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

(Authors' names blinded for peer review)

We consider the appointment scheduling problem, which determines the job allowance over the planning horizon. 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

1. Introduction

Scheduling appointments is commonly used by many agencies to provide consulting services, such as investment, tax, lawsuits, dental care, and medical diagnosis or treatment such as CT scan, magnetic resonance imaging (MRI). 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 general, solving the appointment scheduling problem to optimality is intractable (Robinson and Chen 2003). Even for a high quality solution with SAA method, 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 depicts a typical "dome"-shaped curve, where the height of the curve corresponds to the job allowance under the optimal policy.

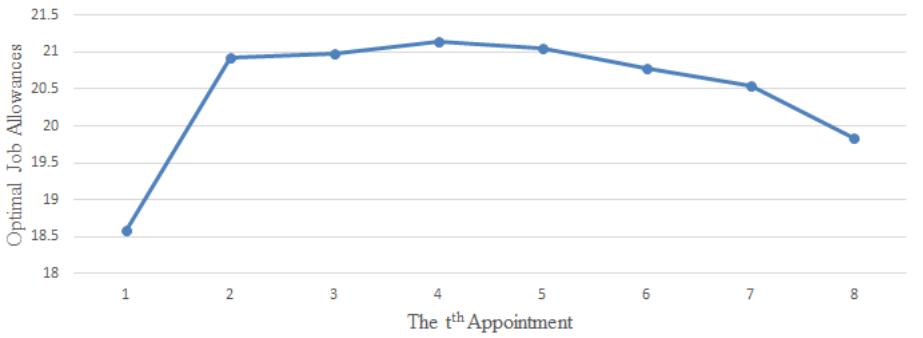

Figure 1 **Dome Shape of Optimal Schedule (service time follows a normal distribution with mean** 20 **and variance** 16**, with unit idle time cost** 3 **and unit waiting time cost** 1**)**

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, the constant policy has been used in hospitals in Ontario which provide 24-hour MRI service to patients who have booked appointments in advance.[1] 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 easiest to implement, and is also fair by assigning an equal service time 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

[1] 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.

target at high-valued 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 schedules 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 services. We also extend this framework to cover 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-folds 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. This upper bound also 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 looks into the performance of the constant policy 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, the methods we used to derive the analytical bounds for constant policy is promising to extend to more general settings and to make more interesting findings. In fact, we have made two relatively easy 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. 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). Some studied aim to minimize the day (session) end time (Klassen and Rohleder 2004, Hassin and Mendel 2008), which is equivalent to minimizing the total idle time. Some researchers assume that the session has a fixed length and aim to minimize the overtime cost as well as the waiting and idling cost (Denton and Gupta 2003, Kaandorp and Koole 2007, De Vuyst et al. 2014, Kong et al. 2013, Mak et al. 2014a). We note that including the overtime cost will make the question very different and our method can no longer apply to. However, the numerical results show that the constant policy performs near optimally even if the overtime cost has been included in the objective function; see Appendix C.

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.

The constant job allowance policy has been studied in a few previous works. Jansson (1966) studied an appointment booking system with constant job allowances and exponentially distributed service durations. He derived 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) studied the steady-state behavior of an appointment system in the presence of customer unpunctuality. Charnetski (1984) used service times of over $2,000$ real surgical procedures to evaluate the performance of a constant job allowance policy in an operating room. Later, Wang (1997) derived 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) studied a new appointment system in which customers are given the earliest possible appointment times under the service level constraints, and derived 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. (2016). 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 $\boldsymbol{A} = (A_1, A_2, \cdots, A_n)$. (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 (12) and (29)) allow the service times to follow any continuous distribution.

We assume that the $n+1$ appointments, indexed by $t = 1, 2, \cdots, n+1$, 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 $\boldsymbol{s} = (s_1, s_2, \cdots, s_n)$ 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 $\boldsymbol{s}$, 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. We 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. We assume that the doctor is always available at time 0, so that the first patient can see the

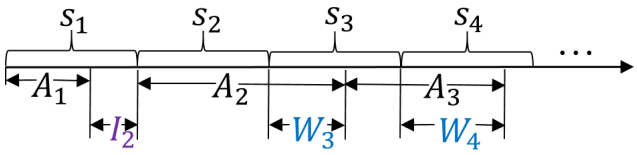

Figure 2 **Variables in Our Model**

doctor immediately, i.e., $W_1 = 0$. A graphical illustration of the variables introduced above is provided in Figure 2.

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 $t = 1, 2, \cdots, n$,

$$W_{t+1} = [W_t + A_t - s_t]^+ \qquad \text{and} \tag{1}$$

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

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,

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

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 $W_{t+1}(\boldsymbol{s})$ and $I_{t+1}(\boldsymbol{s})$ to highlight their dependence on the job allowance vector $\boldsymbol{s}$.

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 $\{A_t\}$, $t = 1, 2, \ldots, n$, are i.i.d. with the common probability density function f and cumulative distribution function F. Let μ and σ 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:

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

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 g denote the minimizer and the minimum value of $\nu(s_1)$ given above, i.e.,

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

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

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 $(t+1)^{th}$ 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., $s_t \le d$ for $t = 1, \ldots, n$. Let ζ denote the skewness of the random service time A and define τ to be a function of ζ and two cost parameters c_W and c_I:

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

For fixed r, we let random variable η_∞^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:

$$\eta_t^r = \min \left\{ j^* \in \{0, 1, \ldots, t\} \;\middle|\; \sum_{i=1}^{j^*} (A_i - r) = \max_{j \in \{0,1,\ldots,t\}} \sum_{i=1}^{j} (A_i - r) \right\} \;. \tag{7}$$

Since the summation in the above equation represents a random walk, an explicit, but still complicated, expression for the probability distribution of η_t^r is given in Theorem 1 of Andersen (1955). Also, by substituting $t = +\infty$, we let η_∞^r follow the limiting distribution of η_t^r as $t \to \infty$. This distribution is well defined provided $r > \mu$ (Spitzer 1956). Then, for each $n \ge 1$, define

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

Also, for $r > \mu$, define

$$\varphi_r \;=\; \inf_{\psi \ge 0} \phi_r(\psi) \quad \text{and} \quad \psi_r \;=\; \arg\inf_{\psi \ge 0} \phi_r(\psi), \qquad \text{where} \qquad \phi_r(\psi) \;=\; E[\exp(\psi(A-r))] \;. \tag{9}$$

These quantities are used in the statement of Theorem 1. In particular, ϵ_n will be used to provide a bound on the performance of the constant policy, and φ_r is used in an upper bound for ϵ_n.

Consider an optimal policy, and let s_t^* denote the job allowance of the t^{th} appointment under this policy. Let

$$r \;=\; \frac{1}{n} \sum_{t=1}^{n} s_t^* \tag{10}$$

denote the average job allowance under the optimal policy. We consider a constant policy r, in which all job allowances are set to the constant value r. We use $\nu(r)$ and ν^* to denote the expected total cost under the constant policy r and the optimal (possibly non-constant) policy, respectively.

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

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

Let $n \geq 1$. *Then, we have* $r > \mu$, *and*

$$\nu^* \;\leq\; \nu(r) \;\leq\; (1+\epsilon_n)\nu^* \;. \tag{11}$$

Furthermore,

$$\epsilon_n \;\leq\; \frac{2c_W r\varphi_r^2}{ng(1-\varphi_r)^2} + \frac{c_I\sigma^2}{2ng(r-\mu)} \;\to\; 0 \qquad \text{as } n \to \infty. \tag{12}$$

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

The relationship in (11) can be written as $0 \leq \nu(r)/\nu^* - 1 \leq \epsilon_n$. We refer to the ratio $\nu(r)/\nu^* - 1$ 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 ϵ_n on the right-hand-side of (8) and the other in (12). The first upper bound can be computed only when A_t follows any 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 (12) shows that a simple constant policy r is asymptotically optimal.

Theorem 1 is about near-optimality of the constant policy r, with r being the average job allowance of the optimal schedule. While constant policies are simple in general, implementing this constant policy however essentially requires solving the optimal schedule to obtain the value of r. This can be challenging for problems of large scale. Alternatively, we can find the optimal constant policy, i.e., the constant policy with the best constant job allowance, by a one-dimensional search. The expected total cost can be evaluated by the sample average approximation. Since the optimal constant policy performs at least as

well as the constant policy r, the upper bound of Theorem 1 applies. The details of the numerical algorithm for computing the optimal constant policy is given in Section 6.

We now elaborate on η_t^r and 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

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

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

$$W_{t+1}^r \;\stackrel{d}{=}\; \max_{j\in\{0,1,\ldots,t\}} \left\{ \sum_{i=1}^{j} (A_i - r) \right\} . \tag{13}$$

We have defined η_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 (13), where ties are broken in favor of the smaller index.

We comment on Conditions 1 and 2 used in Theorem 1. Each of them provides a sufficient condition for the value of $r = (s_1^* + \cdots + s_n^*)/n$ given in (10) to satisfy $r > \mu$. This ensures that r is sufficiently high so that the random walk used in (13) has a negative drift, implying that η_∞^r has a finite expectation.

LEMMA 2. *If at least one of the two conditions in Theorem 1 holds, then* $r > \mu$.

While a rigorous proof for the above lemma is provided in Appendix A.2, 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 $r = (s_1^* + \cdots + s_n^*)/n \geq \underline{s}$, it follows that $r > \mu$. 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 $n \geq 1$, the bound in (11) 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, possibly smaller than μ. 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 $\boldsymbol{s}^* = (s_1^*, s_2^*, \cdots, s_n^*)$ should be large enough so that their average, r, should be at least μ; 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.

Finally, we elaborate on the role of this property, $r > \mu$. Note that ϵ_n is an upper bound on the relative optimality gap. According to classic random walk theory (see e.g., Karlin 2014), η_∞^r, used in the definition of ϵ_n given in (8), has finite first and second moments if and only if $r > \mu$. Thus, $E[(\eta_\infty^r)^2 - \eta_\infty^r] < +\infty$ is finite if and only if $r > \mu$. This property, $r > \mu$, is sufficient and necessary condition for the convergence of ϵ_n to 0.

4.2. Proof of Theorem 1

Now we prove Theorem 1. To prove an upper bound for $(\nu(r) - \nu^*)/\nu^*$, we first derive an alternative expression for $\nu(r) - \nu^*$, and provide an upper bound for the above difference. Then, by lower bounding ν^*, we derive an upper bound for $(\nu(r) - \nu^*)/\nu^*$.

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

LEMMA 3. *For any given schedule* $\boldsymbol{s}$*,*

$$\nu(\boldsymbol{s}) \;=\; E\left[\sum_{t=1}^{n} c_W W_{t+1}\right] \;+\; c_I \sum_{t=1}^{n} s_t \;+\; c_I E\left[W_{n+1}\right] \;-\; c_I n\mu \;.$$

According to Lemma 3, for any constant policy with the constant job allowance r', we have

$$\nu(r') - \nu(\boldsymbol{s}^*) \;=\; c_W \sum_{t=1}^{n} E\left[W_{t+1}^{r'} - W_{t+1}^*\right] + c_I \left(nr' - \sum_{t=1}^{n} s_t^*\right) + c_I E\left[W_{n+1}^{r'} - W_{n+1}^*\right], \quad (14)$$

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

We consider a constant policy with $r' = r$, where $r = (s_1^* + \cdots + s_n^*)/n$. Then, the second term above becomes 0.

To upper bound the first term, consider an alternative expression for $E[W_{t+1}^r]$ using (13), and the definition of η_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 (13):

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

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

$$E\left[W_{t+1}^r - W_{t+1}^*\right] \;\leq\; E\left[\left(\sum_{i=t+1-\eta_t^r}^{t} s_i^*\right) - r\eta_t^r\right] \;. \tag{15}$$

Thus, we have

$$\sum_{t=1}^{n} E\left[W_{t+1}^r - W_{t+1}^*\right] \;\leq\; \sum_{t=1}^{n} E\left[\left(\sum_{i=t+1-\eta_t^r}^{t} s_i^*\right) - r\eta_t^r\right] \;\leq\; rE[(\eta_\infty^r)^2 - \eta_\infty^r] \;, \tag{16}$$

where the second inequality holds provided that the following lemma 4 holds. The proof of the lemma below is in Appendix A.4.

Lemma 4.

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

Hence, with (16), the first term in (14) can be bound as follows

$$c_W \sum_{t=1}^{n} E\left[W_{t+1}^r - W_{t+1}^*\right] \;\leq\; c_W r E[(\eta_\infty^r)^2 - \eta_\infty^r] \;. \tag{17}$$

To upper bound the third term in (14), which is $c_I E[W_{n+1}^r - W_{n+1}^*]$ when $r' = r$, 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 inter-arrival 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

$$E[W_t^r] \;\leq\; E[W_\infty^r] \qquad \text{for all } t \geq 1. \tag{18}$$

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

$$\overline{\Upsilon}_1 \;=\; \frac{\sigma^2}{2r(1-\rho)} \qquad \text{and} \qquad \overline{\Upsilon}_2 \;=\; \frac{4}{e^2(\psi_r)^2}\log\frac{1}{1-\varphi_r} + \left[\frac{\sigma^2}{2r(1-\rho)}\right]^2. \tag{19}$$

Lemma 5. *Consider a $D/G/1$ queue with a deterministic inter-arrival time r and random service time A, with mean μ and standard deviation σ. If $\rho = \mu/r < 1$, then $\psi_r > 0$ holds, and the steady-state waiting distribution W_∞ has the following upper bounds on the first and second moments:*

$$E[W_\infty] \;\leq\; \overline{\Upsilon}_1 \qquad \textit{and} \qquad E[W_\infty^2] \leq \overline{\Upsilon}_2 \;.$$

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 (14). Since $r > \mu$ holds by Lemma 2, we can apply Lemma 5 to obtain

$$c_I E[W_{n+1}^r - W_{n+1}^*] \;\leq\; c_I E[W_{n+1}^r] \;\leq\; c_I \overline{\Upsilon}_1 \;=\; \frac{c_I \sigma^2}{2r(1-\rho)}. \tag{20}$$

Based on the above argument, we are now ready to establish an upper bound on (14), when $r' = r$. From (17) and (20), we can bound $\nu(r) - \nu(\boldsymbol{s}^*)$ in (14) as follows:

$$\nu(r') - \nu(\boldsymbol{s}^*) \;\leq\; c_W r E[(\eta_\infty^r)^2 - \eta_\infty^r] \;+\; 0 \;+\; \frac{c_I \sigma^2}{2r(1-\rho)} \;.$$

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

$$\frac{\nu(r) - \nu^*}{\nu^*} \leq \frac{c_W r E[(\eta_\infty^r)^2 - \eta_\infty^r]}{ng} + \frac{1}{ng} \cdot \frac{c_I \sigma^2}{2r(1-\rho)} \;=\; \epsilon_n \;,$$

where the last equality follows from the definition of ϵ_n in (8). This establishes (11), the first result in the statement of Theorem 1.

Now, we provide an analytical upper bound for ϵ_n, by deriving a bound on $E[(\eta^r_\infty)^2 - \eta^r_\infty]$. Consider the discrete-time random walk defined by $\sum_{i=1}^{j}(A_i - r)$ in $j \in \{0, 1, \ldots, t\}$, and let N_t denote the number of times when this random walk has a positive value. By Andersen (1955) and Bingham (2001), the index η^r_t defined in (7) has the same distribution as the random number N_t for each $t = 0, 1, \ldots, +\infty$. Let N_∞ 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 Lemma 2. By invoking Theorem 5.2 of Spitzer (1956), we derive the following alternative expression for $E[(\eta^r_\infty)^2 - \eta^r_\infty]$:

$$E[(\eta^r_\infty)^2 - \eta^r_\infty] \;=\; E[(N_\infty)^2 - N_\infty] \;=\; E[N_\infty(N_\infty - 1)] \;=\; \left(\sum_{k=1}^{\infty} b_k\right)^2 + \sum_{k=2}^{\infty}(k-1)b_k,$$

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

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

where the last equality follows from the definition of $\phi_r(\psi) = E[\exp(\psi(A - r))]$ in (9). Since the above inequality holds for any $\psi > 0$, it follows from (9) 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 ψ on $(0, \infty)$ from the definition of $\phi_r(\psi)$ in (9). 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,

$$\begin{aligned} E[(\eta^r_\infty)^2 - \eta^r_\infty] &= \left(\sum_{k=1}^{\infty} b_k\right)^2 + \sum_{k=2}^{\infty}(k-1)b_k \\ &\leq \left(\sum_{k=1}^{\infty}(\varphi_r)^k\right)^2 + \sum_{k=2}^{\infty}(k-1)(\varphi_r)^k \;\leq \frac{2\cdot(\varphi_r)^2}{(1-\varphi_r)^2} \,, \end{aligned} \tag{21}$$

where the last inequality directly from the basic manipulation of the geometric series $(\varphi_r)^k$. By substituting the above inequality to the definition of ϵ_n given in (8), we obtain

$$\epsilon_n \;=\; \frac{c_W r E[(\eta^r_\infty)^2 - \eta^r_\infty]}{ng} + \frac{c_I \sigma^2}{2ng(r-\mu)} \;\leq\; \frac{2c_W r \varphi_r^2}{ng(1-\varphi_r)^2} + \frac{c_I \sigma^2}{2ng(r-\mu)} \,,$$

which is the inequality in (12) 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 4 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 η_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, η_t^r used in (A.7) 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).

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 $(\check{A}_1, \check{A}_2, \cdots, \check{A}_n)$ denote the service duration vector in the presence of no-shows. Then, for each $t \in \{1, 2, \cdots, n\}$, we have $\check{A}_t = Z_t A_t$, where $(A_1, \ldots, A_n)$ 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 $t \in \{1, \ldots, n\}$,

$$\begin{aligned} \check{W}_{t+1} &= [\check{W}_t + \check{A}_t - s_t]^+ \qquad \text{and} \\ \check{I}_{t+1} &= [\check{W}_t + \check{A}_t - s_t]^- \ , \end{aligned}$$

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 $\boldsymbol{s}$ can be formulated as, similar to (3),

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

Since Z_{t+1} is independent of $(Z_1, Z_2, \cdots, Z_t)$, 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 $\boldsymbol{s}$, we have

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

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 i.i.d service duration assumption in the basic model (Sections 3 and 4), and show that a generalization of the constant policy achieves asymptotic optimality.

We model the service duration to be piecewisely i.i.d., that is, there can be $M \geq 2$ different types of appointments, and within each type, service duration distributions are identical. Furthermore, we assume that all the appointments of any given type are sequenced consecutively, forming a block in the schedule. In other words, the entire appointment sequence contains M blocks, each consisting of appointments of the same type. 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 μ^m, and σ^m, respectively. 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. All service durations are independent, as before.

The policy that we consider, 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. Let s_t^{m*} denote the optimal job allowance of the t^{th} patient in the m^{th} block under the optimal schedule. Let

$$r^m \;=\; \frac{1}{n^m}\sum_{t=1}^{n^m} s_t^{m*} \tag{22}$$

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

$$\begin{aligned} \underline{s}^m &= \arg\min_{s \in \Re^+} \; c_W \cdot E[A^m - s]^+ + c_I \cdot E[A^m - s]^- \qquad \text{and} \\ g^m &= \min_{s \in \Re^+} \; c_W \cdot E[A^m - s]^+ + c_I \cdot E[A^m - s]^- \; . \end{aligned} \tag{23}$$

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

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

Let $\rho^m = \mu^m / r^m$. Similar to (9), define for $r^m > \mu^m$,

$$\varphi^m_{r^m} \;=\; \inf_{\psi \geq 0} \phi^m_{r^m}(\psi) \quad \text{and} \quad \psi^m_{r^m} \;=\; \arg\inf_{\psi \geq 0} \phi^m_{r^m}(\psi) \quad \text{where} \quad \phi^m_{r^m}(\psi) \;=\; E[\exp(\psi(A^m - r^m))]. \tag{24}$$

Similarly, we can define $\phi^m_{\bar{r}^m}(\psi)$, $\psi^m_{\bar{r}^m}$, $\varphi^m_{\bar{r}^m}$ for a different constant policy $\bar{r}^m$, with

$$\bar{r}^m = r^m - \frac{1}{\bar{K}^m}, \qquad \text{where} \qquad \bar{K}^m = \left\lceil \frac{2}{r^m - \mu^m} \right\rceil \; . \tag{25}$$

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

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

$$\eta_t^{m,r^m} = \min \left\{ j^* \in \{0, 1, \ldots, t\} \;\middle|\; \sum_{i=1}^{j^*} (A_i^m - r^m) = \max_{j \in \{0,1,\ldots,t\}} \sum_{i=1}^{j} (A_i^m - r^m) \right\} \; .$$

Furthermore, let η_∞^{m,r^m} be the limiting distribution of η_t^{m,r^m} as $t \to \infty$. We let, similar to (19),

$$\overline{\Upsilon}_1^m \;=\; \frac{(\sigma^m)^2}{2r^m(1-\rho^m)} \qquad \text{and} \qquad \overline{\Upsilon}_2^m \;=\; \frac{4}{e^2(\psi^m_{r^m})^2} \log \frac{1}{1 - \varphi^m_{r^m}} + \left[\frac{(\sigma^m)^2}{2r^m(1-\rho^m)} \right]^2 \; ,$$

and define

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

where

$$\tilde{\varphi} \;=\; \max_{m \in \{1,2,\ldots,M\}} \varphi^m_{\bar{r}^m} \; . \tag{27}$$

Finally, instead of the definition of ϵ_n given in (8), we define

$$\epsilon_n \;=\; \frac{c_W \sum_{m=1}^{M} \left\{ r^m E\left[\left(\eta_\infty^{m,r^m}\right)^2 - \eta_\infty^{m,r^m} \right] + \Omega^m \right\} \;+\; c_I \sum_{m=1}^{M} \overline{\Upsilon}_1^m}{\sum_{m=1}^{M} n^m g^m} \;. \tag{28}$$

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

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

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

$$\nu^* \;\leq\; \nu(\boldsymbol{r}) \;\leq\; (1+\epsilon_n)\nu^* \;.$$

Furthermore,

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

Similar to Theorem 1, Theorem 2 provides two upper bounds for the relative optimality gap $\nu(\boldsymbol{r})/\nu^* - 1$. The first upper bound, which is ϵ_n 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 ϵ_n, and it converges to zero when $n \to \infty$, proving the asymptotic optimality of the piecewise constant policy $\boldsymbol{r}$.

Condition $1'$ and $2'$ are simply adapting Condition 1 and 2 to a system with only one type of appointment. By Lemma 2, Condition $1'$ and $2'$ provide a sufficient condition for $r^m > \mu^m$ for each m, which gives a sufficient and necessary condition for the first and second moment of η_∞^{m,r^m} to be finite. To achieve this, we consider the waiting-time distributions $\{W_t^{m,r^m}\}$, where $t = 1, 2, \ldots$, in a $D/G/1$ queue with the constant inter-arrival time r^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 5 of Section 4, on the first two moments of the steady-state waiting-time distribution W_∞^{m,r^m} :

$$E\left[W_\infty^{m,r^m}\right] \;\leq\; \overline{\Upsilon}_1^m \qquad \text{and} \qquad E\left[\left(W_\infty^{m,r^m}\right)^2\right] \;\leq\; \overline{\Upsilon}_2^m \;. \tag{30}$$

Earlier in Section 4, recall from (18) 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 ϵ_n in (28) is equivalent to the corresponding definition in (8) of Section 4, and the bound in (29) simplifies to (12).

However, for subsequent blocks ($m \geq 2$), 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 Ω^m term included in the upper bounds in (28) and (29) above.

We now investigate the Ω^m expressions to gain insight into how the order of the blocks affects the overall performance. Recall from the definition of each Ω^m in (26) that Ω^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 $\sum_{m=k+1}^{M}(m-1)\bar{K}^m\overline{\Upsilon}_2^k$. In other words, a block scheduled earlier in the schedule has a smaller k and thus a larger weight, $(M-k)$ or $\sum_{m=k+1}^{M}(m-1)\bar{K}^m$, 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 $\nu(\boldsymbol{r}) - \nu^*$, then find an upper bound for that difference by choosing a particular piecewise constant policy $\boldsymbol{r}$, and finally derive a lower bound for ν^*, leading to an upper bound for the relative optimality gap $\nu(\boldsymbol{r})/\nu^* - 1$. 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 $\nu(\boldsymbol{r}) - \nu^*$.

Throughout the proof, we use $\boldsymbol{s}^* = (s_t^{m*})$ 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 $\boldsymbol{r} = (r^1, \ldots, r^M)$, where each r^m is defined according to (22). In comparison, for fixed m, we use the notation W_t^{m,r^m} to denote the waiting time for the t^{th} patient if the inter-arrival times are deterministically r^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,r^m} is exactly the same as W_t^r used in Section 4. The only difference between $\overline{W}_t^m$ and W_t^{m,r^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,r^m} is initially empty.

For the first step of finding an upper bound for $\nu(\boldsymbol{r}) - \nu^*$, we derive an alternative expression for this quantity, similar to (14).

Lemma 6.

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

A formal proof of Lemma 6 is attached in Appendix B.1. We develop an upper bound for the right-side of (31). From the choice of $r^m = (s_1^{m*} + s_2^{m*} + \cdots + s_{n^m}^{m*})/n^m$ given in (22), r^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 (15). 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 alternate system allows preemption. Figure 3 illustrates both the original system (under "System 1") and the alternate system (under "System 2").

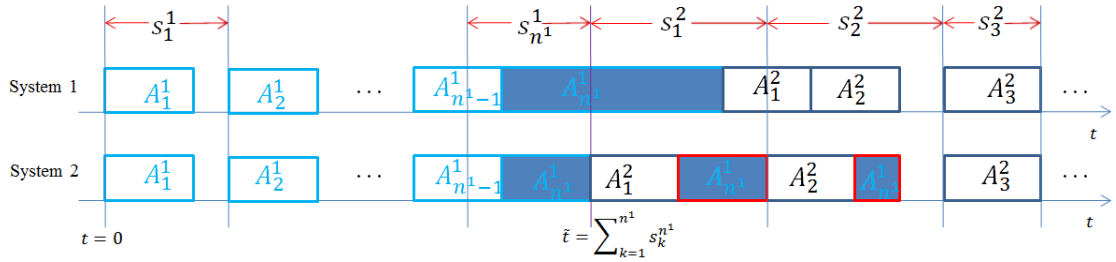

Figure 3 **An Alternative System that Holds the Jobs from the Previous Subsequence tills the Server Is Idle**

It is easy to verify that the server is busy in the original system if and only if the server is busy in the alternate 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 m in either the original system or the alternate system. Recall that the job allowance for each patient in block m is r^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:

$$\Delta^m(w) \;=\; \min\left\{t \in \{1,\ldots\} \;\middle|\; w + (A_1^m + \cdots + A_{t-1}^m) \le (t-1)\cdot r^m\right\} \;. \tag{32}$$

Recall that W_t^{m,r^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 Δ^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, $\overline{W}_t^m = W_t^{m,r^m}$ provided that $t \geq \Delta^m(\overline{W}_1^m)$. 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., $W_t^{m,r^m} \leq \overline{W}_t^m \leq W_t^{m,r^m} + \overline{W}_1^m$. (To see this, note that the original system is busy whenever the system corresponding to W_t^{m,r^m}, i.e., starting with an empty buffer, is busy.) Thus,

$$\overline{W}_t^m \begin{cases} \leq W_t^{m,r^m} + \overline{W}_1^m & \text{if } t < \Delta^m(\overline{W}_1^m) \\ = W_t^{m,r^m} & \text{if } t \geq \Delta^m(\overline{W}_1^m). \end{cases} \tag{33}$$

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 (27). The proof of this lemma appears in Appendix B.2.

LEMMA 7. *For any $w > 0$,*

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

From (33),

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

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 side above using (16):

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

To bound the second term on the right side of (34), we use Lemma 7 to obtain

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

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

$$\overline{W}_1^m \;=\; \overline{W}_{n^{m-1}+1}^{m-1} \;\leq\; W_{n^{m-1}+1}^{m-1,r^{m-1}} + \overline{W}_1^{m-1} \;.$$

By repeatedly applying the above inequality, we obtain

$$\overline{W}_1^m \;\le\; W_{n^{m-1}+1}^{m-1,r^{m-1}} + \overline{W}_1^{m-1} \;\le\; \cdots \;\le\; \sum_{k=1}^{m-1} W_{n^k+1}^{k,r^k} \tag{37}$$

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

$$E\left[\overline{W}_1^m\right] \;\le\; \sum_{k=1}^{m-1} E\left[W_{n^k+1}^{k,r^k}\right] \;\le\; \sum_{k=1}^{m-1} \overline{\Upsilon}_1^k \;. \tag{38}$$

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

$$E\left[\left(\overline{W}_1^m\right)^2\right] \;\le\; E\left[\left(\sum_{k=1}^{m-1} W_\infty^{k,r^k}\right)^2\right] \;\le\; (m-1)\sum_{k=1}^{m-1} E\left[\left(W_\infty^{k,r^k}\right)^2\right] \;\le\; (m-1)\sum_{k=1}^{m-1} \overline{\Upsilon}_2^k \;, \tag{39}$$

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

Putting together, we have

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

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

$$c_W \sum_{m=1}^{M} \sum_{t=1}^{n^m} E\left[\overline{W}_{t+1}^m - \overline{W}_{t+1}^{m*}\right] \;\le\; c_W \sum_{m=1}^{M} \left\{r^m E\left[(\eta_\infty^{m,r^m})^2 - \eta_\infty^{m,r^m}\right] + \Omega^m\right\}.$$

Finally, the second term in (31) satisfies

$$c_I E\left[\overline{W}_{n^M+1}^M - \overline{W}_{n^M+1}^*\right] \;\le\; c_I E\left[\overline{W}_{n^M+1}^M\right] \;=\; c_I E\left[\overline{W}_1^{M+1}\right] \;\le\; c_I \sum_{m=1}^{M} \overline{\Upsilon}_1^m \;,$$

where the last inequality follows from (38). Therefore, from Lemma 6, we conclude

$$\nu(\boldsymbol{r}) - \nu^* \;=\; \nu(\boldsymbol{r}) - \nu(\boldsymbol{s}^*) \;\leq\; c_W \sum_{m=1}^{M} \left\{ r^m E\left[(\eta_\infty^{m,r^m})^2 - \eta_\infty^{m,r^m} \right] + \Omega^m \right\} + c_I \sum_{m=1}^{M} \overline{\Upsilon}_1^m \;.$$

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

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,

$$\frac{\nu(\boldsymbol{r}) - \nu^*}{\nu^*} \;\leq\; \frac{c_W \sum_{m=1}^{M} \left\{ r^m E\left[\left(\eta_\infty^{m,r^m}\right)^2 - \eta_\infty^{m,r^m} \right] + \Omega^m \right\} \;+\; c_I \sum_{m=1}^{M} \overline{\Upsilon}_1^m}{\sum_{m=1}^{M} n^m g^m} \;=\; \epsilon_n \;, \qquad (40)$$

where the equality follows from the definition of ϵ_n in (28). Furthermore, since (21) implies

$$r^m E\left[(\eta_\infty^{m,r^m})^2 - \eta_\infty^{m,r^m} \right] \;\leq\; \frac{2 \cdot r^m \cdot (\varphi_{r^m}^m)^2}{(1 - \varphi_{r^m}^m)^2} \;,$$

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 $\boldsymbol{s}_{SAA}$, 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 $\nu(\boldsymbol{s}_{SAA})$ 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 $(\nu(r_{SAA}) - \nu(\boldsymbol{s}_{SAA}))/\nu(\boldsymbol{s}_{SAA})$ 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. 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. 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 $\nu(\boldsymbol{s}_{SAA})$ and $\nu(r_{SAA})$

Table 1 The gaps (in %) for different service time distributions and ratios of c_I/c_W: i.i.d. service durations

Distribution	μ	σ	c_I/c_W										
			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

to approximate the relative optimality gap, $(\nu(r_{SAA}) - \nu(\boldsymbol{s}_{SAA}))/\nu(\boldsymbol{s}_{SAA})$. The ratios are summarized in Table 1.

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 small-size 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. 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, σ/μ. 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.

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 $\boldsymbol{s}_{SAA}$, we use a smaller $K = 400$ for problems when $n \geq 100$. Then, the standard error has an order of 0.01, and thus at least the first digit after the decimal is of reliable accuracy. We fix $c_I = 3$, and calculate the relative optimality gap for $n = 25, 50, 100, 200$, respectively. The results are summarized in Table 2. 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 three types of service distributions.

Table 2 The gaps (in %) for different problem sizes: i.i.d. service durations

Distribution	μ	σ	n			
			25	50	100	200
Normal	20	4	1.3	0.6	0	0
	20	8	1.5	0.7	0.3	0
Exponential	20	20	2.3	1.4	0.5	0.1

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 four service duration distributions. The first three distributions are the same as those used previously in Section 6.1, and the fourth distribution that we added is an exponential distribution with mean 4. Six combinations of these distributions were tested. Similar to the setting used in Table 1, we use a small problem size with $n = 16$, with 8 patients for each type, and set $K = K' = 2000$. The results are summarized in Table 3.

Table 3 The gaps (in %) for 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	μ	σ	Distribution	μ	σ	0.2	0.4	0.6	0.8	1	1.5	2	2.5	3
Normal	20	4	Normal	20	8	0.20	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	0.90	1.37	1.86	2.25	2.72	2.88	3.01
Exponential	4	4	Exponential	20	20	0.40	0.49	0.83	0.94	1.09	1.29	1.79	1.90	1.93
Exponential	20	20	Exponential	4	4	3.41	3.67	3.90	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	1.50	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	7.50

From Table 3, 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

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 normal 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 and fourth rows of the table (with two exponential distributions, one with standard deviation 4 and another with standard deviation 20). 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.

7. Conclusion

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 several contributions in this paper. 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. The numerical experiments show that the constant policy and the piecewise constant policy perform well, providing support to the implementation of these scheduling polices in practice.

We also establish the explicit relative optimality gap bounds between the constant policy or its extension and the optimal schedule. Further, we prove the asymptotic optimality of our policy, once again advocating for its use in practice. Also, to the best of our knowledge, this is the first proven result to be within $1+\epsilon$ of the optimal schedule, from theoretical

aspect. Furthermore, we have identified several managerial insights – some of which are new and affirm observations made in the literature.

Furthermore, 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. 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. These considerations are beyond the scope of this paper and potential for future research.

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.

Bingham N (2001) Random walk and fluctuation theory. *Handbook of Statistics* 19:171–213.

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 LH, Shao QM (2005) Stein's method for normal approximation. *An introduction to Stein's method* 4:1–59.

Chen RR, Robinson LW (2014) Sequencing and scheduling appointments with potential call-in patients. *Production and Operations Management* 23(9):1522–1538.

De Vuyst S, Bruneel H, Fiems D (2014) Computationally efficient evaluation of appointment schedules in health care. *European Journal of Operational Research* 237(3):1142–1154.

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.

Han Z, Chen Y, Leung E, Xing L (2018) Outpatient appointment scheduling with unpunctual patients. *International Journal of Production Research* (1):1–21.

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 (2016) Appointment scheduling under schedule-dependent patient no-show behavior. Working paper,National University of Singapore, Singapore.

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.

Landsberger M, Meilijson (1993) Mean-preserving portfolio dominance. *Review of Economic Studies* 60(2):479–485.

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.

Samorani M, Ganguly S (2016) Optimal sequencing of unpunctual patients in high-service-level clinics. *Production & Operations Management* 25(2):330–346.

Soriano A (1966) Comparison of two scheduling systems. *Operations Research* 14(3):388–397.

Spitzer F (1956) A combinatorial lemma and its application to probability theory. *Transactions of the American Mathematical Society* 82(2):323–339.

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),

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

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 $t \in \{1, \ldots, n\}$, 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 $\{1, 2, \ldots, t-1\}$. For patients in $\{t+1, \ldots, n\}$, the choice of s_t has an impact on the cost through waiting time distributions $\{W_{t+1}, \ldots, W_n\}$; 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

$$C_t \;=\; c_W \cdot E[W_t + A_t - s_t]^+ + c_I \cdot E[W_t + A_t - s_t]^- \;. \tag{A.2}$$

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

$$-c_W \cdot \mathbb{P}[W_t + A_t > s_t] + c_I \cdot E[W_t + A_t \leq s_t] \;.$$

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 $-c_W \cdot \mathbb{P}[A_t > s_t] + c_I \cdot E[A_t \leq s_t]$, 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 $s_t^* \geq \underline{s}$.

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

$$\begin{aligned}
g &= \min_{s \in \Re^+} \quad c_W \cdot E[A_t - s]^+ + c_I \cdot E[A_t - s]^- \\
&\leq \min_{s \in [-E[W_t], d - E[W_t]]} \quad c_W \cdot E[A_t - s]^+ + c_I \cdot E[A_t - s]^- \\
&\leq \min_{s \in [-E[W_t], d - E[W_t]]} \quad c_W \cdot E[W_t - E[W_t] + A_t - s]^+ + c_I \cdot E[W_t - E[W_t] + A_t - s]^- \\
&= \min_{s \in [0, d]} \quad c_W \cdot E[W_t + A_t - s]^+ + c_I \cdot E[W_t + A_t - s]^- \\
&\leq c_W \cdot E[W_t + A_t - s_t]^+ + c_I \cdot E[W_t + A_t - s_t]^- \;=\; C_t \;,
\end{aligned}$$

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

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 $r = (s_1^* + \cdots + s_n^*)/n \geq \underline{s} > \mu$, as required.

Now, consider Condition 2. This condition is stated in terms of the standard deviation of the service time, σ, and the upper bound d imposed on the choice of each time allowance, as well as τ, 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., $n \geq \max\{2\tau, 6\tau^{\frac{3}{2}}\sigma^{-1}(d - \underline{s})\}$. 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 $1 \leq h \leq n$.

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

$$\gamma_t \;=\; \sum_{k=t+1-h}^{t} s_k^* \;. \tag{A.3}$$

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

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

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

$$E[W_{t+1}^*] \geq \sigma h^{\frac{1}{2}} E\left[\max\left\{0, N + \frac{h\mu - \gamma_t}{\sigma h^{\frac{1}{2}}}\right\}\right] - 3\sigma\zeta \;, \tag{A.4}$$

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

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

where the last equality follows from Lemma 3, and a similar equation holds for the feasible policy given by $s_t = \tilde{r}$ for each t, it follows

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

Above, the second inequality follows from (18) and Lemma 5. Therefore, since $\nu(\boldsymbol{s}^*)$ is at least $c_W\sum_{t=1}^{n}E[W_{t+1}^*]$, which in turn can be bounded below using (A.4), we have

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

which implies

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

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,

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

Therefore,

$$E\left[\max\left\{0,(n+1-h)N + \sum_{t=h}^{n}\frac{h\mu-\gamma_t}{\sigma h^{\frac{1}{2}}}\right\}\right] \le n\left(3\zeta + \sqrt{\frac{2c_I}{c_W}} + \frac{1}{n}\cdot\frac{1}{\sqrt{2}}\cdot(\frac{c_I}{c_W})^{\frac{3}{2}}\right)h^{-\frac{1}{2}} \quad \text{(A.5)}$$

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.3), 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

$$\begin{aligned}\sum_{t=h}^{n}\gamma_t &= \sum_{t=h}^{n}\sum_{k=t+1-h}^{t}s_k^* \\ &= (s_1^*+s_2^*+\cdots+s_h^*) + \left(s_2^*+s_3^*+\cdots+s_{h+1}^*\right) + \cdots + \left(s_{n+1-h}^*+s_{n+2-h}^*+\cdots+s_n^*\right) \\ &\le h\sum_{t=1}^{n}s_t^* - h(h-1)\underline{s}\ ,\end{aligned}$$

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

$$\sum_{t=h}^{n}\gamma_t \;\le\; hnr - h(h-1)\underline{s} \;=\; h(n+1-h)r + h(h-1)(r-\underline{s}) \;\le\; h(n+1-h)r + h(h-1)(d-\underline{s})\;, \tag{A.6}$$

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

Then,

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

where the inequality follows by substituting the upper bound of $\sum_{t=h}^{n}\gamma_t$ given in (A.6) to (A.5), 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,

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

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

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

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

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

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

A.3. Proof of Lemma 3

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

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

This completes the proof. ■

A.4. Proof of Lemma 4

Recall from (7) that η_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 η_∞^r is the limiting distribution of η_t^r as $t \to \infty$. To establish the inequality in the statement of Lemma 4, 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:

$$\mathbb{P}(\eta_t^r = k) \;\geq\; \mathbb{P}(\eta_\infty^r = k) \qquad \text{for any } k \in \{0, 1, \ldots, t\}. \tag{A.7}$$

Intuitively, since η_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.7), fix k and t such that $1 \leq k \leq t$, and define

$$\begin{aligned}
\mathcal{F}_1 &= \left\{ \sum_{i=1}^{k}(A_i - r) > \sum_{i=1}^{j}(A_i - r) \qquad \text{for each j} \in \{0, 1, \ldots, k-1\} \right\} \\
\mathcal{F}_2 &= \left\{ \sum_{i=1}^{k}(A_i - r) \geq \sum_{i=1}^{j}(A_i - r) \qquad \text{for each j} \in \{k+1, \ldots, t\} \right\} \\
\mathcal{F}_3 &= \left\{ \sum_{i=1}^{k}(A_i - r) \geq \sum_{i=1}^{j}(A_i - r) \qquad \text{for each j} \in \{t+1, t+2, \ldots\} \right\} .
\end{aligned}$$

Then, the event $[\eta_t^r = k]$ is in the intersection of $\mathcal{F}_1$ and $\mathcal{F}_2$ whereas the event $[\eta_\infty^r = k]$ is in the intersection of $\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_3$. This completes the proof of claim (A.7).

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 $t \in \{1, \ldots, n\}$,

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

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

$$\begin{aligned}
\sum_{t=1}^{n} E\left[\sum_{i=t+1-\eta_t^r}^{t} s_i^* \right] &= \sum_{t=1}^{n}\sum_{i=1}^{t} s_i^* \mathbb{P}(\eta_t^r \geq t+1-i) \;=\; \sum_{i=1}^{n}\sum_{t=i}^{n} s_i^* \mathbb{P}(\eta_t^r \geq t+1-i) \\
&= \sum_{i=1}^{n}\sum_{t=1}^{n+1-i} s_i^* \mathbb{P}(\eta_{t+i-1}^r \geq t) \;=\; \sum_{i=1}^{n}\sum_{t=1}^{n+1-i} s_i^* \left\{ 1 - \sum_{k=0}^{t-1} \mathbb{P}(\eta_{t+i-1}^r = k) \right\} .
\end{aligned}$$

By applying claim (A.7), we obtain

$$\begin{aligned}
\sum_{t=1}^{n} E\left[\sum_{i=t+1-\eta_t^r}^{t} s_i^* \right] &\leq \sum_{i=1}^{n}\sum_{t=1}^{n+1-i} s_i^* \left\{ 1 - \sum_{k=0}^{t-1} \mathbb{P}(\eta_\infty^r = k) \right\} \;=\; \sum_{i=1}^{n}\sum_{t=1}^{n+1-i} s_i^* \mathbb{P}(\eta_\infty^r \geq t) \\
&\leq \sum_{i=1}^{n}\sum_{t=1}^{\infty} s_i^* \mathbb{P}(\eta_\infty^r \geq t) \;=\; \frac{1}{n}\sum_{i=1}^{n} s_i^* \sum_{j=1}^{n}\sum_{t=1}^{\infty} \mathbb{P}(\eta_\infty^r \geq t) \;=\; r\sum_{j=1}^{n}\sum_{t=1}^{\infty} \mathbb{P}(\eta_\infty^r \geq t),
\end{aligned} \tag{A.8}$$

where the last equality follows from the choice of $r = \sum_{i=1}^{n} s_i^*/n$ given in (10).

Next we find a lower bound for $\sum_{t=1}^{n} rE[\eta_t^r]$. Applying claim (A.7), we have

$$\sum_{t=1}^{n} E[\eta_t^r] \;=\; \sum_{t=1}^{n}\sum_{j=1}^{t} \mathbb{P}(\eta_t^r = j)j \;=\; \sum_{j=1}^{n}\sum_{t=1}^{j} t\mathbb{P}(\eta_j^r = t) \;\geq\; \sum_{j=1}^{n}\sum_{t=1}^{j} t\mathbb{P}(\eta_\infty^r = t)\ .$$

which implies

$$\sum_{t=1}^{n} rE[\eta_t^r] \;\geq\; r\sum_{j=1}^{n}\sum_{t=1}^{j} t\mathbb{P}(\eta_\infty^r = t)\ . \tag{A.9}$$

We now compare the bounds identified in (A.8) and (A.9). For fixed j, note

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

Hence, summing over j, it follows

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

Thus, we conclude

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

completing the proof of Lemma 4. ■

B. Proofs for Section 5

B.1. Proof of Lemma 6

It follows from the proof of Lemma 3 that, for optimal schedule $\boldsymbol{s}^*$, we have

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

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

$$\begin{aligned}\nu(\boldsymbol{r}) - \nu(\boldsymbol{s}^*) &= c_W \sum_{m=1}^{M}\sum_{t=1}^{n^m} E\left[\overline{W}_{t+1}^{m} - \overline{W}_{t+1}^{m*}\right] + c_I E\left[\overline{W}_{n^m+1}^{m} - W_{n^m+1}^{*}\right] + c_I \sum_{m=1}^{M}\left\{n^m r^m - \sum_{t=1}^{n^m} s_t^{m*}\right\} \\ &= c_W \sum_{m=1}^{M}\sum_{t=1}^{n^m} E\left[\overline{W}_{t+1}^{m} - \overline{W}_{t+1}^{m*}\right] + c_I E\left[\overline{W}_{n^m+1}^{m} - W_{n^m+1}^{*}\right] + 0.\end{aligned}$$

This is the required result. ■

B.2. Proof of Lemma 7

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

$$\Delta^m(w) \ = \ \min\left\{t \in \{1,2,\ldots\} \ \middle| \ w + (A_1^m - r^m) + \cdots + (A_{t-1}^m - r^m) \le 0\right\} \ . \tag{B.1}$$

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

$$\begin{aligned}E\left[\Delta^m(w)\middle|w\right] &= \sum_{k=1}^{\infty}\mathbb{P}\left(\Delta^m(w) \ge k\right) \\ &\le \sum_{k=1}^{\infty}\mathbb{P}\left(w + \sum_{t=1}^{k-2}(A_t^m - r^m) > 0\right)\end{aligned}$$

$$= \sum_{k=1}^{\infty} \mathbb{P}\left(\sum_{t=1}^{k-2} \left(r^m - \frac{w}{k-2} - A_t^m \right) < 0 \right)$$
$$= \sum_{k=1}^{\infty} \mathbb{P}\left(\sum_{t=1}^{k} \left(r^m - \frac{w}{k} - A_t^m \right) < 0 \right) .$$

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,

$$\mathbb{P}\left(\sum_{t=1}^{k} \left(r^m - \frac{w}{k} - A_t^m \right) < 0 \right) \;\leq\; \mathbb{P}\left(\sum_{t=1}^{k} \left(r^m - \frac{1}{\bar{K}^m} - A_t^m \right) < 0 \right) \;=\; \mathbb{P}\left(\sum_{t=1}^{k} (\bar{r}^m - A_t^m) < 0 \right) .$$

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

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

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

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

Then, we have

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

where the second inequality follows from $\mathbb{P}\left(\sum_{t=1}^{k} \left(r^m - \frac{w}{k} - A_t^m \right) < 0 \right) \leq 1$ 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 4. 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 4. 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 4 to Table 1, we see that the relative optimality gap of the patient unpunctual case in Table 4 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.

Table 4 The gaps (in %) in presence of patient unpunctuality or no-show; $n = 16$

Scenarios		c_I/c_W								
		0.2	0.4	0.6	0.8	1	1.5	2	2.5	3
Not punctual	$L \sim EXP(10)$	0.17	0.32	0.33	0.44	0.56	0.93	1.08	1.22	1.49
	$L \sim EXP(4)$	0.27	0.45	0.61	0.89	1.19	1.46	1.52	1.81	2.09
I.I.D. No-shows	$p = 0.9$	0.48	0.84	1.24	1.28	1.83	2.32	2.71	3.09	3.38
	$p = 0.8$	0.5	0.93	1.31	1.43	1.85	2.33	3.01	3.35	3.87
	$p = 0.7$	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
Non-I.I.D. No-shows	$\boldsymbol{p1}$	0.93	1.2	1.57	2.12	2.33	2.8	3.2	3.51	3.72
	$\boldsymbol{p2}$	2.32	2.73	3.07	3.27	3.31	3.89	4.15	4.44	4.94

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: $\boldsymbol{p1}$ which alternates between 0.8 and 0.7, and $\boldsymbol{p2}$ alternating between 0.9 and 0.6. From Table 4, 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 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 μ 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 5. 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: The Optimal Constant Policy and the Optimal Schedule

Finally, we compare job allowance between the optimal constant policy and the optimal schedule. Since the optimal schedule has varying job allowance, we compare the average of job allowance durations. We compare them under various values of c_I/c_W and n, and the results are shown in Table 6. We see that both policies have smaller job allowance when c_I/c_W is high. This phenomenon is reasonable because when c_I is relatively small, we should schedule larger job allowances to avoid the building up of patients who are waiting. Also, the job allowance under both policies are larger than the mean of service duration ($\mu = 20$), which is the result of mitigating to the effect of random service durations. When we compare the job allowance between the optimal constant policy and the optimal schedule, the constant policy has a slightly higher job allowance. This may happen

Table 5 **The gaps (in %) for different session length T, ratios of c_I/c_W and c_o; $n = 16$**

c_o	T	c_I/c_W								
		0.2	0.4	0.6	0.8	1	1.5	2	2.5	3
1	$n\mu$	1.83	2.29	2.16	2.17	2.43	3.22	3.33	3.35	3.49
	$1.5n\mu$	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
3	$n\mu$	3	3.27	3.28	3.37	3.49	4.01	3.85	4.33	4.21
	$1.5n\mu$	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
5	$n\mu$	4.17	3.87	3.58	4.35	4.41	4.44	4.76	4.56	4.91
	$1.5n\mu$	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

since the constant policy has less flexibility and needs to allocate extra allowance to mitigate randomness. However, the difference is not substantial indicating that the constant policy is a good heuristic.

Table 6 **Job allowance comparison: Optimal constant policy and the average in the optimal schedule**

n		c_I/c_W								
		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
	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
	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
	Optimal (average)	47.65	40.37	37.1	34.55	33.28	30.95	29.44	28.51	27.71