Operations Research

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

Dynamic Interday and Intraday Scheduling

Christos Zacharias, Nan Liu, Mehmet A. Begen

To cite this article:
Christos Zacharias, Nan Liu, Mehmet A. Begen (2022) Dynamic Interday and Intraday Scheduling. Operations Research

Published online in Articles in Advance 24 Aug 2022

. https://doi.org/10.1287/opre.2022.2342

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

OPERATIONS RESEARCH
Articles in Advance, pp. 1–19
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Dynamic Interday and Intraday Scheduling

Christos Zacharias,[a,*] Nan Liu,[b] Mehmet A. Begen[c]

[a] Miami Herbert Business School, University of Miami, Coral Gables, Florida 33146; [b] Carroll School of Management, Boston College, Chestnut Hill, Massachusetts 02467; [c] Ivey Business School, Western University, London, Ontario N6G 0N1, Canada
*Corresponding author

Contact: czacharias@bus.miami.edu, https://orcid.org/0000-0002-9911-7860 (CZ); nan.liu@bc.edu, https://orcid.org/0000-0001-7644-7341 (NL); mbegen@ivey.uwo.ca, https://orcid.org/0000-0001-7573-0882 (MAB)

Received: June 18, 2021
Revised: March 1, 2022; May 25, 2022
Accepted: May 31, 2022
Published Online in Articles in Advance: August 24, 2022

Area of Review: Stochastic Models

https://doi.org/10.1287/opre.2022.2342

Abstract. The simultaneous consideration of appointment day (interday scheduling) and time of day (intraday scheduling) in dynamic scheduling decisions is a theoretical and practical problem that has remained open. We introduce a novel dynamic programming framework that incorporates jointly these scheduling decisions in two timescales. Our model is designed with the intention of bridging the two streams of literature on interday and intraday scheduling and to leverage their latest theoretical developments in tackling the joint problem. We establish theoretical connections between two recent studies by proving novel theoretical results in discrete convex analysis regarding constrained multimodular function minimization. Grounded on our theory, we develop a practically implementable and computationally tractable scheduling paradigm with performance guarantees. Numerical experiments demonstrate that the optimality gap is less than 1% for practical instances of the problem.

Keywords: dynamic programming • discrete convexity • stochastic models • appointment scheduling

1. Introduction

Appointment scheduling has significant clinical, operational, and economical impact on healthcare systems. An informed scheduling strategy that can effectively match patient demand and service capacity dynamically is vital for the business of medical providers, quality of care, and patient satisfaction. By regulating patient flow via an appointment system, healthcare providers can mitigate arrival process variability and improve operational performance. From the perspective of patients, appointment scheduling provides the ease of knowing in advance when to receive service and planning their visits accordingly.

Scheduling an appointment entails determining the specific date and time of a patient's visit. This decision is made simultaneously in two different timescales: *interday* (i.e., on which day) and *intraday* (i.e., at what time). Respectively, this decision incurs delays for patients in two timescales (Gupta and Denton 2008): *indirect delay* (on the order of days, weeks) and *direct delay* (on the order of minutes, hours). Indirect delay is defined as the time gap between the appointment request and the offered appointment. Direct delay is the physical waiting experienced by patients at the medical facility before they see their provider. Both indirect and direct delays affect access to care and quality of care, and there is a fine trade-off between them (Zacharias and Armony 2017). Moreover, interday and intraday scheduling problems are related and interdependent (for example, the output of the former becomes a dynamic input to the latter). Our study introduces and analyzes the first analytical model to inform a healthcare provider, dynamically, upon a patient's request, an optimal appointment scheduling decision that simultaneously determines on which day and at what time the patient should be served.

The study of appointment scheduling can be traced back to the seminal work by Bailey (1952). Since then, there have been significant developments in the operations research literature on this topic. A substantial portion of this literature is developed in the context of healthcare operations, and thus, in our discussions, the word "patient" represents some customer seeking an appointment-based service. Traditionally, the research on this topic focuses on intraday scheduling, which aims to optimize the scheduled arrival times of patients within a workday, one important performance metric being the direct delays experienced by patients. More recently, a rising stream of literature has been analyzing interday scheduling models, addressing the question of how to dynamically assign appointment requests to future days, taking into consideration the impact of indirect delays.

Despite the tremendous growth of the appointment scheduling literature in the past few decades, no previous study has analytically tackled the joint interday and intraday scheduling problem, which remains open to the best of our knowledge and as indicated in Gupta and Denton (2008) and Feldman et al. (2014). This is due to the highly stochastic nature, complex structure, and large dimensionality of the joint problem. Our research fills this critical gap in the literature and provides the first analytical model and optimization framework to address this problem. In addition to being an interesting and open research question, there are many practical applications of dynamic interday and intraday scheduling, such as elective surgery scheduling, diagnostic testing, and outpatient care as well as in other service industries beyond healthcare (Sauré et al. 2020).

We make contributions to modeling, methodology, and theory of appointment scheduling. We model and analyze the dynamic problem of making joint interday and intraday scheduling decisions as a Markov decision process (MDP). Patients are given an appointment for a specific date (interday scheduling) and time (intraday scheduling) for their service at the time of their requests. Our model captures important features in the complex reality of appointment scheduling, such as stochastic demand for medical services, stochastic consultation times, no-shows, and walk-ins.

The rest of this article is structured as follows. First, we position our contribution within the appointment scheduling literature. Next, we introduce a novel dynamic programming framework that incorporates joint scheduling decisions in two timescales. This model is designed with the intention of bridging the two seemingly independent streams of literature of interday and intraday scheduling and leveraging their latest theoretical developments in tackling the joint problem. Subsequently, we present and characterize theoretically two distinct scheduling paradigms, and we demonstrate how they relate to the optimal dynamic policy. The first paradigm is a methodically crafted heuristic solution. The second paradigm is based on an idealistic solution with intuitive interpretation. Both scheduling paradigms reduce the joint problem to tractable single-variable MDPs with fully characterized and easy-to-compute optimal controls. They bound the optimal value function from above and below, leading to theoretically guaranteed and computationally tractable performance evaluation of our heuristic. Finally, we present computational implementations of our methods and discuss our conclusions.

2. Related Literature

Our study draws upon a broad body of studies related to appointment scheduling developed over decades. We organize this literature into four streams: intraday scheduling, interday and/or allocation scheduling, strategic design of appointment systems (e.g., sizing of patient panel and choice of daily capacity level) via queueing systems in the steady state, and inpatient flow management. We also provide a brief review of the related literature on discrete convexity. We only draw attention to the recent developments in each stream, and we discuss how we build upon the collective knowledge of the field. Interested readers may refer to in-depth literature reviews, such as Cayirli and Veral (2003), Gupta and Denton (2008), Ahmadi-Javid et al. (2017), and Dai and Shi (2020).

2.1. Intraday Scheduling

The intraday scheduling literature seeks to optimize a single day's operations by analyzing the detailed intraday dynamics of the system. More specifically, this literature develops and analyzes mathematical programming models to determine the scheduled arrival times of patients by optimally balancing the trade-off between waiting times and capacity utilization. Patients may be scheduled to arrive at any time during the continuous time spectrum of a workday or assigned to specific discrete time slots. Different sources of uncertainties in the system are considered in these models: stochastic service times, no-shows, and nonpunctuality as well as potential arrivals of walk-in patients. Some recent studies are Hassin and Mendel (2008), Robinson and Chen (2010), Zeng et al. (2010), Begen and Queyranne (2011), LaGanga and Lawrence (2012), Begen et al. (2012), Kong et al. (2013), Chen and Robinson (2014), Zacharias and Pinedo (2014, 2017), Kuiper et al. (2015), Qi (2017), Jiang et al. (2017), Wang et al. (2020), and Zacharias and Yunes (2020). A common theme in this literature is to develop an optimal *static* schedule for a single day, assuming that the set of patients to be scheduled is known in advance and/or can be selected optimally by the scheduler. In contrast, our study explicitly considers the interday dynamics of the system, and the number of patients scheduled in a day is dynamically determined based on the state of the appointment book and in anticipation of stochastic future demand.

2.2. Interday and Allocation Scheduling

More recently, a growing stream of literature considers interday dynamics in appointment scheduling problems. Interday scheduling is concerned with an *online* decision regarding how to schedule patients to future days upon their appointment requests. Therefore, dynamic programming is the primary modeling tool. Interday scheduling is also referred to as *advance scheduling* in the literature. Patrick et al. (2008) consider dynamic scheduling of surgical patients with different priorities and targets on delays. Liu et al. (2010) study how to dynamically schedule patients with delay-dependent no-show and cancellation probabilities. Feldman et al.

(2014) and Liu et al. (2019) extend the work by Liu et al. (2010) and incorporate patient choice behavior in making scheduling decisions. Deo et al. (2013) study how to schedule patients over periods, taking into account their disease progression. As we discuss in more detail in Section 2.6, Truong (2015) derives a characterization of an optimal policy for the dynamic interday problem and an algorithm to compute such a policy exactly and efficiently by considering stochastic daily demand for appointments and stochastic capacity utilization. Carew et al. (2020) examine the sequential capacity planning problem in which a hospital allocates operating room time to different surgical specialties. Keyvanshokooh et al. (2020) consider an online resource allocation problem in which a heterogeneous stream of arrivals that vary in service times and rewards make service requests from multiple providers. Wang et al. (2018, 2019) and Diamant et al. (2018) further study how to dynamically schedule patients in a network structure in which patients may need to visit multiple stations.

In addition to advance scheduling, the literature on *allocation scheduling* is related and noteworthy. In allocation scheduling, all requests for appointments join the same wait-list, and a scheduler sequentially decides how many patients to serve tomorrow, whereas the rest of the patients remain on the wait-list for future service. In other words, the scheduler does not assign any appointments in advance upon request, but only calls the patients the day before their offered appointments. Allocation scheduling is used to manage surgical wait-lists in countries with publicly funded healthcare systems (e.g., Canada and the United Kingdom); see Gerchak et al. (1996) and Huh et al. (2013) for analyses of such models.

Whereas the studies of advance scheduling and allocation scheduling consider indirect delay costs in making scheduling decisions, they often assume a newsvendor-type model for the daily operational costs and do not consider the intraday details. These models cannot directly inform patients, upon request, their appointed date and time of service jointly.

2.3. Queueing Models

There is also literature, usually leveraging queueing models in the steady state to address system-level design questions for appointment scheduling; see, for example, Green and Savin (2008), Liu (2016), and Zacharias and Armony (2017). These studies provide methods and insights to determine the size of the patient base (called panel size in the context of primary care), the level of daily capacity, and the choice of appointment window (i.e., the maximum time allowed for patients to make advance appointments). These queueing studies address the question of how to design and plan an appointment scheduling system at a strategic level, whereas the dynamic intraday and interday scheduling problem we address in this paper is concerned with how to tactically manage an appointment scheduling system.

2.4. Inpatient Flow Management

In the context of hospital inpatient flow management, some studies consider joint interday and intraday operations; see Dai and Shi (2020) for a recent review on this literature. However, the hospital setting leads to a significantly different problem, often with simpler interday or intraday details involved. For instance, Helm and Van Oyen (2014) develop a static optimization model to determine the optimal cyclic schedules for elective hospital admissions, taking into account their impact on the use of hospital beds. One key assumption is that patients are scheduled according to the fixed cyclic schedule and then "passively" flow through the system, whereas in our model the daily schedule is managed dynamically. Sauré et al. (2020) jointly solve advance and intraday scheduling problems in surgical care. However, they only consider idle time and overtime of the surgeon during the day and do not consider patient wait time (hence, all patients are scheduled at the beginning of the day). This approach may be appropriate for filling up an operating room block schedule (Santibáñez et al. 2007), but it falls short to fully capture intraday details when patient waiting is included in the equation.

2.5. Multimodularity and Discrete Convexity

Multimodularity of the intraday cost function is a key assumption of our model, backed theoretically by the recent literature and with intuitive interpretation. Multimodularity is a discrete convexity property, formally and rigorously defined first by Hajek (1985) within the context of optimal admission control to queues. Multimodularity and other notions of discrete convexity are studied and incorporated progressively in various areas/applications of discrete optimization and operations research. Interested readers are referred to Li and Yu (2014), Moriguchi and Murota (2019), and Chen and Li (2021a, b) for expositions of the mathematical definitions, theoretical properties, and practical implications of such functions.

Multimodularity is shown to guarantee global optimality of integer solutions that are optimal within some discrete local neighborhood of exponential size. Murota (2005) provides an algorithm for minimizing an unconstrained multimodular function over $\mathbb{Z}^n$ in polynomial time via unconstrained submodular set-function minimization. Zacharias and Yunes (2020) demonstrate how to minimize an unconstrained multimodular function over $\mathbb{Z}_+^n$ in polynomial time via submodular set-function minimization over ring families of sets.

The objective function in static intraday scheduling is proven to be multimodular in various studies in the literature under different models and assumptions (Zeng et al. 2010, Zacharias and Pinedo 2017, Wang et al. 2020, Zacharias and Yunes 2020). More pertinent to our study, Zacharias and Yunes (2020) prove that the intraday cost function is multimodular under general stochastic service times, no-shows, walk-ins, and heterogeneous waiting cost coefficients.

2.6. Positioning and Contribution

We build upon the results of two recent studies: one from interday scheduling (Truong 2015) and one from intraday scheduling (Zacharias and Yunes 2020). We prove theoretical connections between them to develop an analytically and computationally tractable optimization framework for the joint problem.

Truong (2015) characterizes an optimal policy for the dynamic interday problem and derives an algorithm to compute such a policy exactly and efficiently. The characterization and solution procedure in Truong (2015) relies on the property of *successive refinability*, which elegantly reduces the interday problem (multidimensional dynamic programming) to sequential allocation scheduling (one-dimensional dynamic programming). The intraday cost function in Truong (2015) is assumed to be a convex function of one variable (the total number of patients in the schedule) and does not explicitly consider the detailed intraday dynamics.

Zacharias and Yunes (2020) model and analyze the static intraday problem as an integer nonlinear mathematical program, in which the objective function (intraday cost) depends on a detailed schedule for a workday and is the outcome of stochastic analysis in the transient state. They prove that the problem possesses discrete convexity properties and develop an algorithm that identifies an optimal intraday schedule efficiently.

We extend the dynamic model of Truong (2015) to account for joint interday and intraday decisions so that the intraday cost function depends on a detailed schedule for a workday (i.e., a vector), and no-shows can be incorporated. We provide a characterization of the optimal value function by treating the number of patients in a daily schedule as the connecting link between Truong (2015) and Zacharias and Yunes (2020). In particular, we treat the total number of patients in a daily schedule as a parameter (the right-hand side of an equality constraint) in a mathematical program that involves constrained multimodular function minimization.

We prove that, under two distinct intraday scheduling paradigms, the constrained static intraday problem can be solved in polynomial time. Moreover, the corresponding minimal intraday cost functions are convex in the total number of patients (a scalar that appears in the right-hand side of an equality constraint). To the best of our knowledge, these theoretical results in discrete convex analysis are novel on their own, independent of our model and underlying problem. They relate to the theory of discrete optimization and its applications within and beyond the field of appointment scheduling. They are essential in our development of theoretical lower and upper bounds for the joint interday and intraday problem. Based on these bounds, we develop a computationally efficient heuristic solution with a theoretically guaranteed optimality gap. Extensive numerical experiments indicate that the optimality gap is less than 1% for practical instances of the problem and reveal additional managerial implications.

3. Dynamic Programming Framework

We first introduce a comprehensive dynamic model that accounts for joint scheduling decisions in two different timescales (interday and intraday). Subsequently, we reproduce relevant results from the literature on dynamic allocation and interday scheduling, adjusted accordingly to our modeling framework and assumptions. These reproduced results serve as steppingstones to advancing our understanding of the problem and building our theory.

3.1. Interday and Intraday Scheduling Model

We model the dynamic interday and intraday scheduling problem as an MDP. We consider a horizon of T days, where T is allowed to be infinity, indexed by $t \in \{1, 2, \ldots, T\}$. Future outcomes are discounted by a factor $\gamma \in (0, 1)$. Each workday is partitioned into n time slots of equal duration, for example, slots of 30, 20, or 10 minutes, depending on how refined we want the intraday schedules to be. The state of the appointment system at the beginning of day t, before new demand is realized, is a collection of vectors (daily schedules):

$$\mathbf{X}_t = \begin{bmatrix} \mathbf{x}_{t1} \\ \mathbf{x}_{t2} \\ \vdots \\ \mathbf{x}_{t\tau} \\ \vdots \end{bmatrix} = \begin{array}{c} \\ \text{day } 1 \\ \text{day } 2 \\ \vdots \\ \text{day } \tau \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \text{slot } 1 & \text{slot } 2 & \ldots & \text{slot n} \end{array} \\ \begin{bmatrix} x_{t1}^1 & x_{t1}^2 & \ldots & x_{t1}^n \\ x_{t2}^1 & x_{t2}^2 & \ldots & x_{t2}^n \\ \vdots & \vdots & \ddots & \vdots \\ x_{t\tau}^1 & x_{t\tau}^2 & \ldots & x_{t\tau}^n \\ \vdots & \vdots & & \vdots \end{bmatrix} \end{array} \in \mathcal{X},$$

where $\mathcal{X}$ denotes the set of all admissible appointment books, $\mathbf{x}_{t\tau} = (x_{t\tau}^1, x_{t\tau}^2, \ldots, x_{t\tau}^n) \in \mathbb{Z}_+^n$ is the schedule for day τ in the rolling horizon from day t, and $x_{t\tau}^i = \#$ of patients scheduled to arrive at slot i in τ days from day t. We use $|\cdot|$ to denote the L^1-norm. Thus, the size of the appointment book is $|\mathbf{X}_t| = \sum_\tau \sum_i x_{t\tau}^i$, and the size of

the schedule for a single day is $|\mathbf{x}_{t\tau}| = \sum_i x^i_{t\tau}$. Our model assumes that the sequence of events during period t occurs in the following order:

Step 1. The state of the MDP $\mathbf{X}_t$ at the beginning of day t is observed.

Step 2. Stochastic demand $d_t \sim d$ during day t is realized and observed. We assume that $\{d_t\}_{t=1}^T$ is a sequence of independent and identically distributed (i.i.d.) random variables with support on some subset of $\mathbb{Z}_+$.

Step 3. A decision of how to assign new requests to slots in the horizon is made. The booking decision on day t is

$$\mathbf{B}_t = \begin{bmatrix} \mathbf{b}_{t1} \\ \mathbf{b}_{t2} \\ \vdots \\ \mathbf{b}_{t\tau} \\ \vdots \end{bmatrix} = \begin{array}{c} \\ \text{day } 1 \\ \text{day } 2 \\ \vdots \\ \text{day } \tau \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \text{slot } 1 & \text{slot } 2 & \dots & \text{slot n} \end{array} \\ \begin{bmatrix} b^1_{t1} & b^2_{t1} & \dots & b^n_{t1} \\ b^1_{t2} & b^2_{t2} & \dots & b^n_{t2} \\ \vdots & \vdots & \ddots & \vdots \\ b^1_{t\tau} & b^2_{t\tau} & \dots & b^n_{t\tau} \\ \vdots & \vdots & & \vdots \end{bmatrix} \end{array} \in \mathcal{B}(\mathbf{X}_t, d_t),$$

where $\mathbf{b}_{t\tau} = (b^1_{t\tau}, b^2_{t\tau}, \dots, b^n_{t\tau}) \in \mathbb{Z}^n_+$ is the vector of new patients booked in the schedule for day τ in the scheduling horizon, $b^i_{t\tau} = \#$ of new patients booked to arrive at slot i in τ days from day t, and $\mathcal{B}(\mathbf{X}_t, d_t) = \{\mathbf{B}_t : \mathbf{B}_t \geq 0, |\mathbf{B}_t| = d_t, \mathbf{X}_t + \mathbf{B}_t \in \mathcal{X}\}$. The outcome of the decision is an updated schedule:

$$\mathbf{Z}_t = \mathbf{X}_t + \mathbf{B}_t = \begin{bmatrix} \mathbf{z}_{t1} \\ \mathbf{z}_{t2} \\ \vdots \\ \mathbf{z}_{t\tau} \\ \vdots \end{bmatrix} = \begin{array}{c} \\ \text{day } 1 \\ \text{day } 2 \\ \vdots \\ \text{day } \tau \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} \text{slot 1} & \text{slot 2} & \dots & \text{slot n} \end{array} \\ \begin{bmatrix} z^1_{t1} & z^2_{t1} & \dots & z^n_{t1} \\ z^1_{t2} & z^2_{t2} & \dots & z^n_{t2} \\ \vdots & \vdots & \ddots & \vdots \\ z^1_{t\tau} & z^2_{t\tau} & \dots & z^n_{t\tau} \\ \vdots & \vdots & & \vdots \end{bmatrix} \end{array} \in \mathcal{Z}(\mathbf{X}_t, d_t),$$

where $\mathcal{Z}(\mathbf{X}_t, d_t) = \{\mathbf{Z}_t \in \mathcal{X} : \mathbf{Z}_t \geq \mathbf{X}_t, |\mathbf{Z}_t| = |\mathbf{X}_t| + d_t\}$. The first row of $\mathbf{Z}_t$ (i.e., the vector $\mathbf{z}_{t1}$) is the schedule for day $t+1$.

Step 4. The system incurs the one-period cost $C(\mathbf{Z}_t) = \texttt{inter}(\mathbf{Z}_t) + \texttt{intra}(\mathbf{z}_{t1})$, according to the updated schedule. Consistent with Truong (2015), a daily appointment delay cost c_a is incurred for every patient registered in the appointment book; therefore, $\texttt{inter}(\mathbf{Z}_t) = c_a|\mathbf{Z}_t|$. The intraday cost function $\texttt{intra}(\cdot)$ is assumed to be a nonnegative multimodular function defined on $\mathbb{Z}^n_+$. For completeness and for the readers interested in implementing our scheduling algorithms, in Appendix B, we provide a mathematical definition of multimodularity and present a general class of such functions that capture direct delay, overtime, and idle time costs.

Step 5. The state of the appointment system is updated as $\mathbf{X}_{t+1} = \zeta(\mathbf{Z}_t)$, where the operator $\zeta(\mathbf{Z}_t)$ removes the first row from $\mathbf{Z}_t$. That is, the state of the system "rolls" one day forward.

A scheduling policy is a sequence of controls $\mu = \{\mu_1, \mu_2, \dots, \mu_T\}$, where $\mu_t : \mathcal{X} \times \mathbb{Z}_+ \to \mathcal{X}$ and $\mu_t(\mathbf{X}_t, d_t) \in \mathcal{Z}(\mathbf{X}_t, d_t)$ for all $t = 1, 2, \dots, T$. The total discounted cost incurred from period t and beyond under policy μ, given the then-current schedule $\mathbf{X}_t$ and demand d_t, is defined as

$$V^\mu_t(\mathbf{X}_t, d_t) = \mathbb{E}\left[\sum_{\tau=t}^T \gamma^{\tau-t} C(\mu_\tau(\mathbf{X}_\tau, d_\tau)) \middle| \mathbf{X}_t, d_t\right],$$

and the optimal cost function

$$V_t(\mathbf{X}_t, d_t) = \min_\mu V^\mu_t(\mathbf{X}_t, d_t)$$

satisfies the optimality equation

$$V_t(\mathbf{X}_t, d_t) = \min_{\mathbf{Z}_t \in \mathcal{Z}(\mathbf{X}_t, d_t)} \{C(\mathbf{Z}_t) + \gamma \mathbb{E}[V_{t+1}(\zeta(\mathbf{Z}_t), d_{t+1})]\}, \quad t = 1, 2, \dots, T. \tag{1}$$

In the finite horizon case, we assume that $V_{T+1}(\cdot, \cdot) = 0$ without loss of generality. The infinite horizon problem is well defined because the cost per period $C(\cdot)$ is nonnegative, and therefore, the conditions of proposition 4.1.1 in Bertsekas (2000) are satisfied.

We note that our dynamic programming model and subsequent methods can incorporate patient no-show behavior and walk-ins only through the choice of a suitable $\texttt{intra}(\cdot)$ function (for example, the one presented in Appendix B incorporates no-shows, walk-ins, and general stochastic service times).

3.2. Allocation Scheduling Model

In this section, we consider a simplified version of the problem in which all requests for an appointment join the same wait-list. In this simplified model, a scheduler sequentially decides how many patients to serve in the next period and, consequently, how many patients to keep on the wait-list for future service. In other words, the scheduler does not assign appointments (neither a date nor a time slot) in advance upon request and only notifies the patients the day before their service dates. As we discuss in Section 2, this scheduling model is referred to in the literature as an allocation scheduling model and is used to model the management of surgical wait-lists in publicly funded healthcare systems. Even though the allocation scheduling model cannot be applied in the context of general appointment scheduling, it serves as a stepping-stone to solve the more intricate interday and intraday scheduling problem.

In an allocation scheduling model, the state of the system at the beginning of period t, before new demand is realized, is the size of the wait-list x_t. This model reduces

the state of the system from a matrix to a scalar, and thus, it is analytically and computationally more tractable. The sequence of events and dynamic programming formulation are analogous:

Step 1. At the beginning of day t, the size of the wait-list x_t is observed.

Step 2. Stochastic demand $d_t \sim d$ during day t is realized and observed, and the size of the wait-list becomes $z_t = x_t + d_t$. We assume that $\{d_t\}_{t=1}^T$ is a sequence of i.i.d. random variables with support on some subset of $\mathbb{Z}_+$.

Step 3. A decision is made to book b_t patients from the wait-list to arrive in period $t+1$. The remaining $z_t - b_t$ patients stay on the wait-list.

Step 4. The system incurs the one-period cost $\bar{C}(b_t, z_t) = \overline{\texttt{inter}}(z_t) + \overline{\texttt{intra}}(b_t)$, where $\overline{\texttt{inter}}(z_t) = c_a z_t$, and $\overline{\texttt{intra}}(\cdot)$ is assumed to be a convex function of one variable. We note that, in Truong (2015), the intraday cost function is assumed to be convex and increasing. In our model, it takes a more general format. It has a minimizer $a \geq 0$ and it is decreasing on $[0, a]$ and increasing on $[a, \infty)$.

Step 5. The state of the system is updated as $x_{t+1} = z_t - b_t$.

After observing the wait-list z_t, the maximal optimal control (booking) policy on day t is denoted by $\bar{\pi}_t(z_t)$, and $\bar{V}_t(z_t)$ denotes the optimal discounted cost incurred from period t and beyond and satisfies the optimality equation

$$\bar{V}_t(z_t) = \min_{b_t \in [0, z_t]} \{\bar{C}(b_t, z_t) + \gamma \mathbb{E}[\bar{V}_{t+1}(z_t - b_t + d_{t+1})]\}, \quad t = 1, 2, \ldots, T. \tag{2}$$

In the finite horizon case, $\bar{V}_{T+1}(\cdot) = 0$. For the infinite horizon problem, the conditions of proposition 4.1.5 in Bertsekas (2000) are satisfied, and therefore, there exists a maximal optimal stationary policy $\bar{\pi}$ that does not depend on the time index t.

Lemma 1. *(i) $\bar{V}_t(\cdot)$ is convex for all t. (ii) $\bar{\pi}_t(\cdot)$ is increasing, and $\bar{\pi}_t(z_t + 1) \leq \bar{\pi}_t(z_t) + 1$ for all t.*

The value function is convex in the size of the wait-list. As a result, given also the problem's separable costs and lattice action space, the cost function to be minimized in period t is submodular in the action-state pair (b_t, z_t). Consequently, the optimal allocation scheduling policy is increasing in the size of the wait-list, and when the wait-list increases by one patient, the optimal control increases by at most one patient. The latter property is leveraged to speed up the solution procedure for the dynamic program (2). We note that similar structural properties are obtained in Gerchak et al. (1996), Huh et al. (2013), and Truong (2015). We present the proofs of Lemma 1 and subsequent theoretical results in Appendix A.

3.3. Interday Scheduling Model

As demonstrated in Truong (2015), the interday scheduling problem can be reduced to an iterative sequence of allocation scheduling problems. In interday scheduling, patients are being assigned an appointment for some day in the future but not a particular time slot. The state of the system on day t is a vector $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots)$, where $x_{t\tau}$ is the total number of patients scheduled to arrive in τ days from t. We note that the state of the system $\mathbf{x}_t$ in this model contains more information compared with its counterpart in the allocation scheduling model (2), but not as many details as the one in the interday/intraday scheduling model (1). The model assumes that the sequence of events during period t occurs in the following order:

Step 1. The state of the system $\mathbf{x}_t$ at the beginning of day t is observed.

Step 2. Stochastic demand $d_t \sim d$ during day t is realized and observed. We assume that $\{d_t\}_{t=1}^T$ is a sequence of i.i.d. random variables with support on some subset of $\mathbb{Z}_+$.

Step 3. A decision of how to assign new requests to future days is made. The booking decision on day t is denoted by $\mathbf{b}_t = (b_{t1}, b_{t2}, \ldots.)$ with $|\mathbf{b}_t| = d_t$. The outcome of the decision is an updated schedule $\mathbf{z}_t = (z_{t1}, z_{t2}, \ldots.) = \mathbf{x}_t + \mathbf{b}_t$ with $|\mathbf{z}_t| = |\mathbf{x}_t| + d_t$.

Step 4. The system incurs the one-period cost $\tilde{C}(\mathbf{z}_t) = \overline{\texttt{inter}}(|\mathbf{z}_t|) + \overline{\texttt{intra}}(z_{t1})$, where $\overline{\texttt{inter}}(\cdot)$ and $\overline{\texttt{intra}}(\cdot)$ are as described in Section 3.2.

Step 5. The state of the appointment system is updated as $\mathbf{x}_{t+1} = \eta(\mathbf{z}_t)$ by removing the first element of $\mathbf{z}_t$.

Given the then-current schedule $\mathbf{x}_t$ and demand d_t, $\tilde{V}_t(\mathbf{x}_t, d_t)$ denotes the total discounted cost incurred from t and beyond and satisfies the optimality equation

$$\tilde{V}_t(\mathbf{x}_t, d_t) = \min_{\mathbf{z}_t : \mathbf{z}_t \geq \mathbf{x}_t,\ |\mathbf{z}_t| = |\mathbf{x}_t| + d_t} \{\tilde{C}(\mathbf{z}_t) + \gamma \mathbb{E}[\tilde{V}_{t+1}(\eta(\mathbf{z}_t), d_{t+1})]\}, \quad t = 1, 2, \ldots, T. \tag{3}$$

The next result follows from theorems 5 and 6 in Truong (2015).

Theorem 1 (Truong 2015). *Assume that $\mathbf{x}_1 = \mathbf{0}$ and that the horizon is infinite. There exists an optimal schedule $\mathbf{z}_t^{\bar{\pi}}(\mathbf{x}_t, d_t) = (z_{t1}^{\bar{\pi}}(\mathbf{x}_t, d_t), z_{t2}^{\bar{\pi}}(\mathbf{x}_t, d_t), \ldots)$ for the interday scheduling problem in (3) for every period t satisfying*

$$z_{t\tau}^{\bar{\pi}}(\mathbf{x}_t, d_t) = \begin{cases} \bar{\pi}(|\mathbf{x}_t| + d_t) & \text{for } \tau = 1 \\ \bar{\pi}\left(|\mathbf{x}_t| + d_t - \sum_{k=1}^{\tau-1} z_{tk}^{\bar{\pi}}(\mathbf{x}_t, d_t)\right) & \text{for } \tau = 2, 3, \ldots, \end{cases} \tag{4}$$

where $\bar{\pi}(\cdot)$ is the a maximal optimal stationary policy for the allocation scheduling model (2). Moreover, $z_{t1}^{\bar{\pi}}(\mathbf{x}_t, d_t) \geq z_{t2}^{\bar{\pi}}(\mathbf{x}_t, d_t) \geq \ldots$.

According to Theorem 1, we can construct an optimal solution to the interday scheduling problem based on the tractable optimal control of allocation scheduling. First, we apply the allocation function $\bar{\pi}$ to the total number of outstanding patients $|\mathbf{x}_t| + d_t$ to obtain the number of patients to serve tomorrow. Then, we apply the allocation function $\bar{\pi}$ to the remaining number of outstanding patients to obtain the number of patients to be allocated to the day after tomorrow and so on and so forth until all patients are allocated to some day in the future. The optimality of this policy is based on the elegant successive refinability property proved in Truong (2015). If we relax the constraint in (3) that prior scheduling commitments are binding (i.e., remove the constraint $\mathbf{z}_t \geq \mathbf{x}_t$), we obtain a solution that is a refinement of the existing schedule and, therefore, feasible for the constrained problem. In other words, any changes in the updated schedule can be made with new requests according to the allocation scheduling function, and therefore, the constraint $\mathbf{z}_t \geq \mathbf{x}_t$ is redundant.

4. A Theory-Based Practical Heuristic

The dynamic program (1) is analytically and computationally intractable when relying directly on existing tools from the literature. In this section, we consider a constrained version of the problem, resulting in a reduction to an efficiently solvable dynamic program with a theoretical support.

The intraday cost function in allocation scheduling in Section 3.2 is a general convex function of one variable (the total number of patients in the schedule) as opposed to a detailed schedule for a workday. In order to establish a connection between the dimensionality reduction results in Section 3.3 and the joint interday and intraday problem addressed in Section 3.1, we first need to define a meaningful relationship between $\overline{\texttt{intra}}(\cdot)$ and $\texttt{intra}(\cdot)$. To that end, we introduce a theory-based and practical scheduling paradigm: sequentially refinable intraday scheduling (SRIS).

SRIS has three critical characteristics. First, to the best of our knowledge, it is the first analytical approach that can provide patients with a date (interday timescale) and a time (intraday timescale) simultaneously when they book appointments. It is computationally tractable and fully implementable in practice. Second, it is feasible to our original problem (1), and therefore, it yields a theoretical upper bound for the value function of the interday and intraday scheduling problem. Third, SRIS is shown numerically to be nearly optimal for practical instances of the problem. In the subsequent section, we develop a theoretical lower bound to (1), based on which we bound from above the optimality gap of SRIS.

Definition 1. We say that a sequence of vectors $(\mathbf{y}_b)_{b\in\mathbb{Z}_+}$ is sequentially refinable iff $|\mathbf{y}_b| = b$ for all $b \in \mathbb{Z}_+$ and $\mathbf{0} = \mathbf{y}_0 \leq \mathbf{y}_1 \leq \mathbf{y}_2 \cdots$.

Let $\mathbf{s} = (s^1, s^2, \ldots, s^n) \in \arg\min\{\texttt{intra}(\mathbf{x}) : \mathbf{x} \in \mathbb{Z}_+^n\}$ be an optimal solution to the static intraday problem. Based on the assumption that $\texttt{intra}(\cdot)$ is multimodular, $\mathbf{s}$ can be computed efficiently based on theorem 7 in Zacharias and Yunes (2020). In an idealistic appointment system in which demand for appointments can be perfectly regulated so that a deterministic stream of exactly $\bar{b} = |\mathbf{s}| = \sum_{i=1}^n s_i$ appointment requests arrive at the scheduler on a daily basis, then each daily schedule is equal to $\mathbf{s}$, and thus, no patient experiences indirect delays, and the system consistently incurs the minimal intraday cost. In reality (because of variability in the number of new daily requests arriving at the scheduler and indirect delay costs and potentially other factors) some daily schedules are anticipated to have a different number of patients and/or structure, responding dynamically and optimally to the stochastically evolving state of the system.

We construct a sequence of intraday schedules centered around $\mathbf{s}$ defined as

$$\mathbf{s}_b \triangleq \begin{cases} \mathbf{s} & \text{if} \quad b = \bar{b} \\ \arg\min\{\texttt{intra}(\mathbf{x}) : \mathbf{x} \in \mathbb{Z}_+^n, |\mathbf{x}| = b, \mathbf{x} \leq \mathbf{s}_{b+1}\} & \\ & \text{for} \quad b = \bar{b}-1, \bar{b}-2, \ldots, 2, 1, 0 \\ \arg\min\{\texttt{intra}(\mathbf{x}) : \mathbf{x} \in \mathbb{Z}_+^n, |\mathbf{x}| = b, \mathbf{x} \geq \mathbf{s}_{b-1}\} & \\ & \text{for} \quad b = \bar{b}+1, \bar{b}+2, \ldots. \end{cases} \tag{5}$$

The sequence $(\mathbf{s}_b)_{b\in\mathbb{Z}_+}$ is sequentially refinable because $\mathbf{s}_{b+1} = \mathbf{s}_b + \mathbf{e}_i$ for some $i \in \{1, 2, \ldots, n\}$ for all $b \in \mathbb{Z}_+$, where $\mathbf{e}_i \in \mathbb{Z}_+^n$ is the vector that has zeros everywhere except in the ith component in which it is equal to one. Based on $(\mathbf{s}_b)_{b\in\mathbb{Z}_+}$ we can establish a link between $\texttt{intra}(\cdot)$ and its single-variable counterpart $\overline{\texttt{intra}}(\cdot)$ and thereby leverage the results of Sections 3.2 and 3.3 to address the joint interday/intraday problem. Consider the piecewise linear function with integer break points defined as

$$\overline{\texttt{intra}}^s(b) \triangleq \begin{cases} \texttt{intra}(\mathbf{s}_b) & \text{if } b \in \mathbb{Z}_+ \\ (b - \lfloor b \rfloor)\texttt{intra}(\mathbf{s}_{\lfloor b \rfloor}) + (\lceil b \rceil - b)\texttt{intra}(\mathbf{s}_{\lfloor b \rfloor}) & \\ & \text{if } b \in \mathbb{R}_+ \backslash \mathbb{Z}_+. \end{cases}$$

Lemma 2. $\overline{\texttt{intra}}^s(\cdot)$ *is convex on* $\mathbb{R}_+$.

From Lemma 2, the results of Sections 3.2 and 3.3 hold when we set $\overline{\texttt{intra}}(\cdot) = \overline{\texttt{intra}}^s(\cdot)$. Let $\bar{\pi}^s$ be the corresponding optimal stationary policy for allocation scheduling and consider the special case of (1) defined as

$$V_t^s(\mathbf{X}_t, d_t) = \min_{\mathbf{Z}_t \in \mathcal{Z}(\mathbf{X}_t, d_t)} \{C^s(\mathbf{Z}_t) + \gamma \mathbb{E}[V_{t+1}^s(\zeta(\mathbf{Z}_t), d_{t+1})]\}, \quad t = 1, 2, \ldots, T, \tag{6}$$

where

$$C^s(\mathbf{X}_t) \triangleq \begin{cases} C(\mathbf{X}_t) & \text{if } \mathbf{X}_t \in \mathcal{X}^s \triangleq \{\mathbf{X}_t : \mathbf{x}_{t\tau} = \mathbf{s}_{|\mathbf{x}_{t\tau}|} \text{ for all } \tau\} \subseteq \mathcal{X} \\ \infty & \text{otherwise.} \end{cases}$$

We characterize analytically a computationally tractable solution to (6) in Theorem 2.

Theorem 2. *Assume that $\mathbf{X}_1 = \mathbf{0}$ and that the horizon is infinite. There exists an optimal policy $\mathbf{Z}_t^s$ for the joint interday and intraday scheduling problem in (6) for all t such that $\mathbf{X}_t \in \mathcal{X}^s$ and*

$$\mathbf{Z}_t^s(\mathbf{X}_t, d_t) = \begin{bmatrix} \mathbf{s}_{z_{t1}^{\bar{\pi}^s}(\theta(\mathbf{X}_t), d_t)} \\ \mathbf{s}_{z_{t2}^{\bar{\pi}^s}(\theta(\mathbf{X}_t), d_t)} \\ \vdots \\ \mathbf{s}_{z_{t\tau}^{\bar{\pi}^s}(\theta(\mathbf{X}_t), d_t)} \\ \vdots \end{bmatrix},$$

where $\theta(\mathbf{X}_t) = (|\mathbf{x}_{t1}|, |\mathbf{x}_{t2}|, \ldots)$.

According to Theorem 2, if we restrict the set of admissible appointment books $\mathcal{X}$ in Section 3.1 to be a collection of matrices with sequentially refinable rows, then the state of the appointment system can be reduced from a matrix (a detailed appointment book) to the corresponding vector for interday scheduling. Thus, from Theorem 1, the problem can be reduced further to sequential allocation scheduling with tractable optimal solution. In sum, Theorem 2 completely characterizes a feasible policy SRIS to the joint interday and intraday scheduling problem (1).

We note that the optimal allocation policy $\bar{\pi}^s$ specifies the number of patients to be served in each day, and the sequence $(\mathbf{s}_b)_{b\in\mathbb{Z}_+}$ defined in (5) prescribes, for each day, which time slots to schedule incoming requests for appointments. Because the sequence $(\mathbf{s}_b)_{b\in\mathbb{Z}_+}$ is sequentially refinable, the policy SRIS can schedule new appointment requests online without revoking the scheduling decisions made in previous days. More importantly, we note that SRIS can handle a system in which appointment assignments are made one at a time as soon as individual requests arrive at the scheduler as opposed to handling jointly the total number of requests once the daily demand is realized. In order to support real-time decision making, SRIS restricts the search space of schedules within the class of sequentially refinable ones. Moreover, as we demonstrate next, SRIS is strikingly close to optimal despite its restricted state space.

5. A Theory-Based Idealistic Solution with Practical Implications

In this section, we construct a lower bound for the value function of the joint interday/intraday scheduling problem in (1). This lower bound is based on theory and has an intuitive interpretation. Moreover, it is instrumental in evaluating the performance of SRIS.

We relax the requirement that appointment times, once assigned to patient requests, cannot be changed. Rather, we allow the scheduler to finalize the appointment times of patients on the day right before they receive their services, whereas the assigned appointment days are binding. This may require some "reshuffling" of the existing and dynamically planned intraday schedules, rendering this scheduling paradigm impractical, yet idealistic. We refer to this scheduling paradigm as reoptimized intraday scheduling (ROIS).

ROIS is founded upon the optimal static intraday schedules constrained on the total number of patients. Let $\mathbf{r}_b \in \arg\min\{\texttt{intra}(\mathbf{x}) : \mathbf{x} \in \mathbb{Z}_+^n, |\mathbf{x}| = b\}$ be an optimal intraday schedule with b scheduled patients, $b \in \mathbb{Z}_+$. When $\texttt{intra}(\cdot)$ is multimodular, we can compute $\mathbf{r}_b$ efficiently as shown in our Theorem 3. Theorem 3 builds upon discrete convexity results from Altman et al. (2000), Murota (2004, 2005), Kaandorp and Koole (2007), and Zacharias and Yunes (2020).

Let $g : \mathbb{Z}_+^n \to \mathbb{R}$ be a multimodular function and $\mathcal{M}_b = \{\mathbf{x} \in \mathbb{Z}_+^n : |\mathbf{x}| = b\}$ for some $b \in \mathbb{Z}_+$. According to Kaandorp and Koole (2007), Algorithm 1 terminates with a minimizer of g over $\mathcal{M}_b$.

Algorithm 1 (Kaandorp and Koole 2007; Minimization of a Multimodular Function g on $\mathcal{M}_b$)

1: define $\mathcal{E} \triangleq \{-\mathbf{e}_1, \mathbf{e}_1 - \mathbf{e}_2, \mathbf{e}_2 - \mathbf{e}_3, \ldots, \mathbf{e}_{n-1} - \mathbf{e}_n, \mathbf{e}_n\}$
2: pick an $\mathbf{x} \in \mathcal{M}_b$
3: find $\mathcal{V}^* \in \arg\min\{g(\mathbf{x} + \sum_{\mathbf{v}\in\mathcal{V}}\mathbf{v}) : \mathcal{V} \subseteq \mathcal{E}, \mathbf{x} + \sum_{\mathbf{v}\in\mathcal{V}}\mathbf{v} \in \mathcal{M}_b\}$ and set $\mathbf{x}^* = \mathbf{x} + \sum_{\mathbf{v}\in\mathcal{V}^*}$
4: if $g(\mathbf{x}) \le g(\mathbf{x}^*)$, then stop ($\mathbf{x}$ is a minimizer of g over $\mathcal{M}_b$)
5: set $\mathbf{x} \leftarrow \mathbf{x}^*$ and go to step 3

Step 2 in Algorithm 1 involves exhaustive local search over a discrete neighborhood of size up to 2^n integer vectors; a task with exponential complexity. However, we prove in Theorem 3 that we can perform local search in polynomial time via submodular set-function minimization over ring families of sets. Let $f : \mathbb{Z}_+^n \to \mathbb{R} : \mathbf{x} \longmapsto g(Q\mathbf{x})$, where

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & & & \ddots & \ddots & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

According to lemma 2.1 in Murota (2005), f is $\mathrm{L}^\natural$-convex. The problem $\min_{\mathbf{x}\in\mathcal{M}_b} g(\mathbf{x})$ is equivalent to $\min_{\mathbf{y}\in\mathcal{L}_b} f(\mathbf{y})$, where $\mathcal{L}_b = \{\mathbf{y} \in \mathbb{Z}_+^n : 0 \le y_1 \le y_2 \le \ldots \le y_n = b\}$. In particular, for $b \in \mathbb{Z}_+$ and $\mathbf{x} \in \mathcal{M}_b$, step 2 in Algorithm 1 is

equivalent to

$$\min\{f(\mathbf{y}+\epsilon\mathbf{e}_Y) : \epsilon \in \{-1,1\}, Y \subseteq \{1,2,\ldots,n\}, \quad \mathbf{y}+\epsilon\mathbf{e}_Y \in \mathcal{L}_b\}, \qquad (\mathbf{P}_{\mathbf{y},b}),$$

where $\mathbf{y} = Q^{-1}\mathbf{x} \in \mathcal{L}_b$ and $\mathbf{e}_Y = (\mathbb{1}_{1\in Y}, \mathbb{1}_{2\in Y}, \ldots, \mathbb{1}_{n\in Y}) \in \{0,1\}^n$ is the characteristic vector of some $Y \subseteq \{1,2,\ldots,n\}$. According to Murota (2004), problem $(\mathbf{P}_{\mathbf{y},b})$ involves constrained minimization of two submodular set functions

$$\begin{aligned} \min\ & \rho^+_{\mathbf{y},b}(Y) \triangleq f(\mathbf{y}+\mathbf{e}_Y) \\ \text{s.t. } & Y \subseteq \{1,2,\ldots,n\} \\ & \mathbf{y}+\mathbf{e}_Y \in \mathcal{L}_b \end{aligned} \qquad (\mathbf{P}^+_{\mathbf{y},b})$$

and

$$\begin{aligned} \min\ & \rho^-_{\mathbf{y},b}(Y) \triangleq f(\mathbf{y}-\mathbf{e}_Y) \\ \text{s.t. } & Y \subseteq \{1,2,\ldots,n\} \\ & \mathbf{y}-\mathbf{e}_Y \in \mathcal{L}_b. \end{aligned} \qquad (\mathbf{P}^-_{\mathbf{y},b})$$

The best solution between $(\mathbf{P}^+_{\mathbf{y},b})$ and $(\mathbf{P}^-_{\mathbf{y},b})$ solves $(\mathbf{P}_{\mathbf{y},b})$.

Theorem 3. *Problems $(\mathbf{P}^+_{\mathbf{y},b})$ and $(\mathbf{P}^-_{\mathbf{y},b})$ can be solved in polynomial time via unconstrained submodular set-function minimization for all $b \in \mathbb{Z}_+$ and $\mathbf{y} \in \mathcal{L}_b$.*

Theorem 3 is the constrained counterpart of theorem 7 in Zacharias and Yunes (2020), in which they address unconstrained minimization of multimodular functions over the whole set of vectors in $\mathbb{Z}^n_+$. The number of scheduled patients in Zacharias and Yunes (2020) is an outcome of optimization as opposed to being the fixed parameter b in our problem $(\mathbf{P}_{\mathbf{y},b})$. They prove that the problem can be solved in polynomial time via constrained submodular set-function minimization over two (properly constructed) ring families of subsets, both families stemming from the common ground set $\{1,2,\ldots,n\}$.

In contrast to Zacharias and Yunes (2020), Theorem 3 addresses nonnegative multimodular function minimization over $\mathbb{Z}^+_n$ subject to the additional constraint that $|\mathbf{x}| = b$. This problem is also equivalent to solving two constrained submodular set-function minimization problems. The two constraint sets differ from their counterparts in Zacharias and Yunes (2020), accounting for the fixed number of patients to be scheduled. Moreover, the two families of subsets in the corresponding constrained sets are stemming from two distinct ground sets. The proof of Theorem 3 prescribes how to construct the two appropriate ground sets that serve as the bases to generate the two ring families of subsets for the corresponding constrained problems $(\mathbf{P}^+_{\mathbf{y},b})$ and $(\mathbf{P}^-_{\mathbf{y},b})$. The readers interested in performing computational implementations of our methods are directed to appendix B of the e-companion to Zacharias and Yunes (2020), in which they can follow the detailed mathematical and algorithmic steps, adjusted accordingly to the proper ground sets that we prescribe in our Theorem 3.

Once we identify $\mathbf{r}_b$, we define the corresponding single-variable intraday cost as a piecewise linear function with integer break points

$$\overline{\texttt{intra}}^r(b) \triangleq \begin{cases} \texttt{intra}(\mathbf{r}_b) & \text{if } b \in \mathbb{Z}_+ \\ (b-\lfloor b\rfloor)\texttt{intra}(\mathbf{r}_{\lfloor b\rfloor}) + (\lceil b\rceil - b)\texttt{intra}(\mathbf{r}_{\lceil b\rceil}) & \\ & \text{if } b \in \mathbb{R}_+\backslash\mathbb{Z}_+. \end{cases}$$

As indicated in Corollary 1, $\overline{\texttt{intra}}^r(.)$ is convex. In fact, we establish a more general result in Lemma 3.

Lemma 3. *Let $g : \mathbb{Z}^n_+ \to \mathbb{R}$ be a multimodular function. The function $h(b) = \min_{\mathbf{x}\in\mathcal{M}_b} g(\mathbf{x})$ is discretely convex in b on $\mathbb{Z}_+$.*

The proof of Lemma 3 involves the continuous extension of a multimodular function presented in Altman et al. (2000, section 2.2). This extension is locally linear on the convex hull formed by neighboring extreme-point integer vectors. We leveraged this extension to apply a classic result on minimizing continuous convex functions on $\mathbb{Z}^n_+$ with constrained L^1-norms.

In this section, thus far, we treat the total number of patients in a daily schedule b as a parameter (the right-hand side of an equality constraint) in a static integer program that involves constrained multimodular function minimization, $\min_{\mathbf{x}\in\mathcal{M}_b} g(\mathbf{x})$. According to Theorem 3, the constrained minimization problem can be solved in polynomial time, and according to Lemma 3, the minimal objective function is discretely convex in b. Theorem 3 and Lemma 3 are stand-alone results, independent of our model and underlying problem yet motivated by our pursuit of addressing (1). To the best of our knowledge, these results are novel. They relate to the theory of discrete optimization and its applications within and beyond the field of appointment scheduling. More pertinent to this paper, they are instrumental in our development of a computationally efficient lower bound to (1) as shown subsequently. Based on these results, we establish a theoretically guaranteed optimality gap for SRIS and evaluate its performance.

Corollary 1. $\overline{\texttt{intra}}^r(\cdot)$ *is convex on $\mathbb{R}_+$.*

From Corollary 1, the results of Sections 3.2 and 3.3 hold when we set $\overline{\texttt{intra}}(\cdot) = \overline{\texttt{intra}}^r(\cdot)$. Let $\bar{\pi}^r$ be the corresponding optimal policy for allocation scheduling. If we allow rearrangement of patients within each

intraday schedule as it evolves dynamically with new patients so that each intraday schedule is equal to $\mathbf{r}_b$ for some $b \in \mathbb{Z}_+$, then the problem can be reduced to sequential allocation scheduling. More specifically, consider the relaxation of (1) defined as

$$V_t^r(\mathbf{X}_t, d_t) = \min_{\mathbf{Z}_t \in \mathcal{Z}^r(\mathbf{X}_t, d_t)} \{C(\mathbf{Z}_t) + \gamma \mathbb{E}_{d_{t+1}}[V_{t+1}^r(\zeta(\mathbf{Z}_t), d_{t+1})]\}, \tag{7}$$

where $\mathcal{Z}_t^r(\mathbf{X}_t, d_t) = \{\mathbf{Z}_t \in \mathcal{X} : \theta(\mathbf{Z}_t) \geq \theta(\mathbf{X}_t), |\mathbf{Z}_t| = |\mathbf{X}_t| + d_t\} \supseteq \mathcal{Z}(\mathbf{X}_t, d_t)$ and $\theta(\mathbf{X}_t) = (|\mathbf{x}_{t1}|, |\mathbf{x}_{t2}|, \ldots)$. We note that the difference between (1) and (7) is in the respective constraints $\mathbf{Z}_t \geq \mathbf{X}_t$ and $\theta(\mathbf{Z}_t) \geq \theta(\mathbf{X}_t)$. The relaxed constraint allows the scheduler to reoptimize the existing appointment times (but not days) as the system evolves dynamically, rendering the solution to (7) potentially infeasible for (1). We characterize analytically a computationally tractable solution to (7) in Theorem 4.

Theorem 4. *Assume that $\mathbf{X}_1 = \mathbf{0}$ and that the horizon is infinite. There exists an optimal policy $\mathbf{Z}_t^r$ for the joint interday and intraday scheduling problem in (7) for all t such that*

$$\mathbf{Z}_t^r(\mathbf{X}_t, d_t) = \begin{bmatrix} \mathbf{r}_{z_{t1}^{\tilde{\pi}^r}(\theta(\mathbf{X}_t), d_t)} \\ \mathbf{r}_{z_{t2}^{\tilde{\pi}^r}(\theta(\mathbf{X}_t), d_t)} \\ \vdots \\ \mathbf{r}_{z_{t\tau}^{\tilde{\pi}^r}(\theta(\mathbf{X}_t), d_t)} \\ \vdots \end{bmatrix},$$

where $\theta(\mathbf{X}_t) = (|\mathbf{x}_{t1}|, |\mathbf{x}_{t2}|, \ldots)$.

6. Contrast of SRIS and ROIS, Implications, and Performance Guarantees

By construction and as illustrated in Figure 1, $\overline{\texttt{intra}}^s(\cdot) \geq \overline{\texttt{intra}}^r(\cdot)$, the two functions have the same minimizer $\bar{b}$, and $\overline{\texttt{intra}}^s(\bar{b}) = \overline{\texttt{intra}}^r(\bar{b})$. Both functions are convex (Lemma 2 and Corollary 1), they appear to be almost identical away from their common minimizer $\bar{b}$, and they demonstrate slightly notable differences around $\bar{b}$. Moreover, Figure 1 signifies that the commonalities between the two functions are not sensitive to the variance of the service time distribution or the no-show probability.

In Figure 2, we display a sample comparison between the corresponding sequences of intraday schedules $(\mathbf{s}_b)_{b \in \mathbb{Z}_+}$ and $(\mathbf{r}_b)_{b \in \mathbb{Z}_+}$. In this particular example, the optimal number of patients to accommodate in both sequences is $\bar{b} = 18$. When additional patients need to be accommodated in a schedule because of potential dynamic demand spikes and as prescribed by Theorems 2 and 4, the schedules become denser. In extreme situations of cumbersome appointment backlogs, excessive overbooking appears at the very last slot of an intraday schedule. Such extreme scenarios exist in theory as a means to construct well-defined intraday cost functions with unbounded domains. However, because of the convexity of the intraday cost function and the dynamic containment of the appointment backlog adhering to Lemma 1(b), such extreme schedules are quite unlikely to appear in practice given that the average daily demand volume is within the same order of magnitude as $\bar{b}$. In a system in which the demand volume is an order of magnitude larger than $\bar{b}$, then inevitably, the appointment backlog grows problematically large to overwhelming levels, and as a result, patients endure long appointment delays and/or an overcrowded wait room. Healthcare providers can avoid such situations by regulating their demand volume and/or adjusting their intraday capacity (translated as increased $\bar{b}$ in our model). Green and Savin (2008), Liu (2016), and Zacharias and Armony (2017) are some recent studies that provide methods and insights to inform such strategic-level operational design.

The sequence $(\mathbf{s}_b)_{b \in \mathbb{Z}_+}$ is sequentially refinable and thereby can be used as a basis to generate a feasible heuristic solution (SRIS as described in Theorem 2) to the joint interday and intraday problem. Whereas the sequence $(\mathbf{r}_b)_{b \in \mathbb{Z}_+}$ does not possess this feature, we can leverage its properties to bound from below the value function in (1) and thereby assess the performance of SRIS. We state and prove mathematically this implication in Proposition 1 that follows.

Proposition 1. *Assume that $\mathbf{X}_1 = \mathbf{0}$. Then, $V_t^r(\cdot,\cdot) \leq V_t(\cdot,\cdot) \leq V_t^s(\cdot,\cdot)$ for all t.*

As a direct consequence of Proposition 1, the heuristic solution SRIS has a theoretically guaranteed and computationally efficient upper bound on its optimality gap.

Corollary 2. *Assume that $\mathbf{X}_1 = \mathbf{0}$. The percentage optimality gap of SRIS is bounded from above by $\frac{V_t^s(\cdot,\cdot) - V_t^r(\cdot,\cdot)}{V_t^r(\cdot,\cdot)} \times 100\%$.*

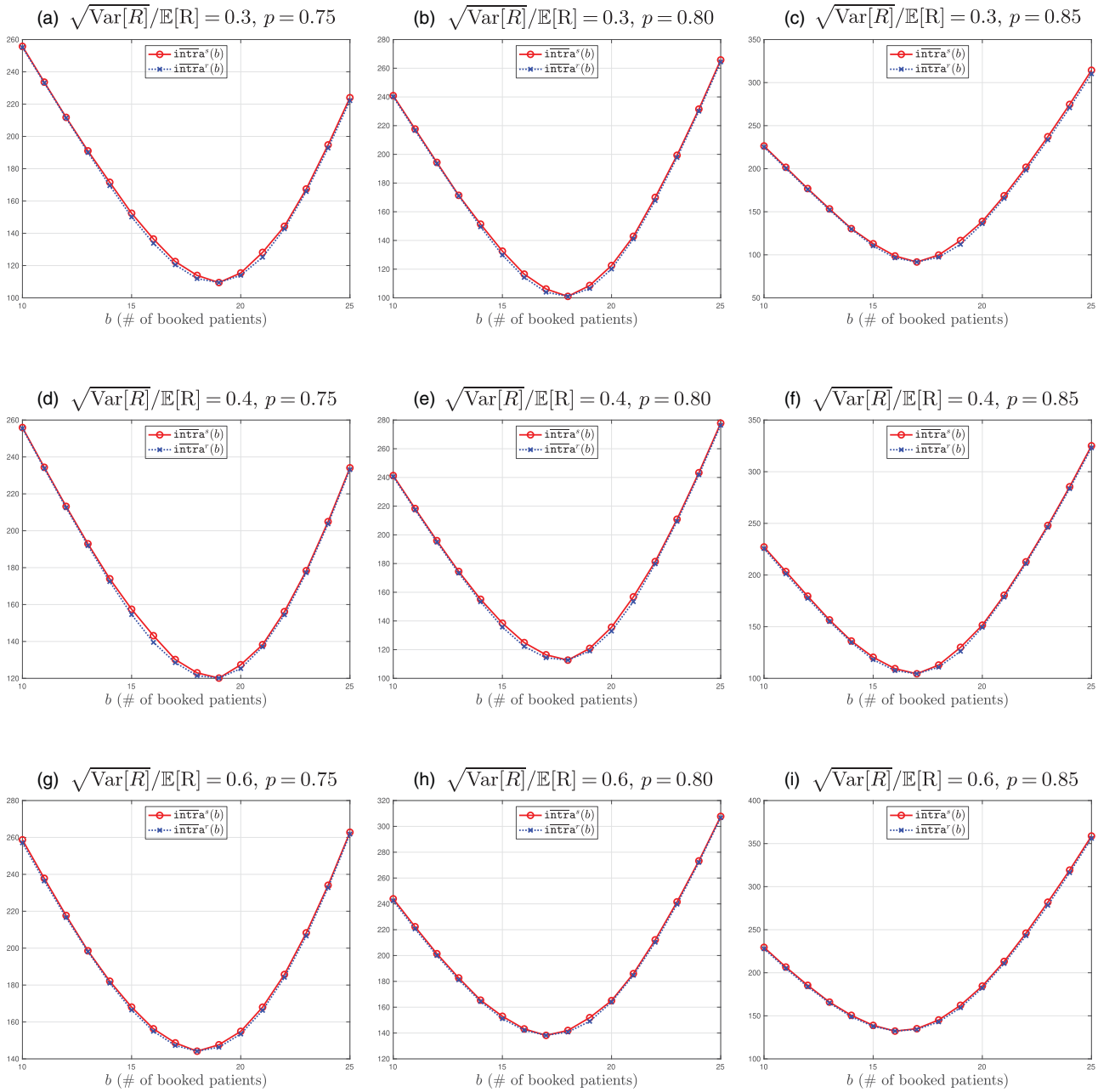
7. Implementation and Computational Experiments

Equipped with tractable analytical tools to tackle the joint dynamic interday and intraday problem, in this section, we present computational implementations of our methods that assess their practical merit and expose additional managerial insights.

7.1. Dynamic Programming Implementation

First, we discuss how we implemented the dynamic scheduling paradigms SRIS and ROIS. The allocation

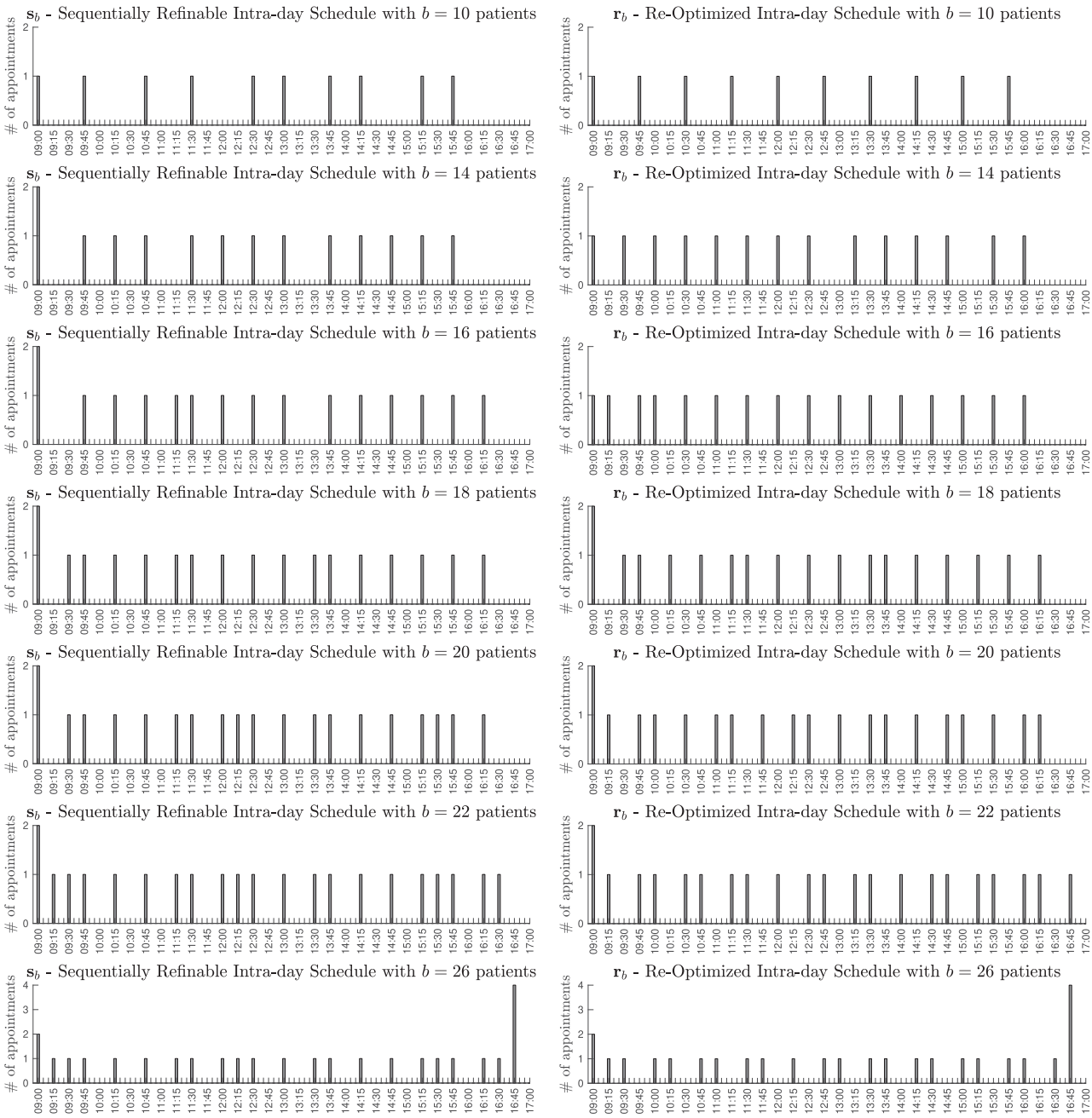
Figure 1. Contrast of $\mathtt{intra}^s(\cdot)$ and $\mathtt{intra}^r(\cdot)$ with Stochastic Service Times R and Show-up Probabilities p



Notes. $\mathtt{intra}(\cdot)$ as in Appendix B. The model input for $\mathtt{intra}(\cdot)$, using the notation in Appendix B, is $N = 480$ minutes (eight hours, 9 a.m.–5 p.m.), $k = 15$ minutes, $c_i = 1$, $c_o = 1$, $c_s = 0.1$, $\mathbf{U} = \mathbf{0}$, $R \sim$ beta-binomial with $\mathbb{E}[R] = 30$ minutes and support on $\{0, 1, 2, \ldots, 90\}$.

scheduling problem has a countably infinite state space. Even though an optimal stationary policy exists (see Bertsekas 2000, section 4.1), it is not possible to determine the optimal value functions and optimal controls via standard dynamic programming methods (see Bertsekas 2000, section 2.7). We solve finite-state approximations of the problem with a wait-list up to 1,000 patients by implementing the *policy-space* method proposed in White (1979, 1982). The policy-space method is an iterative process. In iteration n, the value function is approximated for states $\{0, 1, 2, \ldots, n\}$ based on the results of iteration $n-1$. For more details on this iterative methodology, the reader is referred to the original paper as well as its more recent exposition in Lee et al. (2017). Finally, we note that solving one instance of the interday and intraday scheduling problem in the scale we consider in this section takes only a few minutes. However, solving the original MDP

Figure 2. Sample Comparison Between Intraday Schedules $\mathbf{s}_b$ and $\mathbf{r}_b$



Notes. $\texttt{intra}(\cdot)$ as in Appendix B. The model input for $\texttt{intra}(\cdot)$, using the notation in Appendix B, is $N = 480$ minutes (eight hours, 9 a.m.–5 p.m.), $k = 15$ minutes, $p = 0.8$, $c_i = 1$, $c_o = 1$, $c_s = 0.1$, $\mathbf{U} = \mathbf{0}$, $R \sim$ beta-binomial with $\mathbb{E}[R] = 30$ minutes, $\sqrt{\mathrm{Var}[R]}/\mathbb{E}[R] = 0.4$ and support on $\{0, 1, 2, \ldots, 90\}$.

without resorting to our approaches and theoretical findings is computationally intractable because of the curse of dimensionality.

7.2. Model Input

The intraday cost function $\texttt{intra}(\cdot)$ quantifies the operational cost incurred under some intraday schedule, and it is assumed in our analytical model to be some nonnegative multimodular function. In our computational experiments, we incorporated the class of intraday cost functions introduced in Zacharias and Yunes (2020), defined as a weighted sum of expected direct delays experienced by patients with appointments, expected direct delays experienced by walk-in patients, expected provider overtime, and expected provider idle time. This class of $\texttt{intra}(\cdot)$ functions is the

outcome of stochastic analysis in the transient state and is proven to be multimodular under general stochastic service times, no-shows, walk-ins, and heterogeneous waiting cost coefficients. It is versatile, and it has nine distinct inputs (random variables, parameters, cost coefficients) that can be tailored to describe a variety of clinical environments.

The impact, sensitivity analyses, and managerial implications of static intraday scheduling models are analyzed extensively in the literature (see Wang et al. 2020, Zacharias and Yunes 2020 for two recent studies). In our computational experiments, we focus on the trade-off between direct and indirect delays and the corresponding interplay between interday and intraday scheduling. Accordingly, we consider a fixed input regarding the intraday components of the problem. In particular, consultation times follow a beta-binomial distribution with average 30 minutes, coefficient of variation 0.4, and support on $\{0, 1, \ldots, 90\}$ (see Zacharias and Yunes 2020 for a discussion about the qualities and appropriateness of this distribution). An eight-hour workday is partitioned into 32 15-minute slots. There is a no-show rate of 20%, and we assume no walk-in patients. Regarding the cost coefficients, following the literature, the idle time cost coefficient is normalized to one, overtime is equally as costly as idle time, and the waiting cost coefficient is equal to 0.1. In this environment, the optimal static intraday schedule accommodates 18 patients; see Figures 1 and 2.

In our computational study, the appointment delay cost coefficient c_a takes 49 different values ranging from 0 up to 12 with increments of 0.25. For example, when $c_a = 3$, a patient experiencing a direct delay of 30 minutes is equally as costly as a patient experiencing an indirect delay of one day. In the extreme case of $c_a = 0$, indirect delays are not taken under consideration in the scheduling decisions. The case of $c_a = 12$ corresponds to the other extreme in which a patient experiencing a direct delay of two hours is equally as costly as a patient experiencing an indirect delay of one day. This range of c_a is aligned well with the empirical findings in Liu et al. (2018) that a direct delay of 45 minutes is roughly equivalent to an indirect delay of one week from the patients' perspective.

Daily requests for appointments follow a Poisson(λ) distribution with λ in $\{5, 6, \ldots, 30\}$, a set that contains the total number of patients accommodated by the optimal static intraday schedule. The Poisson distribution is empirically and theoretically justified to be a good model for stochastic arrivals in service systems. In the context of outpatient scheduling, we refer to Green et al. (2007), Robinson and Chen (2010), Liu et al. (2010), and Zacharias and Armony (2017) for justifications about the appropriateness of the Poisson distribution to model daily requests for appointments. A commonly used and intuitive argument is that a panel of patients on the order of 1,000s with each patient requesting daily and independently an appointment with some small probability generates binomially distributed daily demand, which can be approximated well by a Poisson distribution.

To recap, in our numerical experiments, we analyzed 1,274 distinct instances of the problem based on different values of c_a and λ. We focused on the impact of the demand volume and the relative importance of the waiting cost coefficients and the corresponding interplay between inter/intraday scheduling. The discount factor γ was set to 97.5%.

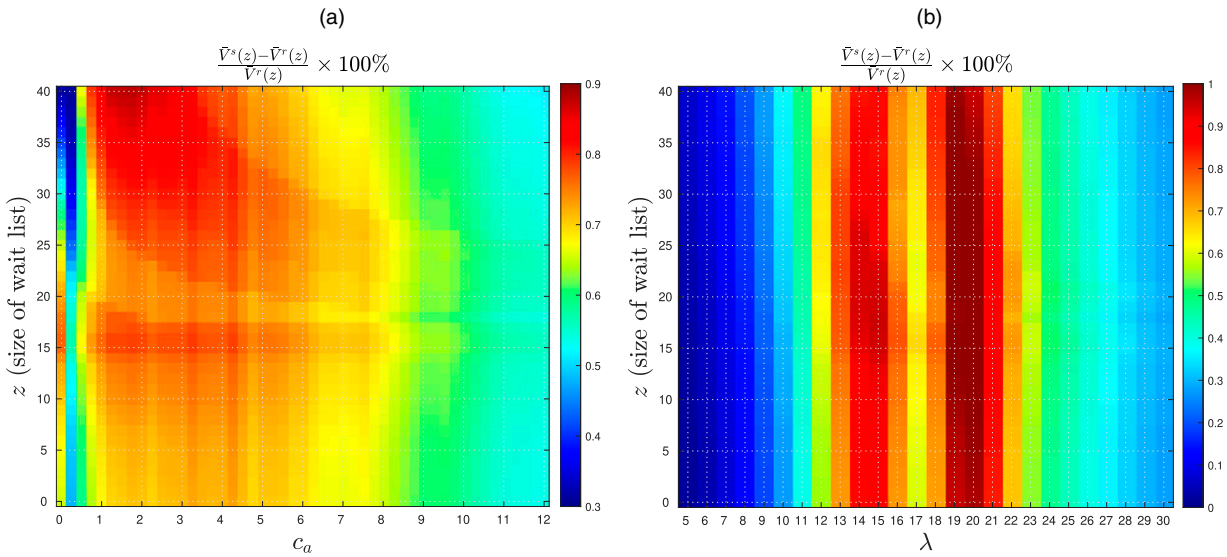
7.3. SRIS Performance Evaluation

Recall that SRIS allows us to obtain a feasible solution to the interday and intraday scheduling problem and that the value functions of ROIS and SRIS (respectively) provide lower and upper bounds on the value function in (1); see Proposition 1. We analyzed numerically the performances of SRIS and ROIS and thereby the optimality gap of SRIS. The optimality gap of SRIS is consistently less than 1% across all 1,274 instances, suggesting that it is an effective (besides practically implementable) heuristic solution. The two heat maps in Figure 3 display a sample of our analysis on the optimality gap. Figure 3, (a) and (b), depicts the optimality gap as a function of the interday cost c_a and arrival rate λ, respectively. Besides the optimality gap being less than 1%, we observe that it does not vary significantly with the state of the system (size of the wait-list), as the heat maps do not exhibit notable vertical variations in color. We point out that the optimality gap is slightly larger for values of λ around 18 (the optimal number of patients for the static intraday problem) in agreement with Figure 1.

Moreover, Figure 3 demonstrates that the idealistic scheduling paradigm ROIS only marginally outperforms SRIS. One key takeaway is that, when equipped with an informed and methodically crafted scheduling template, there is negligible value in maintaining the flexibility to reoptimize and reassign appointment times at the last possible moment to match their optimal static counterpart (besides such practice being impractical).

It is noteworthy to point out that the heat map in Figure 3(a) demonstrates an irregular "discontinuity of colors" for values of c_a between 0 and 0.5. One plausible explanation is the following. When $c_a = 0$, the system prioritizes dynamically the delivery of optimal intraday schedules, disregarding indirect delays imposed on backlogged patients. When c_a transitions from 0 to 0.25, a remarkable paradigm shift takes place, in which the optimal policy takes under consideration the interday dynamics, penalizes appointment delays, and potentially compromises slightly the optimal intraday input resulting from a model in which $c_a = 0$.

Figure 3. Upper Bound on the Percentage Optimality Gap of SRIS



Notes. (a) $\lambda = 18$, c_a varies. (b) $c_a = 3$, λ varies.

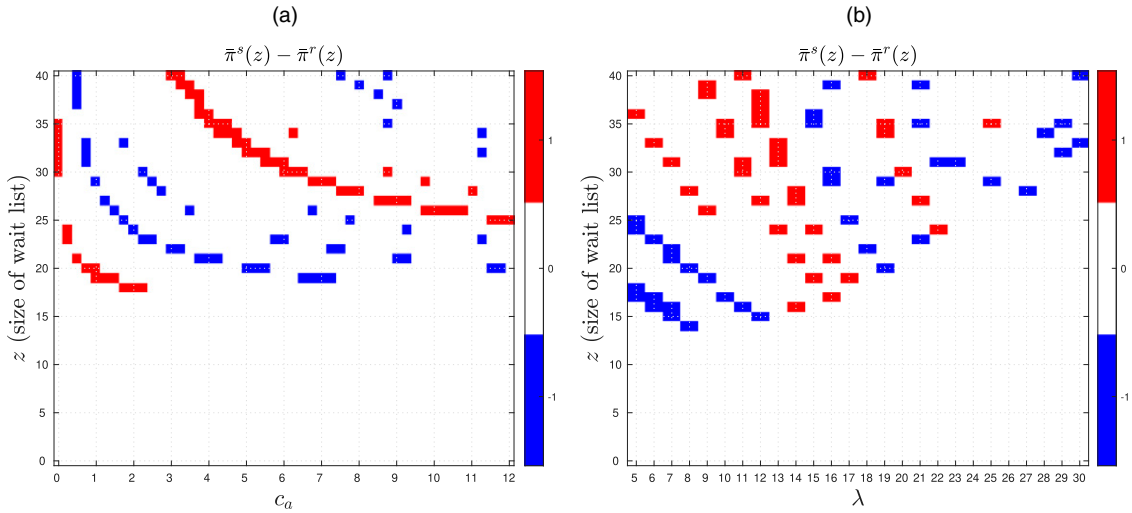
7.4. SRIS vs. ROIS

SRIS and ROIS, by construction, generate different intraday schedules for the joint problem except at $\bar{b}$, where $\mathbf{s}_{\bar{b}} = \mathbf{r}_{\bar{b}}$. At a higher level, we are also interested to see how they differ in solving the corresponding interday problem (3). In our next experiment, we investigate the difference between their optimal controls $\bar{\pi}^s(\cdot)$ and $\bar{\pi}^r(\cdot)$. Figure 4 visualizes a sample of our results: (a) as a function of the size of wait-list z and the interday cost c_a and (b) as a function of the size of wait-list z and the arrival rate λ. We observe that the two policies are identical in the majority of the parameter/state space represented by the white area of the heat maps. For the rest of the parameter/state space, their absolute difference is one. The similarities of the two policies can be attributed to the commonalities between $\overline{\texttt{intra}}^s(\cdot)$ and $\overline{\texttt{intra}}^r(\cdot)$, see Figure 1, and can explain, to a certain extent, the remarkably small optimality gap of SRIS.

7.5. Sensitivity Analysis of SRIS

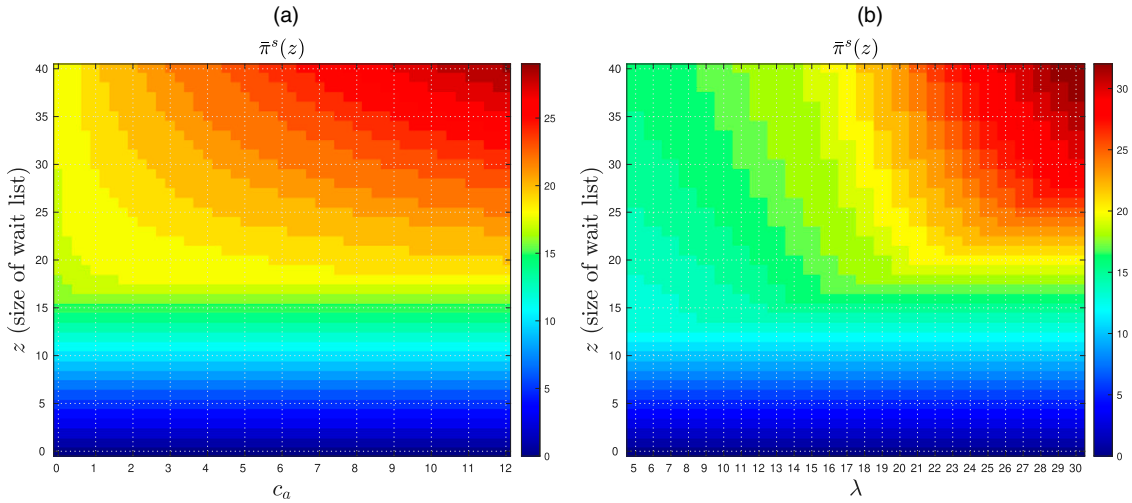
In Figure 5, we visualize with two heat maps what an SRIS optimal allocation policy looks like: (a) as a function of the size of wait-list z and the interday cost c_a and (b) as a function of the size of wait-list z and the arrival rate λ. In both heat maps, we observe that, for

Figure 4. Difference Between Optimal Allocation Policies for SRIS and ROIS



Notes. (a) $\lambda = 18$, c_a varies. (b) $c_a = 3$, λ varies.

Figure 5. Optimal Allocation Policy for SRIS



Notes. (a) $\lambda = 18$, c_a varies. (b) $c_a = 3$, λ varies.

small values of z, the optimal allocation policy offers to all patients a next-day appointment, whereas for larger values of the wait-list, the optimal allocation policy increases in a concave manner (see Lemma 1). We also observe that, besides the size of the wait-list, both the anticipated future demand volume (captured by λ) and the relative importance of indirect delay (captured by c_a) have an impact on the optimal policy and should be taken under consideration. Intuitively, as either c_a or λ increase, the optimal allocation policy increases as well.

8. Conclusion

Appointment scheduling systems are commonly used by healthcare providers to manage capacity and handle patient demand effectively. An informed scheduling strategy, which responds dynamically to fluctuations in patient demand, not only improves access to care, but also reduces variability in day-to-day operations and boosts productivity. Scheduling an appointment typically entails determining dynamically the specific date (interday scheduling) and time (intraday scheduling) of a patient's visit. Interday and intraday scheduling problems are related and interdependent. Despite the tremendous growth of the appointment scheduling literature in the past few decades, no previous study has analytically tackled the joint interday and intraday scheduling problem, which had remained an open area of research with significant practical implications. Our paper fills this critical gap in the literature and provides the first analytical model and optimization platform that can be applied by healthcare professionals to manage their patient scheduling dynamically.

We make contributions to modeling, methodology, and theory of dynamic appointment scheduling. We build theoretical connections between two independently established streams of literature on appointment scheduling (interday and intraday scheduling). We prove novel theoretical results in discrete convex analysis regarding constrained multimodular function minimization. We leverage these results and our dynamic programming framework to develop an informed heuristic solution SRIS. Based on theoretical and computationally tractable performance guarantees, SRIS is shown to be nearly optimal. Besides establishing a rigorous and practically implementable approach to an open problem, our analysis bears important managerial implications. Notably, we demonstrate numerically that a methodically crafted and easy-to-implement scheduling paradigm (such as SRIS) performs nearly as well as an idealistic solution in which the dynamic intraday schedules can be reoptimized at the last moment to match their optimal static counterpart (ROIS).

Our model is capable of handling important practical features of patient scheduling such as stochastic demand for medical services, stochastic consultation times, no-shows, and walk-ins. It captures the tradeoffs involved in utilizing valuable resources efficiently and providing timely access to care. It is versatile and can be tailored to describe a variety of medical practice environments. Our scheduling methods and findings relate to healthcare professionals who manage appointment-based services such as primary care, dental care, pediatrics, diagnostic tests, and surgical departments.

We conclude with future directions of interest and/or limitations of this study. In lieu of known probability distributions for the problem's stochastic elements, one

can make use of observed data and develop a sampling-based approach, such as in Begen et al. (2012). Consideration of heterogeneous patients is another direction for future research, and it bears a challenging aspect to the problem, namely, the sequencing of patients. It is also of interest to consider how no-shows and walk-ins are affected by system congestion as opposed to being homogeneous across all dynamically evolving states. Dynamic optimization of appointment systems that manage schedules for a pool of multiple providers is another future direction of interest (see, e.g., Zacharias and Pinedo 2017, Soltani et al. 2019, Kuiper and Lee 2022 for some recent studies addressing the corresponding static intraday problem). Additional challenging aspects of the problem, not addressed by our methods, include patient preferences, cancellations, nonpunctuality, balking behavior, and strategic interactions between providers and patients.

Acknowledgments

The authors are deeply grateful to the editor in chief, area editor, associate editor, and two referees for a thoughtful and constructive review process.

Appendix A. Proofs

Proof of Lemma 1.

(i) Proof by induction. Let $T < \infty$. Then, $\bar{V}_{T+1}(\cdot) = 0$ is trivially convex. The inductive hypothesis is that $\bar{V}_{t+1}(\cdot)$ is convex, based on which we need to prove that $\bar{V}_t(\cdot)$ is also convex. Let $\bar{G}_t(b_t, z_t) = \bar{C}(b_t, z_t) + \gamma\mathbb{E}[\bar{V}_{t+1}(z_t - b_t + d_{t+1})]$ be the cost function to be minimized in period t. Then, $\bar{V}_t(\cdot) = \min_{b_t \in [0, z_t]} \bar{G}(b_t, z_t)$ is also convex because $\bar{G}(b_t, z_t)$ is jointly convex in (b_t, z_t) and $[0, z_t]$ is a convex set; see Boyd and Vandenberghe (2004, chapter 3). In order to see that $\bar{G}(b_t, z_t)$ is jointly convex in (b_t, z_t) on $\{(b, z) : 0 \le b \le z\}$, first note that $\bar{C}(b_t, z_t) = \overline{\texttt{inter}}(z_t) + \overline{\texttt{intra}}(b_t)$ is separable and both $\overline{\texttt{inter}}(\cdot)$ and $\overline{\texttt{intra}}(\cdot)$ are convex. Then, let $0 \le b \le z$, $0 \le b' \le z'$, $\lambda \in [0, 1]$, and define $f(b, z) = \bar{V}_{t+1}(z - b)$.

$$\begin{aligned} f(\lambda(b,z) + (1-\lambda)(b',z')) &= f(\lambda b + (1-\lambda)b', \lambda z + (1-\lambda)z) \\ &= \bar{V}_{t+1}(\lambda(z-b) + (1-\lambda)(z'-b')) \\ &\le \lambda \bar{V}_{t+1}(z-b) + (1-\lambda)\bar{V}_{t+1}(z'-b') \\ &= \lambda f(b,z) + (1-\lambda) f(b',z'). \end{aligned}$$

The infinite horizon problem has the same properties by taking the limit as $T \to \infty$. □

(ii) From the proof of theorem 3 in Truong (2015), it suffices to show that $\bar{G}_t(\cdot, \cdot)$ is submodular. First note that $\bar{C}(b_t, z_t) = \overline{\texttt{inter}}(z_t) + \overline{\texttt{intra}}(b_t)$ is separable and, therefore, submodular. Moreover, $\bar{V}_{t+1}(z_t - b_t + d)$ is submodular in (b_t, z_t) because $\bar{V}_{t+1}(\cdot)$ is convex. □

Proof of Lemma 2.

First, we show that $\overline{\texttt{intra}}^s(\cdot)$ is discretely convex on $\mathbb{Z}_+$. Let $b \in \mathbb{Z}_+$. Then, $\mathbf{s}_{b+1} = \mathbf{s}_b + \mathbf{e}_i$ and $\mathbf{s}_{b+2} = \mathbf{s}_b + \mathbf{e}_i + \mathbf{e}_j$ for some $i, j \in \{1, 2, \ldots, n\}$. It suffices to show that

$$\begin{aligned} &\overline{\texttt{intra}}^s(b+2) - \overline{\texttt{intra}}^s(b+1) \ge \overline{\texttt{intra}}^s(b+1) - \overline{\texttt{intra}}^s(b) \\ \Leftrightarrow\ &\texttt{intra}(\mathbf{s}_{b+2}) - \texttt{intra}(\mathbf{s}_{b+1}) \ge \texttt{intra}(\mathbf{s}_{b+1}) - \texttt{intra}(\mathbf{s}_b) \\ \Leftrightarrow\ &\texttt{intra}(\mathbf{s}_b + \mathbf{e}_i + \mathbf{e}_j) + \texttt{intra}(\mathbf{s}_b) \ge 2\,\texttt{intra}(\mathbf{s}_b + \mathbf{e}_i), \end{aligned}$$

which is true because, from the directional convexity of $\texttt{intra}(\cdot)$, see Zacharias and Yunes (2020), we get

$$\begin{aligned} \texttt{intra}(\mathbf{s}_b + \mathbf{e}_i + \mathbf{e}_j) + \texttt{intra}(\mathbf{s}_b) &\ge \texttt{intra}(\mathbf{s}_b + \mathbf{e}_i) + \texttt{intra}(\mathbf{s}_b + \mathbf{e}_j) \\ &\ge 2\,\texttt{intra}(\mathbf{s}_b + \mathbf{e}_i). \end{aligned}$$

Then, $\overline{\texttt{intra}}^s(\cdot)$ is convex on $\mathbb{R}_+$ as it is the piecewise linear extension of a discretely convex function. □

Proof of Theorem 2.

Assume that $\mathbf{X}_1 = \mathbf{0}$. First note that, if $\mathbf{X}_t \in \mathcal{X}^s$, then also $\zeta(\mathbf{X}_t) \in \mathcal{X}^s$. Because $\mathbf{0} \in \mathcal{X}$ and $\mathcal{X}^s \subseteq \mathcal{X}$, there exists a policy such that $\mathbf{Z}_t \in \mathcal{Z}(\mathbf{X}_t, d_t) \cap \mathcal{X}^s$, and therefore, $V_t^s(\mathbf{X}_t, d_t)$ is finite for every t. On the other hand, if a policy is such that $\mathbf{Z}_t \notin \mathcal{X}^s$ for some t, then $V_t^s(\mathbf{X}_t, d_t) = \infty$. Therefore, there exists an optimal policy such that $\mathbf{X}_t, \mathbf{Z}_t \in \mathcal{X}^s$ for all t.

Now, let $\mathbf{X}_t \in \mathcal{X}^s$. Row $\mathbf{x}_{t\tau}$ of $\mathbf{X}_t$ is equal to $\mathbf{s}_{|\mathbf{x}_{t\tau}|}$, and the problem

$$\begin{aligned} &V_t^s(\mathbf{X}_t, d_t) \\ &= \min_{\mathbf{Z}_t \in \mathcal{Z}(\mathbf{X}_t, d_t)} \{\texttt{inter}(\mathbf{Z}_t) + \texttt{intra}(\mathbf{z}_{t1}) + \gamma\mathbb{E}[V_{t+1}^s(\zeta(\mathbf{Z}_t), d_{t+1})]\} \end{aligned}$$

is equivalent to

$$\begin{aligned} \tilde{V}_t(\theta(\mathbf{X}_t), d_t) &= \min_{\mathbf{z}_t : \mathbf{z}_t \ge \theta(\mathbf{X}_t),\, |\mathbf{z}_t| = |\theta(\mathbf{X}_t)| + d_t} \{\overline{\texttt{inter}}(|\mathbf{z}_t|) \\ &\qquad + \texttt{intra}(\mathbf{s}_{z_{t1}}) + \gamma\mathbb{E}[\tilde{V}_{t+1}(\eta(\mathbf{z}_t), d_{t+1})]\} \\ &= \min_{\mathbf{z}_t : \mathbf{z}_t \ge \theta(\mathbf{X}_t),\, |\mathbf{z}_t| = |\theta(\mathbf{X}_t)| + d_t} \{\overline{\texttt{inter}}(|\mathbf{z}_t|) \\ &\qquad + \overline{\texttt{intra}}^s(z_{t1}) + \gamma\mathbb{E}[\tilde{V}_{t+1}(\eta(\mathbf{z}_t), d_{t+1})]\}, \end{aligned}$$

where $\theta(\mathbf{X}_t) = (|\mathbf{x}_{t1}|, |\mathbf{x}_{t2}|, \ldots)$. The result follows from the convexity of $\overline{\texttt{intra}}^s(\cdot)$ in Lemma 2 and from Theorem 1. □

Proof of Theorem 3.

Let $\kappa(\mathbf{y}) = \max\{i \in \{1, 2, \ldots, n\} : y^i < b\}$, $N_{\mathbf{y}}^+ = \{1, 2, \ldots, \kappa(\mathbf{y})\}$, $N^- = \{1, 2, \ldots, n-1\}$, and $\mathcal{L} = \{\mathbf{y} \in \mathbb{Z}_+^n : 0 \le y^1 \le y^2 \le \ldots \le y^n\}$. Problem $(\mathbf{P}_{\mathbf{y},\mathbf{b}}^+)$ is equivalent to $\min\{\rho_{\mathbf{y}}^+(Y) : Y \subseteq N_{\mathbf{y}}^+, \mathbf{y} + \mathbf{e}_Y \in \mathcal{L}\}$, which is solvable in polynomial time because the constraint set is a ring family of subsets of $N_{\mathbf{y}}^+$; see Zacharias and Yunes (2020, lemma EC.2). Problem $(\mathbf{P}_{\mathbf{y},\mathbf{b}}^-)$ is equivalent to $\min\{\rho_{\mathbf{y}}^-(Y) : Y \subseteq N^-, \mathbf{y} - \mathbf{e}_Y \in \mathcal{L}\}$, which is solvable in polynomial time because the constraint set is a ring family of subsets of N^-; see Zacharias and Yunes (2020, lemma EC.2). □

Proof of Lemma 3.

Let $g : \mathbb{Z}_+^n \to \mathbb{R}$ be a multimodular function. Consider its continuous extension $\tilde{g}$ on $\mathbb{R}_+^n$ as defined in Altman et al. (2000, section 2.2). Then, $\tilde{g}$ is convex on $\mathbb{R}_+^n$ and agrees with g on $\mathbb{Z}_+^n$. Define $\tilde{h}(b) = \min\{\tilde{g}(\mathbf{x}) : |\mathbf{x}| = b, \mathbf{x} \in \mathbb{R}_+^n\}$ for $b \in \mathbb{R}_+$. Then, $\tilde{h}(b)$ is convex; see Boyd and Vandenberghe (2004, example 3.17).

It remains to show that there exists an integer vector $\mathbf{x}^*$ such that $|\mathbf{x}^*| = b$ and $g(\mathbf{x}^*) = \tilde{h}(b)$ whenever b is in $\mathbb{Z}_+$. When $b = 0$, then trivially $\mathbf{x}^* = \mathbf{0}$. Let $b \ge 1$ in $\mathbb{Z}_+$ and

$\mathbf{y} \in \arg\min\{\tilde{g}(\mathbf{x}) : |\mathbf{x}| = b, \mathbf{x} \in \mathbb{R}^n_+\}$. Consider the atom $S(\mathbf{y})$ as defined in Altman et al. (2000, section 2.2). By construction, $S(\mathbf{y})$ is the convex hull of $n+1$ vectors in $\mathbb{Z}^n_+$, say, the convex hull of $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Also, by construction, the vectors $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$ can be partitioned into two subsets so that vectors in each subset have the same sum of individual components. In particular, the sum of components of the vectors within the two subsets is either b and $b-1$ or b and $b+1$. Therefore, $\mathbf{y}$ must belong to the convex hull of vectors of some subset of $\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with the sum of individual components equal to b, denoted as $\tilde{S}(\mathbf{y}) \subseteq S(\mathbf{y})$. By construction, $\tilde{g}$ is linear on $\tilde{S}(\mathbf{y})$; therefore, one of the extreme points of $\tilde{S}(\mathbf{y})$ minimizes $\tilde{g}$ on $\tilde{S}(\mathbf{y})$. □

Proof of Theorem 4. Assume that $\mathbf{X}_1 = \mathbf{0}$ and let $\mathbf{X}_t \in \mathcal{X}$. The problem

$$V^r_t(\mathbf{X}_t, d_t) = \min_{\mathbf{Z}_t \in \mathcal{Z}^r(\mathbf{X}_t, d_t)} \{\texttt{inter}(\mathbf{Z}_t) + \texttt{intra}(\mathbf{z}_{t1}) + \gamma \mathbb{E}_{d_{t+1}}[V^r_{t+1}(\zeta(\mathbf{Z}_t), d_{t+1})]\}$$

is equivalent to

$$\begin{aligned}\tilde{V}_t(\theta(\mathbf{X}_t), d_t) &= \min_{\mathbf{z}_t : \mathbf{z}_t \geq \theta(\mathbf{X}_t), |\mathbf{z}_t| = |\theta(\mathbf{X}_t)| + d_t} \{\overline{\texttt{inter}}(|\mathbf{z}_t|) \\ &\quad + \texttt{intra}(\mathbf{r}_{z_{t1}}) + \gamma \mathbb{E}_{d_{t+1}}[\tilde{V}_{t+1}(\eta(\mathbf{z}_t), d_{t+1})]\} \\ &= \min_{\mathbf{z}_t : \mathbf{z}_t \geq \theta(\mathbf{X}_t), |\mathbf{z}_t| = |\theta(\mathbf{X}_t)| + d_t} \{\overline{\texttt{inter}}(|\mathbf{z}_t|) \\ &\quad + \overline{\texttt{intra}}^r(z_{t1}) + \gamma \mathbb{E}_{d_{t+1}}[\tilde{V}_{t+1}(\eta(\mathbf{z}_t), d_{t+1})]\}.\end{aligned}$$

The result follows from the convexity of $\overline{\texttt{intra}}^r(\cdot)$ in Corollary 1 and from Theorem 1. □

Proof of Proposition 1. Because $C(\cdot) \leq C^s(\cdot)$ and $\mathcal{Z}(\cdot,\cdot) \subseteq \mathcal{Z}^r_t(\cdot,\cdot)$, it follows that $V^r_t(\cdot,\cdot) \leq V_t(\cdot,\cdot) \leq V^s_t(\cdot,\cdot)$. □

Appendix B. An Intraday Scheduling Model

The intraday cost function $\texttt{intra}(\cdot)$ quantifies the operational cost incurred given a daily appointment schedule, and it is assumed in this paper to be a general multimodular function defined over nonnegative integer vectors.

Definition B.1. A function $g : \mathbb{Z}^n_+ \to \mathbb{R}$ is said to be multimodular if

$$g(\mathbf{x}+\mathbf{v}) - g(\mathbf{x}) \geq g(\mathbf{x}+\mathbf{v}+\mathbf{w}) - g(\mathbf{x}+\mathbf{w})$$

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

$$\mathcal{E} = \{-\mathbf{e}_1, \mathbf{e}_1 - \mathbf{e}_2, \mathbf{e}_2 - \mathbf{e}_3, \ldots, \mathbf{e}_{n-1} - \mathbf{e}_n, \mathbf{e}_n\}.$$

As a companion to our computational experiments and for the reader interested in implementing our dynamic scheduling paradigms based on a class of explicitly defined multimodular intraday cost functions, in this appendix, we recapitulate the stochastic model introduced in Zacharias and Yunes (2020) and its transient analysis.

- Timescale: Time is measured in minutes, and the length of a regular workday is N minutes during which the scheduled appointments are allocated. The provider may continue to operate in overtime as well, beyond N, until all patients are served. Time is continuous, and a workday is partitioned into n discrete time slots of equal duration $k = N/n$. We assume that k is a positive integer such that n is also some positive integer. Assuming that the first slot starts at time zero, then time slot t occupies the time interval $[(t-1)k, tk)$, $t = 1, 2, \ldots, n$.
- Arrival process: There are two arrival streams: one driven by scheduled appointments (dynamic decisions in our model) and one from unscheduled walk-ins (exogenous stochastic events). 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, that is, scheduled to arrive at time $(t-1)k$. Each scheduled patient independently shows up with probability $p \in (0,1]$. The random vector $\mathbf{S}^{\mathbf{x}} = (S^{\mathbf{x}}_1, S^{\mathbf{x}}_2, \ldots, S^{\mathbf{x}}_n) \in \mathbb{Z}^n_+$ denotes the arrival process from scheduled appointments under $\mathbf{x}$, where $S^{\mathbf{x}}_t \sim \text{Binomial}(x_t, p)$ 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 walk-ins may arrive throughout the day. The arrival process from unscheduled walk-ins is an independent sequence (but not necessarily identically distributed) of random variables denoted by $\mathbf{U} = (U_1, U_2, \ldots, U_n) \in \mathbb{Z}^n_+$. The resulting arrival process from both schedule $\mathbf{x}$ and walk-ins is denoted by the random vector $\mathbf{A}^{\mathbf{x}} = (A^{\mathbf{x}}_1, A^{\mathbf{x}}_2, \ldots, A^{\mathbf{x}}_n) \in \mathbb{Z}^n_+$, where $A^{\mathbf{x}}_t = S^{\mathbf{x}}_t + U_t$ is the total number of patients that arrive right at the beginning of slot $t = 1, 2, \ldots, n$. The service discipline is first-in, first-out, and scheduled patients have priority over unscheduled walk-in patients when they show up at the same slot.
- Service times: Service times are i.i.d. random variables following some general distribution (either continuous, discrete, or a mixture) with finite mean and variance. Let R be the random variable representing the consultation duration of one patient. 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 one.
- 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 $W^{\mathbf{x}}_t$. The workload process $\mathbf{W}^{\mathbf{x}} = (W^{\mathbf{x}}_1, W^{\mathbf{x}}_2, \ldots, W^{\mathbf{x}}_n) \in \mathbb{R}^n_+$ satisfies the recursion $W^{\mathbf{x}}_t = \max\{W^{\mathbf{x}}_{t-1} + Y^{\mathbf{x}}_t - k, 0\}$, where $W^{\mathbf{x}}_0 = 0$ with probability one, and $W^{\mathbf{x}}_n$ corresponds to the overtime workload.

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 walk-in patients' aggregate waiting time, respectively, under schedule $\mathbf{x}$. Their expectations can be expressed in terms of the expected workload process. In particular, let $G^{\mathbf{x}}_t(w) \triangleq \mathbb{P}(W^{\mathbf{x}}_t \leq w)$, $w \geq 0$, denote the cumulative distribution function (CDF) of $W^{\mathbf{x}}_t$. As a notational convention, we let $G^{\mathbf{x}}_0(w) = 1$ for all $w \geq 0$. Let also $H^{\mathbf{x}}_t(y) \triangleq \mathbb{P}(R^{(A^{\mathbf{x}}_t)} \leq y)$, $y \geq 0$, denote the CDF of $Y^{\mathbf{x}}_t$. Then, $G^{\mathbf{x}}_t(\cdot)$ can be expressed recursively for $t = 1, 2, \ldots, n$ as $G^{\mathbf{x}}_t(w) = \int_0^{k+w} G^{\mathbf{x}}_{t-1}(w+k-y)\ \mathrm{d}H^{\mathbf{x}}_t(y)$, $w \geq 0$, and the performance measures of interest as $\mathbb{E}[I(\mathbf{x})] = \mathbb{E}[W^{\mathbf{x}}_n] + N - \mathbb{E}[R]\sum_{t=1}^n (px_t + \mathbb{E}[U_t])$, $\mathbb{E}[O(\mathbf{x})] = \mathbb{E}[W^{\mathbf{x}}_n]$, $\mathbb{E}[W_s(\mathbf{x})] = \sum_{t=1}^n \big[px_t \mathbb{E}[W^{\mathbf{x}}_{t-1}] + \mathbb{E}[S_t(S_t-1)]\frac{\mathbb{E}[R]}{2}\big]$, and $\mathbb{E}[W_u(\mathbf{x})] = \sum_{t=1}^n \big[\mathbb{E}[U_t]\mathbb{E}[W^{\mathbf{x}}_{t-1}] + \mathbb{E}[U_t]px_t \mathbb{E}[R] + \mathbb{E}[U_t(U_t-1)]\frac{\mathbb{E}[R]}{2}\big]$, where $\mathbb{E}[W^{\mathbf{x}}_t] = \int_0^\infty w\ \mathrm{d}G^{\mathbf{x}}_t(w)$ for all $t = 1, 2, \ldots, n$.

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 all patients are served. The corresponding intraday cost function is defined as

$$\texttt{intra}(\mathbf{x}) = c_i\mathbb{E}[I(\mathbf{x})] + c_o\mathbb{E}[O(\mathbf{x})] + c_s\mathbb{E}[W_s(\mathbf{x})] + c_u\mathbb{E}[W_u(\mathbf{x})]. \tag{B.1}$$

Theorem B.1. (Zacharias and Yunes 2020). *The function* $\texttt{intra}(\cdot)$ *in* (B.1) *is multimodular on* $\mathbb{Z}_+^n$.

References

Ahmadi-Javid A, Jalali Z, Klassen K (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.

Bailey N (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *J. Roy. Statist. Soc. B* 14(2):185–199.

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

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

Bertsekas D (2000) *Dynamic Programming and Optimal Control*, vol. 2. 4th ed. (Athena Scientific, Belmont, MA).

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Carew S, Nagarajan M, Shechter S, Arneja J, Skarsgard E (2020) Dynamic capacity allocation for elective surgeries: Reducing urgency-weighted wait times. *Manufacturing Service Oper. Management* 23(2):407–424.

Cayirli T, Veral E (2003) Outpatient scheduling in healthcare: A review of literature. *Production Oper. Management* 12(4):519–549.

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

Chen X, Li M (2021a) Discrete convex analysis and its applications in operations: A survey. *Production Oper. Management* 30(6):1904–1926.

Chen X, Li M (2021b) $M^\natural$-convexity and its applications in operations. *Oper. Res.* 69(5):1396–1408.

Dai J, Shi P (2020) Recent modeling and analytical advances in hospital inpatient flow management. *Production Oper. Management* 30(6):1838–1862.

Deo S, Iravani S, Jiang T, Smilowitz K, Samuelson S (2013) Improving health outcomes through better capacity allocation in a community-based chronic care model. *Oper. Res.* 61(6): 1277–1294.

Diamant A, Milner J, Quereshy F (2018) Dynamic patient scheduling for multi-appointment healthcare programs. *Production Oper. Management* 27(1):58–79.

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

Gerchak Y, Gupta D, Henig M (1996) Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Sci.* 42(3):321–334.

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

Green L, Savin S, Murray M (2007) Providing timely access to care: What is the right patient panel size? *Joint Commission J. Quality Patient Safety* 33(4):211–218.

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

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.

Helm J, Van Oyen M (2014) Design and optimization methods for elective hospital admissions. *Oper. Res.* 62(6):1265–1282.

Huh W, Liu N, Truong V (2013) Multiresource allocation scheduling in dynamic environments. *Manufacturing Service Oper. Management.* 15(2):280–291.

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

Keyvanshokooh E, Shi C, Van Oyen M (2020) Online advance scheduling with overtime: A primal-dual approach. *Manufacturing Service Oper. Management* 23(1):246–266.

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

Kuiper A, Lee R (2022) Appointment scheduling for multiple servers. *Management Sci.*, ePub ahead of print February 4, https://doi.org/10.1287/mnsc.2021.4221.

Kuiper A, Kemper B, Mandjes M (2015) A computational approach to optimized appointment scheduling. *Queueing Systems* 79(1): 5–36.

LaGanga L, Lawrence S (2012) Appointment overbooking in healthcare clinics to improve patient service and clinic performance. *Production Oper. Management* 21(5):874–888.

Lee I, Epelman M, Romeijn H, Smith R (2017) Simplex algorithm for countable-state discounted Markov decision processes. *Oper. Res.* 65(4):1029–1042.

Li Q, Yu P (2014) Multimodularity and its applications in three stochastic dynamic inventory problems. *Manufacturing Service Oper. Management* 16(3):455–463.

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

Liu N, van de Ven P, Zhang B (2019) Managing appointment booking under customer choices. *Management Sci.* 65(9): 4280–4298.

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

Liu N, Finkelstein S, Kruk M, 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.

Moriguchi S, Murota K (2019) On fundamental operations for multimodular functions. *J. Oper. Res. Soc. Japan* 62(2):53–63.

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.

Patrick J, Puterman M, 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 L, Chen R (2010) A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing Service Oper. Management* 12(2):330–346.

Santibáñez P, Begen M, Atkins D (2007) Surgical block scheduling in a system of hospitals: An application to resource and wait list management in a British Columbia health authority. *Health Care Management Sci.* 10(3):269–282.

Sauré A, Begen M, Patrick J (2020) Dynamic multi-priority, multi-class patient scheduling with stochastic service times. *Eur. J. Oper. Res.* 280(1):254–265.

Soltani M, Samorani M, Kolfal B (2019) Appointment scheduling with multiple providers and stochastic service times. *Eur. J. Oper. Res.* 277(2):667–683.

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

Wang D, Muthuraman K, Morrice D (2019) Coordinated patient appointment scheduling for a multistation healthcare network. *Oper. Res.* 67(3):599–618.

Wang S, Liu N, Wan G (2020) Managing appointment-based services in the presence of walk-in customers. *Management Sci.* 66(2):667–686.

Wang D, Morrice D, Muthuraman K, Bard J, Leykum L, Noorily S (2018) Coordinated scheduling for a multi-server network in outpatient pre-operative care. *Production Oper. Management* 27(3):458–479.

White D (1979) Finite state approximations for denumerable-state infinite horizon contracted Markov decision processes: The policy space method. *J. Math. Anal. Appl.* 72(2):512–523.

White D (1982) Finite state approximations for denumerable state infinite horizon discounted Markov decision processes with unbounded rewards. *J. Math. Anal. Appl.* 86(1):292–306.

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.

Zacharias C, Yunes T (2020) Multimodularity in the stochastic appointment scheduling problem with discrete arrival epochs. *Management Sci.* 66(2):744–763.

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.

Christos Zacharias is an assistant professor of management science at the Miami Herbert Business School. His areas of expertise are stochastic models, dynamic programming, discrete optimization. His research is motivated by problems relating to healthcare operations.

Nan Liu is an associate professor in the department of business analytics at the Boston College Carroll School of Management. His research focuses on service operations management. One particular application area of his research is healthcare, in which he studies how to match supply and demand for healthcare services in a way that leads to easy access, low cost, and high quality.

Mehmet A. Begen is an associate professor of management science in the Ivey Business School at the Western University. Mehmet's research interests are optimization, data-driven approaches, scheduling, healthcare, and analytics applications.