12-2016

Appointment sequencing: Why the Smallest-Variance-First rule may not be optimal

Qingxia KONG
Adolfo Ibáñez University

Chung-Yee LEE
Hong Kong University of Science & Technology

Chung-Piaw TEO
National University of Singapore

Zhichao ZHENG
Singapore Management University, DANIELZHENG@smu.edu.sg

Appointment Sequencing: Why the Smallest-Variance-First Rule May Not Be Optimal

Qingxia Kong* Chung-Yee Lee† Chung-Piaw Teo‡ Zhichao Zheng§

Abstract

We study the design of a healthcare appointment system with a single physician and a group of patients whose service durations are stochastic. The challenge is to find the optimal arrival sequence for a group of mixed patients such that the expected total cost of patient waiting time and physician overtime is minimized. While numerous simulation studies report that sequencing patients by increasing order of variance of service duration (Smallest-Variance-First or SVF rule) performs extremely well in many environments, analytical results on optimal sequencing are known only for two patients. In this paper, we shed light on why it is so difficult to prove the optimality of the SVF rule in general. We first assume that the appointment intervals are fixed according to a given template and analytically investigate the optimality of the SVF rule. In particular, we show that the optimality of the SVF rule depends on two important factors: the number of patients in the system and the shape of service time distributions. The SVF rule is more likely to be optimal if the service time distributions are more positively skewed, but this advantage gradually disappears as the number of patients increases. These results partly explain why the optimality of the SVF rule can only be proved for a small number of patients, and why in practice, the SVF rule is usually observed to be superior, since most empirical distributions of the service durations are positively skewed, like log-normal distributions. The insights obtained from our analytical model apply to more general settings, including the cases where the service durations follow log-normal distributions and the appointment intervals are optimized.

*School of Business, Universidad Adolfo Ibáñez, Chile. Email: q.kong@uai.cl

†Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Hong Kong. Email: cylee@ust.hk

‡Department of Decision Sciences, NUS Business School, National University of Singapore, Singapore. Email: bizteocp@nus.edu.sg

§Lee Kong Chian School of Business, Singapore Management University, Singapore. Email: danielzheng@smu.edu.sg

1 Introduction

We study the design of a healthcare appointment system with a single physician and a set of patients whose service durations are random. The physician plans to see these patients in an appointment session within a fixed time interval – for example, from 8am to 12pm. Patients arrive punctually at their appointed times. Due to uncertainty in service duration, however, the physician may serve patients later than the appointed time (incurs patient waiting time) and/or the physician may not finish serving all the patients by the end of the appointment session (incurs physician overtime). A typical appointment design problem contains two sets of decisions: order of arrivals and time of arrivals. We refer to the first set as sequencing decisions and the second set as scheduling decisions. The objective is to minimize the expected total cost of patient waiting time and physician overtime.

We focus on the optimal sequencing problem, i.e., we decide the arrival sequence of the patients so that expected total cost to the system is minimized and scheduling decisions – the appointment interval assigned to each patient – will be fixed. We assume that patients are heterogeneous and can be classified according to the mean and variability of their service durations. As reported in the literature (cf. Klassen & Rohleder 1996, Cayirli et al. 2008), it is common policy in many outpatient appointment systems to assign an equal appointment interval to each patient, which is also consistent with our experience in various local clinics. In practice, however, patients fall into different classes, and service duration can vary significantly for each class. Therefore, system performance largely depends on patients' arrival sequence.

In considering the optimal sequencing problem, it has been widely conjectured that the Smallest-Variance-First (SVF) rule is optimal (cf. Weiss 1990, Wang 1999, and Gupta 2007). For small cases with only two patients, the SVF rule has been proved to be optimal under exponential service time distributions (cf. Wang 1999) and convex ordering (cf. Gupta 2007). However, it is only a conjecture that the SVF rule is optimal for more than two patients under general service time distributions. The problem has been investigated over decades and there is still an on-going research that tries to solve the problem. For larger systems, this conjecture is mainly supported by extensive simulation results that compare different sequencing policies (cf. Klassen & Rohleder 1996, Cayirli et al. 2008). The intuition is that a patient with larger variance is more likely to overrun the stipulated session, and by putting him/her last, no later arrivals will be affected by the propagating effect of unpredictable wait times. To

the best of our knowledge, however, no analytical results have been reported on the optimal sequence for more than two patients under stochastic ordering. Recent results reported by Mak et al. (2015) reinforce the superiority of the SVF rule; they demonstrate, using a class of distributionally robust optimization models (in which only the marginal mean and standard deviation of each consultation duration are known), that the SVF rule is in fact optimal in the worst-case distribution under certain technical conditions. However, those conditions may not hold in all settings, and the correlations in the worst-case distribution are quite extreme. Finding the optimal sequence in general settings remains an important open problem.

By analyzing the problem under several assumptions, we shed light on why efforts to prove the optimality of the SVF rule in general settings have been futile. We first introduce a deterministic variant of the appointment sequencing problem (when service durations are known) and prove that the deterministic version of the problem is already $\mathcal{NP}$-hard. We then assume the appointment intervals are fixed according to a given template and the service time distributions are symmetric around the given appointment intervals, and analytically investigate the optimality of the SVF rule. In particular, we identify two important factors that affect the optimality of the SVF rule: the number of patients and the shape of service time distributions. We show that a large number of patients is needed for the SVF rule to be outperformed, which explains why the optimality of the SVF rule can only be proved for a small number of patients. Furthermore, the SVF rule is more likely to be optimal if the service time distributions are positively skewed, which explains why in practice, the SVF rule is usually observed to be superior, since most empirical distributions of the service durations are positively skewed, like log-normal distributions (cf. May et al. 2011). Based on these insights obtained from our analytical study, we construct some counterexamples that the SVF rule is not optimal. Furthermore, our numerical analysis shows that these insights hold even when the service time distributions are not symmetric and the appointment intervals are optimized rather than fixed.

In this paper, we classify patients into distinct groups according to the characteristics of their consultation durations. We focus mainly on the impact of sequencing rules on patient waiting time and physician overtime. Our main contributions in this paper are as follows:

- We study a deterministic variant of the appointment sequencing problem, and show that even if each patient's consultation duration is deterministic (but may not coincide with the allocated appointment interval for that patient), the optimal deterministic sequencing

problem is already $\mathcal{NP}$-hard. This result partly explains why finding an optimal sequence for the stochastic problem is so difficult.

- We use the theory of stochastic ordering to study the appointment sequencing problem. We obtain insights as to why scheduling patients using the SVF rule may not be optimal, and on the other hand, why in practice, the SVF rule is usually observed to be superior.

- We exploit the insights obtained from our analytical model to construct counterexamples showing that the SVF rule is not optimal. We show that these insights hold even when our model assumptions are relaxed, including the case where the appointment intervals are variable and optimized.

- Using likelihood ratio ordering, we show that the SVF rule is optimal for the last two patients. We also obtain several sufficient conditions under which the SVF rule is optimal.

2 Literature Review

In appointment system design, studies tend to examine either scheduling or sequencing rules or the combination of the two. For the sake of brevity, we will not discuss the literature in detail here; interested readers are referred to Cayirli & Veral (2003), Gupta (2007), and Gupta & Denton (2008) for excellent reviews of healthcare appointment scheduling and sequencing problem.

Some studies assume that patients are homogeneous and use the First-Call-First-Appointment (FCFA) rule. Under these assumptions, scheduling rules (i.e., determining appointment intervals) are the main concern. There is extensive research being dedicated to optimal scheduling problem for a given sequence, and we briefly review a few here. Wang (1993) studied static and dynamic scheduling rule using queuing theory. Denton & Gupta (2003) developed a sequential bounding approach to determine the upper bounds of the problem. Begen & Queyranne (2011) showed that the scheduling problem can be solved in polynomial time when the cost function is linear and service durations follow discrete distributions. Ge et al. (2014) extended their work to piecewise linear cost functions that are more practical. Cayirli et al. (2012) introduced a general scheduling rule that incorporates patient no show and walks-in and fits all clinic environments. In our paper, we use the methodology developed in Kong et al. (2013) to solve for the

near-optimal schedules when constructing counterexamples. Detailed discussion the methodology is included in Section 6 when we present our counterexamples. It is worthwhile to note that most methodologies developed for the optimal scheduling problem are capable of handling heterogeneous patients, but limited structural insights on the optimal schedules are available if patients are not homogeneous.

In practice, patients are distinct based on ages, type of procedure, nature of the visit, etc., and we can use variability in service duration to generalize these differences. Higher variance in service duration and/or a larger percentage of new patients will create higher variability in the system, and sequencing patients will be more valuable (cf. Vanden Bosch & Dietz 2000, Robinson & Chen 2003, and Cayirli et al. 2008). Starting with Bailey (1952), this problem has been extensively studied over the past 60 years. In what follows, we will discuss the most relevant research on appointment sequencing problem.

Weiss (1990) was arguably the first to study the optimal sequencing problem analytically. In his 1990 paper, he jointly explored the optimal starting time and sequencing of surgical procedures to best utilize medical resources such as surgeons and operating rooms. Weiss showed that sequencing lower-variance procedures first is optimal in the case of two procedures under exponential or uniform service duration. He also conjectured that the SVF rule is optimal in general. Later on, similar results were reported for location-scale distribution such as normal and uniform distribution (cf. Gupta 2007). Wang (1999) investigated the optimal appointment sequencing of n customers under exponential service distribution and proved the optimality of the SVF rule as Weiss (1990) did for $n = 2$ but used a different method. Wang argued that the result can be generalized to the case of n patients ($n > 2$) using a similar approach without any proof. The intuition is that larger variability will lead to longer waiting time, so sequencing patients with smaller-variance service durations first can reduce waiting time for subsequent patients. Gupta (2007) generalized the two-customer result under convex ordering.

Until now, there have been no analytical results on optimal sequencing for large problems. As reported by Gupta (2007), attempts to establish the optimality of the SVF rule to larger problems "have not been fruitful", leaving this an important open problem in the field of appointment scheduling.

In addition to analytical approaches, another stream of research has used simulation to test the performance of different scheduling and/or sequencing rules. Klassen & Rohleder (1996)

classified patients as "low-" and "high-" variance based on their service time variability, and used simulation to compare alternative ways of sequencing these patients. They found that sequencing low-variance patients at the beginning of the session (the LVBEG rule) performs better than other rules. In a later study, Rohleder & Klassen (2000) considered the possibility that the scheduler might make an error when classifying patients and, furthermore, the scheduler might not be able to sequence patients perfectly if some patients insisted on particular slots. The authors found that the LVBEG rule still performs well under these more realistic assumptions. Vanden Bosch & Dietz (2000) examined scheduling and sequencing policies for a specific primary clinic by classifying patients into three groups based on type of procedure. This was also the first attempt to study the best patient-mix and sequence over several days. Cayirli et al. (2006) classified patients as "new" or "return". In their simulation model, the effects of sequencing rules were investigated with consideration given to patient panel characteristics such as walk-ins, no-shows, and punctuality. They concluded that the "return patients in the beginning" rule performs most efficiently if patient's waiting time cost is large enough. Kolisch & Sickinger (2008) tested sequencing and scheduling rules in a radiology department to dynamically allocate resources to three patient groups: inpatient, outpatient, and emergencies. Cayirli et al. (2008) incorporated patient classification into appointment system design with interval adjustment. They compared the performance of 18 appointment systems, which combined three sequencing rules, three scheduling rules and interval-adjustment condition (with and without), and attempted to identify a robust set of policies specific to the characteristics of a clinical practice. The guidelines that emerged from their simulation study were developed based on the tradeoff between patient waiting time and the physician idle time.

From the literature, sequencing n distinct patients is a difficult problem. Furthermore, very limited complexity results are available and to the best of our knowledge, all $\mathcal{NP}$-hardness results known to date were derived assuming that each patient's costs can be different. For instance, Mancilla & Storer (2012) approached the appointment sequencing problem using a sample average approximation method and proved that the problem is $\mathcal{NP}$-complete with only two scenarios, but with unequal waiting costs for each patient.

Since almost all current literature in the area of appointment scheduling and sequencing focus on the trade-off between waiting time and overtime, the relative cost of waiting time to overtime is thus an important parameter to estimate. In a recent paper, Robinson & Chen

(2011) proposed a queue-based approach to estimate the relative cost of waiting time to overtime by drawing the connection to inventory cost estimation. Furthermore, Turkcan et al. (2011) investigate sequential appointment scheduling with service criteria. They discussed fairness properties of generated schedules and proposed new unfairness measures to capture the inequity among patients assigned to different slots. In this paper, the costs of patient waiting time and physician overtime are set to be equal, because this is the case in which the SVF rule is conjectured to be optimal.

3 Assumptions and Notation

To isolate the impact of sequencing rules on system performance, we make the following assumptions to rule out the presence of other disruptions in our system:

1. The appointment interval for each patient is given.
2. Patients arrive punctually at the scheduled appointment time, and no-shows are not considered.
3. There is a single physician in the system. The physician arrives punctually at the beginning of the session and only serves the scheduled patients during the session. No breaks are taken when the physician is serving a patient.
4. Walk-ins and emergencies are not considered.

Let n be the number of patients and use $i \in \{1, 2, \cdots, n\}$ as the index of all patients. Let u_i be the stochastic service duration of patient i, $i = 1, \ldots, n$. Denote $j \in \{1, 2, \cdots, n\}$ as the index of the slots. Let $\phi(\cdot)$ be a sequence, where $\phi(j)$ denotes the index of the patient who is scheduled to arrive at the beginning the jth slot. For example, $\phi(1) = 4$ denotes that patient 4 is scheduled to arrive first. We assume that service time distributions are independent of the sequence. Let s_i denote the length of the appointment interval allocated to patient i and v_i denote the excess (or redundant) time of patient i, i.e., $v_i = u_i - s_i$, $i = 1, \ldots, n$. Under the schedule $\boldsymbol{s}$, the starting time of the jth slot is $\sum_{k=1}^{j-1} s_{\phi(k)}$. Let $w_{\phi(j)}$ be the waiting time of the jth arrival. It is reasonable to assume that the session starts at time zero, i.e., $w_{\phi(1)} = 0$. Waiting times for subsequent arrivals are given by the following recursion (c.f. Denton & Gupta

2003):

$$\begin{aligned} w_{\phi(j)} &= \max\left\{0,\, w_{\phi(j-1)} + v_{\phi(j-1)}\right\} \\ &= \max\left\{0,\, v_{\phi(j-1)},\, v_{\phi(j-1)} + v_{\phi(j-2)},\, \ldots,\, \sum_{k=1}^{j-1} v_{\phi(k)}\right\},\ j = 2, \ldots, n. \end{aligned} \tag{1}$$

If there is an additional patient $i = n+1$ arriving after the nth patient, then the waiting time for this patient is exactly the physician's overtime, i.e., physician's overtime $= w_{n+1} = \max\{0, w_{\phi(n)} + v_{\phi(n)}\}$. This dummy patient's sequence is fixed as the last, i.e., $\phi(n+1) = n+1$. Let c_i be the unit waiting time cost of patient i. The physician's unit overtime cost is denoted as c_o. All costs are assumed to be nonnegative. The objective of the appointment sequencing problem is to minimize the sum of the expected cost of patient waiting time and physician overtime, i.e.,

$$(S) \qquad \min_{\phi}\ \mathbf{E}\left[\sum_{j=1}^{n} c_{\phi(j)} w_{\phi(j)} + c_o w_{n+1}\right].$$

The objective can easily be extended to include physician idle time if the session length, T, is predetermined. The expected total idle time is given by $T + \mathbf{E}[w_{n+1}] - \mathbf{E}[\sum_{i=1}^{n} u_i]$. Therefore, we only need to adjust the value of the overtime cost c_o to incorporate the cost of physician idle time. In our model analysis, we restrict ourselves to the case that all the appointments have to be scheduled within the session length, i.e.,

$$\sum_{j=1}^{n} s_{\phi(j)} = \sum_{i=1}^{n} s_i = T,\ s_i \geq 0, \forall i = 1, \cdots, n. \tag{2}$$

This constraint is based on the common observation that many service systems, such as a bank, post office, and/or clinic, usually operate within stipulated office hours, and it is natural for these systems to ask customers with appointments to arrive before the end of the office hour, or sometimes even half an hour earlier. The system will continue to serve all the customers waiting in the system even after the end of the office hour, but will not accept any more who arrives later.

In the rest of the paper, we set the costs for both waiting time and overtime equal to one for the following two reasons. First, equal-costs-for-all is the case where the SVF rule is conjectured to be optimal (cf. Gupta 2007). Second, when the costs are different, trade-offs are more obvious and we can construct another counterexample to the conjecture that the SVF

rule is optimal. For the sake of brevity, we move this counterexample to Appendix A.

4 Complexity of the Deterministic Sequencing Problem

In this section, we consider the deterministic problem, in which exact service durations are known before making the sequencing decision. We show that this is already a tough problem to solve. In fact, under this deterministic assumption, Vanden Bosch (1997) has demonstrated that when the objective coefficients c_i's and c_o are allowed to take arbitrary values, determining the optimal sequence is equivalent to solving a nonlinear knapsack problem and is thus $\mathcal{NP}$-hard. Surprisingly, we show next that the problem in fact is strongly $\mathcal{NP}$-hard even when c_i's and c_o are identical. To facilitate better understanding of why this problem is difficult, consider the following example:

Example 1 *Consider an appointment system design problem with 10 patients, with each given a 10-min consultation slot with the physician. The arrival of patients is thus deterministic and 10 min apart. Suppose patients 1 to 5 require service durations of 13 min each, but patients 6 to 10 require service durations of 7 min each. If patients are sequenced to arrive in the order of their indices, then the waiting time of the patients are, respectively, 0, 3, 6, 9, 12, 15, 12, 9, 6, and 3 min, with a total waiting time of 75 min, with no overtime cost for the physician. If the second group of patients arrives before the first group, then the waiting time of the patients will be 0, 0, 0, 0, 0, 0, 3, 6, 9, and 12 min, respectively, with a total waiting time of 30 min, and 15 min of overtime for the physician. A better sequence is to interlace the arrival of the two group of patients, with patient 1 coming before patient 6. In this case, the waiting time will be 0, 3, 0, 3, 0, 3, 0, 3, 0, and 3 min. Total waiting time is 15 min for the patients, and there is no overtime for the physician. Clearly, the third sequence is superior to the first two sequences.*

The above example shows that the optimal sequence requires a careful matching of patients with long and short duration (measured against the slots allocated). This combinatorial problem turns out to be very difficult.

Theorem 1 *The appointment sequencing problem is $\mathcal{NP}$-hard in the strong sense, even if the allocated appointment interval s_j is fixed and $c_o = c_i = 1$, $\forall i = 1, \ldots, n$.*

The proof of the above result is based on a reduction from a well-known $\mathcal{NP}$-complete problem: the numerical 3-dimensional matching problem. We refer readers to Appendix B for details of the proof. This result indicates that finding the optimal sequence when service duration is stochastic is exceedingly challenging.

5 Optimality Conditions of the SVF Rule under Likelihood Ratio Order

In this section, we employ likelihood ratio order to analyze the appointment sequencing problem and gain insights into why the SVF rule is not optimal. In particular, we identify several key influential factors on the optimal appointment sequencing rule. We also provide two sufficient conditions under which the SVF rule is optimal. In what follows, we first describe the model assumptions and then briefly introduce the likelihood ratio order.

Suppose that in a single consultation session, we need to sequence the arrivals of n patients. We assume that the patients' consultation durations are independent of each other and symmetric around their mean. The schedule is set to the mean of the service duration for each patient so that $\mathbf{E}[v_i] = 0$ for all $i = 1, 2, \ldots, n$. We first focus on the case in which all waiting times and overtime costs are the same. Without loss of generality, we assume that $c_o = c_i = 1$ for all $i = 1, 2, \ldots, n$. We relax these assumptions in our numerical analysis in Section 6 and show that the insights obtained from our model analysis still hold and are crucial in constructing counterexamples. We begin with a brief introduction to the likelihood ratio order. To facilitate comparison of different sequences, we need to order the random excess time using the theory of stochastic ordering (cf. Shaked & Shanthikumar 1994).

Definition 1 *Let X and Y be two continuous random variables with density functions f and g, respectively. If $\frac{f(t)}{g(t)}$ decreases over the union of the supports of X and Y, i.e.,*

$$f(t)g(s) \geq g(t)f(s), \forall t \leq s,$$

then X is said to be smaller than Y in the likelihood ratio order, denoted by $X \leq_{lr} Y$.

The likelihood ratio order can be found in many common random variables. For completeness, we present some examples:

Example 2 *Suppose X and Y are normal random variables with means μ_x and μ_y respectively and the same standard deviation. If $\mu_x \leq \mu_y$, then $X \leq_{lr} Y$.*

Example 3 *Suppose X and Y are normal random variables with standard deviations σ_x and σ_y and means μ_x and μ_y, respectively. If $\sigma_x \leq \sigma_y$, then $|X - \mu_x| \leq_{lr} |Y - \mu_y|$.*

Example 4 *Suppose X and Y are exponential random variables with rates λ and μ. If $\lambda \leq \mu$, then $X \geq_{lr} Y$.*

Example 5 *Suppose X and Y are uniform on $[a, b]$ and $[c, d]$, respectively. If $a \leq c$ and $b \leq d$, then $X \leq_{lr} Y$.*

In what follows, we present our analysis based on the assumption of likelihood ratio order on the absolute value of the excess time of the patients, i.e., $|v_i|$. Note that $|v_i| \leq_{lr} |v_j|$ indicates that the absolute value of the excess time of patient i is less than that of patient j in the likelihood ratio sense, which implies that patient j has higher variability. Recall Example 3 and 5 from above.

5.1 Why the SVF Rule May Not Be Optimal?

We first investigate the optimal arrival orders of the first two patients followed by any subsequent sequence and demonstrate that sequencing the patient with smaller variance to arrive first may not be optimal. Note that under our problem settings, if we can show that the optimal order of the first two patients follows the SVF rule, we can use induction to conclude that the optimal sequence is exactly the SVF rule. This is why we begin our investigation with the first two patients. We assume that $|v_1| \geq_{lr} |v_2|$, which implies that patient 1 has higher variability.

Let $TW(v_1, v_2)$ and $TW(v_2, v_1)$ denote the sum of total waiting time and overtime under the two sequences $[1, 2, 3, \ldots, n]$ and $[2, 1, 3, \ldots, n]$, respectively, and $\Delta_{TW} := TW(v_1, v_2) - TW(v_2, v_1)$. Let $w_i(v_1, v_2)$ denote the waiting time of the ith patient under sequence $[1, 2, 3, \ldots, n]$, $i = 1, \ldots, n$, and $w_{n+1}(v_1, v_2)$ denote the physician's overtime under sequence $[1, 2, 3, \ldots, n]$. $w_i(v_2, v_1)$ is defined similarly for sequence $[2, 1, 3, \ldots, n]$ for $i = 1, \ldots, n+1$. Hence, $TW(v_1, v_2) = \sum_{i=1}^{n+1} w_i(v_1, v_2)$, and

$$\mathbf{E}\left[w_1(v_1, v_2)\right] = 0,\ \mathbf{E}\left[w_2(v_1, v_2)\right] = \mathbf{E}\left[\max\{0, v_1\}\right],\ \mathbf{E}\left[w_3(v_1, v_2)\right] = \mathbf{E}\left[\max\{0, v_1, v_1 + v_2\}\right],$$

and for $i = 4, \cdots, n+1$,

$$\begin{aligned}
\mathbf{E}\left[w_i(v_1, v_2)\right] &= \mathbf{E}\left[\max\left\{0, v_1, v_1+v_2, \sum_{j=1}^{3} v_j, \ldots, \sum_{j=1}^{i-1} v_j\right\}\right] \\
&= \mathbf{E}\left[\max\left\{-v_1-v_2, -v_2, 0, v_3, \ldots, \sum_{j=3}^{i-1} v_j\right\}\right] + \mathbf{E}\left[v_1+v_2\right] \\
&= \mathbf{E}\left[\max\left\{\max\left\{0, v_2, v_1+v_2\right\}, \max\left\{0, v_3, \ldots, \sum_{j=3}^{i-1} v_j\right\}\right\}\right].
\end{aligned}$$

Define the partial sum $S_k := \sum_{j=3}^{k+2} v_j$, for $k = 1, \ldots, n-2$, and $S_0 := 0$; we then have a random walk $\{S_k\}$. We can simplify the expression of the expected waiting time for

$$\mathbf{E}\left[w_i(v_1, v_2)\right] = \mathbf{E}\left[\max\left\{\max\left\{0, v_2, v_1+v_2\right\}, \max_{0\le k\le i-3}\left\{S_k\right\}\right\}\right].$$

Similarly, for the second sequence, we have

$$\mathbf{E}\left[w_1(v_2, v_1)\right] = 0, \ \mathbf{E}\left[w_2(v_2, v_1)\right] = \mathbf{E}\left[\max\{0, v_2\}\right], \ \mathbf{E}\left[w_3(v_2, v_1)\right] = \mathbf{E}\left[\max\{0, v_2, v_2+v_1\}\right],$$

and

$$\mathbf{E}\left[w_i(v_2, v_1)\right] = \mathbf{E}\left[\max\left\{\max\left\{0, v_1, v_1+v_2\right\}, \max_{0\le k\le i-3}\left\{S_k\right\}\right\}\right], \ i = 4, \ldots, n+1.$$

We now compute the expected difference between the sum of total waiting time and overtime under two sequences conditional on the realization of v_1 and v_2. Let f and g be the density functions of v_1 and v_2, respectively. Then

$$\begin{aligned}
\mathbf{E}\left[\Delta_{TW}\right] &= \iint_{x<y} \mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y\right] f(x)\, g(y)\, dx\, dy \\
&\quad + \iint_{x>y} \mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y\right] f(x)\, g(y)\, dx\, dy \\
&= \iint_{x>y} \mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y\right] \left\{f(x)\, g(y) - f(y)\, g(x)\right\} dx\, dy.
\end{aligned}$$

The second equality follows from

$$\mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y\right] = -\mathbf{E}\left[\Delta_{TW} \mid |v_1| = y, |v_2| = x\right]$$

and the change in variables.

Since $|v_1| \geq_{lr} |v_2|$, then when $x \geq y$, $f(x)g(y) \leq f(y)g(x)$. We next investigate the sign of $\mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y\right]$ when $x > y$. With some abuse of notation, let v_1 and v_2 also denote the realization of v_1 and v_2, respectively, $|v_1| = x$, and $|v_2| = y$ with $x > y \geq 0$. By symmetry, $v_1 = x$ or $-x$ with equal probability, and $v_2 = y$ or $-y$ with equal probability conditional on $|v_1| = x$ and $|v_2| = y$. We have the following four equally possible cases:

(a) If $v_1 > 0$ and $v_2 > 0$, then $\Delta_{TW} = x - y$.

(b) If $v_1 < 0$ and $v_2 < 0$, then $\Delta_{TW} = 0$.

(c) If $v_1 > 0$ and $v_2 < 0$, then

$$\Delta_{TW} = x + y + \sum_{i=1}^{n-2} \left(\max\left\{ x - y, \max_{0 \leq k \leq i} S_k \right\} - \max\left\{ x, \max_{0 \leq k \leq i} S_k \right\} \right).$$

(d) If $v_1 < 0$ and $v_2 > 0$, then

$$\Delta_{TW} = -2y - \sum_{i=1}^{n-2} \left(\max\left\{ 0, \max_{0 \leq k \leq i} S_k \right\} + \max\left\{ y, \max_{0 \leq k \leq i} S_k \right\} \right).$$

Now we define a function

$$Q\left(t, \{S_k\}\right) := \sum_{i=1}^{n-2} \max\left\{ t, \max_{0 \leq k \leq i} S_k \right\}. \tag{3}$$

The value of the function depends on the difference between t and the random walk $\{S_k\}$. Before the random walk reaches t, the value of each item in the summation is t. After the random walk reaches t for the first time, the value of each item is the maximal value the random walk has ever reached. Therefore, given that the last $(n-2)$ patients have been assigned, and conditional on v_1, v_2, and $\{S_k\}$, the expected difference of the sum of total waiting time and overtime between

the two sequences is

$$\begin{aligned} & \mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y, \{S_k\}\right] \\ = \; & \frac{1}{2}(x-y) + \frac{1}{4}\left[Q\left(y, \{S_k\}\right) + Q\left(x-y, \{S_k\}\right) - Q\left(x, \{S_k\}\right) - Q\left(0, \{S_k\}\right)\right]. \end{aligned} \tag{4}$$

Define

$$\Delta_Q := Q\left(y, \{S_k\}\right) - Q\left(0, \{S_k\}\right) - \left[Q\left(x, \{S_k\}\right) - Q\left(x-y, \{S_k\}\right)\right]. \tag{5}$$

To visualize how Δ_Q can be computed, we give an illustration in Figure 1 under the condition of $x - y > y$, i.e., $x > 2y$. The solid line in Figure 1 depicts a sample path of the random walk $\{S_k\}$. The size of the lower shadow area gives the value of $Q(y, \{S_k\}) - Q(0, \{S_k\})$, and the upper shadow area is $Q(x, \{S_k\}) - Q(x-y, \{S_k\})$. Therefore, Δ_Q is just equal to the lower shadow area minus the upper shadow area.

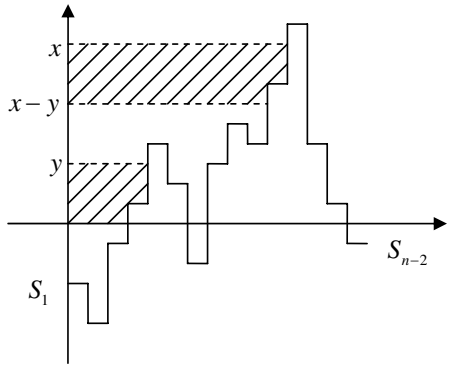


Figure 1: Value of $\triangle_Q$ when $x > 2y$

Given the random walk $\{S_k\}$, $Q(t, \{S_k\})$ is convex in t. Then it is easy to verify that

$$Q\left(y, \{S_k\}\right) + Q\left(x-y, \{S_k\}\right) \leq Q\left(x, \{S_k\}\right) + Q\left(0, \{S_k\}\right),$$

i.e., $\Delta_Q \leq 0$. The zero upper bound of Δ_Q is tight in either the case in which $\{S_k\}$ are all below the horizontal axis or the case in which the first positive rise of $\{S_k\}$ is above x. The lower bound of Δ_Q corresponds to either the case in which S_1 is the highest ladder and equal to $x - y$ when $x \geq 2y$, or the case in which S_1 is the highest ladder and equal to y when $x < 2y$. Note that they are not the unique cases for the lower bound. In general, we know that

$$-(n-2)\min\{y, x-y\} \leq \Delta_Q \leq 0. \tag{6}$$

Equation (4) shows that putting a patient with larger variability in front incurs a positive net waiting time of $(x-y)/2$, which can be offset by a negative term $\Delta_Q/4$. Although $x-y \geq 0$, the value of Equation (4) could be positive, zero, or negative.

From the above analysis, we know that to ensure that sequencing patients in increasing variance is optimal, i.e., to let Δ_Q approach its upper bound, it is required either (1) that $\{S_k\}$ be below the horizontal axile or (2) that the first positive rise of $\{S_k\}$ approach $x > 0$. The first situation will be increasingly unlikely if the number of patients n increases. While for the second case, it is more probably to happen if the steps from the random walk $\{S_k\}$ becomes more positively skewed, which is exactly in the case of log-normal service time distributions. Based on these insights, we construct some counterexamples that the SVF rule is not optimal. We further investigate numerically the situations where our model assumptions are relaxed. We demonstrate that all the insights still hold under more general settings, where the service time distributions are asymmetric and the appointment intervals are optimized rather than fixed. These numerical results are presented in Section 6.

5.2 Optimal Sequence of the Last Two Patients

In the previous section, we investigated the optimality conditions for the first two patients and constructed counterexamples showing that the SVF rule may not always be optimal. Interestingly, we can still show that the SVF rule is optimal for the last two patients, given that the first $(n-2)$ patients have been assigned. Together with results from the previous analysis, we prove that the optimal sequence for three patients follows the SVF rule. We relegate the proofs for results in this subsection and next to Appendix C.

Theorem 2 *For the last two patients in any sequence, if* $|v_{n-1}| \leq_{lr} |v_n|$*, it is optimal to schedule patient* $(n-1)$ *before patient* n *in the sequence.*

From the early discussion on the ordering of the first two patients, when $n \leq 4$, since $-2\min\{y, x-y\} \leq \Delta_Q$, we have

$$\mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y, \{S_k\}\right] \geq 0, \forall x > y.$$

Together with Theorem 2, we obtain the following immediate corollary.

Corollary 1 *When $n = 3$, the optimal sequence of the patients is in increasing likelihood ratio order of $|v_i|$.*

5.3 Sufficient Conditions under Which the SVF Rule Is Optimal

In Section 5.1, based on our assumptions, we conclude that even for the first two patients, given that succeeding patients have been assigned, sequencing patients in increasing variance could impede an appointment system. One of the key assumptions in the previous analysis is that the costs of all the waiting time and overtime are identical. Here, we relax this assumption and provide a sufficient condition under which the SVF rule is optimal. Recall that $c_{\phi(j)}$ is the waiting time cost of the jth patient, for $j = 1, \ldots, n$, and $c_o = c_{\phi(n+1)}$ is the physician's overtime cost.

Proposition 1 *If $|v_{i_1}| \geq_{lr} |v_{i_2}|$ for all $i_1 < i_2$, $i_1, i_2 = 1, \ldots, n$, and the conditions*

$$c_{\phi(j)} \geq \frac{1}{2} \sum_{k=j+2}^{n+1} c_{\phi(k)}, \forall j = 1, \ldots, n-1$$

are satisfied, then the SVF rule for the appointment sequencing problem is optimal.

The intuition of the above sufficient condition is that the cost associated with the earlier time slot is so high that the effect of uncertainty from the random walk $\{S_k\}$ can be eliminated. For the situation in which $c_o = c_i = 1, \forall i = 1, \ldots, n$, we show next that the effect of uncertainty from the random walk can also be eliminated if random service fluctuations are sufficiently far apart by considering a special case of service time distributions.

Proposition 2 *If $v_i = \pm x_i$, $x_i \geq 0$, each with equal probability, for $i = 1, \ldots, n$, and the conditions*

$$\sum_{k=1}^{i-1} x_k \leq x_i, \forall i = 2, \ldots, n$$

are satisfied, then the SVF rule for the appointment sequencing problem is optimal.

6 Counterexamples to the Optimality of the SVF Rule

Based on the insights obtained from our model analysis, we construct several counterexamples to the conjecture about the optimality of the SVF rule in various environments.

6.1 Counterexamples with Fixed Appointment Intervals

In what follows, we first present a counterexample under our model assumptions with symmetric service time distributions and fixed appointment intervals. We assume that the appointment interval assigned to each patient is fixed to the mean of his/her service duration, and focus solely on the sequencing problem to address our question: Is it optimal to sequence patients with smaller variance to arrive earlier in the session? The following example demonstrates that in general, this is not true.

Example 6 *Suppose that there are n patients ($n > 3$) and their service durations $u_i, i = 1, \ldots, n$ are independent and follow uniform distributions. Let u_1 follow uniform distribution on interval $[0,\ 2]$, u_2, u_3 on interval $[0, 4]$ and u_k on interval $[0, 6]$ for all $k \geq 4$. Let $v_i := u_i - \boldsymbol{E}[u_i]$, then v_i follows uniform distribution and is symmetric around 0, for $i = 1, \ldots, n$. Note that v_i is defined as the difference between service duration u_i and appointment interval and we call it "excess time" for ease of exposition throughout the paper. We compare the performance of the sequence $[v_1, v_2, \ldots, v_n]$ with another obtained by switching v_1 and v_2. Patients in the first sequence are put in nondecreasing order of variance and so we call it SVF sequence and the second as Non-SVF sequence. We run a simulation with 2×10^6 sample points and plot the difference in the expected sum of total waiting time and overtime between the two sequences, as a function of n in Figure 2, i.e., the black dotted line. The grey lines above and below dotted line show 95% confidence interval of the mean differences. A positive value shows that the SVF sequence is worse off compared to the sequence obtained from switching v_i and v_2, whereas a negative value shows the contrary.*

In this example, the figure shows that when the number of patients is small (e.g., $n \leq 50$), sequencing patients with smaller variance first is generally better. Surprisingly, however, this behavior changes as n increases, and for a large enough n, putting patient 2 in front of patient 1 actually reduces the sum of total waiting time and overtime. Consequently, sequencing patients by increasing variance is no longer optimal.

The example demonstrates that the optimal sequence is clearly affected by the number of patients in the system, which is exactly the insight we obtained from our analytical study. We made similar observations with other symmetric distributions of the service durations, such as two-point and normal distributions: When the number of patients becomes large enough, the

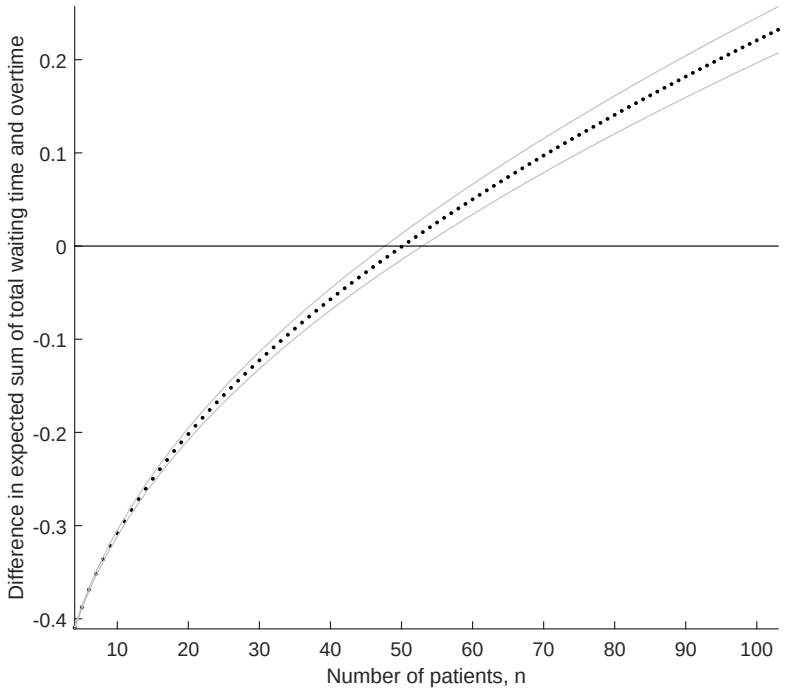


Figure 2: Simulation results of Example 6 with uniform service time distributions

SVF rule performs worse than the Non-SVF rule obtained from simply switching the order of the first two patients. In particular, *the SVF rule fails when the number of patients is greater than 20 for two-point distributions, and around 60 for normal distributions.*

To relax the assumption on service time distributions, we consider an asymmetric distribution: the log-normal distribution. It has been reported frequently in literature that the log-normal distributions tend to have a closer fit to empirical data. For example, Strum et al. (2000) showed that log-normal distribution provides a better fit to historical data on surgical durations than normal distribution. The review of May et al. (2011) concluded that log-normal distributions are the present state of the art. Since the log-normal distribution is positively skewed, from our model analysis, it creates a favorable environment for the SVF rule to be optimal. We confirm this insight with the following example.

Example 7 *Similar to Example 6, we consider n patients ($n > 3$) and their service durations are independent and follow log-normal distributions. All the patients have the same mean service duration of 20, and the variances are the same as in Example 6, where the standard deviation of the service duration of the first patient under the SVF rule is 1, and the second and third*

patients are 2, and the rest is 3. Similarly, the appointment intervals are fixed to the mean service durations. The sample size is 10^7*. The result is plotted in Figure 3.*

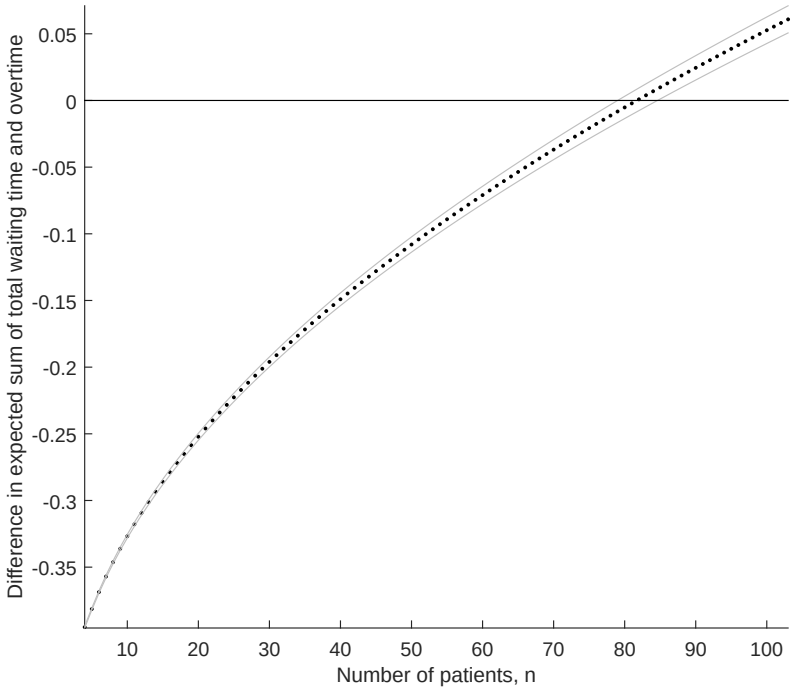


Figure 3: Simulation results of Example 7 with log-normal service time distributions

It is interesting to note that under more realistic distributions, it takes significantly more patients (more than 80) to observe that the SVF rule performs worse than its alternative. This partly explains why it is difficult to detect the suboptimality of the SVF rule in practice.

Next, we consider the case when appointment intervals are determined using information on the second moment, and in particular, "*mean* $+ z \times$ *standard deviation*" schedules, where z is a constant similar to the concept of safety stock in inventory management. We tested Example 6 under these schedules for different values of z under different distributions, including log-normal distributions. Similar results are observed in our numerical analysis for small values of z. When z gets larger, the allocated appointment intervals become larger. As a result, waiting time and overtime decrease, and it becomes much harder to observe the performance difference between different sequencing rules: More patients (i.e., larger n) are required to reach the performance turning point, and more sample points are required in simulation to ensure that the observed differences remain significant as the sum of total waiting time and

overtime gets smaller. Eventually, when z is large enough, sequencing will become unimportant, because there will be sufficient buffer times for all patients and the total waiting time and overtime will approach zero. Nevertheless, we can conclude that the insights obtained from our model and counterexamples still hold if more sophisticated scheduling policies are used, and in particular, "*mean* + $z \times$ *standard deviation*" schedules. Since these results are similar to what we have reported in the fixed-interval case, we omit the detailed numerical results from the paper. Instead, we focus on the situations with optimized appointment intervals in the following subsections.

6.2 Counterexample with Optimal Appointment Intervals under Specific Distributions

In this subsection, we report a counterexample to the SVF rule when the appointment interval are optimized under specific service time distributions. We relax the restriction on arrival times and allow the patients to arrive beyond the session length. We first present the counterexample below followed by the discussion on the logic behind the construction.

Example 8 *Suppose that there are 6 patients to be scheduled for a session of 6 time units, say minutes. Service time durations of the first 5 patients follow binomial distributions* $\mathcal{B}(6, 1/6)$*, and that of the last patient has two possible realizations,* 0 *and* 2*, with equal probability.*[1] *Note that the service time duration of the last patient has the largest variance of 1 as compared to the variance of the first 5 patients, which is* $5/6$*. All the service distributions are independent from each other. The optimal appointment intervals under the SVF sequence are* [1 2 1 2 1]*.*[2] *We simulate the performance of the optimal appointment intervals under the SVF sequence with* 10^7 *sample points, and the 95% confidence interval of the expected total cost is* $[4.6228, 4.6296]$*. We then construct a non-SVF sequence by switching the last patient with the second one in the SVF sequence, and keeping the appointment intervals unchanged.*[3] *The 95% confidence interval of the expected total cost under this new appointment policy is* $[4.2856,\ 4.2926]$*.*

[1] If one requires the service time durations to be strictly positive, we can simply add a positive constant to all the service time realizations and obtain a similar counterexample.

[2] We used the sample average approximation method to solve for the optimal appointment intervals with 10^7 sample points. Since the service time distributions are discrete, it is reasonable to believe that the solutions found are exact optimal.

[3] Note that to construct a counterexample, we only need to show that a non-SVF sequence together with feasible appointment intervals performs better than the SVF sequence under the optimal appointment intervals. In fact, the optimal appointment intervals for the constructed non-SVF sequence turn out to be the same: [1 2 1 2 1].

When constructing the counterexample, we start with independent and identically distributed service time distributions for all patients, and then replace the service distribution of the last patient with one that has larger variance but smaller maximum service time while keeping the mean unchanged. We obtain the optimal appointment intervals and look for an interval that is equal to or larger than the maximum service time of the last patient. If such an interval exists (the second interval in the above counterexample), we put the last patient in the position where the interval is. By doing this, we make sure that the appointment interval is long enough to cover the service duration of the last patient in the SVF sequence with probability one, thus avoiding the waiting time propagation in the new sequence to some extend. Consequently, this new sequence might perform better than the SVF sequence. This is how we construct the above counterexample.

6.3 Counterexamples with Distributionally Robust Appointment Intervals

Mak et al. (2015) showed that the SVF rule is optimal under a class of distributionally robust appointment scheduling problem. They assume that the service durations have known means and variances, and proved that the worst-case performance is minimized when the sequence follows the SVF rule. However, the worst-case distribution for each sequence of patient arrivals has extremal correlations (usually taking values of 1 and -1), and significantly deviates from the independence condition under which the optimality of the SVF rule is often associated with. It is still unknown if the SVF rule will remain optimal for more sophisticated classes of distributionally robust models.

To this end, we use the conic programming approach developed by Kong et al. (2013) to study this problem. They assume that the service durations are positively supported with known means, variances, and covariances, which resembles the practical service time distributions to a larger extend. Kong et al. (2013) showed that the appointment intervals obtained from their method perform on a par with the stochastic programming method under various service time distributions, which means that they also obtained near-optimal schedules.

With these concerns, even only optimizing appointment intervals under a given sequence is a challenging task from the numerical perspective, and we cannot guarantee that the appointment intervals we found using either of the above methods are exactly optimal. Nevertheless, we can still analyze the performance of different sequencing rules under near-optimal appointment

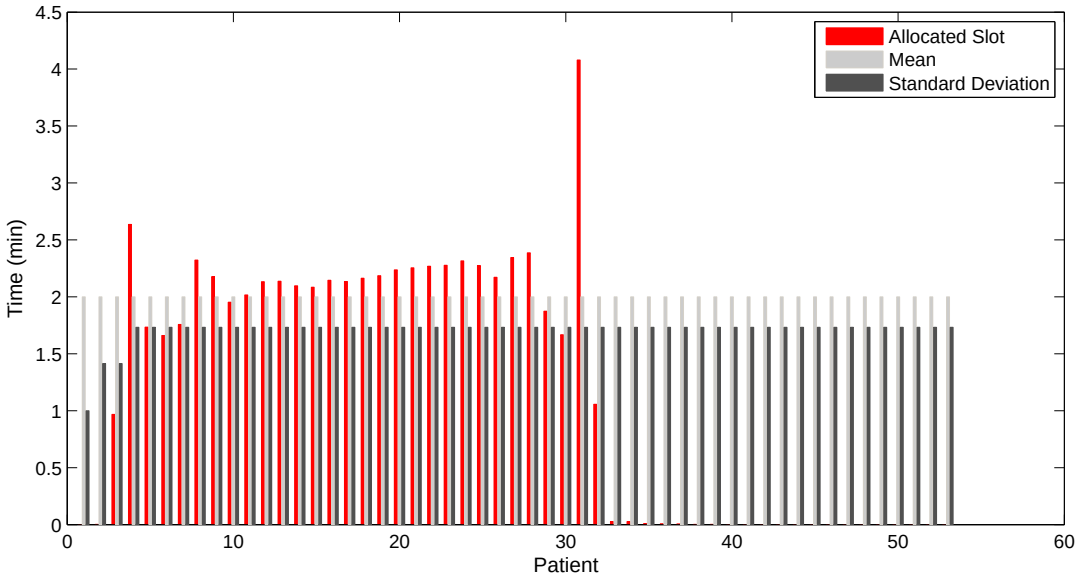
intervals. We manage to find such a counterexample that the SVF rule is not optimal when the appointment intervals are near optimal and all the appointments have to arrive before the session ends, which we present below.

Example 9 *We consider the scheduling problem of 53 patients in a consultation session that lasts for 63.6 min. Suppose that all the appointments have to be scheduled within the session length. The patients' service durations are independent with the same mean of 2 min, and the variances are 1, 2, 2 for the first 3 patients and 3 for the rest of the 50 patients. The structure of this example mimics previous counterexamples and the session time is chosen as 60% of the sum of the mean service durations, which represents the situation in a congested system. We use the methodology developed in Kong et al. (2013) to solve for near-optimal appointment intervals under the SVF and Non-SVF rules as described in Example 6. Figure 4 depicts the optimal schedules under the two sequencing rules. We use the YALMIP interface in MATLAB with MOSEK solver (cf. Löfberg 2004) to solve the conic programming model by Kong et al. (2013), and the relative duality gap of the solution is less than* 10^{-6} *(the absolute duality gap is less than* 10^{-5}*).*

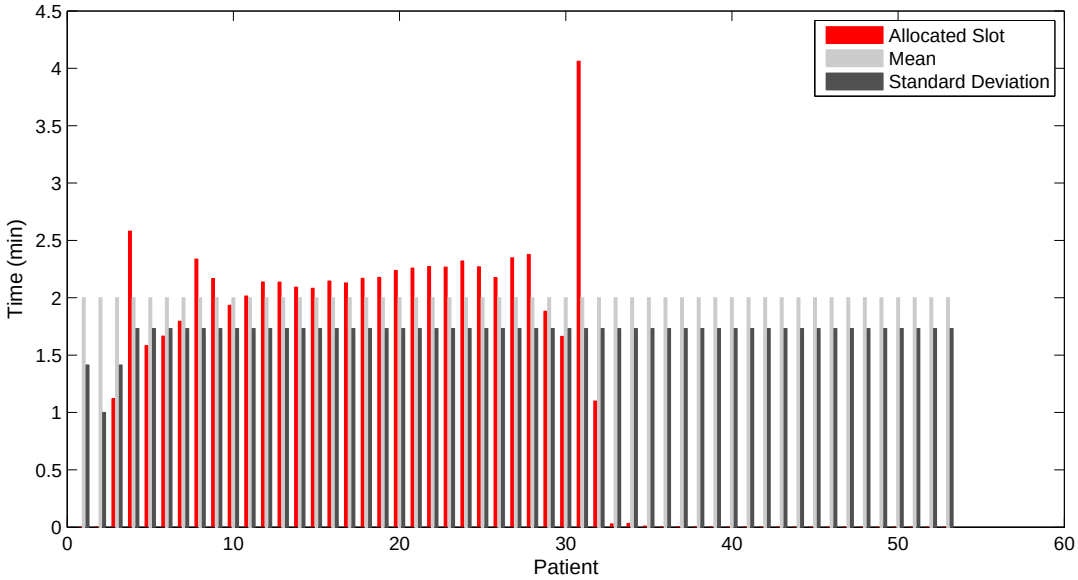
In such a congested system, the optimal schedules assign very tiny slot to the first two patients, which is similar to the Bailey's Rule recommended in Bailey (1952). We run simulation of 10^7 *sample points under log-normal service distribution, and compare the sum of total waiting time and overtime of the near-optimal schedules under these two sequences. The expected sum of total waiting time and overtime and its 95% confidence interval under the SVF rule is 730.2473 min and (730.4105 min, 730.5736 min), respectively. For the Non-SVF rule, they are 729.8339 min and (729.9973 min, 730.1606 min), respectively. The SVF rule performs significantly worse than the Non-SVF rule, though the gap is small.*

7 Conclusion

In this paper, we study a stochastic appointment sequencing problem to identify the optimal arrival order for patients with distinct characteristics. We use the theory of stochastic ordering to study optimal sequencing rules, and in particular, why the SVF rule may not be optimal. By connecting our problem to the random walk, we show that sometimes scheduling patients with larger variance before those with smaller variance outperforms the SVF rule. With the insights



(a) SVF



Non-SVF

Figure 4: Optimal appointment intervals under two sequencing rules in Example 9

obtained from the analytical model, we find several counterexamples under the likelihood ratio order that demonstrate that the optimal sequence depends on patient mix. We then explore the sufficient conditions under which the SVF rule is optimal. In addition, we provide a formal proof that the deterministic variant of the appointment sequencing problem is already strongly $\mathcal{NP}$-hard when both patient waiting time cost and physician overtime cost are one. This result partly explains why finding an optimal sequence for the stochastic problem is difficult.

While the counterexamples demonstrate that the SVF rule is not optimal under certain conditions, it is worth mentioning that such examples are not easy to find and that results are very sensitive to variable parameters. Additionally, even though the SVF rule performs worse than a certain sequence (with a large-variance patient first), the difference in performance is relatively small. In the simulation, huge sample sizes are necessary to significantly differentiate the performance of different sequencing policies. In all of our analyses, we use at least two million sample points. This implies that it is particularly difficult to identify the true optimal sequence through numerical methods, e.g., the sample average approximation approach. These observations illustrate why studies using simulation typically suggest that the SVF rule performs quite well in various settings, although there is a gap in theoretical results for more than two patients. Furthermore, we show that under more realistic service time distributions (like log-normal distributions), the SVF rule is more likely to be optimal and the performance gap between different sequencing rules can be much smaller, which implies that a much larger sample size is necessary to narrow down the confidence interval for the performance gap. On the other hand, all these could be good news for practitioners, as the SVF rule, despite not being strictly optimal might still be close to optimal in more realistic environment, where there are usually less than fifty patients for a doctor within a clinical session (for most clinical data reported in literature).

Our analytical results are based on two assumptions: symmetric service time distributions and fixed appointment intervals. These assumptions are imposed for analytical tractability. In numerical studies, however, we relax both assumptions and the insights obtained from our theoretical analysis can still apply. In particular, we provide counterexamples that the SVF rule is not optimal under asymmetric distributions such as log-normal distribution, and we constructed a counterexample where the appointment intervals are optimized rather than fixed.

Another interesting observation is that the conjecture "SVF rule is optimal" is always made

under independent distributions, and that is why we focus our study on independent service time distributions. We believe that the SVF rule will fail to be optimal if arbitrary correlations on service durations are imposed. For example, if two large-variance patients are strongly negatively correlated, then by sequencing these two patients together, we may get a new "patient" with very low variance in total service time. Consequently, they together can be treated as a single low-variance patient and scheduled upfront, which violates the SVF rule from the perspective of individual service time variances.

In future research, more analytical studies can be pursued by relax the assumptions on symmetric distributions and constant appointment intervals. Furthermore, improving the sufficient conditions for the optimality of the SVF rule or obtaining the optimality gap will be interesting as it provides stronger guarantees when the SVF rule is implemented in practice.

Acknowledgement

We thank the Editor and four referees, for their insightful comments on an earlier version of the paper. This work was partly funded by FUNDECYT grant No.11121634, Chile.

References

Anderson, F. J., P. Nash (1987) *Linear Programming in Infinite-Dimensional Spaces: Theory and Applications*, John Wiley & Sons Ltd.

Bailey, N. T. J. (1952) *A Study of Queues and Appointment Systems in Hospital Outpatient Departments with Special Reference to Waiting Times*, Journal of the Royal Statistical Society, **14**, pp. 185–199.

Berman, A., Shaked-Monderer, N. (2003) *Completely Positive Matrices*, World Scientific.

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

Bomze, I. M., Dür, M., Klerk, E. D., Roos, C., Quist, A. J., Terlaky, T. (2000) *On copositive Programming and Standard Quadratic Optimization Problems*, Journal of Global Optimization, **18**, pp. 301-320.

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

Cayirli, T., E. Veral, H. Rosen (2006) *Designing Appointment Scheduling Systems for Ambulatory Care Services*, Healthe Care Management Science, **9**, pp. 338–353.

Cayirli, T., E. Veral, H. Rosen (2008) *Assessment of Patient Classification in Appointment System Design*, Production and Operations Management, **17**, pp. 47–58.

Cayirli, T., Yang, K. K. and Quek, S. A. (2012) *A Universal Appointment Rule in the Presence of No-Shows and Walk-ins*, Production and Operations Management, **21**, pp. 682–697.

Chen, R. R. and Robinson, L. W. (2014) *Sequencing and Scheduling Appointments with Potential Call-In Patients*, Production and Operations Management, **23**, pp. 1522C-1538.

Conforti, D., F. Guerriero, R. Guido (2008) *Optimization Models for Radiotherapy Patient Scheduling*, 4OR: A Quartly Journal of Operations Research, **6**, pp. 263–278.

Denton, B., D. Gupta (2003) *A Sequential Approach for Appointment Appointment Scheduling*, IIE Transactions, **35**, pp. 1003–1016.

Dür, M. (2009) *Copositive Programming: A Survey*, Available online at http://www.optimization-online.org/DB_HTML/2009/11/2464.html.

Gary, M. R., D. S. Johnson (1979) *Computers and Intractability: A Guide to the Theory of $\mathcal{NP}$-Completeness*, W.H. Freeman and Company, New York.

Grant, M., S. Boyd (2011) *CVX: Matlab Software for Disciplined Convex Programming, version 1.21*, Production and Operations Management, **16**, pp. 689–700.

Gupta, D. (2007) *Surgical Suites' Operations Research*, Production and Operations Management, **16**, pp. 689–700.

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

Gupta, D., B. Denton (2008) *Appointment Scheduling in Health Care: Challenges and Opportunities*, IIE Transactions, **40**, pp. 800–819.

Kemper, B., C. A. J. Klaassen, M. Mandjes (2014) *Optimized Appointment Scheduling*, European Journal of Operational Research, **239(1)**, pp. 243–255.

Harper, P. R., H. M. Gamlim (2003) *Reduced Outpatient Waiting Times with Improved Appiontment Scheduling: A Simulation Modelling Approach*, OR Spectrum, **25**, pp. 207–222.

Klassen, K. J., T. R. Rohleder (1996) *Scheduling Outpatient Appointments in a Dynamic Environment*, Journal of Operations Management, **14**, pp. 83–101.

Klerk, E. de, Pasechnik, D. V. (2002) *Approximation of the Stability Number of a Graph via Copositive Programming*, SIAM Journal on Optimization, **12**, pp. 875–892.

Kolisch, R., S. Sickinger (2008) *Providing Radiology Health Care Services to Stochastic Demand of Different Customer Classes*, OR spectrum, **30**, pp. 375–395.

Kong, Q., C. Y. Lee, C. P. Teo, Z. Zheng (2013) *Scheduling Arrivals to a Stochastic Service Delivery System Using Copositive Cones*, Operations Research, **61(3)**, pp. 711–726.

Lehancy, B., S. A. Clarke, R. L. Paul (1999) *A Case of Intervention in an Outpatients Department*, Journal of the Operational Research Society, **50**, pp. 877–891.

Liang, J. J. (2006) *Intelligent Appointment Scheduling to Reduce Tturnaround Time*, Master Thesis, National University of Singapore.

Löfberg, J. (2004) *YALMIP: A toolbox for modeling and optimization in MATLAB*, Proceedings of the CACSD Conference, Taipei, Taiwan, 2004.

López, M., G. Still (2007) *Semi-infinite Programming*, European Journal of Operational Research, **180**, pp. 491–518.

Mak, H. Y., Y. Rong, J. Zhang (2014) *Sequencing Appointments for Service Systems Using Inventory Approximations*, Manufacturing & Service Operations Management, **16(2)**, pp. 251–262.

Mak, H. Y., Y. Rong, J. Zhang (2015) *Appointment Scheduling with Limited Distributional Information*, Management Science, **61(2)**, pp. 316–334.

Mancilla, C., R. Storer (2012) *A Sample Average Approximation Approach to Stochastic Appointment Sequencing and Scheduling*, IIE Transactions, **44**, pp.655–670.

May, J., W. Spangle, D. Strum, L. Vargas (2011) *The Surgical Scheduling Problem: Current Research and Future Opportunities*, Production and Operations Management, **20(3)**, pp.392–405.

Mosheiov, Gur (1991) *V-Shaped Policies for Scheduling Deteriorating Jobs*, Operations Research, **39(6)**, pp. 979–991.

Murty, K. G., Kabadi, S. N. (1987) *Some $\mathcal{NP}$-Complete Problems in Quadratic and Nonlinear Programming*, Mathematical Programming, **39**, pp. 117–129.

Natarajan, K., M. Sim, J. Uichanco (2010) *Tractable Robust Expected Utility and Risk Models for Portfolio Optimization*, Mathematical Finance, **20**, pp. 695–731.

Natarajan, K., Teo, C. P., Zheng, Z. (2010) *Mixed Zero-One Linear Programs under Objective Uncertainty: A Completely Positive Representation*, Operations Resaerch, **59**, pp. 713–728.

Parrilo, P. A. (2000) *Structured Semidefinite Programs and Semi-algebraic Geometry Methods in Robustness and Optimization*, Ph.D. thesis, California Institute of Technology, Available online at: http://www.cds.caltech.edu/˜pablo/.

Perakis, G., G. Roels (2008) *Regret in the Newsvendor Model with Partial Information*, Operations Research, **56**, pp. 188–203.

Pindeo, M. (1982) *Minimizing the Expected Makespan in Stochastic Flow Shops*, Operations Research, **30**, pp. 148–162.

Robinson, L., Chen, R. (2003) *Scheduling Doctors' Appointments: Optimal and Empirically-Based Heuristic Policies*, IIE Transactions, **35(3)**, pp. 295–307.

Robinson, L., Chen, R. (2011) *Estimating the Implied Value of the Customer's Waiting Time*, Manufacturing & Service Operations Management, **13(1)**, pp. 53–57.

Rohleder, T. R., K. J. Klassen (1996) *Using Client-Variance Informaion to Improve Dynamic Aappointment Scheduling Performance*, Omega, **28**, pp. 293–302.

Ross, S. M. (1996) *Stochastic Processes*, John Wiley & Sons, Inc.

Shaked, M., J. G. Shanthikumar (1994) *Stochastic Orders and Their Applications*, Academic Press, Inc.

Strum, D., J. May, L. Vargas (2000) *Modeling the Uncertainty of Surgical Procedure Times: Comparison of the Log-Normal and Normal Models*, Anesthesiology **92(4)**, pp. 1160–1167.

Suresh, S., Foley, R. D., and Dickey, S. E. (1985). *On Pinedo's Conjecture for Scheduling in a Stochastic Flow Shop*, Operations Research **33**, pp. 1146–1153.

Turkcan, A., Zeng, B., Muthuraman, K., and Lawley, M.A. (2011). *Sequential Clinical Scheduling with Service Criteria*, European Journal of Operational Research **214(3)**, pp. 780–795.

Toh, K. C., M. J. Todd, and R. H. Tutuncu (1999) *SDPT3 — a Matlab software package for semidefinite programming*, Optimization Methods and Software, **11**, pp. 545–581.

Vanden Bosch, P. M. (1997) *Scheduling and sequencing arrivals to a stochastic service system*, PhD Dissertation, Air Force Institute of Technology, Wright Press.

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

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

Wang P. P. (1999) *Sequencing and scheduling N customers for a stochastic server*, European Journal of Operations Research, **119**, pp. 729–738.

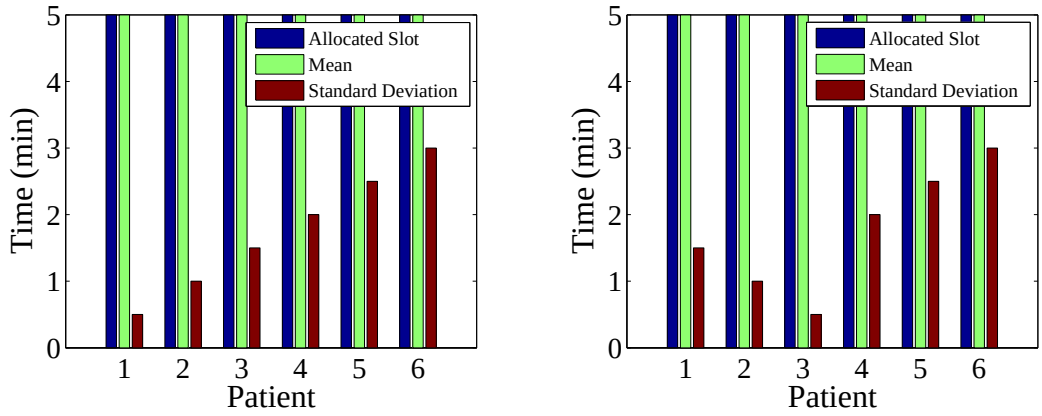
Weiss N. E. (1990) *Models for determining estimated start times and case orderings in hospital operational rooms*, IIE Transactions, **22**, pp. 143–150.

Welch, J. D., N. T. J. Bailey, M. A. Camb (1952) *Appointment systems in hospital outpatient departments*, The Lancet, **259**, pp. 1105–1108.

Appendix A. A Counterexample under Different Cost Structures

Suppose that there are 6 patients to be sequenced. Their mean service durations are all 5 minutes, and the standard deviations range from 0.5 min to 3 min with a step size of 0.5 min.

Assume that their service durations are uncorrelated to each other. The appointment intervals allocated to each patient are equal to their mean service durations. We further assume that the waiting time cost are the same for all patients, i.e., $c_{\phi(j)} = c_w$, $\forall j = 1, 2, \ldots, n$, which could be different from the overtime cost, c_o. We consider two sets of cost structures: (1) $c_w = c_o = 1$; and (2) $c_w = 1$, $c_o = 100$. Two sequencing rules are evaluated: (a) the SVF rule, as depicted in Figure 5(a); (b) the V-Shape rule, where patients with higher variabilities in service durations are scheduled at the beginning and the end of the session but patients with low variabilities in service durations are scheduled in the middle of the session, as depicted in Figure 5(b). In Figure 5, we read the sequences of patients from left to right, with the first patient in the sequence being the leftmost one. The V-Shape rule is named from the shape of standard deviations by visualizing the sequence in such a way.



(a) SVF (b) V-Shape

Figure 5: Optimal sequence under different cost structures for six heterogeneous patients

We simulate the performance of both sequencing rules under the two cost structures discussed above. We consider four possible distributions with the same mean and variance given above: two-point, uniform, normal, and log-normal distributions. The results are summarized in Table 1, and the sample sizes are 10^7 for all simulations. When both waiting time cost and overtime cost are equal to one, the SVF rule performs better under all service time distributions. However, the situations are reversed when the overtime cost increases. When $c_w : c_o = 1 : 100$, the V-Shape rule dominates the SVF rule under all distributions considered. In fact, this V-Shape rule is obtained using the conic programming model presented in Kong et al. (2013).

Cost Structure	Sequences	Two-Point	Uniform	Normal	Log-Normal
$c_w : c_o = 1 : 1$	SVF	8.7893	8.2770	7.8840	7.4931
	V-Shape	9.5632	9.0098	8.5714	8.2253
$c_w : c_o = 1 : 100$	SVF	301.9397	287.7873	276.2796	258.0095
	V-Shape	300.4005	286.8380	275.5392	257.9627

Table 1: Expected total costs under different sequences for six heterogeneous patients with different cost structures

Appendix B. Proof of Theorem 1

To distinguish from the random service durations u_i, we use $\bar{u}_i$ to denote the deterministic service duration of patient i. We begin by defining the feasibility version of the deterministic appointment sequencing problem as follows:

> ***Deterministic Appointment Sequencing (DAS)***
>
> *Given n patients indexed by i with fixed appointment interval s_i and deterministic service duration $\bar{u}_i$, and a budget C, does there exist a sequence for these patients such that the total cost does not exceed C when the unit waiting time costs for all the patients are the same?*

Interestingly, the DAS problem is related to the following well-known strongly $\mathcal{NP}$-complete problem (cf. Gary & Johnson 1979):

> ***Numerical 3-Dimensional Matching (N3DM)***
>
> *Given three disjoint sets X, Y, Z, each containing m elements with size $S(a) \in \mathbb{Z}^+$ for each element $a \in X \cup Y \cup Z$, and a bound $B \in \mathbb{Z}^+$, does there exist a partition of $X \cup Y \cup Z$ into m disjoint sets $A_1, A_2, \ldots, A_m$ such that each A_i contains exactly one element from each of X, Y and Z and such that, for $i = 1, \ldots, m$, $\sum_{a \in A_i} S(a) = B$?*

Step 1. Construction

First, we show that any instance of the N3DM problem can be transformed to an instance of the DAS problem in polynomial time. Given an instance of N3DM, we construct $3m + 1$ patients with $s_i = 5M$ for $i = 1, \ldots, 3m$, $s_{3m+1} = 0$ and

$$\begin{cases} \bar{u}_i = 6M + S(a^X), & \text{if } a^X \text{ is the } i\text{th element in } X, \forall i = 1, \ldots, m, \\ \bar{u}_{m+i} = 4M + S(a^Y), & \text{if } a^Y \text{ is the } i\text{th element in } Y, \forall i = 1, \ldots, m, \\ \bar{u}_{2m+i} = 5M - B + S(a^Z), & \text{if } a^Z \text{ is the } i\text{th element in } Z, \forall i = 1, \ldots, m, \\ \bar{u}_{3m+1} = (m+1)M, & \end{cases}$$

where M is a big number such that $M > (3m+1)B$. Let $c^l = 0$ and $c_i^w = 1$, $\forall i = 1, \ldots, 3m+1$. Denote $S(X) = \sum_{a^X \in X} S(a^X)$, $S(Y) = \sum_{a^Y \in Y} S(a^Y)$ and $S(Z) = \sum_{a^Z \in Z} S(a^Z)$. Define the budget $C = mM + 2S(X) + S(Y)$. For notational convenience, we refer to patients corresponding to set X, Y and Z as X-type, Y-type and Z-type patients, respectively. Without loss of generality, we assume $B = (S(X) + S(Y) + S(Z))/m$.

Step 2. Feasible N3DM $\Longrightarrow$ Feasible DAS

To prove the equivalence of above instances of the two problems, we start by showing that if the N3DM problem has a feasible solution, then it leads to a feasible solution to the constructed DAS problem that meets the budget C. We construct a sequence of $m+1$ consecutive blocks for $3m+1$ patients, where the ith block corresponds to A_i, $i = 1, \ldots, m$ and the $(m+1)$th block contains only patient $3m+1$. Patients within each of first m blocks are sequenced in X, Y, Z order. Hence, we have a complete sequence of all the $3m+1$ patients. Under this sequence, there is no idle time for the physician during actual service. In each of the first m blocks, the waiting time is zero for the X-type patient, $M + S(a^X)$ for Y-type patient and $S(a^X) + S(a^Y)$ for Z-type patient, where $S(a^X)$ corresponds to the X-type patient served in the first position and $S(a^Y)$ corresponds to the Y-type patient served in the second position. The waiting time is zero for patient $3m+1$. Hence, the total waiting time is $mM + 2S(X) + S(Y)$, which equals to C by our construction and we have found a feasible sequence to the constructed DAS problem.

Step 3. Feasible DAS $\Longrightarrow$ Feasible N3DM

Next, we show that any feasible solution to the constructed DAS problem must lead to a feasible solution to the N3DM problem. Suppose that there exists a sequence ϕ to the constructed DAS problem such that $\sum_{i=1}^{3m+1} w_{\phi(i)} \leq C$.

Step 3(a). We first claim that in such a sequence, patient $3m+1$ must be processed in the last position. Otherwise, after patient $3m+1$ there are still other patients left and the waiting time is at least $(m+1)M$, which is larger than C by the definition of M. This is a contradiction to our assumption.

Step 3(b). Then we show that any Y-type patient must be scheduled immediately after a X-type patient. Note that the waiting time of any patient scheduled immediately after a X-type patient is at least $M + S(a^X)$ for some $a^X \in X$. Combining with the result from Step 3(a), we know that the total waiting time for these patient is at least $mM + S(X)$.

Suppose there exists one X-type patient who is not followed by an Y-type patient under the sequence ϕ. Then this X-type patient must be followed by either a X-type or a Z-type patient. Firstly, if it is a X-type patient who is scheduled immediately after, then the total waiting time would be at least $(m+1)M + S(X)$, which is bigger than C. Otherwise, if it is a Z-type patient, then the total waiting time would be at least $(m+1)M - B$, which is also bigger than C. Therefore, we reach a contradiction and conclude that the patient scheduled immediately after a X-type patient must be an Y-type patient. Consequently, the total waiting time is at least $mM + 2S(X) + S(Y)$.

Step 3(c). Similarly, we can show that any Y-type patient must be followed by a Z-type patient in the sequence ϕ. Otherwise, from the previous result, there must exist a Y-type patient who is followed by either a X-type patient or the last patient. In either case, the total waiting time would exceed the budget C.

Step 3(d). From Step 3(a) to 3(c), we show that any feasible sequence ϕ must partition the patients into blocks of three patients with an ordering of X-type Y-type Z-type patients in each block except the last patient, i.e., patient $3m+1$. Such sequence has resulted in a total waiting time of at least $mM+2S(X)+S(Y) = C$. Therefore, we must have $S(a^X)+S(a^Y)+S(a^Z) = B$ for every tuple a^X, a^Y, a^Z that correspond to X-type, Y-type and Z-type patients in every block of the sequence. Thus, we obtain a feasible solution to the N3DM problem and complete the proof.

Appendix C. Proofs of Technical Results in Section 5.2 and 5.3

Proof of Theorem 2

Let t ($t \geq 0$) denote the waiting time of the patient before the last patient. t does not depend on the ordering of the last two patients. Let $W(v_{n-1}, v_n)$ and $W(v_n, v_{n-1})$ denote the total waiting time under the two sequences $[1, 2, \ldots, n-2, n-1, n]$ and $[1, 2, \ldots, n-2, n, n-1]$, respectively. Define

$$\Delta_W := W(v_{n-1}, v_n) - W(v_n, v_{n-1}) = \max\{0, v_{n-1} + t\} - \max\{0, v_n + t\}. \tag{7}$$

Let f and g be the density functions of v_{n-1} and v_n, respectively. Conditional on the service

time realization of the first $n-2$ patients, we have

$$\begin{aligned}
& \mathbf{E}\left[W\left(v_{n-1}, v_n\right)\right] \\
= & \iint\limits_{x<y} \mathbf{E}\left[W\left(v_{n-1}, v_n\right) \mid\left|v_{n-1}\right|=x,\left|v_n\right|=y\right] f(x) g(y)\, dx\, dy \\
& + \iint\limits_{x<y} \mathbf{E}\left[W\left(v_{n-1}, v_n\right) \mid\left|v_{n-1}\right|=y,\left|v_n\right|=x\right] f(y) g(x)\, dx\, dy \\
= & \iint\limits_{x<y} \mathbf{E}\left[W\left(v_{n-1}, v_n\right) \mid\left|v_{n-1}\right|=x,\left|v_n\right|=y\right] f(x) g(y)\, dx\, dy \\
& + \iint\limits_{x<y} \mathbf{E}\left[W\left(v_n, v_{n-1}\right) \mid\left|v_{n-1}\right|=x,\left|v_n\right|=y\right] f(y) g(x)\, dx\, dy.
\end{aligned}$$

Similarly,

$$\begin{aligned}
& \mathbf{E}\left[W\left(v_n, v_{n-1}\right)\right] \\
= & \iint\limits_{x<y} \mathbf{E}\left[W\left(v_n, v_{n-1}\right) \mid\left|v_{n-1}\right|=x,\left|v_n\right|=y\right] f(x) g(y)\, dx\, dy \\
& + \iint\limits_{x<y} \mathbf{E}\left[W\left(v_n, v_{n-1}\right) \mid\left|v_{n-1}\right|=y,\left|v_n\right|=x\right] f(y) g(x)\, dx\, dy \\
= & \iint\limits_{x<y} \mathbf{E}\left[W\left(v_n, v_{n-1}\right) \mid\left|v_{n-1}\right|=x,\left|v_n\right|=y\right] f(x) g(y)\, dx\, dy \\
& + \iint\limits_{x<y} \mathbf{E}\left[W\left(v_{n-1}, v_n\right) \mid\left|v_{n-1}\right|=x,\left|v_n\right|=y\right] f(y) g(x)\, dx\, dy.
\end{aligned}$$

If we can show that

$$\mathbf{E}\left[\Delta_W \mid\left|v_{n-1}\right|,\left|v_n\right|\right] \leq 0, \text{ whenever } \left|v_{n-1}\right| \leq\left|v_n\right|, \tag{8}$$

then by the definition of likelihood ratio order,

$$\begin{aligned}
& \iint\limits_{x<y} \mathbf{E}\left[\Delta_W \mid\left|v_{n-1}\right|=x,\left|v_n\right|=y\right] f(x) g(y)\, dx\, dy \\
\leq & \iint\limits_{x<y} \mathbf{E}\left[\Delta_W \mid\left|v_{n-1}\right|=x,\left|v_n\right|=y\right] f(y) g(x)\, dx\, dy.
\end{aligned}$$

By Equation (7) and rearranging the terms, we have

$$\mathbf{E}\left[W\left(v_{n-1}, v_n\right)\right] \leq \mathbf{E}\left[W\left(v_n, v_{n-1}\right)\right].$$

We verify Equation (8) in the rest of the proof. With some abuse of notation, let v_{n-1} and v_n also denote the realization of v_{n-1} and v_n, respectively, and $|v_{n-1}| = x$, and $|v_n| = y$ with $0 \leq x \leq y$. By symmetry, $v_{n-1} = x$ or $-x$ with equal probability conditional on $|v_{n-1}| = x$. Similarly, $v_n = y$ or $-y$ with equal probability conditional on $|v_n| = y$. We have four possible cases:

(a) If $v_{n-1} \geq 0$ and $v_n \geq 0$, then $\Delta_W = v_{n-1} - v_n = |v_{n-1}| - |v_n| = x - y$.

(b) If $v_{n-1} \leq 0$ and $v_n \leq 0$, then $\Delta_W = 0$ when $t < x$ and $\Delta_W = t - x$ when $x \leq t \leq y$, whereas $\Delta_W = y - x$ when $t > y$.

(c) If $v_{n-1} \leq 0$ and $v_n \geq 0$, then $\Delta_W = -y - t$ when $t < x$ and $\Delta_W = -x - y$ when $t \geq x$.

(d) If $v_{n-1} \leq 0$ and $v_n \geq 0$, then $\Delta_W = x + t$ when $t \leq y$ and $\Delta_W = x + y$ when $t > y$.

To summarize, when $t < x$,

$$\begin{aligned}
\mathbf{E}\left[\Delta_W \mid |v_{n-1}| = x, |v_n| = y\right] &= \frac{1}{4}(x - y) + \frac{1}{4}(-y - t) + \frac{1}{4}(x + t) \\
&= \frac{1}{2}(x - y) \\
&\leq 0.
\end{aligned}$$

When $x \leq t \leq y$,

$$\begin{aligned}
\mathbf{E}\left[\Delta_W \mid |v_{n-1}| = x, |v_n| = y\right] &= \frac{1}{4}(x - y) + \frac{1}{4}(t - x) + \frac{1}{4}(-x - y) + \frac{1}{4}(x + t) \\
&= \frac{1}{2}(t - y) \\
&\leq 0.
\end{aligned}$$

When $t > y$,

$$\begin{aligned}
\mathbf{E}\left[\Delta_W \mid |v_{n-1}| = x, |v_n| = y\right] &= \frac{1}{4}(x - y) + \frac{1}{4}(y - x) + \frac{1}{4}(-x - y) + \frac{1}{4}(x + y) \\
&= 0.
\end{aligned}$$

Thus, Equation (8) is established.

We next consider the difference of the overtime for the physician under the two sequences. The difference is given by

$$\Delta_O = \max\{0, v_n, v_n + v_{n-1} + t\} - \max\{0, v_{n-1}, v_{n-1} + v_n + t\}.$$

We can use a similar argument to analyze the expected value of Δ_O conditional on $|v_{n-1}| = x$ and $|v_n| = y$, with $x \leq y$. It is not difficult to show that $\mathbf{E}[\Delta_O] \leq 0$.

Therefore, we have proved that scheduling the last two patients in increasing likelihood ratio order is optimal. ■

Proof of Proposition 1

We prove the above result for the first two positions in the sequence. The rest of the proof follows using similar arguments.

Referring to the discussion in the previous section, if $c_{\phi(j)}$ is associated with the waiting time of the jth patient, the first term in Equation (4) turns out to be $(1/2)c_{\phi(1)}(x-y)$. If $x \leq 2y$, then the lower bound of the second term in Equation (4) is $-(1/4)(x-y)\sum_{j=3}^{n+1} c_{\phi(j)}$. Therefore,

$$\begin{aligned}
\mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y, \{S_k\}\right] &\geq \frac{1}{2}c_{\phi(1)}(x-y) - \frac{1}{4}(x-y)\sum_{j=3}^{n+1} c_{\phi(j)} \\
&= \frac{1}{2}(x-y)\left(c_{\phi(1)} - \frac{1}{2}\sum_{j=3}^{n+1} c_{\phi(j)}\right) \\
&\geq 0.
\end{aligned}$$

If $x > 2y$, it is easy to verify that

$$\begin{aligned}
\mathbf{E}\left[\Delta_{TW} \mid |v_1| = x, |v_2| = y, \{S_k\}\right] &\geq \frac{1}{2}c_{\phi(1)}(x-y) - \frac{1}{4}y\sum_{j=3}^{n+1} c_{\phi(j)} \\
&\geq \frac{1}{2}y\left(c_{\phi(1)} - \frac{1}{2}\sum_{j=3}^{n+1} c_{\phi(j)}\right) \\
&\geq 0.
\end{aligned}$$

Thus, sequencing the patient with lower variability in the first slot incurs lower total waiting

time and overtime. ■

Proof of Proposition 2

Recall that the waiting time of the jth patient given a sequence ϕ is

$$w_{\phi(j)} = \max\left\{0, v_{\phi(j-1)}, v_{\phi(j-1)} + v_{\phi(j-2)}, \ldots, \sum_{k=1}^{j-1} v_{\phi(k)}\right\}, \; i = 2, \ldots, n+1.$$

Let ϕ^* denote the SVF rule, i.e., $|v_{\phi^*(j-1)}| = x_j$, $\forall j = 1, \ldots, n$. When $v_{\phi^*(j-1)} = -x_{j-1}$, given the condition in the proposition, $w_{\phi^*(j)} = 0$. When $v_{\phi^*(j-1)} = x_{j-1}$,

$$\begin{aligned} w_{\phi^*(j)} &= \max\left\{-x_{j-1}, 0, v_{j-2}, v_{j-2} + v_{j-3}, \ldots, \sum_{k=1}^{j-2} v_k\right\} + x_{j-1} \\ &= \max\left\{0, v_{j-2}, v_{j-2} + v_{j-3}, \ldots, \sum_{k=1}^{j-2} v_k\right\} + x_{j-1} \\ &= w_{\phi^*(j-1)} + x_{j-1}. \end{aligned}$$

Then,

$$\begin{aligned} \mathbf{E}\left[w_{\phi^*(j)}\right] &= \frac{1}{2}\mathbf{E}\left[w_{\phi^*(j-1)}\right] + \frac{1}{2}x_{j-1}. \\ &= \frac{1}{2}\left(\frac{1}{2}\mathbf{E}\left[w_{\phi^*(j-2)}\right] + \frac{1}{2}x_{j-2}\right) + \frac{1}{2}x_{j-1} \\ &= \ldots \\ &= \frac{1}{2^{j-1}}x_1 + \frac{1}{2^{j-2}}x_2 + \cdots + \frac{1}{2}x_{j-1}. \\ &= \sum_{k=1}^{j-1}\frac{1}{2^{j-k}}x_k. \end{aligned}$$

For the SVF rule, the expected total waiting time and overtime is thus

$$\sum_{j=2}^{n+1}\mathbf{E}\left[w_{\phi^*(j)}\right] = \sum_{j=1}^{n} C_j x_j, \text{ where } C_j = \sum_{k=j+1}^{n+1}\frac{1}{2^{k-j}}.$$

Given any other sequence ϕ, when $v_{\phi(j-1)} = -x_{\phi(j-1)}$, we can only conclude that $w_{\phi(j)} \geq 0$.

When $v_{\phi(j-1)} = x_{\phi(j-1)}$, $w_{\phi(j)} = w_{\phi(j-1)} + x_{\phi(j-1)}$. Hence,

$$\begin{aligned}
\mathbf{E}\left[w_{\phi(j)}\right] &\geq \frac{1}{2}\mathbf{E}\left[w_{\phi(j-1)}\right] + \frac{1}{2}x_{\phi(j-1)}. \\
&\geq \frac{1}{2}\left(\frac{1}{2}\mathbf{E}\left[w_{\phi(j-2)}\right] + \frac{1}{2}x_{\phi(j-2)}\right) + \frac{1}{2}x_{\phi(j-1)} \\
&\geq \ldots \\
&\geq \sum_{k=1}^{j-1} \frac{1}{2^{j-k}} x_{\phi(k)},
\end{aligned}$$

and

$$\sum_{j=2}^{n+1} \mathbf{E}\left[w_{\phi(j)}\right] \geq \sum_{j=1}^{n} C_j x_{\phi(j)}, \text{ where } C_j = \sum_{k=j+1}^{n+1} \frac{1}{2^{k-j}}.$$

Since $\{C_j\}_{j=1}^n$ is a decreasing sequence and $\{x_i\}_{i=1}^n$ is increasing, ϕ^* minimizes $\sum_{j=1}^n C_j x_{\phi(j)}$. Thus,

$$\sum_{j=2}^{n+1} \mathbf{E}\left[w_{\phi^*(j)}\right] = \sum_{j=1}^{n} C_j x_{\phi^*(j)} \leq \sum_{j=1}^{n} C_j x_{\phi(j)} \leq \sum_{j=2}^{n+1} \mathbf{E}\left[w_{\phi(j)}\right].$$

Therefore, sequencing patients in increasing variance is optimal. ■