy for the generated sample service times. Note that the theoretical results in Sections 4 and 5 have used a specific constant policy, namely one with the average job allowance under the optimal policy, which requires knowing the optimal schedule. Instead, by using a linear program, we search for the best constant job allowance among all constant policies.

We let  $\nu(\mathbf{s}_{SAA})$  and  $\nu(r_{SAA})$  denote the expected cost associated with the "optimal' policy and the best constant policy, respectively, using sample average approximation. The ratio  $(\nu(r_{SAA}) - \nu(\mathbf{s}_{SAA}))/\nu(\mathbf{s}_{SAA})$  gives the relative gap between the optimal constant policy and the optimal policy solved using the SAA method. Since the SAA method usually solves the problem to optimality when the problem scale is not too large, this ratio provides an estimation of the relative optimality gap for the optimal constant policy. We normalize the waiting time cost parameter to be 1, i.e.,  $c_W = 1$ , and vary the unit idling time cost parameter  $c_I$ , which then corresponds to the ratio  $c_I/c_W$ . We consider the case with i.i.d. service durations in Section 6.1, and discuss the case with piecewise i.i.d. service durations in Section 6.2. In Appendix C, we also compare the constant policy and the optimal policy under several extensions of the model, e.g., when the objective includes an overtime cost, when patients have no-show or arrival unpunctuality.

### 6.1. Homogeneous Customers: I.I.D. Service Durations

The numerical study aims to complement the theoretical results in two aspects. First, our theoretical bound in Theorem 1 shows that the constant policy is near optimal either if  $c_I/c_W$  is sufficiently small (Condition 1), or if the number of appointments, n, is sufficiently large (Condition 2). Our computation focuses on the cases not covered by the theorem, namely when n is small and  $c_I/c_W$  is large. Second, while the theoretical convergence rate identified in Theorem 1 is O(1/n), we learn the actual convergence rate computationally.

To test the performance of the constant policy for small n, we fix n = 16, and choose various ratios of  $c_I/c_W$ . Since appointment scheduling covers a wide range of applications, all the tested parameters can find examples in real life. We let  $d = 5\mu$ , but the results will keep the same when d takes even larger values. In each scenario, we test four service duration distributions: exponential distribution with mean 20, lognormal distribution with mean 20 and standard deviation 4, and two normal distributions with mean 20 and standard deviation 4 and 8. (In the case of normal distributions, we truncate them at 0, ensuring nonnegativity.) In each test, we randomly generate K = 2000 scenarios, solve the sample average approximation linear program to determine the corresponding optimal schedule and best constant policy. Then, we generate another K' = 2000 scenarios to calculate  $\nu(s_{SAA})$  and  $\nu(r_{SAA})$ to approximate the relative optimality gap,  $(\nu(r_{SAA}) - \nu(s_{SAA}))/\nu(s_{SAA})$ . The ratios are summarized in Table 1.

					uura	tions								
Distribution		_	$c_I/c_W$											
	$\mu$	0	0.2	0.4	0.6	0.8	1	1.5	2	2.5	3	10	100	
Normal -	20	4	0.15	0.25	0.43	0.56	0.68	1.03	1.14	1.77	1.89	4.48	25.15	
	20	8	0.18	0.33	0.59	0.67	0.79	1.51	1.87	1.98	2.16	5.06	25.45	
Exponential	20	20	0.43	0.83	1.00	1.27	1.45	2.18	2.63	2.78	3.12	6.48	25.63	
Lognormal	20	4	0.09	0.38	0.42	0.54	0.86	1.23	1.54	1.9	2.09	4.87	25.21	

Table 1Relative optimality gaps (in %) for different service time distributions and ratios of  $c_I/c_W$ : i.i.d. servicedurations

Table 1 provides several insights. First, the constant policy performs well in general under various combinations of  $c_I/c_W$  and service duration distributions. The relative optimality gap under the worst combination is less than 4% when  $c_I/c_W \leq 3$ . Hence, even for smallsize problems (n = 16), the constant policy is able to achieve near-optimal performance as long as the idling cost is not too large compared to the waiting cost. However, if the idling cost is too large, then the constant policy can lose substantially to the optimal schedule. The theoretical upper bound  $\epsilon_n(\bar{s}^*)$  we derived in Theorem 1 can be much larger than the actual optimality gap for small n. Further strengthening the theoretical bound, however, presents a challenge.

Second, the constant policy achieves the best performance when the service duration has a normal distribution with small standard deviation and has the worst performance for exponential distribution. It appears that the constant policy works well when the service duration has a small coefficient of variation,  $\sigma/\mu$ . This is not surprising since the constant job allowance is optimal if all service times are deterministic. Third, for each type of service duration distribution, the relative optimality gap exhibits a clear decreasing trend approaching 0 as  $c_I/c_W$  becomes smaller. This observation is consistent with what has been reported in Denton and Gupta (2003), Hassin and Mendel (2008) – the curve of the job allowance under the optimal schedule, despite still having a dome shape, is more flattened when the unit idling cost is relatively small.

In order to compare optimal constant policy (OC) with optimal schedule (OPT) more vividly, for each combination of service time distribution and ratio of  $c_I/c_W$ , we compare the performance indicators (i.e., total waiting times (WTime) and total idle times(ITime)) and corresponding total idle costs (ICost) for two policies (Note that we set  $c_W=1$ , the waiting time equals to waiting cost). The results are summarized in Table 2.

From Table 2, we can observe that all elements, including total waiting times, total idle times and total idle cost exhibit a small difference between two policies except when the ratio  $c_I/c_W$  is large. More interestingly, both total waiting times and total idle costs witness an increasing trend as the ratio  $c_I/c_W$  increases, while the total idle cost exhibit an opposite trend.

We next consider the relative optimality gap as we vary the number of total appointments (i.e., problem sizes), n, to understand how the relative optimality gap converges. Since the technical difficulty with a large n is that it is very slow to solve the sample average approximation linear program to obtain  $s_{SAA}$ , we use a smaller K = 400 for problems when  $n \ge 100$ . We fix  $c_I = 3$ , and calculate the relative optimality gap for n = 25, 50, 100, 200, respectively. The results are summarized in Table 3. It shows that the relative optimality gap decreases as n increases, as expected. The speed that the gap converges to zero actually depends on the type of service duration distribution. While the actual convergence rates are difficult to estimate with limited n values, it appears that the convergence rate is faster than O(1/n) in all four types of service distributions when  $n \ge 50$ .

### 6.2. Multiple Customers Types: Piecewise I.I.D. Service Durations

We next study the performance of the piecewise constant policy when the service durations are piecewise i.i.d. We consider a system with M = 2 types of patients, with an equal proportion, i.e.,  $q^1 = q^2 = 0.5$ . For each type patients, we assume its service duration can take one of the five service duration distributions. The first four distributions are the same as those used previously in Section 6.1, and the fifth distribution that we added is an exponential distribution with mean 4. Eight combinations of these distributions were tested. Similar to the setting used in Table 1, we use a small problem size with n = 16,

Distribution		~		Policies	$c_I/c_W$																		
Distribution	μ	σ		Policies	0.2	0.4	0.6	0.8	1	1.5	2	2.5	3	10	100								
			WTime	OPT OC	$5\\5$	$\begin{array}{c} 10.1 \\ 10.1 \end{array}$	$\begin{array}{c} 14.5 \\ 14.4 \end{array}$	$\begin{array}{c} 18.7 \\ 18.5 \end{array}$	$22.8 \\ 22.5$	$31.5 \\ 31.1$	$40 \\ 39.6$	$\begin{array}{c} 46.8\\ 46.6\end{array}$	$54.8 \\ 54.2$	$\begin{array}{c} 122.1 \\ 120.5 \end{array}$	$311.4 \\ 336.1$								
	20	4	ITime	OPT OC	$74.2 \\ 74.5$	$58.4 \\ 58.6$	$\begin{array}{c} 48.1 \\ 48.5 \end{array}$	$\begin{array}{c} 43.7\\ 44.4\end{array}$	$38.3 \\ 39.1$	$31.7 \\ 32.5$	$26.5 \\ 27.4$	$23.9 \\ 24.8$	$20.5 \\ 21.5$	$8.1 \\ 9.3$	$0.7 \\ 1.5$								
			ICost	OPT OC	$\begin{array}{c} 14.8\\ 14.9\end{array}$	$23.4 \\ 23.4$	$28.8 \\ 29.1$	$\frac{35}{35.5}$	$38.3 \\ 39.1$	$47.5 \\ 48.8$	$53 \\ 54.7$	$59.7 \\ 62$	$61.4 \\ 64.5$	$81.4 \\ 93$	$70.2 \\ 149.4$								
Normal	20	8	WTime	OPT OC	$\begin{array}{c} 10.2 \\ 10.3 \end{array}$	$\begin{array}{c} 20.1 \\ 19.7 \end{array}$	$28.8 \\ 28.4$	$36.9 \\ 36.5$	$44.5 \\ 43.5$	$\begin{array}{c} 64.2 \\ 63.3 \end{array}$	$79.2 \\ 78.5$	$95.3 \\ 93.9$	$\begin{array}{c} 109.1 \\ 106.9 \end{array}$	$243.9 \\ 240.5$	$\begin{array}{c} 654.9 \\ 674.3 \end{array}$								
			ITime	OPT OC	$\begin{array}{c} 151.2\\ 151.2 \end{array}$	$\begin{array}{c} 117.6\\ 119 \end{array}$	$98.7 \\ 100.1$	$86.6 \\ 87.8$	$77.3 \\ 79.2$	$\begin{array}{c} 62.4 \\ 64.2 \end{array}$	$54.7 \\ 56.3$	$\begin{array}{c} 46.1 \\ 48.1 \end{array}$	$\begin{array}{c} 40.3\\ 42.7\end{array}$	$\begin{array}{c} 16.3 \\ 18.7 \end{array}$	$1.5 \\ 3.3$								
			ICost	OPT OC	$30.2 \\ 30.2$	$\begin{array}{c} 47\\ 47.6\end{array}$	$\begin{array}{c} 59.2\\ 60\end{array}$	$69.2 \\ 70.2$	$77.3 \\ 79.2$	$93.7 \\ 96.3$	$\begin{array}{c} 109.5\\ 112.6 \end{array}$	$\begin{array}{c} 115.2 \\ 120.3 \end{array}$	$\begin{array}{c} 120.9 \\ 128.1 \end{array}$	$\begin{array}{c} 163.1 \\ 187.4 \end{array}$	$152.7 \\ 327$								
	20	20	WTime	OPT OC	$\begin{array}{c} 45.7\\ 45.8\end{array}$	$79.8 \\ 79.2$	$\begin{array}{c} 109.2 \\ 107.5 \end{array}$	$131.9 \\ 130.8$	$\begin{array}{c} 156.4\\ 153.4 \end{array}$	$\begin{array}{c} 206.3\\ 205.1 \end{array}$	$255.9 \\ 254.4$	$284.3 \\ 283.9$	$332.3 \\ 333.2$	$\begin{array}{c} 648.7\\ 650.6\end{array}$	$1362 \\ 1535$								
Exponential			ITime	OPT OC	$\begin{array}{c} 428.9\\ 430.2 \end{array}$	$313.7 \\ 319.3$	$\begin{array}{c} 255.6\\ 261.8 \end{array}$	$\begin{array}{c} 214.6\\ 219.4 \end{array}$	$\begin{array}{c} 192 \\ 198.8 \end{array}$	$\begin{array}{c} 141 \\ 147.8 \end{array}$	$\begin{array}{c} 122.1 \\ 128.8 \end{array}$	$\begin{array}{c} 101.8\\ 108.5 \end{array}$	$91.8 \\ 98.1$	$30.2 \\ 35.9$	$\frac{2}{4}$								
												ICost	OPT OC	$\begin{array}{c} 85.8\\ 86\end{array}$	$\begin{array}{c} 125.5 \\ 127.7 \end{array}$	$153.3 \\ 157.1$	$171.6 \\ 175.5$	$\begin{array}{c} 192 \\ 198.8 \end{array}$	$\begin{array}{c} 211.5\\ 221.6\end{array}$	$244.2 \\ 257.7$	$\begin{array}{c} 254.6\\ 271.2 \end{array}$	$\begin{array}{c} 275.3\\ 294.2 \end{array}$	$302.2 \\ 359.2$
		20 4								WTime	OPT OC	$6.6 \\ 6.5$	$\begin{array}{c} 11.8\\ 11.7\end{array}$	$\begin{array}{c} 17.3 \\ 17.1 \end{array}$	$21.3 \\ 21.1$	$26.2 \\ 25.5$	$35.7 \\ 35.2$	$\begin{array}{c} 43.3\\ 42.4 \end{array}$	$\begin{array}{c} 50.7 \\ 50.3 \end{array}$	$58.9 \\ 58.9$	$\begin{array}{c} 125.5\\ 124.3 \end{array}$	$310.4 \\ 339.1$	
Lognormal	20		ITime	OPT OC	$79.1 \\ 79.4$	$59.2 \\ 59.6$	$50.5 \\ 51.1$	$\begin{array}{c} 43.5\\ 44.3\end{array}$	$38.4 \\ 39.5$	$31.4 \\ 32.3$	$26.1 \\ 27.4$	$22.8 \\ 23.7$	$20 \\ 21.1$	$7.5 \\ 8.7$	$\begin{array}{c} 0.7 \\ 1.3 \end{array}$								
			ICost	OPT OC	15.8 15.9	23.7 23.9	30.3 30.7	34.8 35.5	38.4 39.5	47.1 48.5	52.3 54.8	$56.9 \\ 59.3$	60.1 63.2	75.3 86.8	66.2 133.5								

Table 2Performance indicators comparison (in %) for different service time distributions and ratios of  $c_I/c_W$ :i.i.d. service durations

Table 3 Relative optimality gaps (in %) for different problem sizes: i.i.d. service durations

		$\mu$	n									
Distribution	$\sigma$		25			50		100	200			
			Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	$95\%~{\rm CI}$		
NT 1	20	4	1.60	[1.29, 1.91]	1.06	[0.88, 1.23]	0.29	[0.11, 0.46]	0.00	[0, 0.04]		
Normai	20	8	1.60	[1.26, 1.93]	1.06	[0.85, 1.27]	0.30	[0.14, 0.45]	0.00	[0, 0.04]		
Exponential	20	20	2.65	[2.22, 3.07]	1.78	[1.50, 2.07]	0.63	[0.39, 0.88]	0.10	[0, 0.25]		
Lognormal	20	4	1.90	[1.56, 2.23]	1.25	[1.01, 1.49]	0.35	[0.12, 0.58]	0.00	[0,09]		

with 8 patients for each type, and set K = K' = 2000. The results are summarized in Table 4.

From Table 4, we observe that the relative optimality gap of the piecewise constant policy is less than 9% for all combinations. We also observe that the piecewise constant

Type	1		Type 2			$c_I/c_W$									
Distribution	$\mu$	$\sigma$	Distribution	$\mu$	$\sigma$	0.2	0.4	0.6	0.8	1	1.5	2	2.5	3	
Normal	20	4	Normal	20	8	0.2	0.23	0.26	0.41	0.59	0.86	0.93	0.99	1.06	
Normal	20	8	Normal	20	4	0.34	0.67	0.9	1.37	1.86	2.25	2.72	2.88	3.01	
Exponential	4	4	Exponential	20	20	0.4	0.49	0.83	0.94	1.09	1.29	1.79	1.9	1.93	
Exponential	20	20	Exponential	4	4	3.41	3.67	3.9	4.45	5.57	6.02	7.09	7.92	8.65	
Normal	20	4	Exponential	20	20	0.37	0.53	0.88	1.01	1.22	1.5	1.56	1.63	1.71	
Exponential	20	20	Normal	20	4	4.08	4.65	4.89	5.05	5.25	5.55	6.91	7.12	7.5	
Lognormal	20	4	Exponential	20	20	0.23	0.72	0.92	0.94	1.28	1.32	1.39	1.65	1.87	
Exponential	20	20	Lognormal	20	4	2.42	3.53	4.72	5.07	6.05	6.21	6.47	7.12	8.68	

Table 4Relative optimality gaps (in %) for different service time distributions and ratios of  $c_I/c_W$ : piecewisei.i.d. service durations

policy performs better when the ratio  $c_I/c_W$  is smaller, similar to the i.i.d. service duration case.

More interestingly, when we compare the last two rows of the table, both using the same set of service distributions, it turns out that the performance gap is smaller when the lognormal distribution with standard deviation 4 is used for the first block of patients and the exponential distribution with standard deviation 20 is used for the second block of patients. It suggests that it is better to schedule the patient block with a smaller standard deviation first. A similar observation can be made from the third row to the sixth row of the table (the pair with two exponential distributions, one with standard deviation 4 and another with standard deviation 20; and the pair with one normal distribution and other one with exponential distribution). This is consistent with what we have inferred from the analytical upper bounds – that the performance gap is more sensitive with service distributions in earlier blocks; see Section 5.1.

Finally, we compare the performances of the constant policy and the piecewise constant policy in a simple setting with M = 2 appointment types. We consider scenarios that are described in the last two rows of Table 4. That is, type-1 appointments have a lognormal distribution and type-2 appointments have an exponential distribution. The job allowance of the optimal constant policy  $r_{SAA}$  and that of the optimal piecewise constant policy  $r_{SAA}^{M}$  are computed with the SAA method under different values of  $c_I/c_W$ . Let  $\nu(r_{SAA})$  and  $\nu(r_{SAA}^{M})$  denote the total cost achieved by the constant policy and the piecewise constant policy, respectively. So the ratio  $(\nu(r_{SAA}) - \nu(r_{SAA}^{M}))/\nu(r_{SAA})$  represents the percent-wise reduction in the total cost by the piecewise constant policy compared to the constant policy. Table 5 summarizes the values of  $(\nu(r_{SAA}) - \nu(r_{SAA}^{M}))/\nu(r_{SAA})$  in different parametric settings. We observe all values to be positive, suggesting that the piecewise constant policy outperforms the constant policy in all scenarios. The improvement is more significant when  $c_I/c_W$  is small, because according to Theorem 2, the piecewise constant policy has very small cost when  $c_I/c_W$  is small.

Table 5 $(\nu(r_{SAA}) - \nu(r_{SAA}^M))/\nu(r_{SAA})$  (in %) under different service time distributions and ratios of  $c_I/c_W$ :piecewise i.i.d. service durations

Type 1	Type 2	$c_I/c_W$										
Distribution $\mu \sigma$	Distribution mu sigma	0.2 0.4 0.6	0.8 1 1.5 2	2.5 3								
Lognormal204Exponential2020	Exponential 20 20 Lognormal 20 4	22.981813.9518.114.5311.62	10.848.543.561.788.517.884.873.77	$\begin{array}{ccc} 0.59 & 0.14 \\ 2.3 & 1 \end{array}$								

# 7. Conclusion and Future Work

Prior studies on appointment scheduling report that the optimal schedule has a "dome" pattern for the i.i.d. service durations, in which the service allowances for patients exhibit little variation in the middle of the planning horizon. Inspired by this, we analyze a simple but effective constant scheduling policy for the traditional appointment scheduling problem. The numerical experiments show that the constant policy performs well in various parameters and distributions combinations. Most importantly, we prove that the constant policy is asymptotically optimal for homogeneous patients, and we extend the asymptotic optimality result to heterogeneous patient types.

We make a novel analysis of the easy-to-implement policy for the traditional appointment scheduling problem, which is robust over a range of settings such as i.i.d. service duration distributions and the possibility of no-shows. We establish the explicit relative optimality gap bounds between the constant policy or its extension and the optimal schedule. This bound also leads to asymptotic optimality of the constant and piecewise constant policies, once again advocating for their use in practice. Furthermore, we have identified several managerial insights – some of which are new and affirm observations made in the literature. Our approach of attaining asymptotic optimality results can be extended to other optimization problems that makes a tradeoff between two types of costs. For example, in a slot queue staffing assignment problem, the number of servers scheduled in a slot directly impacts the staffing cost and customer goodwill with respect to waiting. However, for a more general type of service duration distributions beyond the piecewise i.i.d. service duration structure, the optimality structure is still unknown and remains open. Walk-ins may disrupt the appointment system thus deserve for more attention when making a schedule. In addition, patients' behaviors, including non-punctuality and time window preferences, would also affect the performance of appointment system. The case when patients may arrive later than their scheduled times is a possibility for future research. Even though some policies have been shown to be numerically effective in the literature, no theoretical performance bounds have yet been derived for such patient behavior under consideration, as well as in cases with session length constraint and overtime cost. These considerations are beyond the scope of this paper and potential for future research.

# Acknowledgement

Yichuan Ding and Guohua Wan are co-corresponding authors. We thank the department editor, the senior editor, and two anonymous reviewers for their valuable and constructive suggestions, which have improved the paper significantly. We thank Professor Chung Piaw Teo for many of his key suggestions, including focusing the discussion on the appointment scheduling problem. We also thank Professor Avishai Mandelbaum, Professor Yehua Wei, and Professor Linwei Xin for their helpful feedback. Yichuan Ding is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants (RGPIN 2019-05539) and grants from the Key Program of NSFC-FRQSC Joint Project (NSFC No. 72061127002, FRQSC No. 295837). Woonghee Tim Huh acknowledges support from the NSERC Discovery Grants (RGPIN 2020-04213) and the Canada Research Chair Program. Guohua Wan receives financial support from the National Natural Science Foundation of China (NSFC No. 71520107003, 71931008 and 71421002).

### References

- Andersen ES (1955) On the fluctuations of sums of random variables II. *Mathematica Scandinavica* 2:195–223.
- Bailey NT (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. Journal of the Royal Statistical Society. Series B (Methodological) 14(2):185–199.
- Begen MA, Levi R, Queyranne M (2012) A sampling-based approach to appointment scheduling. *Operations Research* 60(3):675–681.
- Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of literature. Production and Operations Management 12(4):519–549.
- Cayirli T, Veral E, Rosen H (2006) Designing appointment scheduling systems for ambulatory care services. Health Care Management Science 9(1):47–58.

- Cayirli T, Veral E, Rosen H (2008) Assessment of patient classification in appointment system design. *Production and Operations Management* 17(3):338–353.
- Cayirli T, Yang KK, Quek SA (2012) A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management* 21(4):682–697.
- Charnetski JR (1984) Scheduling operating room surgical procedures with early and late completion penalty costs. *Journal of Operations Management* 5(1):91–102.
- Chen D, Wang R, Yan Z, Benjaafar S, Jouini O (2016) Appointment systems under service level constraints. Working paper, University of Toronto, Canada.
- Chen RR, Robinson LW (2014) Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management* 23(9):1522–1538.
- Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Transactions* 35(11):1003–1016.
- Denton B, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science* 10(1):13–24.
- Goldberg DA, Katz-Rogozhnikov DA, Lu Y, Sharma M, Squillante MS (2016) Asymptotic optimality of constant-order policies for lost sales inventory models with large lead times. *Mathematics of Operations Research* 41(3):898–913.
- Gupta D, Denton B (2008) Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* 40(9):800–819.
- Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management Science* 54(3):565–572.
- Ho CJ, Lau HS (1992) Minimizing total cost in scheduling outpatient appointments. *Management Science* 38(12):1750–1764.
- Jansson B (1966) Choosing a good appointment system: a study of queues of the type (d, m, 1). Operations Research 14(2):292–312.
- Kaandorp GC, Koole G (2007) Optimal outpatient appointment scheduling. Health Care Management Science 10(3):217–229.
- Karlin S (2014) A first course in stochastic processes (Academic Press), New York.
- Kingman J (1962) Some inequalities for the queue GI/G/1. Biometrika 49(3/4):315–324.
- Klassen KJ, Rohleder TR (2004) Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. International Journal of Service Industry Management 15(2):167–186.
- Klassen KJ, Yoogalingam R (2009) Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management* 18(4):447–458.
- Kleywegt AJ, Shapiro A, Homem-de Mello T (2002) The sample average approximation methoad for stochastic discrete optimization. SIAM Journal on Optimization 12(2):479–502.
- Kong Q, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research* 61(3):711–726.
- Kong Q, Li S, Liu N, Teo CP, Yan Z (2020) Appointment scheduling under time-dependent patient no-show behavior. *Management Science* Online.
- Kuiper A, Kemper B, Mandjes M (2015) A computational approach to optimized appointment scheduling. Queueing Systems 79(1):5–36.
- LaGanga LR, Lawrence SR (2012) Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management* 21(5):874–888.
- Lau HS, Lau AHL (2000) A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. *IIE Transactions* 32(9):833–839.
- Luo J, Kulkarni VG, Ziya S (2012) Appointment scheduling under patient no-shows and service interruptions. Manufacturing & Service Operations Management 14(4):670–684.

- Mak HY, Rong Y, Zhang J (2014a) Appointment scheduling with limited distributional information. *Management Science* 61(2):316–334.
- Mak HY, Rong Y, Zhang J (2014b) Sequencing appointments for service systems using inventory approximations. Manufacturing & Service Operations Management 16(2):251–262.
- Mancilla C, Storer R (2012) A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions* 44(8):655–670.
- Marshall KT (1968) Some inequalities in queuing. Operations Research 16(3):651–668.
- Mercer A (1960) A queueing problem in which the arrival times of the customers are scheduled. Journal of the Royal Statistical Society. Series B (Methodological) 108–113.
- Muthuraman K, Lawley M (2008) A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions* 40(9):820–837.
- Robinson LW, Chen RR (2003) Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions* 35(3):295–307.
- Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. Manufacturing & Service Operations Management 12(2):330–346.
- Sabria F, Daganzo CF (1989) Approximate expressions for queueing systems with scheduled arrivals and established service order. *Transportation Science* 23(3):159–165.
- Soriano A (1966) Comparison of two scheduling systems. Operations Research 14(3):388–397.
- Wang PP (1993) Static and dynamic scheduling of customer arrivals to a single-server system. Naval Research Logistics (NRL) 40(3):345–360.
- Wang PP (1997) Optimally scheduling n customer arrival times for a single-server system. Computers & Operations Research 24(8):703-716.
- Weiss EN (1990) Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions* 22(2):143–150.
- Xin L, Goldberg DA (2016) Optimality gap of constant-order policies decays exponentially in the lead time for lost sales models. *Operations Research* 64(6):1556–1565.
- Yang KK, Lau ML, Quek SA (1998) A new appointment rule for a single-server, multiple-customer service system. Naval Research Logistics (NRL) 45(3):313–326.
- Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. Production and Operations Management 23(5):788–801.
- Zacharias C, Pinedo M (2017) Managing customer arrivals in service systems with multiple identical servers. Manufacturing & Service Operations Management 19(4):639–656.
- Zeng B, Turkcan A, Lin J, Lawley M (2010) Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Annals of Operations Research 178(1):121–144.

## A. Proofs for Section 4

#### A.1. Proof of Lemma 1

From the definition given in (3),

$$\nu(s) = \sum_{t=1}^{n} \left\{ c_W \cdot E\left[W_{t+1}\right] + c_I \cdot E\left[I_{t+1}\right] \right\} = \sum_{t=1}^{n} \left\{ c_W \cdot E\left[W_t + A_t - s_t\right]^+ + c_I \cdot E\left[W_t + A_t - s_t\right]^- \right\}$$
(A.1)

where the second equality follows from (1) and (2). Recall that each  $W_t$  is nonnegative and it is independent of  $A_t$ .

For fixed  $t \in \{1, ..., n\}$ , we consider the impact of  $s_t$ , on the above objective function. The job allowance for the  $t^{th}$  patient does not impact the cost for patients in  $\{1, 2, ..., t-1\}$ . For patients in  $\{t + 1, ..., n\}$ , the choice of  $s_t$  has an impact on the cost through waiting time distributions  $\{W_{t+1}, ..., W_n\}$ ; more specifically, the summation of costs from t + 1 to n decreases in  $s_t$ . This implies that the optimal value of  $s_t$  should be bounded below by the optimizer of the myopic cost (cost in period t) given by

$$C_{t} = c_{W} \cdot E[W_{t} + A_{t} - s_{t}]^{+} + c_{I} \cdot E[W_{t} + A_{t} - s_{t}]^{-} .$$
(A.2)

Note that this is a well-known Newsvendor function, which is a convex function of  $s_t$ , and the derivative of  $C_t$  with respect to  $s_t$  is given by

$$-c_W \cdot \mathbb{P}\left[W_t + A_t > s_t\right] + c_I \cdot E\left[W_t + A_t \le s_t\right]$$

This derivative depends on the distribution of  $W_t$  and can be made the smallest when  $W_t = 0$ , i.e., the above expression is bounded below by  $-c_W \cdot \mathbb{P}[A_t > s_t] + c_I \cdot E[A_t \le s_t]$ , which turns out to be the function inside the minimization operator in (4) used for the definition of  $\underline{s}$ . Thus, we conclude that  $\underline{s}$  is a lower bound for the optimizer of (A.2), which in turn is a lower bound for the optimal value of  $s_t$ , denoted by  $s_t^*$ . This proves  $s_t^* \ge \underline{s}$ .

Now, since  $\underline{s} \ge 0$ , it follows from the definition of g in (5),

$$g = \min_{s \in \Re^+} c_W \cdot E [A_t - s]^+ + c_I \cdot E [A_t - s]^-$$

$$\leq \min_{s \in [-E[W_t], d - E[W_t]]} c_W \cdot E [A_t - s]^+ + c_I \cdot E [A_t - s]^-$$

$$\leq \min_{s \in [-E[W_t], d - E[W_t]]} c_W \cdot E [W_t - E[W_t] + A_t - s]^+ + c_I \cdot E [W_t - E[W_t] + A_t - s]^-$$

$$= \min_{s \in [0,d]} c_W \cdot E [W_t + A_t - s]^+ + c_I \cdot E [W_t + A_t - s]^-$$

$$\leq c_W \cdot E [W_t + A_t - s_t]^+ + c_I \cdot E [W_t + A_t - s_t]^- = C_t ,$$

where the second inequality follows since  $c_W E [A_t - s]^+ + c_I E [A_t - s]^-$  is a convex function of s and  $W_t - E[W_t] + A_t$  is a mean-preserving spread of  $A_t$  (Landsberger and Meilijson 1993). Since it is shown that g is a lower bound for any  $C_t$ , it follows that  $g \leq C_t^*$  also holds.

### A.2. Proof of Lemma 2

We first provide an algorithm to numerically compute  $E[(\eta_n^r)^2 - \eta_n^r]$ , which allows us to compute  $\epsilon_n(r)$ . Consider the discrete-time random walk defined by  $\sum_{i=1}^t (A_i - r)$  in  $t \in \{0, 1, \ldots, t\}$ , and let  $N_t$  denote the number of times when this random walk takes a positive value. By Andersen (1955) and Bingham (2001), the index  $\eta_t^r$  defined in (7) has the same distribution as the random number  $N_n$  for each  $n = 0, 1, \ldots, +\infty$ . By invoking Theorem 1 of Andersen (1955), we have

$$\mathbb{P}(N_n = j) = p_j q_{n-j} \tag{A.3}$$

and

$$P(\alpha \cdot \beta)Q(\beta) = \sum_{i=0}^{\infty} \sum_{j=0}^{i} p_j q_{i-j} \alpha^j \beta^i$$

where  $P(\alpha \cdot \beta) = exp\left(\int_0^{\alpha \cdot \beta} \frac{B(x)}{x} dx\right)$ ,  $Q(\beta) = (1 - \beta)^{-1} exp\left(\int_0^{\beta} \frac{B(x)}{x} dx\right)$ ,  $B(x) = \sum_{k=1}^{\infty} b_k x^k$ ,  $b_k = \mathbb{P}\left[\sum_{i=1}^k (A_i - r) > 0\right]$ ,  $p_j = \mathbb{P}(N_j = j)$ ,  $q_j = \mathbb{P}(N_j = 0)$ , and  $p_0 = q_0 = 1$ . Let  $\mathbb{M}_i(\alpha) = \sum_{j=0}^i p_j q_{i-j} \alpha^j$  denote the coefficient of  $\beta^i$ . Then, we have

$$P(\alpha \cdot \beta)Q(\beta) - \sum_{i=0}^{n-1} \mathbb{M}_i(\alpha)\beta^i = \sum_{j=0}^n p_j q_{n-j}\alpha^j\beta^n + o(\beta^n) = \mathbb{M}_n(\alpha)\beta^n + o(\beta^n)$$

which indicates

$$\mathbb{M}_n(\alpha) = \lim_{\beta \to 0} \frac{P(\alpha \cdot \beta)Q(\beta) - \sum_{i=0}^{n-1} \mathbb{M}_i(\alpha)\beta^i}{\beta^n}$$

Note that from the definition of  $\mathbb{M}_n(\alpha)$  and (A.3), we have

$$\lim_{\alpha \to 1} \frac{\partial^2}{\partial \alpha^2} \mathbb{M}_n(\alpha) = \sum_{j=0}^n j(j-1)p_j q_{n-j} = \sum_{j=0}^n j(j-1)\mathbb{P}(N_n = j) = E[N_n^2 - N_n] = E[(\eta_n^r)^2 - \eta_n^r], \quad (A.4)$$

where the last equality follows from that  $N_n$  and  $\eta_n^r$  have the same distribution. Through L 'Hospitol law, equation (A.4) implies that we can obtain the value of  $E[(\eta_n^r)^2 - \eta_n^r]$  through:

$$E[(\eta_n^r)^2 - \eta_n^r] = \lim_{\alpha \to 1} \frac{\partial^2}{\partial \alpha^2} \mathbb{M}_n(\alpha)$$
  
=  $\lim_{\alpha \to 1} \frac{\partial^2}{\partial \alpha^2} \left( \lim_{\beta \to 0} \frac{P(\alpha \cdot \beta)Q(\beta) - \sum_{i=0}^{n-1} \mathbb{M}_i(\alpha)\beta^i}{\beta^n} \right)$   
=  $\lim_{\alpha \to 1} \frac{\partial^2}{\partial \alpha^2} \left( \frac{\lim_{\beta \to 0} \frac{\partial^n}{\partial \beta^n} P(\alpha \cdot \beta)Q(\beta)}{n!} \right)$  (A.5)

We next derive a uniform upper bound for  $E[(\eta_n^r)^2 - \eta_n^r]$ . To proceed, by substituting  $n = +\infty$ , we let  $\eta_{\infty}^r$  follow the limiting distribution of  $\eta_n^r$  as  $n \to \infty$ . This distribution is well defined provided  $r > \mu$  (Spitzer 1956). Recall that  $N_t$  denotes the number of times when this random walk  $\sum_{i=1}^t (A_i - r)$  takes a positive value. Let  $N_{\infty}$  denote the limiting random variable for  $N_t$  as  $t \to \infty$ , which is

well-defined (Spitzer 1956) since  $\mu = E[A_i] < r$ . By invoking Theorem 5.2 of Spitzer (1956), we derive the following alternative expression for  $E[(\eta_{\infty}^r)^2 - \eta_{\infty}^r]$ :

$$E[(\eta_{\infty}^{r})^{2} - \eta_{\infty}^{r}] = E[(N_{\infty})^{2} - N_{\infty}] = E[N_{\infty}(N_{\infty} - 1)] = \left(\sum_{k=1}^{\infty} b_{k}\right)^{2} + \sum_{k=2}^{\infty} (k - 1)b_{k}$$

where  $b_k = \mathbb{P}\left[\sum_{i=1}^k (A_i - r) > 0\right]$ . By the Chernoff's inequality, for any  $\psi > 0$ , we have

$$b_k = \mathbb{P}\left[\sum_{i=1}^k (A_i - r) > 0\right] \le (E[\exp(\psi(A - r))])^k = (\phi_r(\psi))^k,$$

where the last equality follows from the definition of  $\phi_r(\psi) = E[\exp(\psi(A-r))]$  in (10). Since the above inequality holds for any  $\psi > 0$ , it follows from (10) that we have  $b_k \leq (\varphi_r)^k$ . Note that  $\varphi_r \in [0,1)$  holds (Kingman 1962). (To see this, note that  $\phi_r(\psi)$  is a continuous function of  $\psi$  on  $(0,\infty)$  from the definition of  $\phi_r(\psi)$  in (10). Also,  $\phi_r(\psi)$  is right differentiable at 0 in which the derivative equals to  $E[A] - r = \mu - r < 0$ . Since  $\phi''_r(\psi) < 0$  for all  $\psi > 0$ ,  $\phi_r(\psi)$  has to keep decreasing since 0. Therefore,  $\phi_r(\psi) < \phi_r(0) = 1$  for all  $\psi > 0$ .) Therefore,

$$E[(\eta_{\infty}^{r})^{2} - \eta_{\infty}^{r}] = \left(\sum_{k=1}^{\infty} b_{k}\right)^{2} + \sum_{k=2}^{\infty} (k-1)b_{k}$$
  
$$\leq \left(\sum_{k=1}^{\infty} (\varphi_{r})^{k}\right)^{2} + \sum_{k=2}^{\infty} (k-1)(\varphi_{r})^{k} \leq \frac{2 \cdot (\varphi_{r})^{2}}{(1-\varphi_{r})^{2}}, \qquad (A.6)$$

where the last inequality directly follows from the basic manipulation of the geometric series  $(\varphi_r)^k$ .

We have thus proved a uniform upper bound for  $E[(\eta_{\infty}^r)^2 - \eta_{\infty}^r]$ . This uniform upper bound also applies to  $E[(\eta_{\infty}^r)^2 - \eta_{\infty}^r]$  because of the next proposition.

PROPOSITION 1. If  $r > \mu$ , then  $\eta_{\infty}^r < \infty$ , and

$$E[(\eta_n^r)^2 - \eta_n^r] \le E[(\eta_\infty^r)^2 - \eta_\infty^r]$$

**Proof of Proposition 1** By applying claim (A.16), for equation (A.17), we have

$$r\sum_{j=1}^{n}\sum_{t=1}^{n}\mathbb{P}(\eta_{n}^{r} \ge t) \le r\sum_{j=1}^{n}\sum_{t=1}^{\infty}\mathbb{P}(\eta_{\infty}^{r} \ge t),$$
(A.7)

and for equation (A.18), we have

$$r\sum_{j=1}^{n}\sum_{t=1}^{j}t\mathbb{P}(\eta_{n}^{r}=t) \geq r\sum_{j=1}^{n}\sum_{t=1}^{j}t\mathbb{P}(\eta_{\infty}^{r}=t) .$$
(A.8)

Thus, for equation (A.19), we have

\_\_\_\_\_

$$\sum_{t=1}^{n} E\left[\sum_{i=t+1-\eta_{t}^{r}}^{t} s_{i}^{*}\right] - \sum_{t=1}^{n} r E\left[\eta_{t}^{r}\right] \le r \sum_{j=1}^{n} \sum_{t=1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) - r \sum_{j=1}^{n} \sum_{t=1}^{j} t \mathbb{P}(\eta_{n}^{r} = t)$$
(A.9)

$$\leq r \sum_{j=1}^{n} \sum_{t=1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \geq t) - r \sum_{j=1}^{n} \sum_{t=1}^{j} t \mathbb{P}(\eta_{\infty}^{r} = t) .$$
 (A.10)

As for (A.10), for fixed j, we have

$$\begin{split} &\sum_{t=1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) - \sum_{t=1}^{j} t \mathbb{P}(\eta_{\infty}^{r} = t) \\ &= \sum_{t=j+1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) + \sum_{t=1}^{j} \mathbb{P}(\eta_{\infty}^{r} \ge t) - \sum_{t=1}^{j} t \mathbb{P}(\eta_{\infty}^{r} = t) \\ &= \sum_{t=j+1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) + E[\min(\eta_{\infty}^{r}, j)] - E[\eta_{\infty}^{r} \mathbb{I}(\eta_{\infty}^{r} \le j)] \\ &= \sum_{t=j+1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) + E[\eta_{\infty}^{r} \mathbb{I}(\eta_{\infty}^{r} \le j)] + E[j \mathbb{I}(\eta_{\infty}^{r} \ge j+1)] - E[\eta_{\infty}^{r} \mathbb{I}(\eta_{\infty}^{r} \le j)] \\ &= \sum_{t=j+1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) + E[j \mathbb{I}(\eta_{\infty}^{r} \ge j+1)] = \sum_{t=j+1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) + \sum_{t=1}^{j} \mathbb{P}(\eta_{\infty}^{r} \ge j+1) \;. \end{split}$$

Hence, summing over j, it follows

$$\begin{split} &\sum_{j=1}^{n} \sum_{t=1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) - \sum_{j=1}^{n} \sum_{t=1}^{j} t \mathbb{P}(\eta_{\infty}^{r} = t) \\ &= \sum_{j=1}^{n} \sum_{t=j+1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) + \sum_{j=1}^{n} \sum_{t=1}^{j} \mathbb{P}(\eta_{\infty}^{r} \ge j+1) \\ &\leq \sum_{j=1}^{\infty} \sum_{t=j+1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} \ge t) + \sum_{j=1}^{\infty} \sum_{t=1}^{j} \mathbb{P}(\eta_{\infty}^{r} \ge j+1) \\ &= \sum_{j=1}^{\infty} \sum_{t=j+1}^{\infty} \sum_{\ell=t}^{\infty} \mathbb{P}(\eta_{\infty}^{r} = \ell) + \sum_{j=1}^{\infty} \sum_{t=1}^{j} \sum_{\ell=j+1}^{\infty} \mathbb{P}(\eta_{\infty}^{r} = \ell) \\ &= \sum_{\ell=2}^{\infty} \sum_{t=2}^{\ell} \sum_{j=1}^{t-1} \mathbb{P}(\eta_{\infty}^{r} = \ell) + \sum_{\ell=2}^{\infty} \sum_{j=1}^{\ell-1} \sum_{t=1}^{j} \mathbb{P}(\eta_{\infty}^{r} = \ell) \\ &= \sum_{\ell=2}^{\infty} (1+2+\cdots(\ell-1)) \mathbb{P}(\eta_{\infty}^{r} = \ell) + \sum_{\ell=2}^{\infty} (1+2+\cdots(\ell-1)) \mathbb{P}(\eta_{\infty}^{r} = \ell) \\ &= \sum_{\ell=1}^{\infty} \ell(\ell-1) \mathbb{P}(\eta_{\infty}^{r} = \ell) \\ &= E\left[(\eta_{\infty}^{r})^{2} - \eta_{\infty}^{r}\right] . \end{split}$$

Thus, combine (A.20) and (A.11), we conclude

$$\sum_{t=1}^{n} E\left[\sum_{i=t+1-\eta_t^r}^{t} s_i^*\right] - \sum_{t=1}^{n} r E\left[\eta_t^r\right] \le r E[(\eta_n^r)^2 - \eta_n^r] \le r E[(\eta_\infty^r)^2 - \eta_\infty^r]$$

completing the proof of Lemma 2.

#### A.3. Proof of Lemma 3

Suppose Condition 1 holds, i.e.,  $F(\mu) < c_W/(c_W + c_I)$ . Recalling the definition of  $\underline{s}$  in (4), it follows from the property of the Newsvendor problem that we have  $F(\underline{s}) = c_W/(c_W + c_I)$ . Thus, from Lemma 1, we conclude  $r = (s_1^* + \cdots + s_n^*)/n \ge \underline{s} > \mu$ , as required.

Now, consider Condition 2. This condition is stated in terms of the standard deviation of the service time,  $\sigma$ , and the upper bound d imposed on the choice of each time allowance, as well as  $\tau$ , defined in (6):  $\tau = \lceil \{8(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \sqrt{(\frac{c_I}{2c_W})^3})\}^2 \rceil$  where  $\zeta = E[|A - \mu|^3]/\sigma^3$ . Suppose Condition 2 holds, i.e.,  $n \ge \max\{2\tau, 6\tau^{\frac{3}{2}}\sigma^{-1}(d-\underline{s})\}$ . While this condition was inspired by Goldberg et al. (2016), it requires a modified analysis because of the way the recursive equation  $W_t$  is defined, as discussed in the end of Section 4.2. Recall, from definition,  $\tau = \lceil \{8(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \sqrt{(\frac{c_I}{2c_W})^3})\}^2 \rceil$ . Suppose h satisfies  $1 \le h \le n$ .

For t satisfying  $1 \le h \le t \le n$ , define

$$\gamma_t = \sum_{k=t+1-h}^t s_k^* .$$
 (A.12)

For the waiting time of  $(t+1)^{th}$  patient under the optimal schedule, we have

$$\begin{split} E[W_{t+1}^*] &= E\left[\max_{j=0,1,\cdots,t} \left\{\sum_{i=t+1-j}^t (A_i - s_i^*)\right\}\right] \\ &\geq E\left[\max\left\{0,\sum_{i=t+1-h}^t (A_i - s_i^*)\right\}\right] \\ &= \sigma h^{\frac{1}{2}} E\left[\max\left\{0,\frac{\sum_{i=t+1-h}^t (A_i - \mu)}{\sigma h^{\frac{1}{2}}} + \frac{h\mu - \sum_{i=t+1-h}^t s_i^*}{\sigma h^{\frac{1}{2}}}\right\}\right] \\ &= \sigma h^{\frac{1}{2}} E\left[\max\left\{0,\frac{\sum_{i=t+1-h}^t (A_i - \mu)}{\sigma h^{\frac{1}{2}}} + \frac{h\mu - \gamma_t}{\sigma h^{\frac{1}{2}}}\right\}\right], \end{split}$$

where the last two equalities follow from algebraic manipulation. Then, we can find the lower bound for the rightmost expression using Theorem 3 of Goldberg et al. (2016) and Chen and Shao (2005), and we have

$$E[W_{t+1}^*] \ge \sigma h^{\frac{1}{2}} E\left[\max\left\{0, N + \frac{h\mu - \gamma_t}{\sigma h^{\frac{1}{2}}}\right\}\right] - 3\sigma\zeta , \qquad (A.13)$$

where N denotes a standard normal random variable.

Now, we can bound the cost of the optimal schedule using a feasible schedule given by  $s_t = \tilde{r}$  for each  $t \in \{1, \ldots, n\}$ , where  $\tilde{r} = \mu + \sqrt{c_W/(2c_I)\sigma}$ . Since we have

$$\nu(\mathbf{s}^*) = E\left[\sum_{t=1}^n \left( c_W \cdot [W_t^* + A_t - s_t^*]^+ + c_I \cdot [W_t^* + A_t - s_t^*]^- \right) \right]$$
  
=  $E\left[\sum_{t=1}^n \left( c_W \cdot [W_t^* + A_t - s_t^*]^+ + c_I \cdot s_t^* \right) \right] + c_I \cdot [W_n^* + A_n - s_n^*]^+ - c_I n \mu$ ,

where the last equality follows from Lemma 4, and a similar equation holds for the feasible policy given by  $s_t = \tilde{r}$  for each t, it follows

$$\nu(\boldsymbol{s}^*) \leq E\left[\sum_{t=1}^n \left(c_W \cdot [W_t^* + A_t - \tilde{r}]^+ + c_I \cdot \tilde{r}\right)\right] + c_I \cdot [W_n^* + A_n - \tilde{r}]^+ - c_I n \mu$$
  
$$\leq n\left(c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + c_I \tilde{r}\right) + c_I \frac{\sigma^2}{2(\tilde{r} - \mu)} - c_I n \mu$$
  
$$= n\left(c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + c_I (\tilde{r} - \mu) + \frac{c_I}{n} \cdot \frac{\sigma^2}{2(\tilde{r} - \mu)}\right).$$

Above, the second inequality follows from (19) and Lemma 6. Therefore, since  $\nu(s^*)$  is at least  $c_W \sum_{t=1}^n E[W_{t+1}^*]$ , which in turn can be bounded below using (A.13), we have

$$c_W \sigma h^{\frac{1}{2}} \sum_{t=h}^{n} E\left[ \max\{0, N + \frac{h\mu - \gamma_t}{\sigma h^{\frac{1}{2}}}\} \right] - 3c_W \sigma \zeta n \le \nu(\boldsymbol{s}^*) \le n \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + c_I(\tilde{r} - \mu) + \frac{c_I}{n} \cdot \frac{\sigma^2}{2(\tilde{r} - \mu)} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{1}{2}}} \right) + \frac{c_I}{\sigma h^{\frac{1}{2}}} \left( c_W \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{\sigma h^{\frac{$$

which implies

$$\begin{split} \sum_{t=h}^{n} E\left[\max\{0, N + \frac{h\mu - \gamma_t}{\sigma h^{\frac{1}{2}}}\}\right] &\leq n\left(3\sigma\zeta + \frac{\sigma^2}{2(\tilde{r} - \mu)} + \frac{c_I}{c_W}(\tilde{r} - \mu) + \frac{c_I}{nc_W} \cdot \frac{\sigma^2}{2(\tilde{r} - \mu)}\right)\sigma^{-1}h^{-\frac{1}{2}} \\ &= n\left(3\sigma\zeta + \sqrt{\frac{2c_I}{c_W}}\sigma + \frac{c_I\sigma}{2nc_W} \cdot \sqrt{\frac{2c_I}{c_W}}\right)\sigma^{-1}h^{-\frac{1}{2}} \\ &= n\left(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \frac{1}{n} \cdot \frac{1}{\sqrt{2} \cdot (\frac{c_I}{c_W})^{\frac{3}{2}}}\right)h^{-\frac{1}{2}} ,\end{split}$$

where the first equality follows from  $\tilde{r} = \mu + \sqrt{c_W/(2c_I)\sigma}$ . Since  $E[\max\{0, N+y\}]$  is a convex function of y, we have, by Jensen's inequality,

$$E\left[\max\left\{0,(n+1-h)N+\sum_{t=h}^{n}\frac{h\mu-\gamma_{t}}{\sigma h^{\frac{1}{2}}}\right\}\right] \leq \sum_{t=h}^{n}E\left[\max\{0,N+\frac{h\mu-\gamma_{t}}{\sigma h^{\frac{1}{2}}}\}\right]$$

Therefore,

$$E\left[\max\left\{0, (n+1-h)N + \sum_{t=h}^{n} \frac{h\mu - \gamma_{t}}{\sigma h^{\frac{1}{2}}}\right\}\right] \le n\left(3\zeta + \sqrt{\frac{2c_{I}}{c_{W}}} + \frac{1}{n} \cdot \frac{1}{\sqrt{2}} \cdot (\frac{c_{I}}{c_{W}})^{\frac{3}{2}}\right) h^{-\frac{1}{2}} \quad (A.14)$$

Now, we find a lower bound for the left-side expression above using an expression involving r. Since we have defined  $\gamma_t = \sum_{k=t+1-h}^t s_k^*$  in (A.12), which is a sum of h distinct  $s_k^*$  values,  $\sum_{t=h}^n \gamma_t$  is the sum of  $(n-h+1) \cdot h$  number of  $s_k^*$  values, and for fixed k,  $s_k^*$  appears at most h times in the sum. Thus, we must have

$$\sum_{t=h}^{n} \gamma_t = \sum_{t=h}^{n} \sum_{k=t+1-h}^{t} s_k^*$$
  
=  $(s_1^* + s_2^* + \dots + s_h^*) + (s_2^* + s_3^* + \dots + s_{h+1}^*) + \dots + (s_{n+1-h}^* + s_{n+2-h}^* + \dots + s_n^*)$   
 $\leq h \sum_{t=1}^{n} s_t^* - h(h-1)\underline{s}$ ,

since  $\underline{s}$  is a lower bound for each  $s^*$ . Thus,

$$\sum_{t=h}^{n} \gamma_t \leq hnr - h(h-1)\underline{s} = h(n+1-h)r + h(h-1)(r-\underline{s}) \leq h(n+1-h)r + h(h-1)(d-\underline{s}) , \quad (A.15)$$

where the last inequality follows from  $r \leq d$ .

Then,

$$E\left[\max\left\{0, N + \frac{h^{\frac{1}{2}}}{\sigma}\left(\mu - r - \frac{(h-1)(d-\underline{s})}{n+1-h}\right)\right\}\right] \le \frac{n}{n+1-h}\left(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \frac{1}{n} \cdot \frac{1}{\sqrt{2}} \cdot (\frac{c_I}{c_W})^{\frac{3}{2}}\right)h^{-\frac{1}{2}} ,$$

where the inequality follows by substituting the upper bound of  $\sum_{t=h}^{n} \gamma_t$  given in (A.15) to (A.14), and then both dividing by n+1-h. Recall that Condition 2 is  $n \ge \max\{2h, 6h^{\frac{3}{2}}\sigma^{-1}(d-\underline{s})\}$ . Note that  $n \ge 2h$  implies  $\frac{n}{n+1-h} \le 2$ , and  $n \ge 6h^{\frac{3}{2}}\sigma^{-1}(d-\underline{s})$  implies  $\frac{(h-1)(d-\underline{s})}{n+1-h} < \frac{h(d-\underline{s})}{n+1-h} \le \frac{1}{3}\sigma h^{-\frac{1}{2}}$ . Hence,

$$E\left[\max\left\{0, N + \frac{h^{\frac{1}{2}}}{\sigma}(\mu - r - \frac{1}{3}\sigma h^{-\frac{1}{2}})\right\}\right] \le 2\left(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \frac{1}{n} \cdot \frac{1}{\sqrt{2}} \cdot (\frac{c_I}{c_W})^{\frac{3}{2}}\right)h^{-\frac{1}{2}}.$$

The condition  $n \ge 2h$  implies  $n \ge 2$ , hence we have

$$\begin{split} E\left[\max\left\{0, N + \frac{h^{\frac{1}{2}}}{\sigma}(\mu - r - \frac{1}{3}\sigma h^{-\frac{1}{2}})\right\}\right] &\leq 2\left(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \frac{1}{n} \cdot \frac{1}{\sqrt{2}} \cdot (\frac{c_I}{c_W})^{\frac{3}{2}}\right)h^{-\frac{1}{2}} \\ &\leq 2\left(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \frac{1}{2} \cdot \frac{1}{\sqrt{2}} \cdot (\frac{c_I}{c_W})^{\frac{3}{2}}\right)h^{-\frac{1}{2}} \,. \end{split}$$

It can be verified that  $E[\max\{0, N-0.34\}] \ge \frac{1}{4}$ . Let  $h = \tau = \lceil \{8(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \sqrt{(\frac{c_I}{2c_W})^3})\}^2 \rceil$ . Then, we have

$$E\left[\max\left\{0, N + \frac{\tau^{\frac{1}{2}}}{\sigma}(\mu - r - \frac{1}{3}\sigma\tau^{-\frac{1}{2}})\right\}\right] \le 2\left(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \frac{1}{2} \cdot \frac{1}{\sqrt{2}} \cdot (\frac{c_I}{c_W})^{\frac{3}{2}}\right)\tau^{-\frac{1}{2}} \le \frac{1}{4} \le E[\max\{0, N - 0.34\}].$$

Therefore, it follows  $\frac{\tau^{\frac{1}{2}}}{\sigma}(\mu - r - \frac{1}{3}\sigma\tau^{-\frac{1}{2}}) \leq -0.34$ , which implies  $r - \mu \geq 0.006\sigma\tau^{-\frac{1}{2}}$ . Thus, we complete the proof of  $r > \mu$ .

### A.4. Proof of Lemma 4

From the definition of  $[\cdot]^+$  and  $[\cdot]^-$ , as well as the recursive definitions (1) and (2),

$$W_{t-1} + A_{t-1} - s_{t-1} = [W_{t-1} + A_{t-1} - s_{t-1}]^{+} - [W_{t-1} + A_{t-1} - s_{t-1}]^{-} = W_{t} - I_{t}$$

which implies  $\sum_{t=2}^{n+1} I_t = W_{n+1} - \sum_{t=2}^{n+1} A_{t-1} + \sum_{t=2}^{n+1} s_{t-1}$ . Then, from the definition given in (3), the objective function satisfies

$$\nu(s) = E\left[\sum_{t=2}^{n+1} \{c_W \cdot W_t + c_I \cdot I_t\}\right] = E\left[\sum_{t=2}^{n+1} c_W W_t\right] + E\left[c_I W_{n+1}\right] - E\left[c_I \sum_{t=2}^{n+1} A_{t-1}\right] + c_I \sum_{t=2}^{n+1} s_{t-1}$$
$$= E\left[\sum_{t=1}^n c_W W_{t+1}\right] + c_I \sum_{t=1}^n s_t + c_I E\left[W_{n+1}\right] - c_I n\mu .$$

This completes the proof.

#### A.5. Proof of Lemma 5

Recall from (7) that  $\eta_t^r$  corresponds to the index k in which the random walk  $\sum_{i=1}^k (A_i - r)$  attains its maximum, where ties are broken in favor of the smaller index, and  $\eta_{\infty}^r$  is the limiting distribution of  $\eta_t^r$  as  $t \to \infty$ . To establish the inequality in the statement of Lemma 5, we find an upper bound for  $\sum_{t=1}^n E[\sum_{i=t+1-\eta_t^r}^t s_i^*]$ , and then a lower bound for  $\sum_{t=1}^n rE[\eta_t^r]$ . Before finding these bounds, we state and prove the following claim:

$$\mathbb{P}(\eta_t^r = k) \ge \mathbb{P}(\eta_n^r = k) \ge \mathbb{P}(\eta_\infty^r = k) \quad \text{for any } k \in \{0, 1, \dots, t\} \text{ and } t < n.$$
(A.16)

Intuitively, since  $\eta_t^r$  represents the index of the random walk that achieves its maximum, the probability that the random walk attains its maximum with index k is lower with a larger value of t since it would increase the set from which the maximum operator is applied. More formally, to prove (A.16), fix k and t such that  $1 \le k \le t$ , and define

$$F_{1} = \left\{ \sum_{i=1}^{k} (A_{i} - r) > \sum_{i=1}^{j} (A_{i} - r) \quad \text{for each } j \in \{0, 1, \dots, k-1\} \right\}$$

$$F_{2} = \left\{ \sum_{i=1}^{k} (A_{i} - r) \ge \sum_{i=1}^{j} (A_{i} - r) \quad \text{for each } j \in \{k+1, \dots, t\} \right\}$$

$$F_{3} = \left\{ \sum_{i=1}^{k} (A_{i} - r) \ge \sum_{i=1}^{j} (A_{i} - r) \quad \text{for each } j \in \{t+1, \dots, n\} \right\}.$$

Then, the event  $[\eta_t^r = k]$  is in the intersection of  $F_1$  and  $F_2$  whereas the event  $[\eta_n^r = k]$  is in the intersection of  $F_1$ ,  $F_2$  and  $F_3$ . This completes the proof of claim (A.16).

We now provide an upper bound for  $\sum_{t=1}^{n} E[\sum_{i=t+1-\eta_t^r}^{t} s_i^*]$ . To do so, for any  $t \in \{1, \ldots, n\}$ ,

$$E\left[\sum_{i=t+1-\eta_t^r}^t s_i^*\right] = \sum_{k=1}^t \mathbb{P}(\eta_{t+i-1}^r = k) \cdot \left\{\sum_{i=t+1-k}^t s_i^*\right\} = \sum_{i=1}^t s_i^* \mathbb{P}(\eta_t^r \ge t+1-i) \ .$$

Summing over all  $t \in \{1, \ldots, n\}$ ,

$$\begin{split} \sum_{t=1}^{n} E\left[\sum_{i=t+1-\eta_{t}^{r}}^{t} s_{i}^{*}\right] &= \sum_{t=1}^{n} \sum_{i=1}^{t} s_{i}^{*} \mathbb{P}(\eta_{t}^{r} \ge t+1-i) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{n+1-i} s_{i}^{*} \mathbb{P}(\eta_{t+i-1}^{r} \ge t) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{n} \sum_{t=1}^{n} s_{i}^{*} \mathbb{P}(\eta_{t+i-1}^{r} \ge t) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{n} \sum_{t=1}^{n} s_{i}^{*} \mathbb{P}(\eta_{t+i-1}^{r} \ge t) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{n} \sum_{t=1}^{n} s_{i}^{*} \mathbb{P}(\eta_{t+i-1}^{r} \ge t) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{n} \sum_{t=1}^{n} \sum_{t=1}^{n} s_{i}^{*} \mathbb{P}(\eta_{t+i-1}^{r} \ge t) \\ &= \sum_{i=1}^{n} \sum_{t=1}^{n} \sum_{t=1}^{n}$$

By applying claim (A.16), we obtain

$$\sum_{t=1}^{n} E\left[\sum_{i=t+1-\eta_{t}^{r}}^{t} s_{i}^{*}\right] \leq \sum_{i=1}^{n} \sum_{t=1}^{n+1-i} s_{i}^{*} \left\{1 - \sum_{k=0}^{t-1} \mathbb{P}(\eta_{n}^{r} = k)\right\} = \sum_{i=1}^{n} \sum_{t=1}^{n+1-i} s_{i}^{*} \mathbb{P}(\eta_{n}^{r} \ge t) \leq \sum_{i=1}^{n} \sum_{t=1}^{n} s_{i}^{*} \mathbb{P}(\eta_{n}^{r} \ge t)$$
$$= \frac{1}{n} \sum_{i=1}^{n} s_{i}^{*} \sum_{j=1}^{n} \sum_{t=1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) = r \sum_{j=1}^{n} \sum_{t=1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t),$$
(A.17)

where the last equality follows from the choice of  $r = \sum_{i=1}^{n} s_i^*/n$  given in (11) and the last inequality follows from claim (A.16).

Next we find a lower bound for  $\sum_{t=1}^{n} r E[\eta_t^r]$ . Applying claim (A.16), we have

$$\sum_{t=1}^{n} E[\eta_t^r] = \sum_{t=1}^{n} \sum_{j=1}^{t} \mathbb{P}(\eta_t^r = j)j = \sum_{j=1}^{n} \sum_{t=1}^{j} t \mathbb{P}(\eta_j^r = t) \ge \sum_{j=1}^{n} \sum_{t=1}^{j} t \mathbb{P}(\eta_n^r = t) .$$

which implies

$$\sum_{t=1}^{n} rE[\eta_t^r] \ge r \sum_{j=1}^{n} \sum_{t=1}^{j} t\mathbb{P}(\eta_n^r = t) .$$
(A.18)

With relations in (A.17) and (A.18), we can conclude

$$\sum_{t=1}^{n} E\left[\sum_{i=t+1-\eta_{t}^{r}}^{t} s_{i}^{*}\right] - \sum_{t=1}^{n} r E\left[\eta_{t}^{r}\right] \le r \sum_{j=1}^{n} \sum_{t=1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) - r \sum_{j=1}^{n} \sum_{t=1}^{j} t \mathbb{P}(\eta_{n}^{r} = t)$$
(A.19)

As for (A.19), for fix j, we have

$$\begin{split} &\sum_{t=1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) - \sum_{t=1}^{j} t \mathbb{P}(\eta_{n}^{r} = t) \\ &= \sum_{t=j+1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) + \sum_{t=1}^{j} \mathbb{P}(\eta_{n}^{r} \ge t) - \sum_{t=1}^{j} t \mathbb{P}(\eta_{n}^{r} = t) \\ &= \sum_{t=j+1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) + E[\min(\eta_{n}^{r}, j)] - E[\eta_{n}^{r} \mathbb{I}(\eta_{n}^{r} \le j)] \\ &= \sum_{t=j+1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) + E[\eta_{n}^{r} \mathbb{I}(\eta_{n}^{r} \le j)] + E[j \mathbb{I}(\eta_{n}^{r} \ge j+1)] - E[\eta_{n}^{r} \mathbb{I}(\eta_{n}^{r} \le j)] \\ &= \sum_{t=j+1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) + E[j \mathbb{I}(\eta_{n}^{r} \ge j+1)] = \sum_{t=j+1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) + \sum_{t=1}^{j} \mathbb{P}(\eta_{n}^{r} \ge j+1) \ . \end{split}$$

Hence, summing over j, it follows

$$\begin{split} \sum_{j=1}^{n} \sum_{t=1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) &- \sum_{j=1}^{n} \sum_{t=1}^{j} t \mathbb{P}(\eta_{n}^{r} = t) \\ &= \sum_{j=1}^{n} \sum_{t=j+1}^{n} \mathbb{P}(\eta_{n}^{r} \ge t) + \sum_{j=1}^{n} \sum_{t=1}^{j} \mathbb{P}(\eta_{n}^{r} \ge j+1) \\ &= \sum_{j=1}^{n} \sum_{t=j+1}^{n} \sum_{\ell=t}^{n} \mathbb{P}(\eta_{n}^{r} = \ell) + \sum_{j=1}^{n} \sum_{t=1}^{j} \sum_{\ell=j+1}^{n} \mathbb{P}(\eta_{n}^{r} = \ell) \\ &= \sum_{\ell=2}^{n} \sum_{t=2}^{\ell} \sum_{j=1}^{t-1} \mathbb{P}(\eta_{n}^{r} = \ell) + \sum_{\ell=2}^{n} \sum_{j=1}^{\ell-1} \sum_{t=1}^{j} \mathbb{P}(\eta_{n}^{r} = \ell) \\ &= \sum_{\ell=2}^{n} (1+2+\cdots(\ell-1)) \mathbb{P}(\eta_{n}^{r} = \ell) + \sum_{\ell=2}^{n} (1+2+\cdots(\ell-1)) \mathbb{P}(\eta_{n}^{r} = \ell) \\ &= \sum_{\ell=1}^{n} \ell(\ell-1) \mathbb{P}(\eta_{n}^{r} = \ell) \\ &= E\left[(\eta_{n}^{r})^{2} - \eta_{n}^{r}\right] \,. \end{split}$$
(A.20)

completing the proof of Lemma 5.

## B. Proofs for Section 5

### B.1. Proof of Lemma 7

It follows from the proof of Lemma 4 that, for optimal schedule  $s^*$ , we have

$$\nu(s^*) = E\left[\sum_{t=2}^{n+1} c_W W_t^*\right] + E\left[c_I W_{n+1}^*\right] - E\left[c_I \sum_{t=2}^{n+1} A_{t-1}\right] + c_I \sum_{t=2}^{n+1} s_{t-1}^*$$
$$= \sum_{m=1}^M \sum_{t=1}^{n^m} c_W E\left[\overline{W}_{t+1}^{m*}\right] + c_I E\left[W_{n+1}^*\right] - c_I \sum_{t=1}^n E\left[A_t\right] + c_I \sum_{m=1}^M \sum_{t=1}^{n^m} s_t^{m*}$$
$$= \sum_{m=1}^M \sum_{t=1}^{n^m} c_W E\left[\overline{W}_{t+1}^{m*}\right] + c_I E\left[W_{n+1}^*\right] - c_I \sum_{m=1}^M n^m \mu^m + c_I \sum_{m=1}^M \sum_{t=1}^{n^m} s_t^{m*} .$$

Similarly, we can also derive an alternative expression for  $\nu(\bar{s}^*)$ . Since the piecewise constant policy  $\bar{s}^* = (\bar{s}^{1*}, \dots, \bar{s}^{M*})$  satisfies  $\bar{s}^{m*} = \sum_{t=1}^{n^m} s_t^{m*} / n^m$  by (22), it follows that

$$\nu(\overline{s}^{*}) - \nu(s^{*}) = c_{W} \sum_{m=1}^{M} \sum_{t=1}^{n^{m}} E\left[\overline{W}_{t+1}^{m} - \overline{W}_{t+1}^{m*}\right] + c_{I} E\left[\overline{W}_{n^{m}+1}^{m} - W_{n^{m}+1}^{*}\right] + c_{I} \sum_{m=1}^{M} \left\{n^{m} \overline{s}^{m*} - \sum_{t=1}^{n^{m}} s_{t}^{m*}\right\}$$
$$= c_{W} \sum_{m=1}^{M} \sum_{t=1}^{n^{m}} E\left[\overline{W}_{t+1}^{m} - \overline{W}_{t+1}^{m*}\right] + c_{I} E\left[\overline{W}_{n^{m}+1}^{m} - W_{n^{m}+1}^{*}\right] + 0.$$

This is the required result.

### B.2. Proof of Lemma 8

Recall from the definition of  $\Delta^m(w)$  given in (32):

$$\Delta^{m}(w) = \min\left\{t \in \{1, 2, \ldots\} \mid w + (A_{1}^{m} - \overline{s}^{m*}) + \dots + (A_{t-1}^{m} - \overline{s}^{m*}) \le 0\right\} .$$
(B.1)

Then, since the event  $\Delta^m(w) \ge k$  implies that  $w + (A_1^m - \overline{s}^{m*}) + \dots + (A_{k-2}^m - \overline{s}^{m*})$  should be strictly positive, we obtain

$$\begin{split} E\left[\Delta^{m}(w)\middle|w\right] &= \sum_{k=1}^{\infty} \mathbb{P}\left(\Delta^{m}(w) \ge k\right) \\ &\leq \sum_{k=1}^{\infty} \mathbb{P}\left(w + \sum_{t=1}^{k-2} \left(A_{t}^{m} - \overline{s}^{m*}\right) > 0\right) \\ &= \sum_{k=1}^{\infty} \mathbb{P}\left(\sum_{t=1}^{k-2} \left(\overline{s}^{m*} - \frac{w}{k-2} - A_{t}^{m}\right) < 0\right) \\ &= \sum_{k=1}^{\infty} \mathbb{P}\left(\sum_{t=1}^{k} \left(\overline{s}^{m*} - \frac{w}{k} - A_{t}^{m}\right) < 0\right) \,. \end{split}$$

Now, we break the summation in the rightmost expression above into two parts based on whether k is at most  $\bar{K}^m w$  or not. Suppose we have  $k > \bar{K}^m w$ , in which case, we have  $w/k < 1/\bar{K}^m$ . Hence,

$$\mathbb{P}\left(\sum_{t=1}^{k} \left(\overline{s}^{m*} - \frac{w}{k} - A_{t}^{m}\right) < 0\right) \leq \mathbb{P}\left(\sum_{t=1}^{k} \left(\overline{s}^{m*} - \frac{1}{\overline{K}^{m}} - A_{t}^{m}\right) < 0\right) = \mathbb{P}\left(\sum_{t=1}^{k} \left(\overline{r}^{m} - A_{t}^{m}\right) < 0\right) .$$
where the last equality follows from the definition of  $\overline{x}^{m} - \overline{z}^{m*} - \frac{1}{\overline{K}^{m}}$  given in (25). By applying

where the last equality follows from the definition of  $\bar{r}^m = \bar{s}^{m*} - 1/K^m$  given in (25). By applying the Chernoff bound, we obtain that the rightmost expression above is bounded above by

$$\begin{cases} \inf_{\psi > 0} \quad E\left[\exp\left(-\psi \cdot (\bar{r}^m - A^m)\right)\right] \end{cases}^{\mu}$$

which is  $\{\varphi_{\bar{r}^m}^m\}^k$  since we have  $\varphi_{\bar{r}^m}^m = \inf_{\psi \ge 0} \phi_{\bar{r}^m}^m(\psi)$  where  $\phi_{\bar{r}^m}^m(\psi) = E[\exp(\psi(A^m - \bar{r}^m))]$ . Also, since we have defined  $\tilde{\varphi} = \max_{i \in \{1,2,\dots,M\}} \varphi_{\bar{r}^m}^m$  in (27), we have  $\varphi_{\bar{r}^m}^m \le \tilde{\varphi}$ . In summary, if  $k > \bar{K}^m w$ , then

$$\mathbb{P}\left(\sum_{t=1}^{k} \left(\overline{s}^{m*} - \frac{w}{k} - A_t^m\right) < 0\right) \leq \left\{\tilde{\varphi}\right\}^k . \tag{B.2}$$

Then, we have

$$\begin{split} &E\left[\Delta^{m}(w)\big|w\right] \\ &\leq \sum_{k=1}^{\lfloor\bar{K}^{m}w\rfloor} \mathbb{P}\left(\sum_{t=1}^{k} \left(\bar{s}^{m*} - \frac{w}{k+2} - A_{t}^{m}\right) < 0\right) + \sum_{k=\lceil\bar{K}^{m}w\rceil}^{\infty} \mathbb{P}\left(\sum_{t=1}^{k} \left(\bar{s}^{m*} - \frac{w}{k+2} - A_{t}^{m}\right) < 0\right) \\ &\leq \lfloor\bar{K}^{m}w\rfloor + \sum_{k=\lceil\bar{K}^{m}w\rceil}^{\infty} \{\tilde{\varphi}\}^{k} \leq \lfloor\bar{K}^{m}w\rfloor + \frac{\{\tilde{\varphi}\}^{\lceil\bar{K}^{m}w\rceil}}{1 - \tilde{\varphi}} \leq \bar{K}^{m}w + \frac{\tilde{\varphi}}{1 - \tilde{\varphi}} \;, \end{split}$$

where the second inequality follows from  $\mathbb{P}\left(\sum_{t=1}^{k} \left(\overline{s}^{m*} - \frac{w}{k} - A_t^m\right) < 0\right) \le 1$  and the inequality given in (B.2). This completes the proof.

# C. Supplementary Numerical Analysis

In this appendix, we conduct additional supplementary numerical experiments to illustrate the robustness of constant scheduling under various settings. Our purpose is to test the relative optimality gaps under different  $c_I/c_W$  values and other parameters, and see under which conditions the constant policy performs well. We fix the service durations distribution as an i.i.d. exponential distribution with mean 20, and set the number of appointments at n = 16 except for the last experiment.

### C.1. Patient No-Show and Unpunctuality

Since the case of unpunctual arrivals has been studied in the literature (Samorani and Ganguly 2016, Han et al. 2018). We also examine the performance of the constant policy in the presence of patient unpunctuality or no-shows. The numerical results are shown in Table 6. We start with the patients' unpunctual arrival. Specifically, we suppose that each customer arrives late by L time units, where L follows one of the two exponential distributions with mean 4 and 10, which we denote  $L \sim EXP(4)$  and  $L \sim EXP(10)$ . We report the relative optimality gap in Table 6. It is easy to see that the constant policy performs well when  $c_I/c_W$  is small, and the relative optimality gap exhibits an increasing trend as  $c_I/c_W$  increases. The changing trend is similar to Table 1. In addition, we particularly note that the relative performance of the constant policy is better with EXP(10) than with EXP(4). As patient arrivals are more variable, it is more difficult for the optimal policy to anticipate when exactly patients arrive, and the simplicity of the constant policy helps it to perform better. This observation is consistent with another observation that when we compare the first two lines of Table 6 to Table 1, we see that the relative optimality gap of the patient unpunctual case in Table 6 is smaller than the corresponding quantity in Table 1. This means that the relative performance of the constant policy is better when patients have a later arrival pattern. It seems that the later arrival behavior would enlarge the job allowances for jobs at the beginning and ending of the planning horizon, which makes the optimal schedule "flatter" than those in the case without later arrival behavior. As a result, the constant policy is a better approximation of the optimal policy when customers may not be punctual.

Next, we examine the impact of the no-show probability on the relative optimality gap. To do this, we test different  $c_I/c_W$  for both i.i.d. and non-i.i.d no-show cases. For the i.i.d. no-show case, we test four different show-up probabilities: p = 0.6, 0.7, 0.8, 0.9. For the non-i.i.d. no-show case, we also test two different sets of show-up probabilities: p1 which alternatives between 0.8 and 0.7, and p2 alternating between 0.9 and 0.6. From Table 6, we can find that both i.i.d. and non-i.i.d no-show cases exhibit a clear increasing trend as  $c_I/c_W$  increases, similar to the patient unpunctuality case discussed above. In addition, the relative optimality gap exhibits a clear increasing trend as the

Seenani	Scenarios			$ $ $c_I/c_W$										
Scenario	05	0.2	0.4	0.6	0.8	1	1.5	3						
Not pupetual	$L \sim EXP(10)$	0.17	0.32	0.33	0.44	0.56	0.93	1.08	1.22	1.49				
Not punctual	$L \sim EXP(4)$	0.27	0.45	0.61	0.89	1.19	1.46	1.52	1.81	2.09				
	p = 0.9	0.48	0.84	1.24	1.28	1.83	2.32	2.71	3.09	3.38				
LLD No al anna	p = 0.8	0.5	0.93	1.31	1.43	1.85	2.33	3.01	3.35	3.87				
I.I.D. NO-Shows	p = 0.7	0.56	1.06	1.52	1.72	2.06	2.75	3.17	3.72	4.09				
	p = 0.6	0.59	1.18	1.69	2.21	2.43	3.1	3.59	4.63	5.06				
New LLD, New horses	p1	0.93	1.2	1.57	2.12	2.33	2.8	3.2	3.51	3.72				
	p2	2.32	2.73	3.07	3.27	3.31	3.89	4.15	4.44	4.94				

Table 6 Relative optimality gaps (in %) in presence of patient unpunctuality or no-show; n = 16

show-up probability decreases for the i.i.d. no-show case. This result is reasonable. As derived in the section 4.3, a higher show-up probability p results in a higher  $\check{c}_W$  (i.e., the virtual unit waiting time cost in the presence of no-show), thus leading to a lower optimality gap. For the non-i.i.d. no-show case, the results are similar. We particularly note that the constant policy performs better when the fluctuation of show-up probabilities is smaller. This makes sense since the constant policy ignores this fluctuation, and it does not adapt to the changing environment. In summary, even though our theoretical results can not cover some cases, the constant scheduling policy still achieves near optimal performance in the presence of patient unpunctuality or no-show.

### C.2. The Case with Overtime Cost

Next, we test the performance of the constant policy when we consider the overtime cost. We set the session length T at three different level, i.e.,  $T = n\mu$ ,  $T = 1.5n\mu$  and  $T = 2n\mu$ , where n is the number of appointments and  $\mu$  is the mean of service duration ( $\mu = 20$ ). We test different combinations of  $c_I/c_W$  and unit overtime cost, which we denote by  $c_o$ . Following Zacharias and Pinedo (2017), we set the unit overtime cost at 1.5 and 3 times of unit idling cost. We made the following observations from Table 7. First, for small  $c_o$ , the relative optimality gap exhibits an increasing trend as  $c_I/c_w$  increases, which is consistent with what we observed from the no-overtime-cost case. Second, the optimality gap is generally smaller when the session time limit is larger or when the unit overtime cost is smaller, which brings the overtime cost problem closer to the original problem.

#### C.3. Job Allowance Comparison

Using the SAA method, we calculate the job allowance under the optimal constant policy,  $r_{SAA}$ , and the average job allowance under the optimal (non-constant) schedule  $\bar{s}_{SAA} := \sum_{t=1}^{n} s_{SAA,t}$ . We compare these values under different parameters,  $c_I/c_W$  and n. The results are summarized in Table

0	T		$c_I/c_W$											
$c_o$	1	0.2	0.4	0.6	0.8	1	1.5	2	2.5	3				
	$n\mu$	1.83	2.29	2.16	2.17	2.43	3.22	3.33	3.35	3.49				
1	$1.5n\mu$	2.19	2.19	1.97	1.82	2	2.42	2.92	3.07	3.02				
	$2n\mu$	0.55	0.76	0.97	1.21	1.3	2.04	2.56	2.75	3.06				
	$n\mu$	3	3.27	3.28	3.37	3.49	4.01	3.85	4.33	4.21				
3	$1.5n\mu$	2.24	2.37	2.16	2.24	2.52	2.88	3.13	3	3.41				
	$2n\mu$	0.6	0.83	1.35	1.45	1.47	1.8	2.42	2.78	2.98				
	$n\mu$	4.17	3.87	3.58	4.35	4.41	4.44	4.76	4.56	4.91				
5	$1.5n\mu$	2.53	2.66	3.02	3.08	2.88	2.97	3.18	3.51	3.72				
	$2n\mu$	0.78	0.75	0.99	1.19	1.72	1.98	2.58	2.73	3				

Table 7 Relative optimality gap (in %) for different session lengths T, ratios of  $c_I/c_W$  and  $c_o$ ; n = 16

8, from which we observe the follows. First, in all parametric settings,  $r_{SAA}$  is slightly larger than  $\bar{s}_{SAA}$ . This is because a constant policy has less flexibility and therefore calls for extra allowance to mitigate the impact of randomness. Second, the differences between  $r_{SAA}$  and  $\bar{s}_{SAA}$  are small, suggesting that the average job allowance in the optimal policy provides a close approximation to the job allowance that is used by the constant policy. This also implies that we did not relax too much in deriving the performance bounds in Theorem 1. Recall that a key step our analysis is to compare the optimal policy to a constant policy with job allowance  $\bar{s}^*$  (See Equation (11)). Finally, the gap between  $r_{SAA}$  and  $\bar{s}_{SAA}$  is smaller for larger values of n. Intuitively, the optimal policy takes a "plateau-dome" shape (Klassen and Yoogalingam 2009), and the "plateau" part dominates when n is large.

Table 8 Job allowance comparison: Optimal constant policy and the average in the optimal schedule

~						$c_I/c_W$				
$\overline{n}$		0.2	0.4	0.6	0.8	1	1.5	2	2.5	3
16	Constant	46.93	39.49	35.58	33.29	31.27	28.51	26.95	25.28	24.01
10	Optimal (average)	46.69	39.01	34.93	32.68	30.52	27.79	26.22	24.59	23.34
50	Constant	46.94	40.09	37.08	34.70	33.04	30.65	29.07	27.8	27.11
50	Optimal (average)	46.88	39.81	36.81	34.33	32.55	30.14	28.55	27.27	26.53
200	Constant	47.78	40.50	37.37	34.72	33.44	31.06	29.67	28.74	27.95
200	Optimal (average)	47.65	40.37	37.1	34.55	33.28	30.95	29.44	28.51	27.71

#### References

- Andersen ES (1955) On the fluctuations of sums of random variables II. *Mathematica Scandinavica* 2:195–223.
- Bingham N (2001) Random walk and fluctuation theory. Handbook of Statistics 19:171–213.
- Chen LH, Shao QM (2005) Stein's method for normal approximation. An introduction to Stein's method 4:1–59.
- Goldberg DA, Katz-Rogozhnikov DA, Lu Y, Sharma M, Squillante MS (2016) Asymptotic optimality of constant-order policies for lost sales inventory models with large lead times. *Mathematics of Operations Research* 41(3):898–913.
- Han Z, Chen Y, Leung E, Xing L (2018) Outpatient appointment scheduling with unpunctual patients. International Journal of Production Research (1):1–21.
- Kingman J (1962) Some inequalities for the queue GI/G/1. Biometrika 49(3/4):315-324.
- Klassen KJ, Yoogalingam R (2009) Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management* 18(4):447–458.
- Landsberger M, Meilijson (1993) Mean-preserving portfolio dominance. *Review of Economic Studies* 60(2):479–485.
- Samorani M, Ganguly S (2016) Optimal sequencing of unpunctual patients in high-service-level clinics. Production & Operations Management 25(2):330–346.
- Spitzer F (1956) A combinatorial lemma and its application to probability theory. Transactions of the American Mathematical Society 82(2):323–339.
- Zacharias C, Pinedo M (2017) Managing customer arrivals in service systems with multiple identical servers. Manufacturing & Service Operations Management 19(4):639–656.