Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Multimodularity in the Stochastic Appointment Scheduling Problem with Discrete Arrival Epochs

Christos Zacharias, Tallys Yunes

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 66, No. 2, February 2020, pp. 744–763
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Multimodularity in the Stochastic Appointment Scheduling Problem with Discrete Arrival Epochs

Christos Zacharias,[a] Tallys Yunes[a]

[a] Department of Management Science, University of Miami Business School, Coral Gables, Florida 33146

Contact: czacharias@bus.miami.edu, http://orcid.org/0000-0002-9911-7860 (CZ); tallys@miami.edu, http://orcid.org/0000-0002-8308-7812 (TY)

Received: October 17, 2017
Revised: May 11, 2018; August 13, 2018
Accepted: October 12, 2018
Published Online in Articles in Advance: July 18, 2019

https://doi.org/10.1287/mnsc.2018.3242

Abstract. We address the problem of designing appointment scheduling strategies in a stochastic environment accounting for patient no-shows, nonpunctuality, general stochastic service times, and unscheduled emergency walk-ins. A good appointment schedule seeks to help outpatient clinics to utilize their resources efficiently while containing patients' waiting times. The task of identifying an optimal schedule is modeled as a nonlinear integer program, where the objective function is the outcome of stochastic analysis in transient state. We maintain the discrete nature of the appointment scheduling problem by considering arrival epochs with discrete supports. By looking at discrete-time snapshots of the random evolution of a single-server queueing model, we characterize probabilistically the system's workload over time as a function of an appointment schedule, and we derive recursive expressions for the performance measures of interest. Subsequently, we unfold discrete convexity properties of the optimization problem. We prove that under general conditions the objective function is supermodular and componentwise convex. Under assumptions on patient punctuality, we prove that the optimal scheduling strategy minimizes a multimodular function, a property which guarantees that a locally optimal schedule is also globally optimal. The size of the local neighborhood, however, grows exponentially with the dimension of the problem. To the best of our knowledge, this study is the first to develop and implement an algorithm for minimizing locally a multimodular function over nonnegative integer vectors via submodular set-function minimization over ring families, a task that can be performed in polynomial time.

History: Accepted by David Simchi-Levi, operations management.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2018.3242.

Keywords: transient queues • discrete optimization • multimodularity • appointment scheduling

1. Introduction

Appointment scheduling is a ubiquitous endeavor across outpatient clinics in an effort to optimally manage their patient arrivals. From an operational standpoint, a good schedule strikes the right balance between efficient resource utilization and short waiting times. Outpatient clinics often face various sources of variability that disrupt their daily operations, and which add layers of complexity to the appointment scheduling problem. For example, there is variability and seasonality in the daily demand for outpatient care, it is common for patients to not show up for their scheduled services, patients who show up are not necessarily punctual, consultation times are stochastic, and unscheduled emergency patients may need to be seen by their provider on short notice.

We address the problem of designing optimal appointment scheduling strategies so that outpatient clinics utilize their resources efficiently, while containing patients' waiting times, in a stochastic environment. By looking at discrete-time snapshots of the random evolution of a single-server queueing model, we characterize probabilistically the system's workload over time as a function of an appointment schedule and we derive recursive expressions for the performance measures of interest. We maintain the discrete nature of the appointment scheduling problem by considering arrival epochs with discrete supports. The task of finding an optimal schedule is modeled as a nonlinear integer program. We identify conditions under which the problem possesses discrete convexity properties and propose an algorithm that solves the combinatorial problem effectively and efficiently. Our three main contributions can be summarized as follows.

(i) **Queueing:** Our queueing model provides a unifying platform for addressing the well-studied appointment scheduling problem, as many single-server models in the literature can be considered as its special cases (e.g., Kaandorp and Koole 2007, Robinson and Chen 2010, Zeng et al. 2010, LaGanga and Lawrence 2012, and Zacharias and Pinedo 2017 with $s = 1$). This is a combined outcome from considering general stochastic service times, nonpunctuality, and our approach to characterize the random evolution of the workload process of the system (as opposed to the queue-length process).

(ii) **Discrete optimization:** We prove that, under general conditions, the objective function is supermodular and componentwise convex. While these structural properties have an intuitive interpretation about the interactions between the decision variables, and have served as the basis for developing heuristic solutions in the literature (for example in LaGanga and Lawrence 2012), they are not enough to ensure that the problem can be solved to exact optimality. When patients promptly arrive for their scheduled appointments with the same show-up probability, we prove that the optimal scheduling strategy minimizes a multimodular function. This property guarantees that a locally optimal schedule is also globally optimal. The size of the local neighborhood, however, grows exponentially with the dimension of the problem. To the best of our knowledge, this study is the first to develop and implement an algorithm for minimizing locally (and eventually globally) a multimodular function over nonnegative integer vectors in polynomial time. This allowed us to solve large problems up to 96 slots. Our optimization framework bridges recent advances in discrete convex analysis and submodular set-function minimization over ring families, and has the potential to help address problems in other areas within and beyond the field of healthcare operations.

(iii) **Managerial insights:** Our stochastic model, in conjunction with our efficient optimization procedure, allowed us to gain some novel insights into optimal appointment scheduling. For example, how should outpatient clinics adjust their scheduling strategies to deal with the various sources of uncertainty? What is the right timescale for an appointment schedule so that it is practical and achieves our operational goals?

2. Related Literature

The appointment scheduling problem has received much attention in the operations management literature, and it is still an active area of academic research. Ahmadi-Javid et al. (2017) provide a comprehensive review of analytical and numerical optimization studies (published between 2003 and 2016) for designing and planning outpatient appointment systems. Because of the inherent complexity of the problem, and for the sake of analytical tractability, the existing literature often relies on stylized assumptions. For example, it is common to assume that patients who show up are punctual, and/or to assume that service times are either deterministic or follow an exponential distribution (and thus leveraging its memoryless property), and/or to assume that there are no emergency walk-ins. However, consultation times in practice are neither deterministic nor exponential (Cayirli et al. 2006), patients are not always punctual (Kim et al. 2018), and unscheduled emergency walk-ins are often significant (Cayirli et al. 2012). From the patient's point of view, Liu et al. (2018) study patient preferences and choice behavior to better understand how patients value the various operational attributes when scheduling appointments.

2.1. Inter- and Intraday Models

Some papers in the literature seek to optimize the interday operations of outpatient clinics over some finite scheduling horizon, without capturing the detailed intraday dynamics. For example, Patrick et al. (2008), Liu et al. (2010), Feldman et al. (2014), and Truong (2015) develop and analyze dynamic models that deal with daily demand uncertainty and take into account the costs of indirect delay (i.e., the time gap between the request for an appointment and the actual offered appointment) in the scheduling decisions. Green and Savin (2008), Liu and Ziya (2014), Liu (2016), and Zacharias and Armony (2017) analyze the appointment backlog and indirect delay with queueing systems in steady state.

Another stream of literature seeks to optimize a single day's operations by analyzing the detailed intraday dynamics. For example, in Green et al. (2006), Hassin and Mendel (2008), Robinson and Chen (2010), Zeng et al. (2010), LaGanga and Lawrence (2012), Zacharias and Pinedo (2014), Chen and Robinson (2014), and Zacharias and Pinedo (2017), it is assumed that there is enough demand to fill any daily schedule with appointments, and the objective is to design an optimal static schedule for a single day by properly managing patients' appointment times. Green et al. (2006) further establish dynamic priority rules that deal with unscheduled emergency patients. Our work belongs to this stream of literature, and to the best of our knowledge, our stochastic model is the first to consider jointly all main sources of variability: general stochastic service times, no-shows, nonpunctuality, and unscheduled emergency walk-ins.

Feldman et al. (2014) have pointed out that the simultaneous consideration of appointment day (interday) and time of day (intraday) in the scheduling decisions is an important problem that has not been studied adequately in the literature because of its large dimensionality.

2.2. Continuous and Discrete Static Models

The single day static appointment scheduling problem has been analyzed in the literature in a variety of modeling approaches and optimization procedures. One way to classify the various models is with respect to the type of decision variables: discrete, continuous, or a mixture. In continuous settings (e.g., Wang 1997, Lau and Lau 2000, Hassin and Mendel 2008, Luo et al. 2012, and Kuiper et al. 2015) the main decision variables are the interarrival times of a fixed set of patients. Luo et al. (2012) also consider the setting where the number of patients to be scheduled is a decision variable. The optimization problem in these continuous models is solved via nonlinear programming techniques (e.g., sequential quadratic programming, quasi-Newton method, interior point methods) that identify local optima. If the objective function is assumed/believed/proved to be convex, then local optima should also be global optima. Luo et al. (2012) have, in fact, shown numerically that there might be multiple local optima. They mitigate this issue by choosing the best solution for different starting points of interior-point algorithms. Kong et al. (2013), Mak et al. (2015), Qi (2017), and Jiang et al. (2017) use robust optimization techniques to provide tractable and equivalent conic programming and linear programming formulations with exact solutions.

In discrete settings, typically (e.g., Kaandorp and Koole 2007, Zeng et al. 2010, Robinson and Chen 2010, LaGanga and Lawrence 2012, and Zacharias and Pinedo 2017) a workday is partitioned into an integer number of slots, and the decision variables become how many patients to schedule in each slot (potentially zero). In some discrete models the total number of patients is further allowed to be the outcome of optimization, as opposed to a fixed problem parameter. The discrete nonlinear optimization problem is solved via either some local search procedure, complete enumeration on a combinatorially large solution space, or heuristics, depending on the theoretical findings. For example, Kaandorp and Koole (2007) and Zeng et al. (2010) proved that the objective function is multimodular in a single-server system with exponential service times. Zacharias and Pinedo (2017) proved multimodularity of the objective for the multiserver system with deterministic service times. Multimodularity guarantees the global optimality of solutions that are optimal within some discrete local neighborhood. Wang et al. (2019) proved multimodularity of the objective for the single-server system with deterministic service times and general walk-in arrivals, and developed an equivalent two-stage stochastic linear programming formulation. Our work is the first study to identify conditions on the arrival process (as a function of scheduled traffic) under which the objective is multimodular for general stochastic service times, and to identify conditions under which the multimodularity property collapses.

A distinctive case of a static model is the one in Begen and Queyranne (2011), where the optimization problem is to identify on a continuous space the optimal interarrival times of a given set of heterogeneous patients. By considering independent discrete probability distributions for the service times, they prove the existence of an integer optimal schedule and, under a mild condition on the objective coefficients, that the objective function is L-convex. Begen et al. (2012) extend the results of Begen and Queyranne (2011) to consider joint discrete distributions based on independent random samples, and determine bounds on the number of independent samples required to obtain a near-optimal solution with high probability. Ge et al. (2014) extend the results of Begen and Queyranne (2011) and Begen et al. (2012) to account for piecewise linear cost functions with integer break points.

2.3. Discrete Optimization

In discrete optimization it is pivotal to identify structures that guarantee the success of some efficient optimization procedure; either to exact optimality or within some good approximation factor. Various researchers have proposed discrete analogues of convex functions, all possessing the following property: a local optimum is also a global optimum. Examples include "discretely convex" functions by Miller (1971), "multimodular" functions by Hajek (1985), "integrally convex" functions by Favati and Tardella (1990), and "M/M$^\natural$-convex" and "L/L$^\natural$-convex" functions by Murota (1998). Locality, though, is defined in different ways, according to the type of discrete convexity considered. Multimodular functions and their properties were introduced by Hajek (1985) in the context of optimal admission control to queues with no state information. Altman et al. (2000) generalize some of the results of Hajek (1985) and present additional properties of multimodular functions and asymptotically optimal admission control policies with no state information. In the static appointment scheduling problem, an optimal schedule is established in advance without any prior knowledge about the resulting arrival process or the workload of the system over time (i.e., no future state information). Therefore, intuition suggests that the appointment scheduling problem possesses multimodularity as well, under certain assumptions.

It is demonstrated in Murota (2005) that multimodularity and L$^\natural$-convexity are related through a linear transformation. Both properties guarantee the global optimality of local optima. Locality, though, is defined differently under these two notions of discrete convexity. Both local neighborhoods of a given vector $\mathbf{x} \in \mathbb{Z}^n$ contain $2^{n+1} - 2$ neighbors, a number that grows exponentially with the dimension of the problem, but they are not identical. As an example, Figure 1 illustrates what the local neighborhoods look like in three dimensions. Murota (2005) provides an algorithm for minimizing an unconstrained L$^\natural$-convex function over $\mathbb{Z}^n$ in polynomial time. The same algorithm can be readily adjusted to minimize an unconstrained multimodular function over $\mathbb{Z}^n$ through a linear transformation. However, as we demonstrate in Section 4.2, minimizing a multimodular function over $\mathbb{Z}^n_+$ requires a more careful treatment, since the linear transformation results in minimization of an L$^\natural$-convex function subject to nontrivial constraints. By combining the results of Murota (2005) and Schrijver (2000), we propose and implement the first polynomial time algorithm for minimizing locally a multimodular function over nonnegative integer vectors via unconstrained submodular set-function minimization.

3. Queueing Model

In this section a single-server queueing model is introduced, based on which we characterize the random evolution of the system's workload over time and we derive recursive expressions for the performance measures of interest. The model and assumptions are as follows:

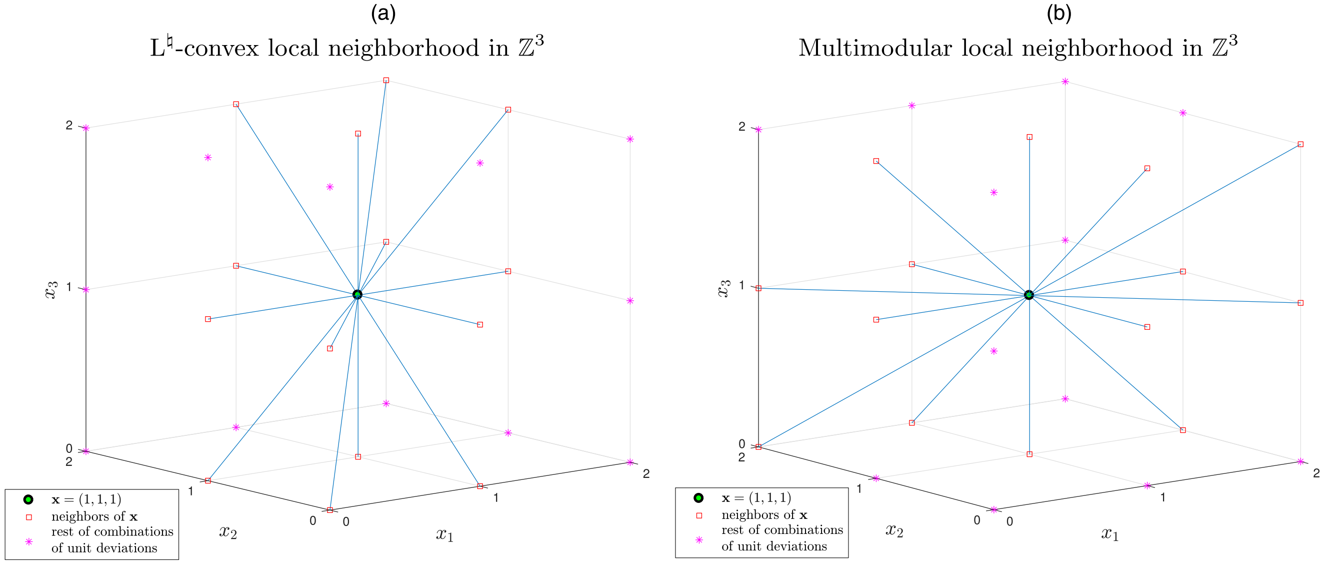
- **Queueing system:** Single-server queue in transient state with discrete arrival epochs and general stochastic service times. The queue is work conserving and the service discipline is first-in, first-out (FIFO).
- **Timescale:** Time is measured in minutes, and the length of a regular workday is T minutes, during which the scheduled appointments are allocated. The clinic, however, continues to work overtime as well, beyond T, until the queue empties out. Time is continuous, yet a workday is partitioned into n discrete time slots of equal duration $d = T/n$—for example, slots of 30 minutes, 15 minutes, 10 minutes, 5 minutes, 1 minute, and so forth—depending on how refined we would like a schedule to be. We assume that d is a positive integer such that n is also some positive integer. The discrete time slots are denoted by $t = 1, 2, \ldots, n, n+1$, where slot $n+1$ is the first overtime slot. Assuming that the first regular slot starts at time zero, then time slot t occupies the time interval $[(t-1)d, td)$.
- **Arrival process:** There are two arrival streams: one driven by scheduled appointments and one from unscheduled emergency walk-ins. An appointment schedule is denoted by a vector

$$\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{Z}^n_+,$$

where x_t is the number of patients assigned to slot t—i.e., scheduled to arrive at time $(t-1)d$. We denote the partial schedule up to slot t with $\mathbf{x}_t = (x_1, x_2, \ldots, x_t) \in \mathbb{Z}^t_+$, $t = 1, 2, \ldots, n$.

Patients are not necessarily punctual and may arrive on time, earlier or later than their scheduled appointment time, or not at all. We consider that no-show probabilities are slot-dependent. In particular, we consider that each patient assigned to slot t independently shows up with probability $p_t \in (0, 1]$—i.e., a patient assigned to slot t showing up or not is an independent Bernoulli(p_t) trial. We also assume that the actual arrival slots of patients assigned to some slot t, given that they show up, are i.i.d. random variables with discrete support on some subset of $\{1, 2, \ldots, n\}$. We denote with $q_{t\tau}$ the probability that a patient assigned to slot t will arrive at the beginning

Figure 1. (Color online) Local Neighborhoods Under Discrete Convexity in Three Dimensions


(a)
L$^\natural$-convex local neighborhood in $\mathbb{Z}^3$
x_3
0
1
2
x_2
x_1
● $\mathbf{x} = (1,1,1)$
□ neighbors of $\mathbf{x}$
∗ rest of combinations of unit deviations
(b)
Multimodular local neighborhood in $\mathbb{Z}^3$
x_3
0
1
2
x_2
x_1
● $\mathbf{x} = (1,1,1)$
□ neighbors of $\mathbf{x}$
∗ rest of combinations of unit deviations

of slot τ, and for notational convenience we denote the no-show probabilities as

$$q_{tn+1} = 1 - \sum_{\tau=1}^{n} q_{t\tau} = 1 - p_t, \quad t = 1, 2, \ldots, n.$$

The random vector $\mathbf{S}^{\mathbf{x}} = (S_1^{\mathbf{x}}, S_2^{\mathbf{x}}, \ldots, S_n^{\mathbf{x}}) \in \mathbb{Z}_+^n$ denotes the arrival process from scheduled appointments under $\mathbf{x}$, where $S_t^{\mathbf{x}}$ is the number of scheduled patients that arrive right at the beginning of slot $t = 1, 2, \ldots, n$. Independently from the schedule $\mathbf{x}$ and the system's workload, unscheduled emergency walk-ins may arrive throughout the day. The random vector $\mathbf{U} = (U_1, U_2, \ldots, U_n) \in \mathbb{Z}_+^n$ denotes the unscheduled arrivals and follows some multivariate distribution with finite mean. The resulting arrival process from both schedule $\mathbf{x}$ and walk-ins is denoted by the random vector $\mathbf{A}^{\mathbf{x}} = (A_1^{\mathbf{x}}, A_2^{\mathbf{x}}, \ldots, A_n^{\mathbf{x}}) \in \mathbb{Z}_+^n$, where $A_t^{\mathbf{x}} = S_t^{\mathbf{x}} + U_t$ is the total number of patients that arrive right at the beginning of slot $t = 1, 2, \ldots, n$. We denote the partial arrival process up to slot t with $\mathbf{A}_t^{\mathbf{x}} = (A_1^{\mathbf{x}}, A_2^{\mathbf{x}}, \ldots, A_t^{\mathbf{x}}) \in \mathbb{Z}_+^t$, $t = 1, 2, \ldots, n$.

- **Service times:** Service times are i.i.d. random variables following some general distribution (either continuous, discrete, or a mixture) with finite mean μ and variance σ^2. Let R be the random variable representing one service time. We denote the k-fold convolution of R as

$$R^{(k)} = \sum_{i=1}^{k} R_i,$$

where $R_i \sim R$, and as a notational convention we consider that $R^{(0)} = 0$ with probability 1.

- **New-workload process:** Let the random vector $\mathbf{Y}^{\mathbf{x}} = (Y_1^{\mathbf{x}}, Y_2^{\mathbf{x}}, \ldots, Y_n^{\mathbf{x}}) \in \mathbb{R}_+^n$ denote the *new-workload* process under schedule $\mathbf{x}$—that is, $Y_t^{\mathbf{x}}$ new workload arrives at the beginning of slot t with distribution

$$Y_t^{\mathbf{x}} \sim R^{(A_t^{\mathbf{x}})}, \quad t = 1, 2, \ldots, n.$$

We denote the new-workload process up to slot t with $\mathbf{Y}_t^{\mathbf{x}} = (Y_1^{\mathbf{x}}, Y_2^{\mathbf{x}}, \ldots, Y_t^{\mathbf{x}}) \in \mathbb{R}_+^t$, $t = 1, 2, \ldots, n$.

- **Workload process:** The workload of the system right at the end of slot t—that is, the unfinished workload carried forward from slot t to slot $t+1$—is denoted by $Z_t^{\mathbf{x}}$. The workload process $\mathbf{Z}^{\mathbf{x}} = (Z_1^{\mathbf{x}}, Z_2^{\mathbf{x}}, \ldots, Z_n^{\mathbf{x}}) \in \mathbb{R}_+^n$ satisfies the Lindley recursion

$$Z_t^{\mathbf{x}} = \max\{Z_{t-1}^{\mathbf{x}} + Y_t^{\mathbf{x}} - d, 0\}, \quad \text{for } t = 1, 2, \ldots, n, \qquad (1)$$

where $Z_0^{\mathbf{x}} = 0$ with probability 1, and $Z_n^{\mathbf{x}}$ corresponds to the overtime workload. We denote the partial workload process up to slot t with $\mathbf{Z}_t^{\mathbf{x}} = (Z_1^{\mathbf{x}}, Z_2^{\mathbf{x}}, \ldots, Z_t^{\mathbf{x}}) \in \mathbb{R}_+^t$, $t = 1, 2, \ldots, n$.

- **Idle-time process:** Finally, we define the idle-time process as the vector $\mathbf{L}^{\mathbf{x}} = (L_1^{\mathbf{x}}, L_2^{\mathbf{x}}, \ldots, L_n^{\mathbf{x}}) \in [0, d]^n$, where $L_t^{\mathbf{x}}$ denotes the idle time during slot t and

$$L_t^{\mathbf{x}} = \max\{d - Z_{t-1}^{\mathbf{x}} - Y_t^{\mathbf{x}}, 0\}, \quad \text{for } t = 1, 2, \ldots, n. \qquad (2)$$

From (1), (2), and the identity $x^+ - x^- = x \; \forall x \in \mathbb{R}$, we also get the relationship

$$Z_t^{\mathbf{x}} - L_t^{\mathbf{x}} = Z_{t-1}^{\mathbf{x}} + Y_t^{\mathbf{x}} - d, \quad \text{for } t = 1, 2, \ldots, n. \qquad (3)$$

3.1. Transient Analysis

Variants of the recursion in (1) have been analyzed in the literature for performance analysis of either a workload process or a queue-length process in transient state. For example, Janssen and van Leeuwaarden (2005) analyze the expected waiting time for the discrete $D/G/1$ queue in transient state and its rate of convergence to steady state. Zeng et al. (2010), Robinson and Chen (2010), and Zacharias and Pinedo (2017), by assuming that patients are punctual, make use of the dynamics captured in a Lindley recursion to provide recursive expressions for the probability distribution of a queue-length process in transient state, and consequently recursive expressions for the performance measures of interest. In our setting, a recursive derivation for the probability distribution of $Z_t^{\mathbf{x}}$ based on (1) requires that $\mathbf{Y}_t^{\mathbf{x}}$ be a sequence of independent random variables; a condition that does not hold when patients are nonpunctual and/or when the walk-in arrival process follows some multivariate distribution. We overcome this obstacle as follows:

(a) First, we derive recursively the conditional cumulative distribution function (CDF) of $Z_t^{\mathbf{x}}$ given a realization of $\mathbf{A}_t^{\mathbf{x}}$ (the random vector $\mathbf{Y}_t^{\mathbf{x}} | \mathbf{A}_t^{\mathbf{x}}$ is indeed a sequence of independent random variables) for all $t = 1, 2, \ldots, n$.

(b) Second, we derive the conditional expectation of $Z_t^{\mathbf{x}}$ given a realization of $\mathbf{A}_t^{\mathbf{x}}$.

(c) Third, we express the conditional expectation of all performance measures (idle time, overtime, waiting time) in terms of the conditional expectation of $Z_t^{\mathbf{x}}$.

(d) Finally, we apply the law of total expectation to get the unconditional expectation of all performance measures of interest.

Let

$$G_{\mathbf{a}_t}(z) \triangleq F_{Z_t^{\mathbf{x}} | \mathbf{A}_t^{\mathbf{x}}}(z | \mathbf{a}_t) = \mathbb{P}(Z_t^{\mathbf{x}} \le z | \mathbf{A}_t^{\mathbf{x}} = \mathbf{a}_t), \quad z \ge 0$$

denote the conditional CDF of $Z_t^{\mathbf{x}}$, given a partial realization of the arrival process $\mathbf{a}_t = (a_1, a_2, \ldots, a_t) \in \mathbb{Z}_+^t$ up to some slot $t = 1, 2, \ldots n$. As a notational convention we let $G_{\mathbf{a}_0}(z) = 1$ for all $z \ge 0$, by assuming that the system is empty at the beginning of the workday. Let also

$$H_{a_t}(y) \triangleq F_{Y_t^{\mathbf{x}} | A_t^{\mathbf{x}}}(y | a_t) = F_{R^{(a_t)}}(y) = \mathbb{P}(R^{(a_t)} \le y), \quad y \ge 0$$

denote the conditional CDF of the new workload $Y_t^{\mathbf{x}}$, given a realization a_t of the number of new arrivals at slot $t = 1, 2, \ldots n$. Then $G_{\mathbf{a}_t}(z)$ can be expressed recursively for $t = 1, 2, \ldots, n$ as

$$\begin{aligned}
G_{\mathbf{a}_t}(z) &= \mathbb{P}(Z_t^{\mathbf{x}} \le z | \mathbf{A}_t^{\mathbf{x}} = \mathbf{a}_t)\\
&= \int_0^{d+z} \mathbb{P}(Z_t^{\mathbf{x}} \le z | \mathbf{A}_t^{\mathbf{x}} = \mathbf{a}_t, Y_t^{\mathbf{x}} = y)\ \mathrm{d}H_{a_t}(y)\\
&= \int_0^{d+z} \mathbb{P}(\max\{Z_{t-1}^{\mathbf{x}} + Y_t^{\mathbf{x}} - d, 0\}\\
&\le z | \mathbf{A}_t^{\mathbf{x}} = \mathbf{a}_t, Y_t^{\mathbf{x}} = y)\ \mathrm{d}H_{a_t}(y)\\
&= \int_0^{d+z} \mathbb{P}(\max\{Z_{t-1}^{\mathbf{x}} + y - d, 0\}\\
&\le z | \mathbf{A}_{t-1}^{\mathbf{x}} = \mathbf{a}_{t-1})\ \mathrm{d}H_{a_t}(y)\\
&= \int_0^{d+z} \mathbb{P}(Z_{t-1}^{\mathbf{x}} \le z + d - y | \mathbf{A}_{t-1}^{\mathbf{x}} = \mathbf{a}_{t-1})\ \mathrm{d}H_{a_t}(y)\\
&= \int_0^{d+z} G_{\mathbf{a}_{t-1}}(z + d - y)\ \mathrm{d}H_{a_t}(y),\ z \ge 0. \qquad (4)
\end{aligned}$$

The conditional expected workload over time, given a realization of the arrival process, is

$$\mathbb{E}[Z_t^{\mathbf{x}} | \mathbf{A}_t^{\mathbf{x}} = \mathbf{a}_t] = \int_0^{\infty} z\ \ \mathrm{d}G_{\mathbf{a}_t}(z). \qquad (5)$$

Based on (5) we can now express the performance measures of interest. Let the random variables $I(\mathbf{x})$, $O(\mathbf{x})$, $W_s(\mathbf{x})$, and $W_u(\mathbf{x})$ denote the system's total idle time, overtime, scheduled patients' aggregate waiting time, and unscheduled patients' aggregate waiting time, respectively, associated with schedule $\mathbf{x}$.

The overtime workload is equal to the workload at the end of the regular workday, and therefore

$$\begin{aligned}
\mathbb{E}[O(\mathbf{x})] &= \mathbb{E}_{\mathbf{A}^{\mathbf{x}}}\big[\mathbb{E}[O(\mathbf{x}) | \mathbf{A}^{\mathbf{x}}]\big]\\
&= \mathbb{E}_{\mathbf{A}^{\mathbf{x}}}\big[\mathbb{E}[Z_n^{\mathbf{x}} | \mathbf{A}^{\mathbf{x}}]\big]. \qquad (6)
\end{aligned}$$

By rearranging terms in (3) we get

$$\begin{aligned}
&L_t^{\mathbf{x}} + Y_t^{\mathbf{x}} = Z_t^{\mathbf{x}} - Z_{t-1}^{\mathbf{x}} + d, \quad \text{for } t = 1, 2, \ldots, n\\
&\Rightarrow \sum_{t=1}^{n} (L_t^{\mathbf{x}} + Y_t^{\mathbf{x}}) = \sum_{t=1}^{n} (Z_t^{\mathbf{x}} - Z_{t-1}^{\mathbf{x}} + d)\\
&\Rightarrow I(\mathbf{x}) + \sum_{t=1}^{n} Y_t^{\mathbf{x}} = O(\mathbf{x}) + nd. \qquad (7)
\end{aligned}$$

We note that (7) can be interpreted intuitively as

(total idle time) + (total workload) = (overtime workload) + (length of regular workday).

Now the total expected idle time can be expressed as

$$\begin{aligned}
\mathbb{E}[I(\mathbf{x})] &= \mathbb{E}[O(\mathbf{x})] + nd - \sum_{t=1}^{n} \mathbb{E}[Y_t^{\mathbf{x}}]\\
&= \mathbb{E}[O(\mathbf{x})] + nd - \sum_{t=1}^{n} \mu \mathbb{E}[A_t^{\mathbf{x}}]\\
&= \mathbb{E}[O(\mathbf{x})] + nd - \mu \sum_{t=1}^{n} (p_t x_t + \mathbb{E}[U_t]). \qquad (8)
\end{aligned}$$

Patients' waiting times require a more careful treatment. While the derivation of system's idle time and overtime are based only on the assumption of a work-conserving queue, we need additional assumptions on the service discipline to express the waiting time for the two patient groups (scheduled and unscheduled walk-in patients). We assume that the service discipline is FIFO, and that scheduled patients have priority over unscheduled walk-in patients when they show up at the same time slot (we reckon that unscheduled walk-in patients are less sensitive to waiting, given the opportunity to see their provider on short notice). Under these two assumptions, we can express the waiting times for the two patient groups based on the workload process and the number of new arrivals. In our setting, a service discipline that prioritizes scheduled patients over all unscheduled patients (e.g., Wang et al. 2019) would require keeping track of two separate queue length processes and two separate workload processes (one for each patient group), a task that would have complicated the analysis significantly and potentially led to an intractable model. In Wang et al. (2019) this service discipline is tractable, since queue length implies workload when service times are deterministic.

Let $\mathbf{s}_t$ and $\mathbf{u}_t$ denote a partial realization of the arrival process from scheduled appointments and from emergency walk-ins, respectively, up to some slot t. Let also

$$\begin{aligned}
W_{t,i}^{\mathbf{x}}(\mathbf{s}_{t-1}, \mathbf{u}_{t-1}) &\overset{\mathrm{d}}{=} \big[Z_{t-1}^{\mathbf{x}} + R^{(i-1)} | \mathbf{S}_{t-1}^{\mathbf{x}} = \mathbf{s}_{t-1}, \mathbf{U}_{t-1}^{\mathbf{x}} = \mathbf{u}_{t-1}\big]\\
&= \big[Z_{t-1}^{\mathbf{x}} + R^{(i-1)} | \mathbf{A}_{t-1}^{\mathbf{x}} = \mathbf{s}_{t-1} + \mathbf{u}_{t-1}\big]
\end{aligned}$$

denote the waiting time of the i^{th} patient that receives service who arrives at slot t, given a realization of the arrival process, $1 \le t \le n$ and $1 \le i \le s_t + u_t$. Based on the assumption that service times are i.i.d. random variables and independent of the arrival process, we get

$$\begin{aligned}
\mathbb{E}[W_{t,i}^{\mathbf{x}}(\mathbf{s}_{t-1}, \mathbf{u}_{t-1})] &= \mathbb{E}[Z_{t-1}^{\mathbf{x}} | \mathbf{A}_{t-1}^{\mathbf{x}} = \mathbf{s}_{t-1} + \mathbf{u}_{t-1}] + \mathbb{E}[R^{(i-1)}]\\
&= \mathbb{E}[Z_{t-1}^{\mathbf{x}} | \mathbf{A}_{t-1}^{\mathbf{x}} = \mathbf{s}_{t-1} + \mathbf{u}_{t-1}] + (i-1)\mu.
\end{aligned}$$

Therefore, the expected aggregate sum of all scheduled patients' waiting times across all slots is

$$\begin{aligned}
\mathbb{E}[W_s(\mathbf{x})] &= \mathbb{E}_{\mathbf{S}^{\mathbf{x}},\mathbf{U}}\big[\mathbb{E}[W_s(\mathbf{x}) | \mathbf{S}^{\mathbf{x}}, \mathbf{U}]\big]\\
&= \mathbb{E}_{\mathbf{S}^{\mathbf{x}},\mathbf{U}}\left[\sum_{t=1}^{n} \sum_{i=1}^{S_t^{\mathbf{x}}} \mathbb{E}[W_{t,i}^{\mathbf{x}}(\mathbf{S}_{t-1}^{\mathbf{x}}, \mathbf{U}_{t-1})]\right]\\
&= \mathbb{E}_{\mathbf{S}^{\mathbf{x}},\mathbf{U}}\left[\sum_{t=1}^{n} \sum_{i=1}^{S_t^{\mathbf{x}}} \big[\mathbb{E}[Z_{t-1}^{\mathbf{x}} | \mathbf{A}_{t-1}^{\mathbf{x}}] + (i-1)\mu\big]\right]\\
&= \mathbb{E}_{\mathbf{S}^{\mathbf{x}},\mathbf{U}}\left[\sum_{t=1}^{n} \Big[S_t^{\mathbf{x}} \mathbb{E}[Z_{t-1}^{\mathbf{x}} | \mathbf{A}_{t-1}^{\mathbf{x}}] + \tfrac{S_t^{\mathbf{x}}(S_t^{\mathbf{x}}-1)}{2}\mu\Big]\right], \qquad (9)
\end{aligned}$$

and, the expected aggregate sum of all unscheduled patients' waiting times across all slots is

$$\begin{aligned}\mathbb{E}[W_u(\mathbf{x})] &= \mathbb{E}_{\mathbf{S}^{\mathbf{x}},\mathbf{U}}\big[\mathbb{E}[W_u(\mathbf{x})|\mathbf{S}^{\mathbf{x}},\mathbf{U}]\big]\\ &= \mathbb{E}_{\mathbf{S}^{\mathbf{x}},\mathbf{U}}\left[\sum_{t=1}^{n}\sum_{i=S_t^{\mathbf{x}}+1}^{S_t^{\mathbf{x}}+U_t}\mathbb{E}[W^{\mathbf{x}}_{t,i}(\mathbf{S}^{\mathbf{x}}_{t-1},\mathbf{U}_{t-1})]\right]\\ &= \mathbb{E}_{\mathbf{S}^{\mathbf{x}},\mathbf{U}}\left[\sum_{t=1}^{n}\sum_{i=S_t^{\mathbf{x}}+1}^{S_t^{\mathbf{x}}+U_t}\big[\mathbb{E}[Z^{\mathbf{x}}_{t-1}|\mathbf{A}^{\mathbf{x}}_{t-1}]+(i-1)\mu\big]\right]\\ &= \mathbb{E}_{\mathbf{S}^{\mathbf{x}},\mathbf{U}}\left[\sum_{t=1}^{n}\Big[U_t\mathbb{E}[Z^{\mathbf{x}}_{t-1}|\mathbf{A}^{\mathbf{x}}_{t-1}]+S_t^{\mathbf{x}}U_t\mu\right.\\ &\qquad\left.+\frac{U_t(U_t-1)}{2}\mu\Big]\right]. \qquad (10)\end{aligned}$$

We point out that for the special case of exponentially distributed service times, the recursive distribution of the workload process in (4) can be simplified so that integration is no longer involved—see for example Zeng et al. (2010). This task is feasible because of the memoryless property of the exponential distribution.

3.2. Punctual Patients and Random Emergency Walk-ins with No Covariates

The computation of the expected system's workload as a function of a schedule in Section 3.1 relies on taking expectation over all possible realizations of a discrete arrival process, a task that has exponential complexity (exponential in the number of patients in a schedule). We assume in this section that patients who do show up are punctual, and that **U** is a sequence of independent random variables. Under these two assumptions we can adjust the recursive distribution in (4) to be dependent on **x** itself (as opposed to a realization of the arrival process) and, as a result, compute all performance measures of interest much more efficiently.

In particular, let

$$\tilde{G}_{\mathbf{x}_t}(z) \triangleq F_{Z_t^{\mathbf{x}_t}}(z) = \mathbb{P}(Z_t^{\mathbf{x}_t} \le z), \quad z \ge 0$$

denote the CDF of $Z_t^{\mathbf{x}_t}$ under a partial schedule $\mathbf{x}_t = (x_1, x_2, \ldots, x_t) \in \mathbb{Z}_+^t$ up to some slot $t = 1, 2, \ldots, n$. As a notational convention we let $\tilde{G}_{\mathbf{x}_0}(z) = 1$ for all $z \ge 0$. The new workload that enters the system over time depends only on the current number of patients scheduled to arrive, and follows a random sum of the service times distribution. In particular, $Y_t^{x_t} \sim R^{(S_t+U_t)}$, where $S_t \sim \text{Binomial}(x_t, p_t)$ for all $t = 1, 2, \ldots, n$. Let also

$$\tilde{H}_{x_t}(y) \triangleq F_{Y_t^{x_t}}(y) = F_{R^{(S_t+U_t)}}(y) = \mathbb{P}(R^{(S_t+U_t)} \le y), \quad y \ge 0$$

denote the CDF of $Y_t^{x_t}$. Similarly to (4), $\tilde{G}_{\mathbf{x}_t}(z)$ can be expressed recursively for $t = 1, 2, \ldots, n$ as

$$\tilde{G}_{\mathbf{x}_t}(z) = \int_0^{d+z} \tilde{G}_{\mathbf{x}_{t-1}}(z+d-y)\ \mathrm{d}\tilde{H}_{x_t}(y), \quad z \ge 0. \qquad (11)$$

Finally, (6), (8), (9), and (10) can be simplified, respectively, as

$$\begin{aligned}\mathbb{E}[O(\mathbf{x})] &= \mathbb{E}[Z_n^{\mathbf{x}}],\\ \mathbb{E}[I(\mathbf{x})] &= \mathbb{E}[O(\mathbf{x})] + nd - \mu\sum_{t=1}^{n}(p_t x_t + \mathbb{E}[U_t]),\\ \mathbb{E}[W_s(\mathbf{x})] &= \left[\sum_{t=1}^{n} p_t x_t \mathbb{E}[Z_{t-1}^{\mathbf{x}}] + \mathbb{E}[S_t(S_t-1)]\frac{\mu}{2}\right],\\ \text{and } \mathbb{E}[W_u(\mathbf{x})] &= \sum_{t=1}^{n}\Big[\mathbb{E}[U_t]\mathbb{E}[Z_{t-1}^{\mathbf{x}}] + \mathbb{E}[U_t]p_t x_t \mu\\ &\qquad + \mathbb{E}[U_t(U_t-1)]\frac{\mu}{2}\Big],\end{aligned}$$

where

$$\mathbb{E}[Z_t^{\mathbf{x}}] = \int_0^{\infty} z\ \mathrm{d}\tilde{G}_{\mathbf{x}_t}(z) \text{ for all } t = 1, 2, \ldots., n.$$

4. Discrete Optimization

Consistently with the literature, we consider three costs (penalties) associated with a scheduling strategy: patients' waiting costs, system's idle time, and overtime costs. A waiting cost c_s (c_u) is incurred for each minute that a scheduled (unscheduled) patient has to wait before starting service. There is an idle time cost c_i per minute of idle time, and an overtime cost c_o is incurred for each minute the system has to operate overtime until the queue empties out. We normalize the objective function with respect to c_i—that is, $c_i = 1$—and we consider the following nonlinear integer program:

$$\begin{aligned}\min\{V(\mathbf{x}) \triangleq \mathbb{E}[I(\mathbf{x})] &+ c_o\mathbb{E}[O(\mathbf{x})] + c_s\mathbb{E}[W_s(\mathbf{x})]\\ &+ c_u\mathbb{E}[W_u(\mathbf{x})] : \mathbf{x} \in \mathbb{Z}_+^n\}. \qquad (\mathbf{P})\end{aligned}$$

We denote the optimal solution to (**P**) with $\mathbf{x}^*$ and the optimal objective value with $V^* = V(\mathbf{x}^*)$.

4.1. Supermodularity, Componentwise Convexity, and Multimodularity

In this section we unfold discrete convexity properties of the optimization problem. These properties guarantee that our search for an optimal schedule can be performed effectively and efficiently. In preparation for our results, we present the following definition.

Definition 1.

(a) A function $g : \mathbb{Z}_+^n \to \mathbb{R}$ is said to be *supermodular* if

$$g(\mathbf{x}) + g(\mathbf{u}) \le g(\mathbf{x} \wedge \mathbf{u}) + g(\mathbf{x} \vee \mathbf{u}) \text{ for all } \mathbf{x}, \mathbf{u} \in \mathbb{Z}_+^n,$$

where

$$\mathbf{x} \wedge \mathbf{u} = (\min(x_1, u_1), \min(x_2, u_2), \ldots., \min(x_n, u_n)) \qquad (12)$$

$$\text{and } \mathbf{x} \vee \mathbf{u} = (\max(x_1, u_1), \max(x_2, u_2), \ldots., \max(x_n, u_n)). \qquad (13)$$

Equivalently, $g : \mathbb{Z}^n_+ \to \mathbb{R}$ is said to be supermodular if

$$g(\mathbf{x} + \mathbf{e}^i_n + \mathbf{e}^j_n) - g(\mathbf{x} + \mathbf{e}^i_n) \geq g(\mathbf{x} + \mathbf{e}^j_n) - g(\mathbf{x}) \qquad (14)$$

for all $\mathbf{x} \in \mathbb{Z}^n_+$ and for all $1 \leq i < j \leq n$, where $\mathbf{e}^k_n \in \mathbb{Z}^n_+$ is the vector which has zeros everywhere, except in the k^{th} component where it is 1, $1 \leq k \leq n$. If $-g$ is supermodular, then g is said to be *submodular*.

(b) A function $g : \mathbb{Z}^n_+ \to \mathbb{R}$ is said to be *componentwise convex* if inequality (14) holds for all $\mathbf{x} \in \mathbb{Z}^n_+$ and for all $1 \leq i = j \leq n$.

(c) A function $g : \mathbb{Z}^n_+ \to \mathbb{R}$ is said to be *directionally convex* if it is supermodular and componentwise convex.

(d) A function $g : \mathbb{Z}^n_+ \to \mathbb{R}$ is said to be *multimodular* if

$$g(\mathbf{x} + \mathbf{u}) - g(\mathbf{x}) \geq g(\mathbf{x} + \mathbf{v} + \mathbf{u}) - g(\mathbf{x} + \mathbf{v}) \qquad (15)$$

for all $\mathbf{x} \in \mathbb{Z}^n_+$ and all $\mathbf{u} \neq \mathbf{v} \in \mathscr{E}_n$ such that $\mathbf{x} + \mathbf{u}, \mathbf{x} + \mathbf{v} \in \mathbb{Z}^n_+$, where

$$\mathscr{E}_n = \{-\mathbf{e}^1_n, \mathbf{e}^1_n - \mathbf{e}^2_n, \mathbf{e}^2_n - \mathbf{e}^3_n, \ldots, \mathbf{e}^{n-1}_n - \mathbf{e}^n_n, \mathbf{e}^n_n\}.$$

Multimodularity is a property stronger than directional convexity (see Altman et al. (2000) Lemma 2.2(b.iii)), which guarantees that a local optimum is also a global optimum. Murota (2005) provides the optimality criterion for minimizing a multimodular function over $\mathbb{Z}^n$, as well as how to obtain a neighbor of a vector $\mathbf{x}$ by adding an alternating sequence of positive and negative unit directions.

Theorem 1. (Murota 2005) *For a multimodular function* $g : \mathbb{Z}^n \to \mathbb{R}$ *we have*

$$g(\mathbf{x}) \leq g(\mathbf{y}) \text{ for all } \mathbf{y} \in \mathbb{Z}^n \iff g(\mathbf{x}) \leq g(\mathbf{x} \pm \mathbf{d}) \text{ for all } \mathbf{d} \in \mathscr{D}, \qquad (16)$$

where $\mathscr{D}$ *is the set of vectors of the form* $\mathbf{e}^{i_1}_n - \mathbf{e}^{i_2}_n + \ldots + (-1)^{k-1}\mathbf{e}^{i_k}_n$ *for some increasing sequence of indices* $1 \leq i_1 < i_2 < \ldots < i_k \leq n$.

The following lemma provides an equivalent characterization of a multimodular function, which we find to be more intuitive and facilitates a better exposition of our proofs.

Lemma 1. *A function* $g : \mathbb{Z}^n_+ \to \mathbb{R}$ *is multimodular if and only if the following four properties hold for all* $\mathbf{x} \in \mathbb{Z}^n_+$:

(i) $g(\mathbf{x} + \mathbf{e}^1_n + \mathbf{e}^n_n) - g(\mathbf{x} + \mathbf{e}^1_n) \geq g(\mathbf{x} + \mathbf{e}^n_n) - g(\mathbf{x})$;

(ii) $g(\mathbf{x} + \mathbf{e}^{i+1}_n + \mathbf{e}^j_n) - g(\mathbf{x} + \mathbf{e}^{i+1}_n + \mathbf{e}^{j+1}_n) \geq g(\mathbf{x} + \mathbf{e}^i_n + \mathbf{e}^j_n) - g(\mathbf{x} + \mathbf{e}^i_n + \mathbf{e}^{j+1}_n)$ $\forall i \neq j \in \{1, 2, \ldots, n-1\}$;

(iii) $g(\mathbf{x} + \mathbf{e}^1_n + \mathbf{e}^j_n) - g(\mathbf{x} + \mathbf{e}^1_n + \mathbf{e}^{j+1}_n) \geq g(\mathbf{x} + \mathbf{e}^j_n) - g(\mathbf{x} + \mathbf{e}^{j+1}_n)$ $\forall j \in \{1, 2, \ldots, n-1\}$;

(iv) $g(\mathbf{x} + \mathbf{e}^{i+1}_n + \mathbf{e}^n_n) - g(\mathbf{x} + \mathbf{e}^{i+1}_n) \geq g(\mathbf{x} + \mathbf{e}^i_n + \mathbf{e}^n_n) - g(\mathbf{x} + \mathbf{e}^i_n)$ $\forall i \in \{1, 2, \ldots, n-1\}$.

As a first step to explore discrete convexity properties of (**P**), we focus on the conditional expectation of the workload process given a realization of the arrival process. In other words, we isolate the variability stemming from service times.

Theorem 2.

(a) $\mathbb{E}[Z^{\mathbf{x}}_t | \mathbf{A}^{\mathbf{x}}_t = \mathbf{a}_t]$ *is directionally convex in* $\mathbf{a}_t$ *on* $\mathbb{Z}^t_+$ *for all* $t = 1, 2, \ldots, n$.

(b) $\mathbb{E}[O(\mathbf{x}) | \mathbf{A}^{\mathbf{x}} = \mathbf{a}]$ *is directionally convex in* $\mathbf{a}$ *on* $\mathbb{Z}^n_+$.

(c) $\mathbb{E}[I(\mathbf{x}) | \mathbf{A}^{\mathbf{x}} = \mathbf{a}]$ *is directionally convex in* $\mathbf{a}$ *on* $\mathbb{Z}^n_+$.

(d) $\mathbb{E}[W_s(\mathbf{x}) | \mathbf{S}^{\mathbf{x}} = \mathbf{s}, \mathbf{U} = \mathbf{u}]$ *is directionally convex in* $\mathbf{s}$ *on* $\mathbb{Z}^n_+$ *for all* $\mathbf{u}$ *in* $\mathbb{Z}^n_+$.

(e) $\mathbb{E}[W_u(\mathbf{x}) | \mathbf{S}^{\mathbf{x}} = \mathbf{s}, \mathbf{U} = \mathbf{u}]$ *is directionally convex in* $\mathbf{s}$ *on* $\mathbb{Z}^n_+$ *for all* $\mathbf{u}$ *in* $\mathbb{Z}^n_+$.

Theorem 2 serves as an intermediate step for investigating the existence of multimodularity in our performance measures. It provides us with insights into how to establish a proof/disproof of the more elaborate multimodularity property.

Theorem 3.

(a) $\mathbb{E}[Z^{\mathbf{x}}_t | \mathbf{A}^{\mathbf{x}}_t = \mathbf{a}_t]$ *is multimodular in* $\mathbf{a}_t$ *on* $\mathbb{Z}^t_+$ *for all* $t = 1, 2, \ldots, n$.

(b) $\mathbb{E}[O(\mathbf{x}) | \mathbf{A}^{\mathbf{x}} = \mathbf{a}]$ *is multimodular in* $\mathbf{a}$ *on* $\mathbb{Z}^n_+$.

(c) $\mathbb{E}[I(\mathbf{x}) | \mathbf{A}^{\mathbf{x}} = \mathbf{a}]$ *is multimodular in* $\mathbf{a}$ *on* $\mathbb{Z}^n_+$.

(d) $\mathbb{E}[W_s(\mathbf{x}) | \mathbf{S}^{\mathbf{x}} = \mathbf{s}, \mathbf{U} = \mathbf{u}]$ *is multimodular in* $\mathbf{s}$ *on* $\mathbb{Z}^n_+$ *for all* $\mathbf{u}$ *in* $\mathbb{Z}^n_+$.

(e) $\mathbb{E}[W_u(\mathbf{x}) | \mathbf{S}^{\mathbf{x}} = \mathbf{s}, \mathbf{U} = \mathbf{u}]$ *is multimodular in* $\mathbf{s}$ *on* $\mathbb{Z}^n_+$ *for all* $\mathbf{u}$ *in* $\mathbb{Z}^n_+$.

Theorem 3 implies that when patients deterministically do show up and on time for their scheduled appointment, then all performance measures of interest are multimodular in $\mathbf{x}$. This result serves as a stepping stone toward investigating the conditions on patient punctuality under which the multimodularity property holds.

Before we proceed with the analysis of the uncoditional performance measures of interest, we would like to contrast our results with Zacharias and Pinedo (2017), where they prove discrete convexity properties of a multiserver model with punctual patients and deterministic service times. The proofs of our Theorems 2 and 3 rely on sample path analyses, by considering realizations of the workload process given an arrival process; a technique similar to the one used to prove Theorems 2 and 3 in Zacharias and Pinedo (2017). However, our proofs contain some novel elements, which enable us to deal with general stochastic service times and heterogeneous patient groups (scheduled and unscheduled). In Zacharias and Pinedo (2017) the workload process can only take values that are integer multiples of one slot's duration, and their proofs rely on the argument that the additional patient from a unit perturbation "can be pushed toward the end of the queue of customers whenever the queue is nonempty (i.e., the service discipline does not affect the cost function, as long as the queue is work conserving)" (p. ec7). This argument cannot be applied in our setting since our workload process takes values on a continuous space and our timescale is more refined, and becuase of our treatment of the two separate patient groups. Moreover, we prove a result more general than the one in Zacharias and Pinedo (2017). We prove multimodularity of the *workload process* (Theorem 3(a)), based on which we prove multimodularity of all of the components of the objective function. Finally, each sample path analysis is complemented with a novel coupling argument, to establish comparisons of the marginal differences of the workload process with respect to the unit perturbations in Definition 1 (see, for example, Equation (EC.34)–(EC.36) in Lemma EC.3 and monotonicity properties in Lemma EC.4).

Theorem 4.

(a) $\mathbb{E}[Z_t^{\mathbf{x}}]$ *is directionally convex in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$ *for all* $t = 1, 2, \ldots, n$.

(b) $\mathbb{E}[O(\mathbf{x})]$ *is directionally convex in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$.

(c) $\mathbb{E}[I(\mathbf{x})]$ *is directionally convex in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$.

(d) $\mathbb{E}[W_s(\mathbf{x})]$ *is directionally convex in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$.

(e) $\mathbb{E}[W_u(\mathbf{x})]$ *is directionally convex in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$.

All performance measures of interest possess directional convexity under general distributions of patient punctuality, under any discrete-time arrival process from emergency walk-ins, and under general stochastic service times. Directional convexity implies that the components of a schedule are operational substitutes, since the marginal cost (benefit) from adding a patient in the schedule increases (decreases) as we add more patients into the schedule. Directional convexity alone, however, does not imply global optimality of locally optimal solutions, under any notion of locality established in the literature. To the best of our knowledge, there is no algorithm in the literature that guarantees to approximate the minimum of a directionally convex function to within any factor. As we demonstrate in what follows, the multimodularity property of the objective as a function of $\mathbf{x}$ only holds under additional assumptions on patient punctuality.

Definition 2.

(a) We say that *show-up patients are punctual* if

$$q_{t\tau} = \begin{cases} p_t & \text{if } \tau = t \\ 1 - p_t & \text{if } \tau = n+1 \\ 0 & \text{otherwise.} \end{cases}$$

(b) We say that *show-up probabilities are slot-homogeneous* if $p_t = p \in (0, 1]$ for all $t = 1, 2, \ldots, n$.

(c) *Timeliness* of a patient who is scheduled to arrive at slot t and shows up at slot τ is defined as the difference $\tau - t$, where positive timeliness corresponds to lateness, negative timeliness corresponds to earliness, and zero timeliness corresponds to promptness.

Theorem 5. *Assume that show-up probabilities are slot-homogeneous and that show-up patients are punctual. Then,*

(a) $\mathbb{E}[Z_t^{\mathbf{x}_t}]$ *is multimodular in* $\mathbf{x}_t$ *on* $\mathbb{Z}_+^t$ *for all* $t = 1, 2, \ldots, n$;

(b) $\mathbb{E}[O(\mathbf{x})]$ *is multimodular in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$;

(c) $\mathbb{E}[I(\mathbf{x})]$ *is multimodular in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$;

(d) $\mathbb{E}[W_s(\mathbf{x})]$ *is multimodular in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$;

(e) $\mathbb{E}[W_u(\mathbf{x})]$ *is multimodular in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$.

When patients promptly arrive for their scheduled appointments with the same show-up probability throughout the workday, then the objective function is multimodular. To the best of our knowledge, Theorem 5 is the first result in the literature to prove multimodularity of the objective function under general stochastic service times. However, as we demonstrate below, under a family of uniform distributions regarding the arrival process, the multimodularity property collapses. Under the following set of assumptions on the arrival process, some of the properties of the equivalent definition of a multimodular function from Lemma 1 are violated, while some others hold.

Assumption 1. *Assume the following setting regarding patient adherence to the schedule:*

(a) *Show-up probabilities are slot-homogeneous.*

(b) *All patients' timelinesses, given that they show up, are i.i.d. random variables uniformly distributed on* $\{-e, \ldots, -1, 0, 1, \ldots, l\}$, *where* e *(the maximum possible earliness) and* l *(the maximum possible lateness) are nonnegative integers such that* $l \le n - 3$ *and* $e + l \neq 0$.

(c) *Patients who arrive too early, before slot 1, do not receive service until the beginning of the first slot. Waiting costs are incurred only for the waiting time patients experienced during or after the first slot. Patients who arrive too late, beyond slot* n, *do not receive service at all and they are accounted as no-shows.*

Theorem 6. *Under Assumption 1,*

(a) $\mathbb{E}[O(\mathbf{x})]$ *is not multimodular in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$;

(b) $\mathbb{E}[I(\mathbf{x})]$ *is not multimodular in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$;

(c) $\mathbb{E}[W_s(\mathbf{x})]$ *is not multimodular in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$;

(d) $\mathbb{E}[W_u(\mathbf{x})]$ *is not multimodular in* $\mathbf{x}$ *on* $\mathbb{Z}_+^n$.

Nonpunctuality defeats the purpose of a schedule to a certain extent. A good appointment schedule aims to consistently match capacity and demand throughout the workday by properly managing patient arrivals. Patients not arriving on their designated appointment times make the existence of a good schedule less valuable.

4.2. Minimizing a Multimodular Function on $\mathbb{Z}_+^n$ via Sequential Local Search in Polynomial Time

In this section we demonstrate how to minimize efficiently a multimodular function over nonnegative integer vectors based on the results of Murota (2004) and Schrijver (2000).

Multimodular and L$^\natural$-convex functions are related through a unimodular coordinate transformation. In particular, let $g : \mathbb{Z}^n \to \mathbb{R}$ be a multimodular function. According to Lemma 2.1 in Murota (2005), the function $f : \mathbb{Z}^n \to \mathbb{R} : \mathbf{x} \mapsto g(B\mathbf{x})$ is L$^\natural$-convex, where $B = (b_{ij})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$ is a bidiagonal matrix with

$$b_{ij} = \begin{cases} 1 & \text{if } j = i \\ -1 & \text{if } j = i-1 \\ 0 & \text{otherwise.} \end{cases}$$

For the rest of this section, to avoid repetition, g will denote a multimodular function and f an L$^\natural$-convex function.

For a subset X of a ground set $N = \{1,2,\ldots,n\}$, let $\mathbf{e}_X = (e_{X1}, e_{X2}, \ldots, e_{Xn}) \in \{0,1\}^n$ be the characteristic vector of X—i.e., $e_{Xi} = \mathbb{1}_{i\in X}$ for all $i = 1,2,\ldots,n$. Murota (2004) provides the following algorithm that minimizes locally, and eventually globally, an L$^\natural$-convex function defined on $\mathbb{Z}^n$ in polynomial time.

Algorithm 1 (Steepest Descent Algorithm for an L$^\natural$-Convex Function f on $\mathbb{Z}^n$ (Murota 2004))

```
1: Pick an $\mathbf{x} \in \mathbb{Z}^n$.
2: Find $(\epsilon^*, X^*) = \arg\min\{f(\mathbf{x}+\epsilon\mathbf{e}_X) : \epsilon \in \{-1,1\}, X \subseteq N\}$.
3: If $f(\mathbf{x}) \le f(\mathbf{x}+\epsilon^*\mathbf{e}_{X^*})$, then stop ($\mathbf{x}$ is a minimizer of $f$).
4: Set $\mathbf{x} \leftarrow \mathbf{x}+\epsilon^*\mathbf{e}_{X^*}$ and go to Step 2.
```

Definition 3. For a finite ground set N, a set function $\rho : 2^N \to \mathbb{R}$ is said to be *submodular* if

$$\rho(U \cup u) - \rho(U) \ge \rho(V \cup u) - \rho(V) \qquad (17)$$

for all $U \subseteq V \subseteq N$ and $u \in N\backslash V$.

Step 2 in Algorithm 1 relies on unconstrained minimization of two submodular set functions

$$\min\{\rho^+_{\mathbf{x}}(X) \triangleq f(\mathbf{x}+\mathbf{e}_X) : X \subseteq N\}$$
$$\text{and } \min\{\rho^-_{\mathbf{x}}(X) \triangleq f(\mathbf{x}-\mathbf{e}_X) : X \subseteq N\};$$

a task that involves $\mathcal{O}(n^5)$ function evaluations and $\mathcal{O}(n^6)$ arithmetic operations by implementing the algorithm of Orlin (2009), and can be readily adjusted to minimize a multimodular function over $\mathbb{Z}^n$. However, minimizing a multimodular function g on $\mathbb{Z}^n_+$ requires a more careful treatment. Note that the optimization problem

$$\min\{g(\mathbf{x}) : \mathbf{x} \in \mathbb{Z}^n_+\}$$

is equivalent to

$$\min\{f(\mathbf{y}) \triangleq g(B\mathbf{y}) : B\mathbf{y} \in \mathbb{Z}^n_+\},$$

which in turn can be written as

$$\min\{f(\mathbf{y}) : 0 \le y_1 \le y_2 \ldots \le y_n, \mathbf{y} \in \mathbb{Z}^n_+\},$$

since for every $\mathbf{x} \in \mathbb{Z}^n_+$ there exists a unique $\mathbf{y}$ such that $B\mathbf{y} = \mathbf{x}$, and vice versa. In particular, $B\mathbf{y} = \mathbf{x}$ if and only if $\mathbf{y} = (x_1, x_1+x_2, x_1+x_2+x_3, \ldots, x_1+\ldots+x_n)$. Therefore, the problem of minimizing a multimodular function over nonnegative integer vectors can be translated into constrained minimization of an L$^\natural$-convex function.

We adjust the algorithm of Murota (2004) to minimize efficiently a multimodular function g on $\mathbb{Z}^n_+$. Let $\mathcal{A} = \{\mathbf{y} \in \mathbb{Z}^n_+ : 0 \le y_1 \le y_2 \cdots \le y_n\}$. It is straightforward to verify that the partially ordered set $(\mathcal{A}, \le)$, with meet and join as defined in (12) and (13), respectively, is a distributive lattice but not a finite one (the reader is referred to Chapter I of Fujishige (2005) and Chapter 5 of Davey and Priestley (2002) for the mathematical preliminaries on lattices and orders). Based on the definition of L$^\natural$-convexity (see Murota (2005)), f is submodular on $\mathcal{A}$. Due to Birkhoff's representation theorem (see Birkhoff (1937)), a finite *integer* lattice can be represented as a finite *set* lattice, with set inclusion $\subseteq$ as the partial order, and set intersection $\cap$ and set union $\cup$ as meet and join, respectively. Consequently, the problem of minimizing a submodular function defined over a finite integer lattice can be solved in polynomial time (with respect to $|\mathcal{A}|$). The distributive lattice $\mathcal{A}$ is not finite. Even if we force $\mathcal{A}$ to be finite (by imposing for example an upper bound on the number of patients in the schedule), the size of $\mathcal{A}$ grows exponentially with the number of decision variables n. However, we can exploit the fact that a local minimum is also a global minimum for an L$^\natural$-convex function, to minimize f sequentially over finite set lattices in polynomial time (with respect to n), until a local minimum is identified. Below we expose in detail how this procedure works.

Algorithm 2 (Steepest Descent Algorithm for a Multimodular Function g on $\mathbb{Z}^n_+$)

```
1: Define $f : \mathcal{A} \to \mathbb{R} : \mathbf{x} \mapsto g(B\mathbf{x})$.
2: Pick an $\mathbf{x} \in \mathcal{A}$.
3: Find $(\epsilon^*, X^*) = \arg\min\{f(\mathbf{x}+\epsilon\mathbf{e}_X) : \epsilon \in \{-1,1\}, X \subseteq N, \mathbf{x}+\epsilon\mathbf{e}_X \in \mathcal{A}\}$.
4: If $f(\mathbf{x}) \le f(\mathbf{x}+\epsilon^*\mathbf{e}_{X^*})$, then stop ($\mathbf{x}$ is a minimizer of $f$ and $B\mathbf{x}$ is a minimizer of $g$).
5: Set $\mathbf{x} \leftarrow \mathbf{x}+\epsilon^*\mathbf{e}_{X^*}$ and go to Step 3.
```

Step 3 in Algorithm 2 involves the *constrained* minimization of two submodular set functions

$$\begin{aligned} \min \quad & \rho^+_{\mathbf{x}}(X) \\ \text{s.t.} \quad & X \subseteq N \\ & \mathbf{x}+\mathbf{e}_X \in \mathcal{A} \end{aligned} \qquad (\mathbf{P}^+_{\mathbf{x}})$$

and

$$\begin{aligned} \min \quad & \rho^-_{\mathbf{x}}(X) \\ \text{s.t.} \quad & X \subseteq N \\ & \mathbf{x}-\mathbf{e}_X \in \mathcal{A}. \end{aligned} \qquad (\mathbf{P}^-_{\mathbf{x}})$$

In general, exact solution to constrained (e.g., cardinality constraints) minimization of submodular set function is NP-hard (Svitkina and Fleischer 2011). In Online Appendix B we demonstrate how to transform ($\mathbf{P}^{+}_{\mathbf{x}}$) and ($\mathbf{P}^{-}_{\mathbf{x}}$) into unconstrained minimization of submodular set-functions based on the results of Schrijver (2000).

Theorem 7. *Problems* ($\mathbf{P}^{+}_{\mathbf{x}}$) *and* ($\mathbf{P}^{-}_{\mathbf{x}}$) *can be solved in polynomial time* via *unconstrained submodular set-function minimization for all* $\mathbf{x} \in \mathcal{A}$.

5. Computational Experiments

Our transient analysis in Section 3 and theoretical investigation of discrete convexity properties in Section 4.1 were carried out based on the assumption of general stochastic service times (either continuous, discrete, or a mixture). For the case of a continuous service time distribution (beyond the exponential) or a mixture, the computation of the recursive workload distribution in (4) requires integration with both symbolic and numerical methods; a task which is computationally slow, and/or inaccurate, and/or not feasible. Begen et al. (2012) also discuss the limitations and computational difficulties in the evaluation of the objective function under continuous distributions. In our computational experiments we consider stochastic service times with discrete supports. In Online Appendix A we demonstrate how to simplify the analysis in Section 3 for such a purely discrete setting.

We consider a Beta-Binomial family of distributions; a three-parameter family of distributions with finite and discrete supports. When the service time distribution is $R \sim \text{BetaBin}(m, \alpha, \beta)$, then R has a support on $\{0, 1, 2, \ldots, m\}$, a probability mass function

$$f_R(r) = \binom{m}{r} \frac{\mathrm{B}(r + \alpha, m - s + \beta)}{\mathrm{B}(\alpha, \beta)},$$

mean

$$\mathbb{E}[R] = \mu = \frac{m\alpha}{\alpha + \beta},$$

and variance

$$\text{Var}[R] = \sigma^2 = \frac{m\alpha\beta(m + \alpha + \beta)}{(\alpha + \beta)^2(\alpha + \beta + 1)},$$

where

$$\mathrm{B}(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1} \ \mathrm{d}t$$

is the beta function. As shown in Figure 2, a Beta-Binomial distribution with fixed support on $\{0, 1, 2, \ldots, m\}$ and fixed mean μ may take a variety of shapes depending on its coefficient of variation $\text{CoV}(R) = \sigma\mu^{-1}$. This is a useful feature when it comes to investigating the impact of service time variability on the optimal scheduling strategies. Another useful feature of the Beta-Binomial family of distributions is its finite supports, which eliminate the need for numerical truncation, and thus reduce the computational error.

First, in Section 5.1–Section 5.4, we consider the setting where no-show probabilities are slot-homogeneous and all patients are punctual when they show up. Under these assumptions, from Theorem 5, the objective function is multimodular and therefore our Algorithm 2 terminates with an optimal schedule. Local search for the best neighbor of a vector $\mathbf{x}$ in step 3 of Algorithm 2 is conducted in polynomial time based on Theorem 7. In particular, the constrained problems ($\mathbf{P}^{+}_{\mathbf{x}}$) and ($\mathbf{P}^{-}_{\mathbf{x}}$) are solved via unconstrained submodular set-function minimization, according to the transformations in (EC.2) and (EC.3) appearing in Online Appendix B. For unconstrained submodular set-function minimization we used the *minimum norm point algorithm* of Fujishige (2005), as implemented in sfo_min_norm_point.m of Krause (2010) with an arithmetic tolerance of 10^{-5}. In Section 5.5 we consider the setting where patients are not necessarily punctual and we develop heuristic solutions.

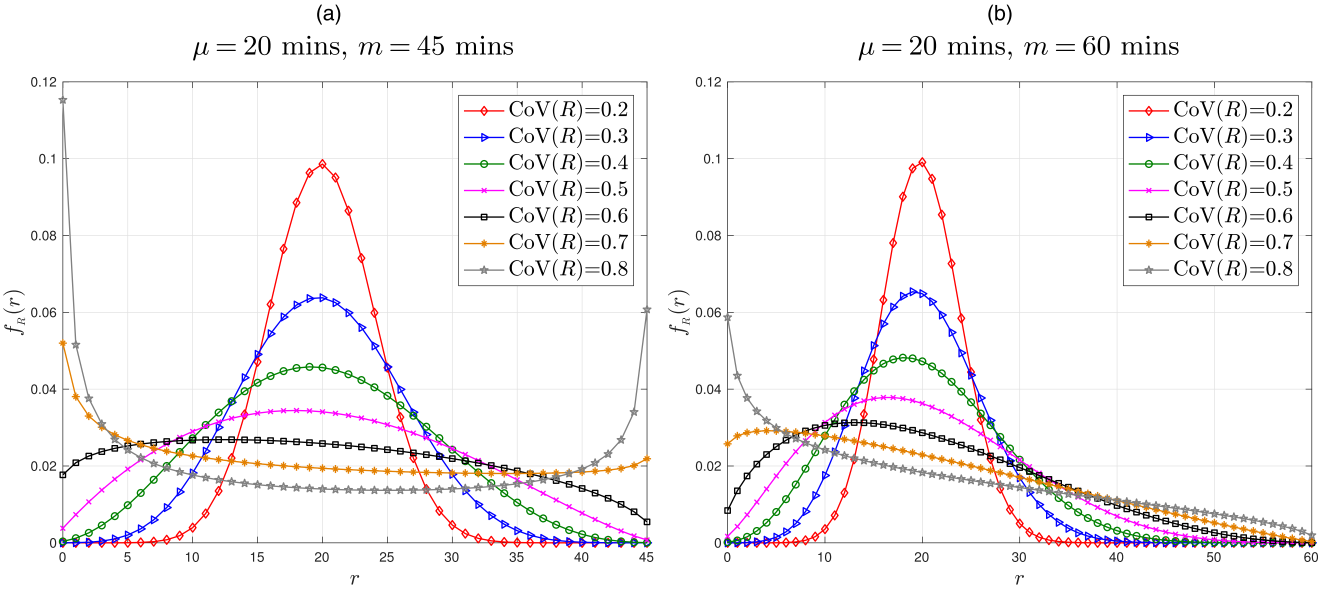
In what follows we display and discuss the optimal scheduling strategies as a function of the various model inputs: $\mathbf{U}$, c_i, c_o, c_s, c_u, p, μ, σ^2, m, T, d (and consequently $n = T/d$). In our experiments c_i is normalized to 1 and we consider an 8-hour workday from 9:00 am to 5:00 pm. In preparation for our analysis and discussions, we introduce the following partial order on $\mathbb{Z}^n_+$ regarding the comparisons of two schedules.

Definition 4. Let $\mathbf{x}, \hat{\mathbf{x}} \in \mathbb{Z}^n_+$. We say that $\mathbf{x}$ is smaller than $\hat{\mathbf{x}}$ in the *front-loaded* order, denoted as $\mathbf{x} \leq_{\text{fl}} \hat{\mathbf{x}}$, iff $\sum_{\tau=1}^{t} x_\tau \leq \sum_{\tau=1}^{t} \hat{x}_\tau$ for all $t = 1, 2, \ldots, n$.

It is straightforward to verify that $\leq_{\text{fl}}$ is indeed a partial order (reflexive, antisymmetric, and transitive) on $\mathbb{Z}^n_+$. We note that the front-loaded order is related to the *majorization order* (see Marshall and Olkin 1979), one main difference being that the latter only applies to vectors of the same sum. Our comparisons based on $\leq_{\text{fl}}$ in what follows are grounded on computational experiments, not on theoretical findings.

Another remark, before we discuss our computational experiments, is that it is often optimal to overbook certain parts of the schedule to hedge against probable patient no-shows, overtime workload, and potentially other factors. Overbooking might come in two types: (i) by double-booking certain slots or (ii) by scheduling certain interarrival times to be less than the average service time. Type (ii) of overbooking is only feasible when we refine the timescale of the schedule so that $d < \mu$—that is, when a slot's duration is less than the average service time.

Figure 2. (Color online) Shapes of Beta-Binomial Distribution with Fixed Mean and for Different Values of CoV



5.1. Impact of Objective Coefficients

Figures 3 and 4 demonstrate the impact of the objective coefficients c_s and c_o on the optimal schedule. First, we place the focus on Figure 3, where the average service time is equal to one slot's duration. One intuitive observation is that as either or both coefficients increase, the optimal objective value increases and the expected patient throughput decreases. Moreover, the optimal schedules become less and less front-loaded as c_s increases, to the point that we might have to compromise patient throughput to contain patients' waiting times.

Interestingly, however, sacrificing patient throughput as c_s increases does not necessarily come along with decreased overtime costs. For example, the optimal schedule in Figure 3(g) has the last slot empty, to absorb any unfinished workload left toward the end of the workday and to avoid high overtime costs. When we double the waiting cost coefficient from 0.05 to 0.10, we observe in Figure 3(h) that the optimal schedule becomes less front-loaded and patient throughput decreases, while overtime increases.

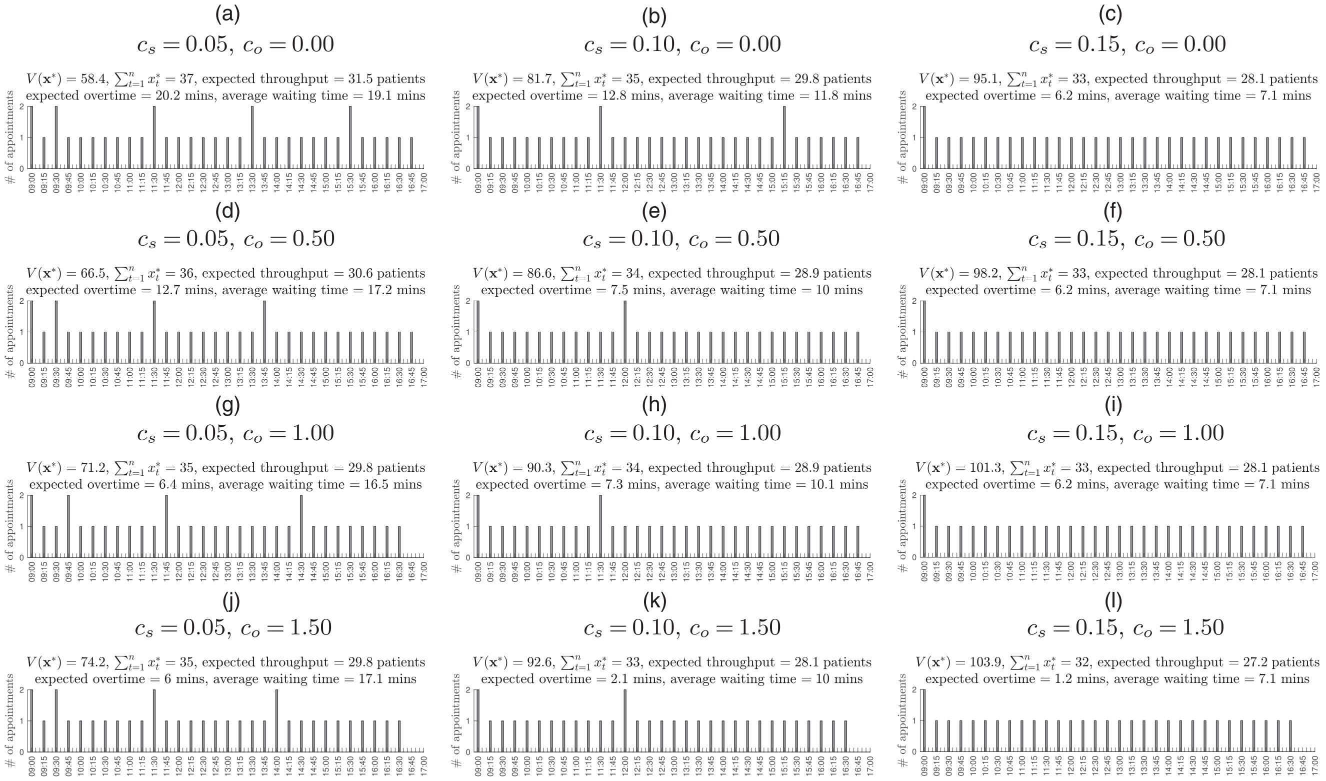
In contrast to the effect of c_s, we do not observe monotonicity properties with respect to the front-loaded order as c_o increases. In Figure 3(b), when the overtime cost is set to zero, the optimal schedule has three double-booked slots spread out throughout the workday to offset a 15% no-show rate. When we increase c_o by 0.5 in Figure 3(e), only two slots are double-booked (to avoid high overtime costs) and the optimal schedule becomes less front-loaded. However, when we double c_o in Figure 3(h), the optimal schedule becomes more front-loaded, while we maintain the same patient throughput, and as a result overtime decreases at the expense of an increase in the average waiting time.

In Figure 4 the average service time is 30 minutes, double the slot's duration, and both types of overbooking appear throughout the workday. The conclusions discussed above regarding Figure 3 still hold in this setting. When contrasting Figures 3 and 4 we observe that, in optimality, longer service times come along with longer waiting times and overtime workload. Furthermore, in Figure 4, double-booking only occurs on the first slot, so that we get an empty system up and running. The rest of the optimal schedule, beyond the first slot, consists of a sequence of scheduled interarrival times that are equal to either 30 minutes (with no overbooking) or, occasionally, 15 minutes. The lack of a clear repetitive pattern in the optimal schedule may be attributed to two reasons: (i) 8 hours is not enough time for the system to reach some sort of steady state or (ii) the discrete nature of the problem offers limited flexibility to consistently match capacity with scheduled appointments throughout the workday.

5.2. Impact of Service Time Variability and No-show Rate

In Figure 5 we demonstrate the impact of variability stemming from service times and no-shows. One clear observation is that as either or both service time variability and no-show rate increase, the optimal objective value increases as well; as the operational environment becomes more and more stochastic, it becomes harder to strike the right balance between efficient resource utilization and contained waiting times. The latter remark suggests that outpatient clinics

Figure 3. Optimal Schedules as a Function of c_s and c_o with $\mu = d$



Note. $T = 8$ hours, $p = 0.85$, $d = 15$ minutes, $R \sim \text{BetaBin}(45, \alpha, \beta)$ with α and β such that $\mu = 15$ minutes and $\sigma\mu^{-1} = 0.3$, $\mathbf{U} = \mathbf{0}$.

should take measures to contain variability emerging from these two sources to better achieve their operational goals.

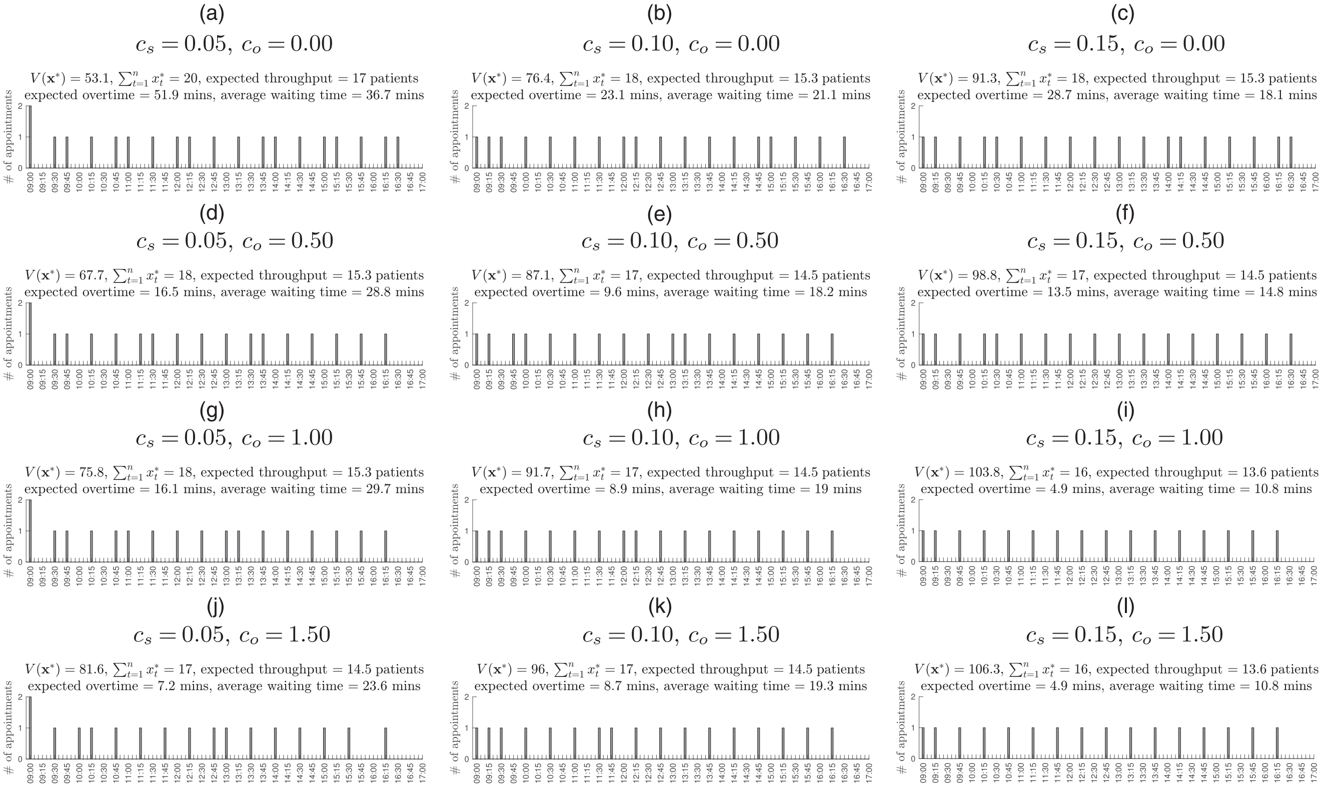
Intuitively, as the no-show rate decreases, the optimal schedules become less front-loaded. Interestingly, however, patient throughput does not necessarily improve as the no-show rate decreases: as we transition from Figure 5(g) with a no-show rate of 15% to Figure 5(j) with a no-show rate of 10%, we slightly compromise patient throughput to the benefit of achieving decreased overtime and waiting time. In contrast to the effect of the no-show rate, we do not observe monotonicity properties with respect to the front-loaded order as σ^2 increases. For example, if we take a careful look at the first row of Figure 5, the optimal schedule in (b) with a CoV(R) = 0.4 is more front-loaded when compared with both schedules from (a) and (c) with CoV(R) = 0.3 and CoV(R) = 0.5, respectively.

Another observation worth noting, probably one that bears limited managerial insight yet stimulates academic curiosity, is the following. The set of nonnegative integer vectors $\mathbb{Z}_+^n$ endowed with the partial order $\leq_{\text{fl}}$ is not a totally ordered set—that is, if we pick any two nonnegative integer vectors they are not necessarily comparable. However, for either a fixed p or a fixed CoV(R), the resulting subset of schedules from Figure 5 is in fact a totally ordered set. As discussed above, monotonicity properties might not necessarily hold, but every pair of schedules is comparable with respect to $\leq_{\text{fl}}$. This observation holds also for Figures 3 and 4 for either a fixed c_o or a fixed c_s. That being the case, we reckon that $\leq_{\text{fl}}$ is a more meaningful tool to contrast different appointment schedules, when compared with the standard order $\leq$ on $\mathbb{Z}_+^n$.

5.3. Impact of Timescale

In the literature (e.g., in Robinson and Chen 2010, LaGanga and Lawrence 2012, Zacharias and Pinedo 2017, and Wang et al. 2019) it is often assumed that a slot's duration d is equal to the average service time μ, and therefore patients are scheduled to arrive only at integer multiples of μ. In our next experiment we place the focus on the impact of a schedule's timescale d. Consider for example the setting where the average service time is equal to 30 minutes. Should a clinic offer appointment times at 9:00 am, 9:30 am, 10:00 am, and so forth? Is there any benefit in refining the timescale of a schedule so that appointment times are potentially offered at 9:00 am, 9:15 am, 9:30 am, and so forth? An appointment schedule in the former setting is also a feasible schedule in the latter, but not the other way around. Therefore, in the latter setting we expect to obtain an optimal schedule at least as good as the one in

Figure 4. Optimal Schedules as a Function of c_s and c_o with $\mu > d$



Note. $T = 8$ hours, $p = 0.85$, $d = 15$, $R \sim \text{BetaBin}(90, \alpha, \beta)$ with α and β such that $\mu = 30$ minutes and $\sigma\mu^{-1} = 0.3$, $\mathbf{U} = \mathbf{0}$.

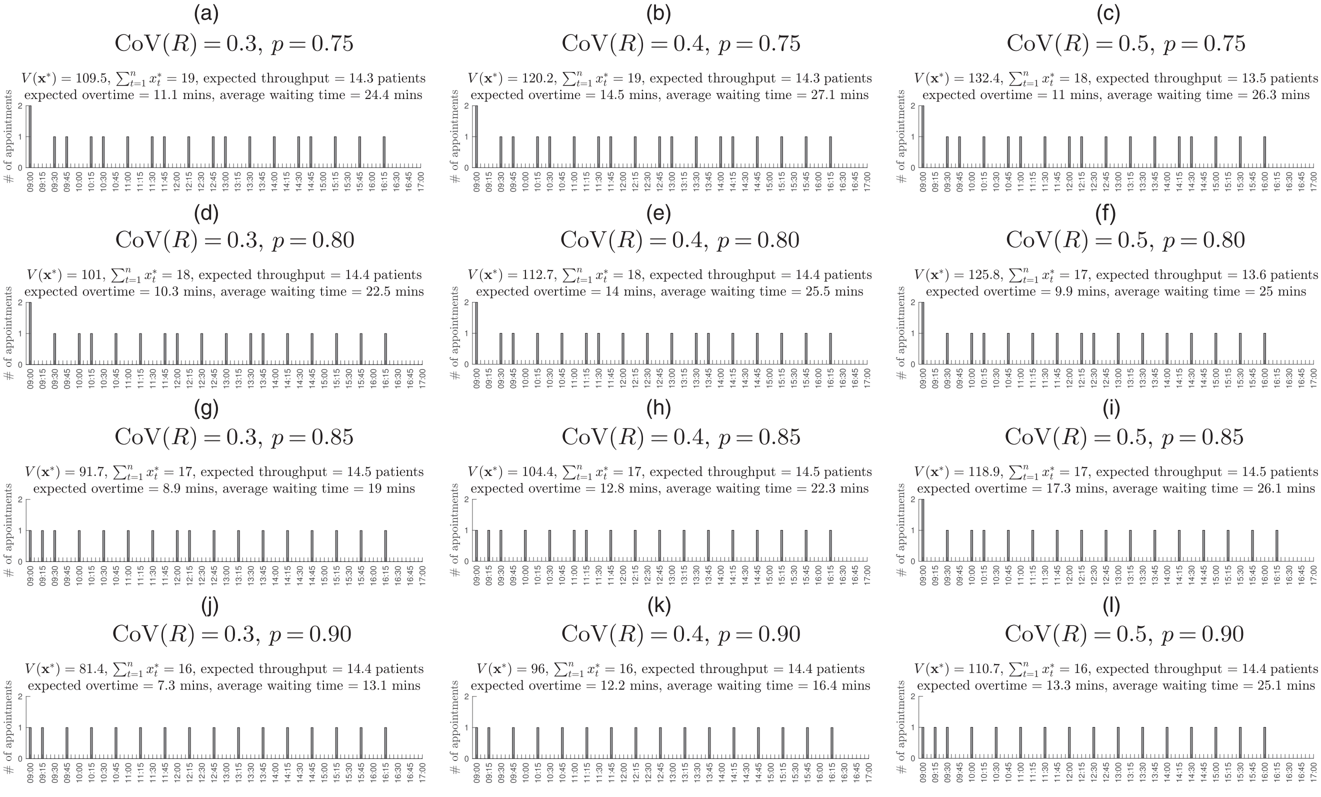
the former setting. On the other hand, a more refined timescale comes along with more decision variables, something that increases the computational complexity of the problem. Is there a significant benefit from sacrificing computational complexity to achieve better operational outcomes?

In Figure 6 the average service time μ is 30 minutes and we display the optimal schedules for $d = 30, 20, 15, 10, 5$ minutes and for different no-show probabilities. Figure 7 shows the percent improvement in the optimal objective value relative to the one when $d = \mu$. We observe that, indeed, when we allow more flexibility in our scheduling decisions by letting d be a proper divisor of μ, the optimal objective value improves. For example, lets take a look at the middle column of Figure 6 (or the blue dotted line in Figure 7). When we double the decision variables from 16 to 32 as we transition from Figure 6(b) to 6(h), the improvement in the optimal objective value is 3.4%. When we triple the decision variables from 16 to 48 as we transition from Figure 6(b) to 6(k), the improvement is 3.8%. When we sextuple the decision variables from 16 to 96 from Figure 6(b) to 6(n), the improvement is 4.3%. Moreover, a more refined timescale might result in higher patient throughput in optimality, as the option of letting d be less than μ provides the flexibility to strategically fit an extra patient into the schedule (compare Figure 6(b) with the rest of the column). Next, we place the focus on the special case where d is not a proper divisor of $\mu = 30$. When $d = 20$, the objective value in fact worsens when the no-show rate is 10%, while the objective value improves for higher no-show rates. This suggests that outpatient clinics can indeed benefit form a more refined timescale in their scheduling decisions when d is a proper divisor of μ, while otherwise we might end up with a worse schedule. Finally, when d is a proper divisor of μ, there is evidence of decreasing marginal differences in the objective as the number of variables n increases.

5.4. Impact of Unscheduled Emergency Walk-ins

In our next computational experiment we investigate the impact of emergency walk-ins on the optimal scheduling strategy. We consider a setting where the average service time μ is 30 minutes, one slot's duration d is 15 minutes, and one unscheduled emergency patient may arrive at each one of the 32 slots with some probability p_u (in agreement with Green et al. 2006). In particular, we let $U_t \sim \text{Bernoulli}(p_u)$ for all $t = 1, 2, \ldots, n$, where $p_u \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. The extreme case where $p_u = 0.5$ corresponds to a scenario were the expected number of all unscheduled emergency walk-ins is equal to the daily capacity to serve patients.

Figure 5. Optimal Schedules as a Function of CoV(R) and p


(a) $\mathrm{CoV}(R) = 0.3,\ p = 0.75$

(b) $\mathrm{CoV}(R) = 0.4,\ p = 0.75$

(c) $\mathrm{CoV}(R) = 0.5,\ p = 0.75$

(d) $\mathrm{CoV}(R) = 0.3,\ p = 0.80$

(e) $\mathrm{CoV}(R) = 0.4,\ p = 0.80$

(f) $\mathrm{CoV}(R) = 0.5,\ p = 0.80$

(g) $\mathrm{CoV}(R) = 0.3,\ p = 0.85$

(h) $\mathrm{CoV}(R) = 0.4,\ p = 0.85$

(i) $\mathrm{CoV}(R) = 0.5,\ p = 0.85$

(j) $\mathrm{CoV}(R) = 0.3,\ p = 0.90$

(k) $\mathrm{CoV}(R) = 0.4,\ p = 0.90$

(l) $\mathrm{CoV}(R) = 0.5,\ p = 0.90$

Note. $T = 8$ hours, $c_o = 1$, $c_s = 0.1$, $d = 15$ minutes, $R \sim \mathrm{BetaBin}(90, \alpha, \beta)$ with α and β such that $\mu = 30$ minutes, $\mathbf{U} = \mathbf{0}$.

Figure 8 displays the optimal schedules for a combination of no-show rates and walk-in rates. We do observe that as the walk-in rate increases, the optimal schedules become less front-loaded, to the point that only the first slot is booked with one scheduled patient, and essentially the outpatient clinic operates under an open access regime (see Green and Savin 2008, Liu and Ziya 2014, Izady 2015, and Zacharias and Armony 2017). Moreover, the higher the walk-in rate, the less control we have over the arrival process and inevitably the optimal objective value increases as well.

Next, we place the focus on the upper right corner of Figure 8, where there is no variability in the arrival process (since $p = 1.0$ and $p_u = 0$). Because of service time variability, the first two patients are scheduled to arrive 15 minutes apart. For the rest of the day, patients are scheduled to arrive deterministically every 30 minutes with two gaps of 15 minutes in the middle and toward the end of the workday, and the last two slots remain empty. As a result, the patient throughput is 15 patients, falling one patient short of the clinic's daily capacity. Even in an ideal scenario where there is no randomness in the arrival process, outpatient clinics might have to sacrifice some of their patient throughput to hedge against service time variability and its impact on waiting times and overtime workload.

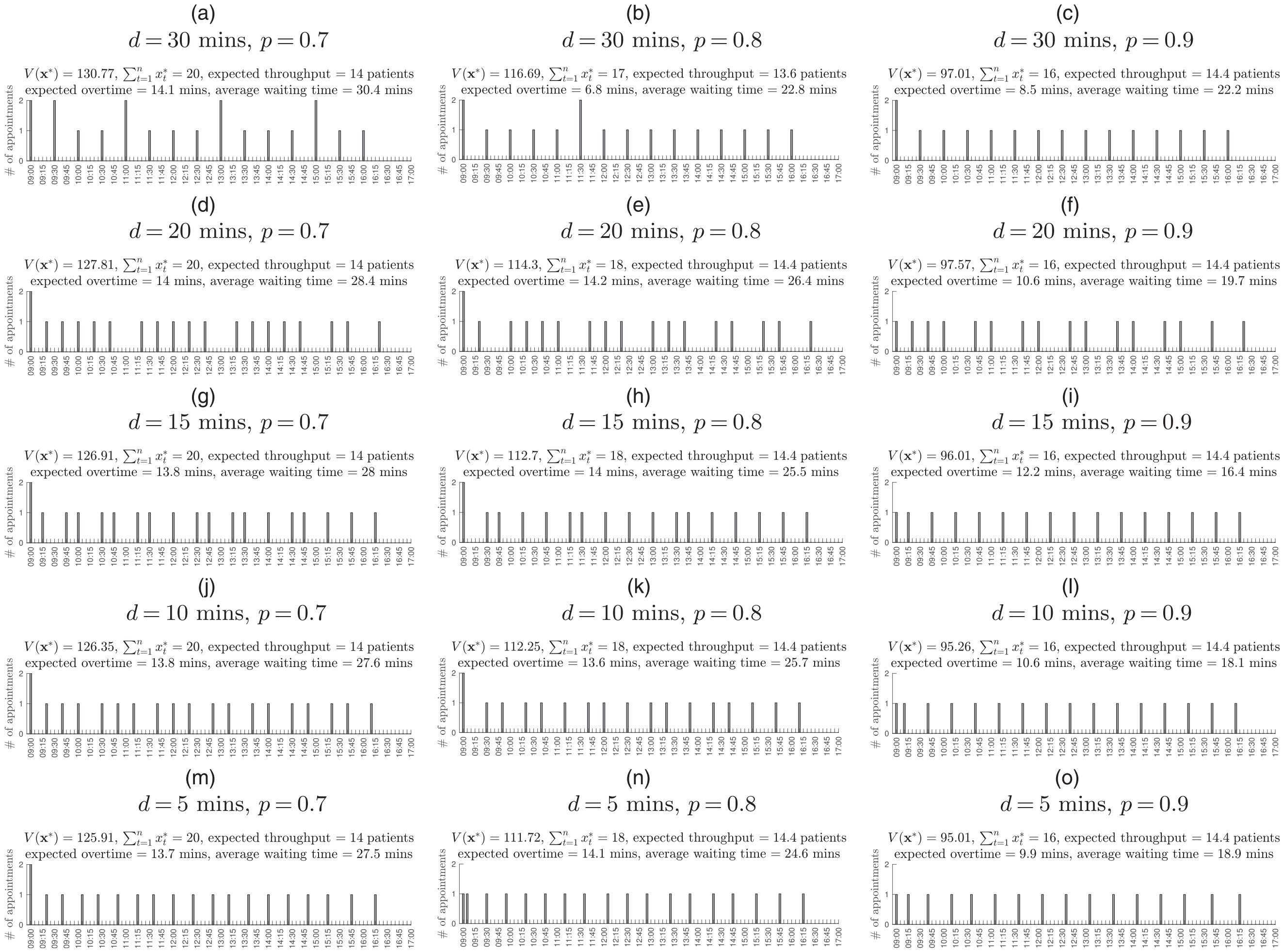
5.5. Impact of Nonpunctuality

The computational analyses in Sections 5.1–5.4 were carried out under the assumption of punctual patients. In our last experiment we investigate the impact of nonpunctuality on the optimal scheduling strategy. We know from Theorem 4 that the objective function is directionally convex under general conditions. However, under the assumptions of Theorem 6, the objective function is not multimodular. To the best of our knowledge, there is no algorithm in the literature that guarantees to approximate the minimum of a directionally convex function to within any factor. Moreover, objective function evaluations have exponential complexity under the assumption of nonpunctual patients (as discussed in Section 3.2). We propose the following two heuristic procedures:

(i) **Exhaustive local search (els):** The first heuristic incorporates nonpunctuality in the objective function evaluations and performs exhaustive local search (locality as defined in Theorem 1). It terminates with a schedule $\mathbf{x}^{els}$ that is optimal within its exponentially large neighborhood.

(ii) **Submodular set-function minimization (ssm):** The second heuristic assumes that patients are punctual (and therefore the objective function is assumed to be multimodular), and identifies a locally optimal schedule $\mathbf{x}^{ssm}$ in polynomial time via Algorithm 2.

Figure 6. Optimal Schedules as a Function of d and p



(a) $d = 30$ mins, $p = 0.7$

(b) $d = 30$ mins, $p = 0.8$

(c) $d = 30$ mins, $p = 0.9$

(d) $d = 20$ mins, $p = 0.7$

(e) $d = 20$ mins, $p = 0.8$

(f) $d = 20$ mins, $p = 0.9$

(g) $d = 15$ mins, $p = 0.7$

(h) $d = 15$ mins, $p = 0.8$

(i) $d = 15$ mins, $p = 0.9$

(j) $d = 10$ mins, $p = 0.7$

(k) $d = 10$ mins, $p = 0.8$

(l) $d = 10$ mins, $p = 0.9$

(m) $d = 5$ mins, $p = 0.7$

(n) $d = 5$ mins, $p = 0.8$

(o) $d = 5$ mins, $p = 0.9$

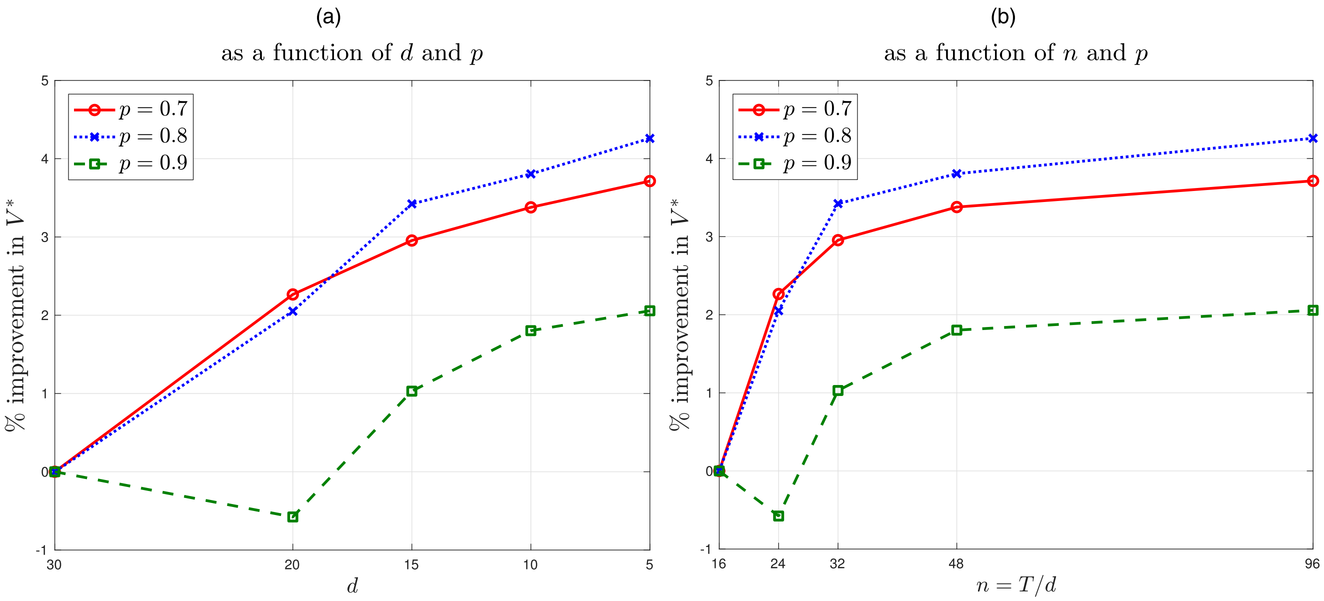
Note. $T = 8$ hours, $c_o = 1$, $c_s = 0.1$, $R \sim \text{BetaBin}(90, \alpha, \beta)$ with α and β such that $\mu = 30$ minutes and $\sigma\mu^{-1} = 0.4$, $\mathbf{U} = \mathbf{0}$.

In Figure 9 we contrast $\mathbf{x}^{els}$ and $\mathbf{x}^{ssm}$ under the assumptions of Theorem 6 with $d = 15$ and $e = l = 1$ (i.e., patients might arrive either on time, 15 minutes early, 15 minutes late, or not show up at all) and for different values of CoV(R). We observe that $\mathbf{x}^{els}$ outperforms $\mathbf{x}^{ssm}$ by 4.5%, 6.3%, and 3.1% when CoV(R) is equal to 0.3, 0.4, and 0.5, respectively. These percent cost differences can be viewed as a lower bound for the "value" of explicitly accounting for patient nonpunctuality in our scheduling decisions. We also observe that $\mathbf{x}^{els}$ is more front-loaded than $\mathbf{x}^{ssm}$. Even though our analysis is not comprehensive (due to the computational challenges involved), it provides evidence that nonpunctuality has a significant impact on the optimal scheduling strategy and should be taken under consideration.

6. Conclusion

We address the problem of designing optimal appointment scheduling strategies so that outpatient clinics can utilize their resources efficiently, while containing patients' waiting times, in a stochastic environment. Our stochastic model provides a unifying platform for addressing the well-studied appointment scheduling problem, as many models in the literature can be considered as its special cases. To the best of our knowledge, this is the first study to unfold discrete convexity properties of the static appointment scheduling problem under general stochastic service times and/or nonpunctual patients. We prove that under general conditions the objective function is directionally convex. While directional convexity deems the components of a schedule to be operational substitutes, it is not enough to ensure that the problem can be solved to exact optimality. We prove that the objective function is multimodular when patients promptly arrive for their scheduled appointments with the same show-up probability, and we identify conditions under which the multimodularity property collapses. This study is the first to develop and implement an algorithm for minimizing locally (and eventually globally) a multimodular function over nonnegative integer

Figure 7. (Color online) Percent Improvement in the Optimal Objective Value When $d \leq \mu$



(a)

as a function of d and p

(b)

as a function of n and p

Notes. $T = 8$ hours, $c_o = 1$, $c_s = 0.1$, $R \sim \text{BetaBin}(90, \alpha, \beta)$ with α and β such that $\mu = 30$ minutes and $\sigma\mu^{-1} = 0.4$, $\mathbf{U} = \mathbf{0}$. Percent improvement is calculated relative to the optimal objective value when $d = \mu$.

vectors in polynomial time; a task that bridges recent advances in discrete convex analysis and submodular set-function minimization over ring families.

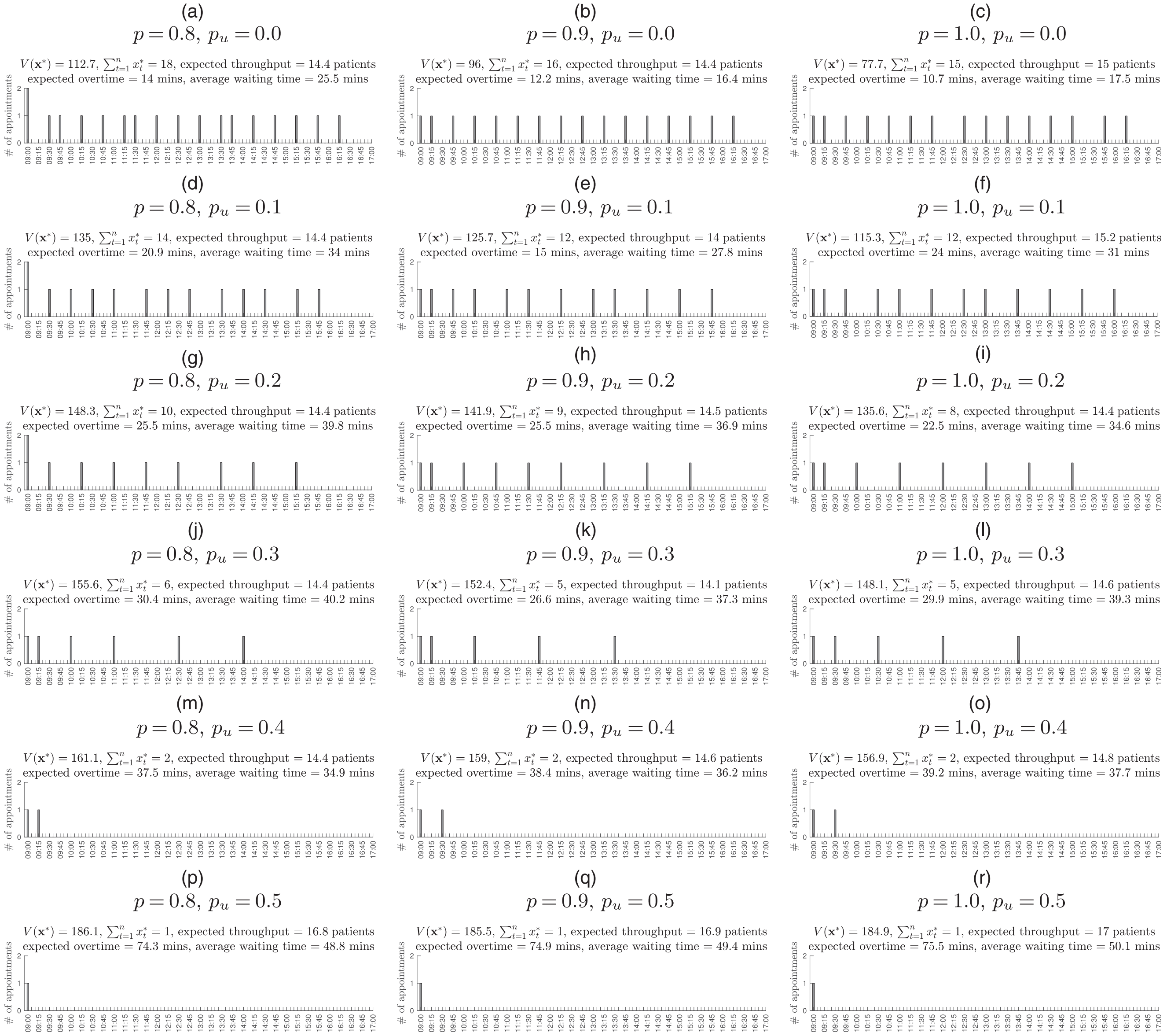
6.1. Managerial Implications and Practical Insights

Below we summarize some insights from our analysis for practical implementation. The reader is referred to Section 5 for a more thorough discussion.

- **Addressing variability in outpatient processes:** Increased variability stemming from all sources (emergency walk-ins, consultation times, no-shows) is associated with worsened operational outcomes, as measured by a weighted sum of the typical performance measures (waiting time, overtime, and idleness). As the operational environment becomes more and more stochastic, it becomes harder to strike the right balance between efficient resource utilization and short waiting times. In a highly stochastic environment, optimal schedules tend to be front-loaded and often with the last slot empty, to hedge against probable patient no-shows and overtime workload. Our model, by properly balancing the involved tradeoffs, can provide precise scheduling guidelines tailored to stochastic clinical environments. On a broader level, we demonstrate that outpatient clinics should take into account all sources of variability in their scheduling decisions and take measures to contain this variability to better achieve their operational goals. For example, our computational analysis indicates that, even in an ideal scenario where the no-show rate is zero and there are no walk-ins, outpatient clinics might have to sacrifice some of their patient throughput to handle service time variability and its impact on waiting times and overtime workload.
- **Choosing the right timescale:** It is common for schedulers to use a slot duration (or timescale) d roughly equal to the average consultation time μ; is this practice ideal? Consider for example the setting where the average consultation time is equal to 30 minutes. Should a clinic offer appointment times at 9:00 am, 9:30 am, 10:00 am, and so forth? Is there a significant benefit in refining the timescale of a schedule so that appointment times are potentially offered at 9:00 am, 9:15 am, 9:30 am, and so forth? An appointment schedule in the former setting is also a feasible schedule in the latter, but not the other way around. Therefore, in the latter setting we expect to obtain an optimal schedule at least as good as the one in the former setting. On the other hand, a more refined timescale comes along with more decision variables, something that increases the computational complexity of the problem. Our efficient optimization procedure enabled us to refine the timescale down to increments of five minutes and to provide an answer to this question.

We studied the impact of using different values of d on the system's overall performance. When d is a proper divisor of μ (e.g., when $\mu = 30$ and $d = 15, 10, 5$), we found that, indeed, outpatient clinics can benefit from a more refined timescale in their scheduling decisions by improving their daily patient flow. With a more refined timescale we can reduce waiting times or achieve a higher patient throughput in optimality (by having the flexibility to strategically fit an extra patient into the schedule). Achieving these benefits,

Figure 8. Optimal Schedules as a Function of p and $\mathbf{U}$



(a) $p = 0.8,\ p_u = 0.0$
$V(\mathbf{x}^*) = 112.7$, $\sum_{t=1}^n x_t^* = 18$, expected throughput = 14.4 patients
expected overtime = 14 mins, average waiting time = 25.5 mins

(b) $p = 0.9,\ p_u = 0.0$
$V(\mathbf{x}^*) = 96$, $\sum_{t=1}^n x_t^* = 16$, expected throughput = 14.4 patients
expected overtime = 12.2 mins, average waiting time = 16.4 mins

(c) $p = 1.0,\ p_u = 0.0$
$V(\mathbf{x}^*) = 77.7$, $\sum_{t=1}^n x_t^* = 15$, expected throughput = 15 patients
expected overtime = 10.7 mins, average waiting time = 17.5 mins

(d) $p = 0.8,\ p_u = 0.1$
$V(\mathbf{x}^*) = 135$, $\sum_{t=1}^n x_t^* = 14$, expected throughput = 14.4 patients
expected overtime = 20.9 mins, average waiting time = 34 mins

(e) $p = 0.9,\ p_u = 0.1$
$V(\mathbf{x}^*) = 125.7$, $\sum_{t=1}^n x_t^* = 12$, expected throughput = 14 patients
expected overtime = 15 mins, average waiting time = 27.8 mins

(f) $p = 1.0,\ p_u = 0.1$
$V(\mathbf{x}^*) = 115.3$, $\sum_{t=1}^n x_t^* = 12$, expected throughput = 15.2 patients
expected overtime = 24 mins, average waiting time = 31 mins

(g) $p = 0.8,\ p_u = 0.2$
$V(\mathbf{x}^*) = 148.3$, $\sum_{t=1}^n x_t^* = 10$, expected throughput = 14.4 patients
expected overtime = 25.5 mins, average waiting time = 39.8 mins

(h) $p = 0.9,\ p_u = 0.2$
$V(\mathbf{x}^*) = 141.9$, $\sum_{t=1}^n x_t^* = 9$, expected throughput = 14.5 patients
expected overtime = 25.5 mins, average waiting time = 36.9 mins

(i) $p = 1.0,\ p_u = 0.2$
$V(\mathbf{x}^*) = 135.6$, $\sum_{t=1}^n x_t^* = 8$, expected throughput = 14.4 patients
expected overtime = 22.5 mins, average waiting time = 34.6 mins

(j) $p = 0.8,\ p_u = 0.3$
$V(\mathbf{x}^*) = 155.6$, $\sum_{t=1}^n x_t^* = 6$, expected throughput = 14.4 patients
expected overtime = 30.4 mins, average waiting time = 40.2 mins

(k) $p = 0.9,\ p_u = 0.3$
$V(\mathbf{x}^*) = 152.4$, $\sum_{t=1}^n x_t^* = 5$, expected throughput = 14.1 patients
expected overtime = 26.6 mins, average waiting time = 37.3 mins

(l) $p = 1.0,\ p_u = 0.3$
$V(\mathbf{x}^*) = 148.1$, $\sum_{t=1}^n x_t^* = 5$, expected throughput = 14.6 patients
expected overtime = 29.9 mins, average waiting time = 39.3 mins

(m) $p = 0.8,\ p_u = 0.4$
$V(\mathbf{x}^*) = 161.1$, $\sum_{t=1}^n x_t^* = 2$, expected throughput = 14.4 patients
expected overtime = 37.5 mins, average waiting time = 34.9 mins

(n) $p = 0.9,\ p_u = 0.4$
$V(\mathbf{x}^*) = 159$, $\sum_{t=1}^n x_t^* = 2$, expected throughput = 14.6 patients
expected overtime = 38.4 mins, average waiting time = 36.2 mins

(o) $p = 1.0,\ p_u = 0.4$
$V(\mathbf{x}^*) = 156.9$, $\sum_{t=1}^n x_t^* = 2$, expected throughput = 14.8 patients
expected overtime = 39.2 mins, average waiting time = 37.7 mins

(p) $p = 0.8,\ p_u = 0.5$
$V(\mathbf{x}^*) = 186.1$, $\sum_{t=1}^n x_t^* = 1$, expected throughput = 16.8 patients
expected overtime = 74.3 mins, average waiting time = 48.8 mins

(q) $p = 0.9,\ p_u = 0.5$
$V(\mathbf{x}^*) = 185.5$, $\sum_{t=1}^n x_t^* = 1$, expected throughput = 16.9 patients
expected overtime = 74.9 mins, average waiting time = 49.4 mins

(r) $p = 1.0,\ p_u = 0.5$
$V(\mathbf{x}^*) = 184.9$, $\sum_{t=1}^n x_t^* = 1$, expected throughput = 17 patients
expected overtime = 75.5 mins, average waiting time = 50.1 mins

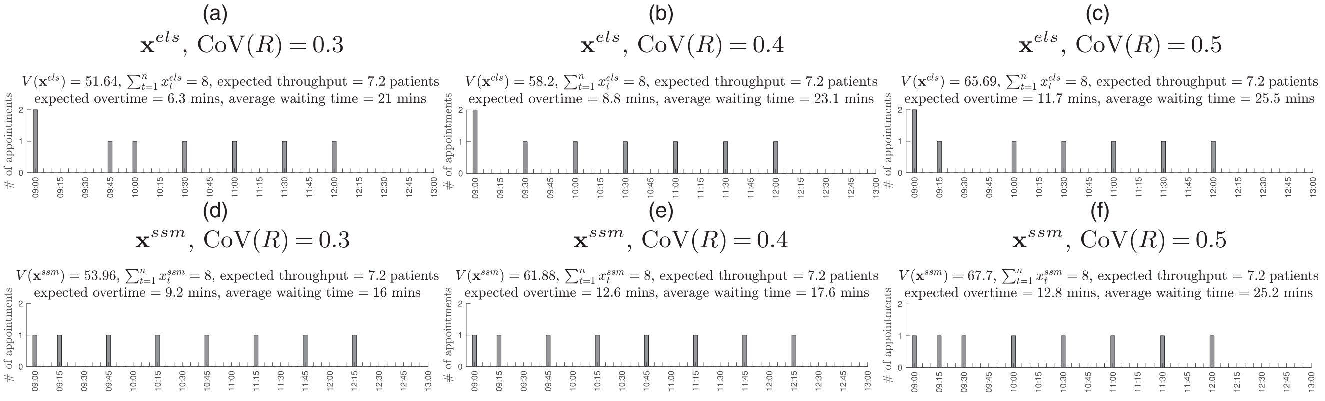
Note. $T = 8$ hours, $c_o = 1$, $c_s = 0.1$, $c_u = 0.075$, $d = 15$ minutes, $R \sim \text{BetaBin}(90, \alpha, \beta)$ with α and β such that $\mu = 30$ minutes and $\sigma\mu^{-1} = 0.4$, $U_t \sim \text{Bernoulli}(p_u)$ for all $t = 1, 2, \ldots, 32$.

however, requires patient adherence to the schedule. We also found that the marginal benefit decreases as the timescale of the schedule becomes more and more refined. Thus, by taking into account the aforementioned observations, for practical implementation we recommend that outpatient clinics consider a timescale equal to half of the average consultation time. Moreover, we should use caution when considering a schedule where d is not proper divisor of μ (e.g., when $\mu = 30$ and $d = 20$), as this practice might in fact lead to adverse operational outcomes.

- **Overbooking:** It is often optimal to overbook certain parts of the schedule to address variability stemming from all sources and its impact on the clinic's productivity. Overbooking can be done in two ways: by double-booking certain slots or by separating consecutive patients' appointments by less than the average consultation time. The former type of overbooking should mainly be used earlier in the day, whereas the latter can be spread later throughout the day. Moreover, in the presence of no-shows and stochastic consultation times, the total number of overbooked patients should be such that the expected patient throughput is slightly below the daily capacity to see patients.

- **Trade-off between productivity and waiting times:** In the presence of variability, elimination of idle time (thereby improving overall productivity and

Figure 9. Hueristic Solutions When Patients Are Not Necessarily Punctual (Under the Assumptions of Theorem 6)



(a) $\mathbf{x}^{els}$, $\text{CoV}(R) = 0.3$

(b) $\mathbf{x}^{els}$, $\text{CoV}(R) = 0.4$

(c) $\mathbf{x}^{els}$, $\text{CoV}(R) = 0.5$

(d) $\mathbf{x}^{ssm}$, $\text{CoV}(R) = 0.3$

(e) $\mathbf{x}^{ssm}$, $\text{CoV}(R) = 0.4$

(f) $\mathbf{x}^{ssm}$, $\text{CoV}(R) = 0.5$

Note. $T = 4$ hours, $c_o = 1$, $c_s = 0.1$, $p = 0.9$, $d = 15$ minutes, $R \sim \text{BetaBin}(90, \alpha, \beta)$ with α and β such that $\mu = 30$ minutes, $\mathbf{U} = \mathbf{0}$, $e = l = 1$.

throughput) and containment of waiting times are conflicting goals. Our weighted objective approach can capture the relative importance between these goals. As waiting time weighs more in our scheduling decisions, we might have to compromise patient throughput and/or idle time and/or overtime. We recommend that practitioners try different configurations on the objective coefficients and pick the one that strikes the right balance between their operational goals.

6.2. Limitations and Future Research

Solving the appointment scheduling problem under nonpunctual patients remains an open problem. We have shown that the multimodularity property collapses under a general class of uniform distributions regarding patient punctuality, and we proposed heuristic solutions. Even though our analysis is not comprehensive (due to the computational challenges involved), it provides the first platform to address nonpunctuality, and demonstrates evidence that nonpunctuality has a significant impact on the optimal scheduling strategy and should not be ignored. Further research is needed in this direction in terms of efficient objective function evaluations and exact optimization algorithms. Moreover, under our FIFO queueing discipline, patients are queued according to their arrival time, as opposed to their scheduled time. Samorani and Ganguly (2016) provide a solution to this problem by analyzing a stylized analytical model.

Our theoretical findings are true under general stochastic service times, yet our numerical analysis was restricted to distributions with discrete supports. This was in part due to the computational inefficiencies involved in the task of integrating with symbolic and numerical methods discussed in Section 5. However, we can readily redefine the timescale of the service times so that one time unit corresponds to one second, and therefore maintain a purely discrete setting where symbolic and numerical integrations are no longer necessary, while time is practically almost continuous. A precise computational analysis based on continuous distributions remains an open direction, though we believe it will not provide any additional insights.

Finally, we demonstrated that the front-loaded order $\leq_{\text{fl}}$ introduced in Section 5 is a more meaningful tool to contrast different scheduling strategies, when compared with the standard order $\leq$ on $\mathbb{Z}_+^n$. Our comparisons and monotonocity patterns based on $\leq_{\text{fl}}$ are grounded on our computational experiments, not on theoretical findings. A development of a theoretical foundation for our observations is a promising future direction. A good starting point is to establish stochastic comparisons of the workload process based on $\leq_{\text{fl}}$.

References

Ahmadi-Javid A, Jalali Z, Klassen KJ (2017) Outpatient appointment systems in healthcare: A review of optimization studies. *Eur. J. Oper. Res.* 258(1):3–34.

Altman R, Gaujal B, Hordijk A (2000) Multimodularity, convexity, and optimization properties. *Math. Oper. Res.* 25(2):324–347.

Begen MA, Queyranne M (2011) Appointment scheduling with discrete random durations. *Math. Oper. Res.* 36(2):240–257.

Begen MA, Levi R, Queyranne M (2012) Technical note—A sampling based approach to appointment scheduling. *Oper. Res.* 60(3): 675–681.

Birkhoff G (1937) Rings of sets. *Duke Math. J.* 3(3):443–454.

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

Cayirli T, Yang KK, Quek SA (2012) A universal appointment rule in the presence of no-shows and walk-ins. *Production Oper. Management* 21(4):682–697.

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

Davey BA, Priestley HA (2002) *Introduction to Lattices and Order* (Cambridge University Press, Cambridge, UK).

Favati P, Tardella F (1990) Convexity in nonlinear integer programming. *Ricerca Oper.* 53:3–44.

Feldman J, Liu N, Topaloglu H, Ziya S (2014) Appointment scheduling under patient preference and no-show behavior. *Oper. Res.* 62(4): 794–811.

Fujishige S (2005) *Submodular Functions and Optimization*, Annals of Discrete Mathematics, vol. 58 (Elsevier, Amsterdam).

Ge D, Wan G, Wang Z, Zhang J (2014) A note on appointment scheduling with piecewise linear cost functions. *Math. Oper. Res.* 39(4):1244–1251.

Green LV, Savin S (2008) Reducing delays for medical appointments: A queueing approach. *Oper. Res.* 56(6):1526–1538.

Green LV, Savin S, Wang B (2006) Managing patient service in a diagnostic medical facility. *Oper. Res.* 54(1):11–25.

Hajek B (1985) Extremal splittings of point processes. *Math. Oper. Res.* 10(4):543–556.

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

Izady N (2015) Appointment capacity planning in specialty clinics: A queueing approach. *Oper. Res.* 63(4):916–930.

Janssen AJEM, van Leeuwaarden JSH (2005) Relaxation time for the discrete D/G/1 queue. *Queueing Systems* 50(1):53–80.

Jiang R, Shen S, Zhang Y (2017) Integer programming approaches for appointment scheduling with random no-shows and service durations. *Oper. Res.* 65(6):1638–1656.

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

Kim SH, Whitt W, Cha WC (2018) A data-driven model of an appointment-generated arrival process at an outpatient clinic. *INFORMS J. Comput.* 30(1):181–199.

Kong Q, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Oper. Res.* 61(3):711–726.

Krause A (2010) SFO: A toolbox for submodular function optimization. *J. Machine Learn. Res.* 11:1141–1144.

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 Oper. Management* 21(5):874–888.

Lau HS, Lau AHL (2000) A fast procedure for computing the total system cost of an appointment schedule for medical and kindred facilities. *IIE Trans.* 32(9):833–839.

Liu N (2016) Optimal choice for appointment scheduling window under patient no-show behavior. *Production Oper. Management* 25(1):128–142.

Liu N, Ziya S (2014) Panel size and overbooking decisions for appointment-based services under patient no-shows. *Production Oper. Management* 23(12):2209–2223.

Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing Service Oper. Management* 12(2):347–364.

Liu N, Finkelstein SR, Kruk ME, Rosenthal D (2018) When waiting to see a doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Sci.* 64(5):1975–1996.

Luo J, Kulkarni VG, Ziya S (2012) Appointment scheduling under patient no-shows and service interruptions. *Manufacturing Service Oper. Management* 14(4):670–684.

Mak HY, Rong Y, Zhang J (2015) Appointment scheduling with limited distributional information. *Management Sci.* 61(2): 316–334.

Marshall AW, Olkin I (1979) *Inequalities: Theory of Majorization and its Applications* (Academic Press, New York).

Miller BL (1971) On minimizing nonseparable functions defined on the integers with an inventory application. *SIAM J. Appl. Math.* 21(1):166–185.

Murota K (1998) Discrete convex analysis. *Math. Programming* 83(1): 313–371.

Murota K (2004) On steepest descent algorithms for discrete convex functions. *SIAM J. Optim.* 14(3):699–707.

Murota K (2005) Note on multimodularity and L-convexity. *Math. Oper. Res.* 30(3):658–661.

Orlin JB (2009) A faster strongly polynomial time algorithm for submodular function minimization. *Math. Programming* 118(2): 237–251.

Patrick J, Puterman ML, Queyranne M (2008) Dynamic multipriority patient scheduling for a diagnostic resource. *Oper. Res.* 56(6): 1507–1525.

Qi J (2017) Mitigating delays and unfairness in appointment systems. *Management Sci.* 63(2):566–583.

Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing Service Oper. Management* 12(2):330–346.

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

Schrijver A (2000) A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *J. Combin. Theory Ser. B* 80(2):346–355.

Svitkina Z, Fleischer L (2011) Submodular approximation: Sampling-based algorithms and lower bounds. *SIAM J. Comput.* 40(6): 1715–1737.

Truong VA (2015) Optimal advance scheduling. *Management Sci.* 61(7):1584–1597.

Wang PP (1997) Optimally scheduling N customer arrival times for a single-server system. *Comput. Oper. Res.* 24(8):703–716.

Wang S, Liu N, Wan G (2019) Managing appointment-based services in the presence of walk-in customers. *Management Sci.* Forthcoming.

Zacharias C, Armony M (2017) Joint panel sizing and appointment scheduling in outpatient care. *Management Sci.* 63(11): 3978–3997.

Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production Oper. Management* 23(5): 788–801.

Zacharias C, Pinedo M (2017) Managing customer arrivals in service systems with multiple identical servers. *Manufacturing Service Oper. 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. *Ann. Oper. Res.* 178(1):121–144.