
Management Science

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

# Optimal Advance Scheduling

Van-Anh Truong

To cite this article:
Van-Anh Truong (2015) Optimal Advance Scheduling. Management Science 61(7):1584-1597. http://dx.doi.org/10.1287/mnsc.2014.2067

MANAGEMENT SCIENCE
Vol. 61, No. 7, July 2015, pp. 1584–1597
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.2067


# Optimal Advance Scheduling

Van-Anh Truong
Department of Industrial Engineering and Operations Research, Columbia University, New York, New York 10027,
vt2196@columbia.edu

The dynamic assignment of patients to exam days in order to manage daily variations in demand and capacity is a long-standing open research area in appointment scheduling. In particular, the dynamic assignment of advance appointments has been considered to be especially challenging because of its high dimensionality. We consider a canonical model of dynamic advance scheduling with two patient classes: an urgent demand class, which must be served on the day of arrival, and a regular demand class, which can be served at a future date. Patients take the earliest appointments offered and do not differentiate among providers. We derive a surprising characterization of an optimal policy and an algorithm to compute the policy exactly and efficiently. These are, to our knowledge, the first analytical results for the dynamic advance assignment of patients to exam days. We introduce the property of successive refinability, which allows advance schedules to be easily computable and under which there is no cost to the system to making advance commitments to patients. We allow multiple types of capacity to be considered and both demand and capacity to be nonstationary and stochastic.

*Keywords*: dynamic programming; decision analysis: theory; healthcare: hospitals; production scheduling
*History*: Received August 20, 2013; accepted August 4, 2014, by Martin Lariviere, operations management.
Published online in *Articles in Advance* February 11, 2015.

## 1. Introduction

A visit to a healthcare facility, whether for a specialist exam, surgical procedure, or diagnostic screen, usually requires an appointment. The appointment-scheduling process allocates service times at healthcare facilities to individual patients. From a systems perspective, this process is the main mechanism for setting the pace of operations and regulating access to healthcare (Chakraborty et al. 2010). As such, it is widely considered to be both a "primary means of improving resource utilization" and a primary means of reducing patient wait times (Lowery and Martin 1989, p. 11).

Poor scheduling is "a major source of operational inefficiency and patient dissatisfaction," according to Chakraborty et al. (2010, p. 354). "In the short term, [it] diverts potentially productive time of clinicians, receptionists, and others into the non-care processes of decision making and triage. In the longer term, it diverts staff into the time-consuming process of managing an intentionally created delay, such as making reminder calls or filing and retrieving lists for future appointments." Inadequate access to healthcare, manifesting as waits and delays, has been attributed by most analyses to problems in matching daily provider capacity with patient demand, rather than a true shortage of capacity (Murray and Berwick 2003).

Historically, the literature on appointment scheduling has focused on two forms of waiting: waiting within a day and waiting across days. Gupta and Denton (2008) define a *direct wait* as approximately the interval between the scheduled appointment and the actual service time on the appointment day and an *indirect wait* as the interval between the patient's first request for an appointment and the actual appointment. They also emphasize the relative importance of the indirect wait on the quality of care, since "excessive indirect wait[s] can pose a serious safety concern" (p. 801).

Most of the literature on appointment scheduling focuses on *intraday* scheduling—namely, the sequencing and timing of appointments within a day that primarily control the direct wait. The assumption is that the number of patients served per day has been predetermined by an external source or that it is fixed as part of a static strategy. Very few papers have tried to make express, active decisions about the indirect wait. In fact, Gupta and Denton (2008) have identified the modeling and control of indirect waiting as an open challenge. Indirect waits can be controlled via *multiday* scheduling where patients are dynamically assigned to appointment days. These *dynamic multiday assignments* are the main way in which a system can cope with the day-to-day variability of demand and capacity. Daily demand variability alone can be extremely high (McManus et al. 2003).

*Allocation scheduling* and *advance scheduling* are the main paradigms for dynamic multiday scheduling. In allocation scheduling, a wait list is maintained and patients are notified on the day of their appointments.

In advance scheduling, patients are given appointments in the future at the time of request. The latter is the predominant paradigm in practice.

The standard dynamic programming formulation of advance scheduling generally cannot be solved easily because of the "curse of dimensionality." When a new appointment is made, optimal advance-scheduling policies must consider all outstanding appointments. A description of these typically includes hundreds of patients and several months' worth of appointment days. It has been observed by Gupta and Denton (2008, p. 813) that, with the added complexity of scheduling for multiple providers, "there is no obvious decomposition that can be applied to simplify the problem."

In this paper, we study an advance-scheduling problem that is motivated by the scheduling of magnetic resonance imaging (MRI) exams at Morgan Stanley Children's Hospital in New York City. This problem has been described by Truong and Ruzal-Shapiro (2013). The hospital operates one MRI machine. It serves three types of patients. Emergency patients must be served immediately. Inpatients coming from within the hospital wards can technically wait beyond a day but are generally served within 24 hours to minimize backups in the wards. Finally, outpatients from various affiliated clinics are served by advance appointments. It is difficult to further classify the patients for planning purposes using information such as medical history, physical characteristics, and the types of scans they require. The duration of exams for patients who require anesthesia, for example, largely depends on the time required to administer anesthesia and stabilize the patients' vital signs, which is very difficult to predict. The challenge faced by the hospital is how to allocate exam time efficiently among these patient classes.

We consider a canonical model of dynamic advance scheduling with two patient classes: an *urgent* demand class, which must be served on the day of arrival, and a *regular* demand class, which can be served at a future date. The urgent class models the emergency and inpatients above, and the regular class models the outpatients. We assume that patients take the earliest appointments offered and do not differentiate among providers. There are costs for patient waiting and overutilization, or overtime use, of available resources. We allow multiple types of capacity corresponding to multiple resources to be considered and both demand and capacity to be nonstationary and stochastic. This model is a simplified version of the model in Patrick et al. (2008) with two instead of $n$ demand classes, random rather than deterministic service times, multiple resources rather than one resource, and nonstationary demands and capacity rather than stationary demand and deterministic capacity. Our model can also be considered an extension of the model of Gerchak et al. (1996), with multiple resources rather than a single resource and nonstationary rather than stationary demands. Whereas Gerchak et al. study allocation scheduling, we study advance scheduling.

Our model is essentially an aggregate-planning model in the same spirit as that of Gerchak et al. (1996). The model can be used as part of a two-stage planning process. In stage 1, the number of patients to be served on any given day is determined. In stage 2, the sequence and timing of individual exams are determined. Our model addresses stage 1 decisions, in which the cost of overutilization of resources is a trade-off with the cost of making patients wait over multiple days.

The multiresource, highly constrained environments that we consider in this paper are one of the "open challenges" described by Gupta and Denton (2008) in their influential review of medical scheduling literature. Healthcare services typically command many different resources. For example, according to Gupta and Denton (2008, p. 816), surgeries require "clinical assistants, nurses, anesthesiologists and surgeons, as well as . . . operating rooms, diagnostic devices, surgical tools and other equipment." A shortage of any of these resources can aggravate waiting time and worsen outcome: the high level of utilization sought for expensive resources creates "resource-use conflicts that are exacerbated by tight schedules[,] . . . patient no-shows, tardiness of providers and staff absences" (p. 816). There are no known characterizations of optimal policies or algorithms with performance guarantees for advance scheduling under these conditions.

The nonstationarity of resource availability and demand that we consider here is another common feature of healthcare scheduling, but it is challenging to model and solve. In surgical scheduling, for example, staffing fluctuates greatly with the time of week and the time of year (Dexter et al. 1999). Demand is also generally nonstationary, exhibiting both seasonality and growth trends (Moore et al. 2008). Policies for stationary environments, therefore, would either perform poorly or simply be unimplementable in practice. The difficulty of finding optimal decisions in these environments comes from the need for such policies to adapt to changing conditions over time and to account for usage across many different resources, especially when both resource availability and demands are stochastic. There are a few published papers on dynamic allocation-scheduling policies in these environments, for example, Ayvaz and Huh (2010) and Huh et al. (2013), but no work on advance scheduling.

For the MRI-scheduling problem that motivates this research, we show that there is an easy-to-compute optimal policy. The optimal policy does not schedule the same number of regular patients for each day but dynamically increases this daily "regular workload" as the total number of regular patients in the system increases. Further, with each unit increase of the number of regular patients in the system, the optimal regular workload for each day only increases rather than decreases, and it increases by at most one patient. In stationary settings, the optimal policy schedules fewer regular patients further into the future. In both stationary and nonstationary settings, there is no loss of flexibility to the system in scheduling patients using advance scheduling rather than allocation scheduling, because under advance scheduling it is always possible to incrementally revise the schedule by following an optimal policy.

On the theory side, we show that there is an intimate connection between allocation scheduling and advance scheduling. Thus, this work connects two main independently analyzed streams of literature. We exhibit a method to construct an optimal solution to a high-dimensional advance-scheduling problem from a solution to an associated, simple-to-compute, allocation-scheduling problem. This construction leads to a full characterization of the advance-scheduling solution and allows it to be computed exactly and efficiently. These are, to our knowledge, the first analytical results for the dynamic advance assignment of patients to exam days. We also introduce the property of *successive refinability*, which allows advance schedules to be easily computable and under which advance scheduling is as efficient as allocation scheduling, so that there is no cost to the system to make advance commitments to patients. We are not aware of any analogous result in this domain.

The rest of this paper is organized as follows. In §2, we discuss advance scheduling in the context of existing literature. In §3, we specify our model precisely. We take a detour in §4 to study properties of a corresponding allocation-scheduling problem. In §5, we study an intermediate model, which serves as a bridge between allocation and advance scheduling. We also introduce the property of successive refinability. We derive an optimal policy for advance scheduling in §6. We discuss extensions in §7. In §8, we provide a numerical example illustrating our optimal advance-scheduling policy. We conclude in §9. We relegate all formal proofs to the appendix.

## 2. Related Literature

Appointment scheduling is a well-established area with connections to many topics, such as machine scheduling, capacity allocation in production and service systems, admission control for queues, and revenue management.

### 2.1. Machine Scheduling

There is an extensive literature on scheduling jobs on machines. These problems are similar to intraday scheduling problems with some key differences. First, jobs do not incur any waiting cost until past their due date; second, machine capacity is usually fixed, rather than extendable; and third, unlike patient service times, job-processing times tend to be deterministic rather than random (Gupta and Denton 2008). Leung (2004) provides a detailed review of this literature.

### 2.2. Capacity Allocation in Production and Service Systems

Dynamic multiday medical scheduling problems share some characteristics with the sale of inventory to multiple customer classes whose arrivals are uncertain. A major difference is that inventory can be stocked, whereas service capacity is lost if it is not used (Ayvaz and Huh 2010). See, for example, Topkis (1968), Ha (1997a, b), Carr and Duenyas (2000), Sobel and Zhang (2001), Duran et al. (2008), and Zhou and Zhao (2010). Several papers study pricing decisions, which are not relevant in healthcare settings (Maglaras and Zeevi 2005, Ding et al. 2006).

### 2.3. Revenue Management

The literature on revenue management shares a similar concern with allocating a finite, perishable amount of capacity among customer classes who arrive randomly over time. In revenue management, demand is for a particular resource at a particular time. Demand cannot be satisfied by supplying the resource at a later date. In our setting, however, demand from regular patients arriving on a particular day can always be satisfied on another day. Another difference is that high-priority demand is lost if not accommodated in revenue management, whereas in healthcare capacity must usually be "extended" at a cost to serve urgent demand (Gupta and Denton 2008). Talluri and Van Ryzin (2005) provide a detailed review of this literature.

### 2.4. Admission Control for Queues

Appointment scheduling shares many similarities with admission control for queues. In queuing models, patients are served in a greedy manner, whereas in appointment scheduling, patients are given appointments at specific times (Gupta and Denton 2008). In addition, the type of stationary analysis employed on queuing models might not be suitable when the number of patients that can be seen per day is small. Queuing models, therefore, tend to be used to study strategic decisions, such as capacity planning and reservation, and physician–panel design. See Young (1963), Esogbue and Singh (1976), Huang (1995), Green and Savin (2008), and Liu and Ziya (2014).

### 2.5. Appointment Scheduling

Many papers in this literature focus on intraday scheduling. Examples of this stream of research include Wang (1999), Rohleder and Klassen (2000), Denton and Gupta (2003), Robinson and Chen (2003), Klassen and Rohleder (2004), Green et al. (2006), Kaandorp and Koole (2007), Hassin and Mendel (2008), Jouini and Benjaafar (2010), Cayirli et al. (2012), and Luo et al. (2012). Some of these papers focus on finding the optimal number of patients to overbook in a day to reduce the impact of no-shows (Kim and Giachetti 2006, LaGanga and Lawrence 2007, Muthuraman and Lawley 2008, Zeng et al. 2010, Huang and Zuniga 2011, LaGanga and Lawrence 2012). Others study the optimal percentage of same-day appointments to offer (Qu et al. 2007, Robinson and Chen 2010).

Some of the literature models two classes of patients, one that must be served on the day of arrival and another, regular class that can be served at a later date. For example, Dobson et al. (2011) fix the number of regular appointments that can be scheduled for each day and study various service measures as a function of the number of slots reserved for same-day patients. Qu et al. (2007) also fix the number of appointment slots available on each day and focus on finding the optimal constant number of slots to open for regular patients. A distinguishing feature of our model is that we allow the number of regular appointments that can be scheduled on each day to be a dynamic decision, which helps the planner to trade off the cost of overtime work with the cost of patient waiting. Gupta and Wang (2008) model patient choice for physician–slot combinations and allow clinics to make acceptance/rejection decisions in response to patient requests. They essentially allow the number of slots given to regular patients to depend dynamically on the set of already accepted requests. However, since their model focuses on a single day, it does not capture the impact of scheduling decisions across days, which is an important feature of our model.

There is a growing literature on allocation-scheduling models. Mullen (2003) surveys approaches for prioritizing waiting lists. Swain et al. (1977) develop heuristics to predict hospital census and determine elective-admission policies. George et al. (1983) formulate a linear-programming model to plan the aggregate throughput of the general surgical department. Gerchak et al. (1996) characterize an optimal policy for allocating capacity to regular and emergency cases. Min and Yih (2010) and Ayvaz and Huh (2010) extend this work by considering multiple patient classes. Min and Yih (2014) go further by modeling time-dependent priorities for patients. Huh et al. (2013) develop policies for allocation scheduling under correlated demands and multiple constrained resources.

Dynamic advance-scheduling problems have generally been deemed intractable. To our knowledge, there has been no direct characterization of optimal policies in the published literature. Patrick et al. (2008) develop approximate dynamic programming-based heuristics for scheduling patients of different priorities. Gocgun and Ghate (2012) also develop an approximate dynamic programming method that uses Lagrangian relaxation and constraint generation to solve a similar problem. Liu et al. (2010) propose heuristics taking into account cancellations and no-shows. Feldman et al. (2014) develop heuristics assuming that patients' preferences for appointment times follow the multinomial logit model. Truong and Ruzal-Shapiro (2013) characterize the optimal policy and propose effective heuristics for a two-class advance-scheduling problem in which patients may be expedited after they are first assigned advance appointments. This systematic expediting of patients might not always be implementable in practice. Patrick (2012) studies a dynamic multiday scheduling problem with no-shows and demonstrates by simulation that using a short booking window might be more advantageous than using an open-access policy.

## 3. Advance-Scheduling Model

The reader can refer to Table 1 for a summary of notation. Consider an horizon of $T$ periods indexed by $t \in \{1, 2, \ldots, T\}$ in a forward manner. We use the terms "period" and "day" interchangeably, as a period normally corresponds to a day in reality. We allow $T = \infty$.

In each period $t$ in the horizon, demand arises for urgent exams and for regular exams. Urgent exams must be performed immediately in period $t$, whereas regular exams can be performed some time in the future. The number of regular exams requested by patients arriving in period $t$ is a random variable, $\delta^t$. The decision in period $t$ is how to allocate these demand requests to periods $\{t, \ldots, T\}$.

There are $R$ resources that are used by urgent and regular exams. The amount of resource $r = 1, \ldots, R$ consumed by each regular exam is a random variable that is independent and identically distributed (i.i.d.) across patients and days. For each $r = 1, \ldots, R$, let $S_r(n)$ represent the total amount of resource $r$ that is consumed by $n$ regular exams. Then $S_r(n)$ is a convolution of $n$ i.i.d. random variables. Let the amount of resource $r$ that is used by all urgent exams on day $t$ be a random variable $\epsilon_r^t$.

There is a random *utilization cost* $u_r^t(\epsilon_r^t + S_r(n))$ to fulfill all urgent and $n$ regular exams scheduled on day $t$. The utilization cost models regular and overtime use of the resource, allowing the quantity of

the resource to be both random and time varying. We assume that $u_r^t(\cdot)$ is convex, increasing, and independent over time. The cost of resource utilization in healthcare facilities is often dominated by staffing costs. Our convex utilization cost captures a common pay structure whereby staffs are compensated at a higher rate for each hour of overtime work. It does not capture fixed setup costs.

Note that since our model focuses on the trade-off between the cost of overtime work and the cost of making patients wait over multiple days, similar to Gerchak et al. (1996), we have approximated the total resource usage per day as a simple sum of individual exam usages. This approximation ignores within-day wait time by patients and idle time by providers.

We assume that all patients prefer to have their exams sooner rather than later. In other words, they will take the earliest appointments offered. This assumption follows Patrick et al. (2008) and Liu et al. (2010). There is a *waiting cost* $W$ for each day that a patient is made to wait for an exam after first making a request for an appointment. The waiting cost captures both the inconvenience of waiting and any loss in productivity to the patient and to society that is caused by delays in treatment. In practice, the real cost of such a wait in terms of health outcomes and perceived inconvenience might be very complicated and might increase steeply with the wait. Our unit waiting cost can be considered an average daily cost of waiting. This model of the waiting cost follows Gerchak et al. (1996) and Ayvaz and Huh (2010).

Let the discount factor in each period be $\gamma \in (0, 1)$. We would like to determine the scheduling policy to minimize the total expected discounted cost incurred over the horizon.

We will consider two types of scheduling environments. In a *nonstationary environment*, the sequences $\{\delta^t\}_t$, $\{(\epsilon_1^t, \ldots, \epsilon_R^t)\}_t$, and $\{(u_1^t(\cdot), \ldots, u_R^t(\cdot))\}_t$ are independent but not necessarily identically distributed over time. In a *stationary environment*, the sequences are i.i.d. In both environments, the distributions of these sequences of exogenous random variables are known at time 1. We sometimes omit the time superscripts when discussing stationary environments. When the horizon is infinite, we shall always assume that the environment is stationary.

The events in each period $t$ follow the following sequence.

*Step* 1. The numbers $x^t = (x_1^t, \ldots, x_{T-t+1}^t)$ of regular exams already scheduled for periods $\{t, \ldots, T\}$ are observed.

*Step* 2. The new random demand $\delta^t$ arises for regular exams. A decision is made to determine in which of the periods $\{t, \ldots, T\}$ these exams will be performed. We call the periods $\{t, \ldots, T\}$ the *booking window*. The outcome of the decision is the new schedule $z^t = (z_1^t, \ldots, z_{T-t+1}^t)$, where $z_i^t$ represents the number of regular exams scheduled for period $i + t - 1$.

*Step* 3. The new demand for urgent exams arising at $t$ arises and is satisfied together with the $z_1^t$ regular exams scheduled for period $t$.

*Step* 4. A waiting cost $W$ is incurred by the system for each patient who is still on the schedule (i.e., as yet unserved).

The state of the system at the beginning of period $t$ is the *schedule* for periods $t,\ t+1, \ldots, T$. The schedule is described by the vector $x^t = (x_1^t, x_2^t, \ldots, x_{T-t+1}^t)$. The component $x_i^t$ specifies the number of regular exams scheduled for period $t + i - 1$ as of the beginning of period $t$. We use $|\cdot|$ to denote the number of patients in the schedule, or the *size* of the schedule. That is, $|x| = \sum_{k=1}^{T-t+1} x_k$. We shall always work with finite-size schedules. We also define $\mathbf{s}(x)$ to be

**Table 1 Summary of Notation**

| | |
|---|---|
| $T$ | The length of the horizon |
| $t$ | The index of the current day or period |
| $R$ | The number of resources used by regular and urgent patients |
| $r$ | The index of a particular resource, of which there are $R$ |
| $\delta^t$ | The number of regular patients arriving in period $t$, a random variable |
| $\epsilon_r^t$ | The total amount of resource $r$ used by all urgent patients in period $t$, a random variable |
| $S_r(n)$ | The amount of resource $r$ used by $n$ regular patients, a convolution of $n$ i.i.d. random variables |
| $u_r^t(\cdot)$ | The utilization cost in period $t$ as a function of the total amount of resource $r$ consumed in period $t$ |
| $W$ | The waiting cost per day per patient |
| $\gamma$ | The discount factor |
| $x^t$ | The schedule at the beginning of period $t$, a vector with $T-t+1$ components, with each component $i$ specifying the number of regular patients scheduled for period $t+i-1$ |
| $z^t$ | The schedule at the end of period $t$, a vector with $T-t+1$ components, with each component $i$ specifyiing the number of regular patients scheduled for period $t+i-1$ |
| $\mathbf{s}(\cdot)$ | The left-shift operator |
| $\|x\|$ | The size of a schedule $x$, which is defined to be the number of people on the schedule $x$ |
| $w^t$ | The size of the wait list at the beginning of period $t$ in the allocation-scheduling model |
| $\bar{w}^t$ | The size of the wait list after demand arises for regular exams but before a decision is made in period $t$ in the allocation-scheduling model |
| $V^t(\delta^t, x^t)$ | The total discounted cost charged to period $t$ and beyond in advance scheduling, given that the schedule at the beginning of period $t$ is $x^t$ and the demand for regular exams observed in period $t$ is $\delta^t$ |
| $\tilde{V}^t(\delta^t, x^t)$ | The total discounted cost charged to period $t$ and beyond in noncommittal advance scheduling, given that the schedule at the beginning of period $t$ is $x^t$ and the demand for regular exams observed in period $t$ is $\delta^t$ |
| $\underline{V}^t(\bar{w}^t)$ | The total expected discounted cost charged to period $t$ and beyond in allocation scheduling, given that the number of regular patients in the system just after the observation of regular demand in period $t$ is $\bar{w}^t$ |

the result of the left-shift operator acting on $x$. That is, $\mathbf{s}(x) = (x_2, x_3, \ldots, x_{T-t+1})$.

Note that our waiting cost is also simpler than that of Patrick et al. (2008). Whereas we model a constant waiting cost $W$ per patient per day of wait, Patrick et al. allow the waiting cost to be 0 if a patient is served within the target deadline for his priority class.

An issue with the above model is that patients might be served out of first-in-first-out (FIFO) order. Regular patients arriving later might be scheduled ahead of those arriving earlier if the total number of patients in the system is high, forcing the system to increase its "rate of service." However, enforcing fairness reduces system flexibility because patients can only be added to the far end of a schedule, rather than anywhere in the schedule. In a real system, fairness can be enforced by expediting early arrivals, rather than scheduling late arrivals ahead of them. Truong and Ruzal-Shapiro (2013) study a model of advance scheduling in which expediting is performed systematically. More justifications of our model can be found in Patrick et al. (2008).

We formulate the problem of determining an optimal advance-scheduling policy as follows. Let $V^t(\delta^t, x^t)$ denote the total discounted cost charged to period $t$ and beyond, given that the schedule at the beginning of period $t$ is $x^t$ and the demand for regular exams observed in period $t$ is $\delta^t$:

$$V^t(\delta^t, x^t) = \min\Bigg\{ W|z^t| + \sum_{r=1}^{R} \mathbf{E}[u_r^t(\epsilon_r^t + S_r(z_1^t))] + \gamma \mathbf{E}_{\delta^{t+1}}[V^{t+1}(\delta^{t+1}, \mathbf{s}(z^t))] \Bigg\} \quad (1)$$

$$\text{s.t. } z^t \geq x^t,$$
$$|z^t| = \delta^t + |x^t|.$$

Above, the first constraint ensures that exams cannot be removed from a day once booked. The second constraint ensures that $z^t$ contains additions to $x^t$ of $\delta^t$ newly arrived regular exam requests. In the nonstationary environment, we also define $V^{T+1}(\cdot, \cdot) = 0$.

Note that since we assume that the distribution of $\epsilon_r^t$ and the distribution of exam times for regular patients are known, the convolution $S_r(n)$ of $n$ regular exam times, thus the function $\mathbf{E}[u_r^t(\epsilon_r^t + S_r(n))]$, can be computed easily by Monte Carlo simulation. In fact, $\mathbf{E}[u_r^t(\epsilon_r^t + S_r(n))]$ can be precomputed easily for each $r$ and each $n = 0, 1, 2, \ldots$, so that for each $r$, $\mathbf{E}[u_r^t(\epsilon_r^t + S_r(n))]$ can be considered as a simple deterministic function of $n$.

The advance-scheduling formulation (1) is challenging in several ways. First, the state and decision space are high dimensional. Indeed, their dimension is equal to the size of the booking window, which is very large in reality. Second, the feasible region lacks the lattice structure that often helps to confer nice monotonicity properties. We will find avenues for solving this problem by examining several related formulations, which seem to be better structured.

## 4. Insights from Allocation Scheduling

Before we attempt to solve (1), let us take a detour into allocation scheduling. The insights gleaned from this model will be useful in constructing optimal policies for advance scheduling.

### 4.1. Allocation-Scheduling Model

In allocation scheduling, a running wait list is maintained, and the manager decides on the number $q^t$ of regular patients from the wait list to serve in each period $t$. The costs are analogous. The events in each period $t$ follow the following sequence.

*Step* 1. The size $w^t$ of the wait list from the previous period is observed.

*Step* 2. The new random demand $\delta^t$ arises for regular exams and is added to the wait list, bringing the total to $\bar{w}^t = w^t + \delta^t$.

*Step* 3. A waiting cost $W$ is incurred by the system for each patient on the wait list.

*Step* 4. The manager makes a decision to serve $q^t$ patients from the wait list in period $t$.

*Step* 5. The random demand for urgent exams at $t$ arises and is satisfied together with the $q^t$ regular exams scheduled for period $t$. There is a random utilization cost $u_r^t(\epsilon_r^t + S_r(q^t))$ to fulfill all urgent and $q^t$ regular exams scheduled on day $t$.

The state of the system at the beginning of period $t$, after the observation of regular demand, is the size $\bar{w}^t$ of the wait list.

We formulate the problem of determining an optimal allocation-scheduling policy as follows. Let $\underline{V}^t(\bar{w}^t)$ denote the total expected discounted cost incurred by the system in period $t$ and beyond, given that the state in period $t$ is $\bar{w}^t$:

$$\underline{V}^t(\bar{w}^t) = \min_{\bar{w}^t \geq q^t \geq 0} \Bigg\{ W\bar{w}^t + \sum_{r=1}^{R} \mathbf{E}[u_r^t(\epsilon_r^t + S_r(q^t))] + \gamma \mathbf{E}_{\delta^{t+1}}[\underline{V}^{t+1}(\bar{w}^t - q^t + \delta^{t+1})] \Bigg\}. \quad (2)$$

In the nonstationary environment, we also define $\underline{V}^{T+1}(\cdot) = 0$. This problem is relatively easy to solve, especially in the stationary setting, because the state and action spaces are both single dimensional. Below, we characterize an optimal policy for this problem.

### 4.2. Allocation-Scheduling Properties

The theorems in this section generalize similar results by Gerchak et al. (1996) for a stationary, single-resource allocation-scheduling model.

First, we establish some basic properties of the value function in allocation scheduling. The convexity and monotonicity of the value function follow easily from the problem's convex cost structure, linear state transitions, and convex action space.

THEOREM 1. *For each $t$, $\underline{V}^t(\cdot)$ is convex and increasing.*

Let the cost function to be minimized in each period $t$ be $G^t(q, \bar{w}) = W\bar{w} + \sum_{r=1}^{R} \mathbf{E}[u_r^t(\epsilon_r^t + S_r(q))] + \gamma \mathbf{E}[\underline{V}^{t+1}(\bar{w} - q + \delta^{t+1})]$.

The submodularity of $G^t(\cdot, \cdot)$ follows from the problem's separable costs, convex value function, and lattice action space. The following theorem uses standard results from supermodularity theory.

THEOREM 2. *For each $t$, $G^t(\cdot, \cdot)$ is submodular.*

A simple policy that is often used in practice is to allocate a fixed number of regular patients to each day. This daily number is called a threshold, and this policy is called a *threshold policy*. Using the above basic properties, we show below that the optimal policy is not a threshold policy as might be surmised. The optimal number $q^t$ of patients served in period $t$ is an increasing function of the size $\bar{w}^t$ of the wait list. This is because the longer the wait list becomes, the more pressure there is to increase the time $t$ workload. At a cost of paying higher utilization costs, the system can check the growth of the wait list and avoid paying higher waiting costs in the future. This property of the optimal policy implies that it is suboptimal to serve the same number of regular patients in each period. Systems should respond to longer wait lists by increasing the workload appropriately.

A second property of the optimal policy, which will later be crucial to our development of a "successive refinability" property, is that when the wait list $\bar{w}^t$ increases by 1, the workload $q^t$ always increases by at most 1. This property is natural in view of the fact that waiting costs, which incentivize the system to serve patients sooner than later, increase linearly in the number of patients who are made to wait. On the other hand, utilization costs, which incentivize the system to spread out the workload by possibly delaying service, increase convexly in the number of patients who are to be served. When the wait list is short, the optimal policy serves every patient on the wait list immediately. The wait list size $\bar{w}^t$ and the workload $q^t$ increase at the same rate. As the wait list grows, and with it the workload, it becomes increasingly expensive in terms of utilization costs to serve an additional patient. The utilization cost gradually dominates the waiting cost, so that we are more likely to make the additional patient wait rather than to serve him immediately.

THEOREM 3. *Let $q^{t*}(\bar{w})$ be the maximum optimal number of regular patients served at $t$ in allocation scheduling, given that the total number of regular patients in the system is $\bar{w}$. Then $q^{t*}(\bar{w})$ is increasing in $\bar{w}$. Moreover, $q^{t*}(\bar{w}+1) \le q^{t*}(\bar{w}) + 1$.*

We call $q^{t*}(\cdot)$ an *allocation function*. We will return to these functions later, in the construction of an optimal advance-scheduling policy.

A reasonable assumption about the system is that at least one patient is served in each period if the wait list is nonempty. We will make this assumption throughout the rest of the paper.

ASSUMPTION 1. *If $\bar{w} \ge 1$, then $q^{t*}(\bar{w}) \ge 1$.*

This assumption is satisfied if for period $t$ and every resource $r$,

$$\mathbf{E}[u_r^t(\epsilon_r^t + S_r(1))] - \mathbf{E}[u_r^t(\epsilon_r^t)] \le \frac{1}{1-\gamma} W. \tag{3}$$

In other words, it is cheaper for the system to serve the one patient on the wait list than to let the patient wait forever.

Armed with insights about optimal allocation scheduling, we will return to advance scheduling. But first, we will examine below a model that is a compromise between allocation and advance scheduling—namely, noncommittal advance scheduling.

## 5. Noncommittal Advance Scheduling

Noncommittal advance scheduling is a seemingly idiosyncratic modification of advance scheduling. It is an advance-scheduling model in which advance commitments do not have to be upheld. In this section we will show that this model serves as a critical link between allocation and advance scheduling.

### 5.1. Noncommittal Advance-Scheduling Model

We obtain the noncommittal advance-scheduling model by taking the advance-scheduling model and dropping the requirement that precommitments must be upheld in subsequent periods, i.e., by dropping the first inequality in (1). We use $\tilde{V}^t(\delta^t, x^t)$ to denote the optimal expected cost on $[t, \infty)$ under noncommittal advance scheduling. Compare the following to the definition of $\underline{V}^t(w^t)$ for allocation scheduling in §4:

$$\tilde{V}^t(\delta^t, x^t) = \min \Bigg\{ W|z^t| + \sum_{r=1}^{R} \mathbf{E}[u_r^t(\epsilon_r^t + S_r(z_1))] + \gamma \mathbf{E}_{\delta^{t+1}}[\tilde{V}^{t+1}(\delta^{t+1}, \mathbf{s}(z^t))] \Bigg\}$$

$$\text{s.t.} \quad |z^t| = \delta^t + |x^t|.$$

This relaxed formulation turns out to be very well structured. It can be observed that the noncommittal

formulation above has a state space that is collapsible to one dimension. Since prior commitments can be ignored, the only information that matters is the number of patients outstanding.

Thus, the noncommittal model is nearly identical to the easier allocation-scheduling model. The main difference is that the outcome of the decision at $t$ in the noncommittal model is an entire schedule (a vector $z^t$) rather than just a single number (an integer $q^t$). Indeed, the two problems have the same optimal cost, since the constraints, the transitions, and the cost structures are the same. Below, we show that this optimal cost is less than or equal to the corresponding cost for the advance-scheduling model.

Proposition 1. *For all periods $t$, schedules $x$, and demands $\delta$, $\tilde{V}^t(\delta, x) = \underline{V}^t(\delta + |x|) \le V^t(\delta, x)$.*

In other words, both noncommittal advance scheduling and allocation scheduling can be viewed as relaxations of advance scheduling. We shall later make use of this interpretation.

### 5.2. Successive Refinability

We call a schedule for the noncommittal advance-scheduling model a *preallocation* since it is not binding in subsequent periods. We will construct optimal preallocations and prove properties of these preallocations, particularly the property of "successive refinability," that make them particularly useful for advance scheduling.

Recall the allocation function $q^{t*}(\bar{w})$ from §4, which specifies an optimal number of regular patients to serve in period $t$ in an allocation-scheduling model, given that the size of the wait list is $\bar{w}$. We will use allocation functions to construct optimal preallocations.

Define $q^{t*}(\bar{w}) = 0$ for $\bar{w} \le 0$. Define a policy $\pi$ for the noncommittal advance-scheduling problem, which determines preallocations as follows:

$$z_1^{t\pi}(\delta, x) = q^{t*}(\delta + |x|), \tag{4}$$

$$z_k^{t\pi}(\delta, x) = q^{*t+k-1}\left(\delta + |x| - \sum_{i=1}^{k-1} z_i^{t\pi}(\delta, x)\right), \quad k = 2, \ldots, \infty. \tag{5}$$

That is, $\pi$ applies the function $q^{t*}(\cdot)$ to the total number of regular patients outstanding to obtain the number to be served on day $t$. Then it applies $q^{t*}(\cdot)$ to the remaining number to obtain the number to be served on day $t+1$. Then it applies $q^{t*}(\cdot)$ to the remaining number to obtain the number to be served on day $t+2$, and so on. Note that in the stationary model $q^{t*}(\cdot)$ is the same for all periods $t$.

As a policy, $\pi$ has several nice properties, each of which we will discuss in turn.

We will first show that $\pi$ is a valid policy for noncommittal advance scheduling in that it allocates some service day to each of the regular patients outstanding. That is, it eventually gets to all of the regular patients in the manner described above. This is because by Assumption 1, the function $q^{t*}(\cdot)$, when applied to a positive number, will return a positive value. Each time that $\pi$ applies the function $q^{t*}(\cdot)$, it strictly reduces the number remaining to be allocated. Thus, it eventually succeeds in allocating all outstanding regular patients to future days. The proposition below states that, given an initial schedule $x$ and a number $\delta$ of new regular arrivals, the schedule $z^{t\pi}(\delta, x)$ constructed by $\pi$ contains all of these $\delta + |x|$ regular patients.

Proposition 2. *For all schedules $x$ and demands $\delta$, $|z^{t\pi}(\delta, x)| = \delta + |x|$.*

Next is a useful technical reinterpretation of the construction of $\pi$. The number of patients allocated to period $t+k$ is the result of the allocation function applied to the number of patients eventually allocated (at the end of period $t$) to period $t+k$ or later. Another way to view it is that the workloads for periods $t+k$ and later are oblivious (via the construction of $\pi$) of the workloads for periods strictly earlier than $t+k$.

Proposition 3. *For all schedules $x$, demands $\delta$, and $k = 1, 2, \ldots,$*

$$z_k^{t\pi}(\delta, x) = q^{*t+k-1}(|\mathbf{s}^{k-1}(z^{t\pi}(\delta, x))|).$$

Finally, we will not find it difficult to see now that $\pi$ is, in fact, an optimal policy for the noncommittal advance-scheduling model. This result is easy to argue. Note that by its construction, $\pi$ serves the same number of regular patients in each period as an optimal allocation-scheduling policy. Thus, $\pi$ has the same cost as that of an optimal allocation-scheduling policy. Indeed, the number of patients served in each period completely determines the cost of a policy. Since the cost of an optimal allocation-scheduling policy is the same as that of an optimal noncommittal advance-scheduling policy (Proposition 1), the result follows.

Proposition 4. *Policy $\pi$ is optimal for the noncommittal advance-scheduling problem.*

A priori, it is entirely possible that a policy for noncommittal advance scheduling might change its allocation from period to period. It might allocate a patient to a particular future period at $t$, then change this allocation to a strictly earlier or later period at $t+1$. This is because the noncommittal advance-scheduling formulation does not require decisions to honor prior commitments. Therefore it is remarkable and surprising that $\pi$ turns out to *always* honor

prior commitments. That is, in period $t+1$, when we repeat the preallocation procedure under $\pi$, none of the patients preallocated in period $t$ needs to be moved. New patients arriving in period $t+1$ are simply added to the previous preallocation. The intuition is that the optimal allocation functions $q^{t*}(n)$, which are used to construct $\pi$, change "slowly" with the state information $n$ (Theorem 3). Therefore, as more patients arrive and the total number of patients changes, these optimal allocations can always be updated in time.

The above property implies that $\pi$ is not merely a valid policy for noncommittal advance scheduling but that it is also a valid policy for advance scheduling. Thus, we will potentially be able to make use of $\pi$ in our search for an optimal policy for advance scheduling.

THEOREM 4. *Assume that* $x^1 = 0$. *Then* $\mathbf{s}(z^{t\pi}) \le z^{t+1\pi}$ *in the noncommittal model in each period* $t$.

We say, therefore, that the preallocation defined by $\pi$ is *successively refinable*. New preallocations are simply refinements of previous preallocations that leave intact all previously preallocated scheduling. The term is named after a loosely analogous concept (Equitz and Cover 1991) in rate-distortion theory (Berger 1971).

The intuition behind the successive refinability property comes from the fact that policy $\pi$ is constructed from an optimal allocation function. By Theorem 3, this allocation function is increasing in the total number of regular patients outstanding. Moreover, when the total number of regular patients increases by 1, the allocation increases by at most 1. Consider any period $t$. We would only want to revise the number to be served at $t$ (after new regular demand arrives in period $t$) if the total number of regular patients in the system changes. Since the total number of patients can only increase with new demand arrivals, we would only revise this number upward. Additionally, since the number of revisions is always less than or equal to the number of new arrivals, it will always be possible to increase the number served at $t$ by allocating some new arrivals to period $t$, rather than by moving some patients to $t$ who have previously been allocated to later periods. As we noted in Proposition 3, the workloads for periods strictly later than $t$ are independent of the workload for $t$ under $\pi$. Therefore, we can repeat this logic to show that, with the new demand arrivals, policy $\pi$ will increase the number of patients allocated to each period in the future, thus keeping intact all of its previous allocations.

Next we will use successive refinability to solve the advance-scheduling problem in §3.

## 6. Optimal Advance Scheduling via Successive Refinability

We are now ready to characterize an optimal policy for advance scheduling.

We say that an advance-scheduling problem is *successively refinable* if there is an optimal preallocation that is successively refinable. From §5.2, we see that the advance-scheduling formulation given in §3 is successively refinable. Because of the successive refinability of the noncommittal model, if we now make preallocations binding, we will not change the optimal value. However, making preallocations binding gives us (1). In other words, if $x^1 = 0$, so that $\pi$ is feasible in period 1, then $\pi$ is feasible for (1) in every period. Since $\pi$ is also optimal for noncommittal advance scheduling, it is optimal for advance scheduling by Proposition 1. Theorem 5 then follows.

THEOREM 5. *Assume that* $x^1 = 0$. *Then* $\pi$ *is an optimal policy for the advance-scheduling problem. That is, there is an optimal schedule* $z^t$ *in each period* $t$ *satisfying*

$$z_1^t(\delta, x) = q^{t*}(\delta + |x|), \tag{6}$$

$$z_k^t(\delta, x) = q^{*t+k-1}\left(\delta + |x| - \sum_{i=1}^{k-1} z_i^t(\delta, x)\right), \quad k = 2, \ldots, \infty. \tag{7}$$

Thus, we have a complete characterization of an optimal policy for advance scheduling. From the properties of the allocation function proved in Theorem 3, we can deduce that the optimal schedule dynamically allocates more regular patients to each day as the total number of regular patients outstanding increases. Further, with each increase of a regular patient to the number outstanding, the optimal regular workload for each day only increases, and it increases by at most one patient.

In the stationary model, we can infer additional structure for an optimal schedule. The theorem below shows that the optimal policy $\pi$ schedules fewer regular patients further into the future.

THEOREM 6. *Assume that* $x^1 = 0$. *In the stationary model, there is an optimal schedule* $z^t$ *in each period* $t$ *satisfying* $z_1^t \ge z_2^t \ge \cdots$.

The proof of the above theorem is easy to see. In the stationary model, the allocation function $q^*(\cdot)$ is time independent. Since it decreases monotonically in the number of patients remaining to be "allocated" by Theorem 2, the theorem follows from the construction of an optimal schedule $z^t$ given in Theorem 5.

It might not always be feasible to follow policy $\pi$. For example, in period 1, the schedule $x^1$ might be nonzero and might violate $x^1 \le z^{1\pi}$. In the theorem

below, we show that we can still compute an optimal advance-scheduling policy by modifying our calculation of the allocation function. The result shows that the optimal policy is highly robust and can be calculated easily in a variety of circumstances. Further, since the result is an application of Theorem 5 to the nonstationary advance-scheduling model, it also showcases the usefulness of this model.

THEOREM 7. *Let the schedule at the beginning of period $t$ be $x \neq 0$. Modify the allocation-scheduling problem in §4 by setting $u^s(y) = u^s(x_{s-t+1} + y)$ for every $y$ and every $s \geq t$. Let $q_x^{t*}$ be the corresponding optimal allocation function. Then an optimal schedule in period $t$ is given by*

$$z_1^t(\delta, x) = x_1 + q_x^{*t}(\delta), \tag{8}$$

$$z_k^t(\delta, x) = x_k + q_x^{*t+k-1}\left(\delta - \sum_{i=1}^{k-1} z_i^t(\delta, x)\right), \quad k = 2, \ldots, \infty. \tag{9}$$

To see the above result, notice that if the starting schedule in period $t$ is $x$, then in each period $s \geq t$, at least $x_{s-t+1}$ patients must be served. We make use of the fact that the optimal policy given in Theorem 5 holds for nonstationary, time-dependent utilization costs. We modify the advance-scheduling problem by computing the extra number of patients to serve in period $s$ in addition to the quantity $x_{s-t+1}$.

## 7. Extensions

Although our model is motivated by the MRI-scheduling problem, it is general enough to apply to the scheduling of other types of diagnostic tests, such as x-rays, ultrasounds, fluoroscopy, etc.; to surgical scheduling, as shown, for example, by Gerchak et al. (1996) and Huh et al. (2013); to general outpatient-care scheduling; and to other types of service scheduling.

All of the results established in this paper extend straightforwardly to models with positive scheduling lead times. That is, the decision in each period $t$ is how to assign to patient appointments in period $t + L$, $t + L + 1, \ldots$, for some positive integer $L$.

Revenue per patient treated can also be included in the model without changing any of our results. Similarly, although we have defined the end-of-horizon penalty to be $V^{T+1}(\cdot, \cdot) = 0$, any linear penalty can be used without changing any of our results.

Whether these results extend to more general models—for example, those featuring no-shows and cancellations, an arbitrary number of patient classes, and a more general cost structure—is an open question at present.

## 8. Numerical Example

We illustrate the optimal advance-scheduling policy $\pi$ with a numerical example. For convenience, we will build on the computational examples provided by Gerchak et al. (1996). Note that Gerchak et al. model revenue per patient treated, but we do not. However, as we have pointed out in §7, all of our results hold with the addition of this revenue.

Consider the stationary setting. Assume that time is the only resource needed. Each day has two eight-hour shifts, for a total of 16 available hours. The total time used to treat urgent patients each day is normally distributed with mean of 400 minutes and standard deviation of 80 minutes. The time required to treat a regular patient is normally distributed with a mean of 60 and standard deviation of 10. The daily regular demand is Poisson distributed with mean $\mathbf{E}[\delta] = 8$ or 10. The revenue earned per regular patient is 600. The daily discount rate is 0.99. The utilization cost as a function of the total number of hours $y$ used in a day is given by $u_1(y) = (y - 16)^+ \times 15$. The unit waiting cost is $W = 2.99$.

The allocation function $q^*(\cdot)$ is shown in Figure 1 for the cases where $\mathbf{E}[\delta] = 8$ and $\mathbf{E}[\delta] = 10$. Note that the function is increasing in each case.

Now fix the mean demand to be 8. Assume that we start in period 1 with an empty schedule. Suppose that 35 patients arrive in the first period. The total number of patients outstanding is therefore 35. Since $q^*(35) = 9$, we allocate 9 patients to period 1, so that 26 patients remain. Since $q^*(26) = 9$, we allocate 9 patients to period 2, so that 17 patients remain. Since $q^*(17) = 8$, we allocate eight patients to period 3, so that nine patients remain. Since $q^*(9) = 8$, we allocate eight patients to period 4, so that one patient remains. Since $q^*(1) = 1$, we allocate this last patient

**Figure 1** **Allocation Function When $\mathbf{E}[\delta] = 8$ (Solid Line) and $\mathbf{E}[\delta] = 10$ (Dashed Line)**

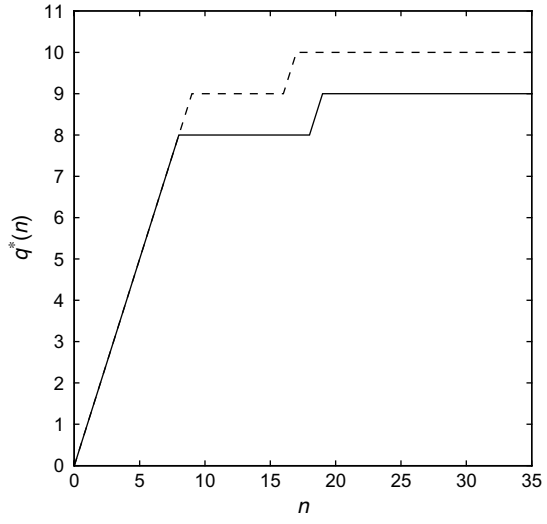
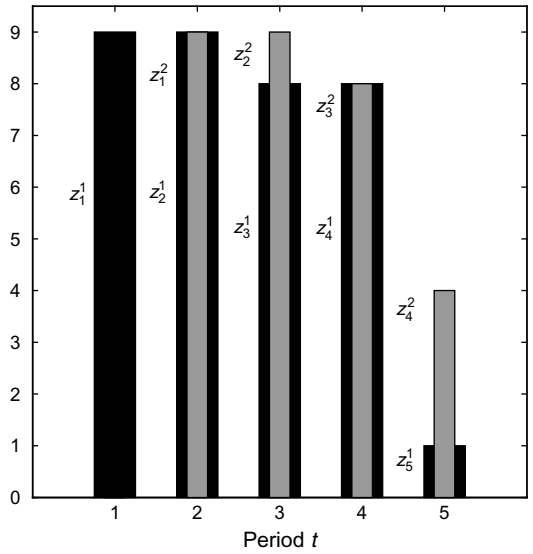


**Figure 2** **Optimal Schedules at the End of Period 1 (Black Bars) and Period 2 (Gray Bars) When $\mathbf{E}[\delta] = 8$, $\delta^1 = 35$, and $\delta^2 = 4$**




to period 5. This schedule, which is the schedule chosen at the end of period 1, is depicted by the black bars in Figure 2.

By the time period 2 arrives, we have already served nine patients in period 1. The remaining schedule allocates nine, eight, eight, and one patients to periods 2, 3, 4, and 5, respectively. Suppose that four new patients arrive, bringing the total number of patients outstanding to $9+8+8+1+4=30$. Using the procedure described above, we will find that our new schedule allocates nine, nine, eight, and four patients to periods 2, 3, 4, and 5, respectively. This new schedule is depicted by the gray bars in Figure 2. Note that the new schedule is a successive refinement of the old schedule, in the sense that it only incrementally adds patients to the old schedule.

## 9. Conclusions

### 9.1. Comparison with Threshold Policies

Note that for any fixed threshold, allocation scheduling under the threshold policy is identical in cost to advance scheduling under the threshold policy. Thus, the optimal threshold policy for allocation scheduling is identical in cost to the optimal threshold policy for advance scheduling. In this paper, we show that the optimal (general) policy for allocation scheduling is also identical in cost to the optimal policy for advance scheduling.

Therefore, to explore the performance gap between the optimal policy and the optimal threshold policy for advance scheduling, it suffices to examine the analogous gap in allocation scheduling. This latter gap has been studied extensively in stationary settings. Erdelyi and Topaloglu (2009) find that the gap is at most 4.08% in the problems they study. On a different set of problems, Gerchak et al. (1996) find that the gap is "situation dependent," ranging between 0.2% and 1.5%. They note, however, that even when the gap is a small fraction of total expected cost it can be "sometimes significant in absolute terms" (pp. 330–331).

### 9.2. Implementation Issues

We assume in our model that the decision maker has full knowledge at the beginning of the horizon of the distributions of the daily regular and emergency demand for all periods. In practice, the distribution of daily regular and emergency arrivals can be inferred from historical data. This distribution can be used unmodified for all periods in stationary settings. The distribution can be scaled to reflect growth and seasonality in nonstationary settings.

The utilization function $u^t(\,)$ captures the cost of providing service to a given number of patients. For each resource used, the utilization function can be estimated from the amount of the resource available on a given day, the cost of providing service using the resource in regular and overtime, and the random amount of the resource consumed by each patient. The nonstationary setting is the more relevant setting in healthcare. In these nonstationary settings, resource availability is usually determined a few months in advance through tactical or strategic planning. Usually, the most relevant resource is provider time. The amount of provider time consumed by each patient can be estimated from historical data. The amount of provider time available is usually determined by staff schedules that are usually known at least three months in advance, with some variations provided for unplanned absences. Provider costs are reflected in wages.

Given the utilization function and the arrival distribution, the allocation function $q^t(\,)$, and thus the optimal schedule, can be easily computed by using backward recursion with $T = 30$ (capturing a horizon of one month) or larger. The value function (for allocation scheduling) in each period is a univariate function with domain $\{0, 1, 2, \ldots, M\}$ for a sufficiently large constant $M$ capturing the maximum possible number of regular patients outstanding at any time.

Since the optimal policy specifies the optimal workload for each day, rather than the specific patients to be served, there is flexibility in implementation to honor a FIFO order. If FIFO service is desired, patients can be asked at the time of appointment whether they are willing to be expedited and, if so, about the range of days within which their appointments can be moved. If a day's workload is increased under the

optimal policy, patients with flexibility can be expedited before new patients are added to the schedule ahead of them.

### 9.3. Insights and Future Directions

For the MRI-scheduling problem that motivates this research and similar problems in diagnostic-imaging scheduling, surgical scheduling, and outpatient scheduling, we show that there is an easy-to-compute optimal policy. The optimal policy does not schedule the same number of regular patients for each day but dynamically increases the regular workload for each day as the total number of regular patients outstanding increases. Further, with each increase of a regular patient to the number outstanding, the optimal regular workload for each day only increases rather than decreases, and it increases by at most one patient. In stationary settings, the optimal policy schedules fewer regular patients further into the future. Finally, there is no loss of flexibility to the system in scheduling patients via advance scheduling rather than allocation scheduling because under advance scheduling, it is always possible to incrementally revise the schedule following an optimal policy.

More generally, we have connected the two main, independently analyzed paradigms in appointment scheduling, advance scheduling and allocation scheduling, by constructing an optimal solution for the one from the other. We have obtained a simple algorithm that is both efficient and optimal for a high-dimensional advance-scheduling problem. We have introduced the fundamental property of successive refinability under which advance scheduling is as efficient as allocation scheduling and where the cost of making advance commitments to patients is zero. It is a promising direction for future research to derive necessary and sufficient conditions for successive refinability and to identify these conditions in more general settings.

### Acknowledgments

The author gratefully acknowledges the helpful comments and constructive feedback provided by the anonymous referees.

## Appendix. Proofs

Proof of Theorem 1. Assume that $T<\infty$, and $\underline{V}^{T+1}(\cdot)$ is convex and increasing trivially. Assume that the theorem holds for $t+1$. For each fixed $q$,

$$W\bar{w}+\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q))]+\gamma\mathbf{E}[\underline{V}^{t+1}(\bar{w}-q+\delta^{t+1})]$$

is jointly convex in $(q,\bar{w})$ and increasing in $\bar{w}$, since each function $S_r(q)$ is increasing and linear in $q$ in the sample-path sense (see Shaked and Shanthikumar 1988, Example 4.3). Moreover, the feasible region $[0,\bar{w}]$ is also convex. Therefore, $\underline{V}^t(\cdot)$ is convex by induction.

The proof for the stationary case is similar. □

Proof of Theorem 2. Assume that $T<\infty$. Note that $W\bar{w}+\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q))]$ is separable and therefore submodular. Moreover, $\underline{V}^{t+1}(\bar{w}-q+\delta^{t+1})$ is submodular in $(q,\bar{w})$ by the convexity of $\underline{V}(\cdot)$. Therefore, $G^t(q,\bar{w})$ is submodular.

The proof for the stationary case is similar. □

Proof of Theorem 3. Assume that $T<\infty$. That $q^{t*}(\cdot)$ is increasing follows from Theorem 2. Let

$$G^t(q,\bar{w})=W\bar{w}+\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q))]+\gamma\mathbf{E}[\underline{V}^{t+1}(\bar{w}-q+\delta^{t+1})],$$

$$F^t(q,\bar{w})=W\bar{w}+\gamma\mathbf{E}[\underline{V}^{t+1}(\bar{w}-q+\delta^{t+1})].$$

Fix $t$ and $\bar{w}$, and let $q=q^{t*}(\bar{w})$. We have

$$\begin{aligned}
&G^t(q+2,\bar{w}+1)-G^t(q+1,\bar{w}+1)\\
&\quad=F^t(q+2,\bar{w}+1)-F^t(q+1,\bar{w}+1)+\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q+2))]\\
&\qquad-\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q+1))]\\
&\quad=F^t(q+1,\bar{w})-F^t(q,\bar{w})+\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q+2))]\\
&\qquad-\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q+1))]\\
&\quad\geq F^t(q+1,\bar{w})-F^t(q,\bar{w})+\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q+1))]\\
&\qquad-\sum_{r=1}^{R}\mathbf{E}[u_r^t(\epsilon_r^t+S_r(q))]\\
&\quad=G^t(q+1,\bar{w})-G^t(q,\bar{w})\\
&\quad\geq 0.
\end{aligned}$$

Therefore, $q^{t*}(\bar{w}+1)\leq q+1$.

The proof for the stationary case is similar. □

Proof of Proposition 2. For $k=1,\ldots,T-t+1$, if $\delta+|x|-\sum_{i=1}^{k-1}z_i^{t\pi}(\delta,x)\geq 1$, then, since $q^{*t+k-1}(\delta+|x|-\sum_{i=1}^{k-1}z_i^{t\pi}(\delta,x))\geq 1$ by Assumption 1, we must have $z_k^{t\pi}(\delta,x)\geq 1$. Thus, there is a finite $K$ such that $\delta+|x|-\sum_{i=1}^{K-1}z_i^{t\pi}(\delta,x)\leq 0$ for the first time. Thus, $z_k^{t\pi}(\delta,x)=0$ for all $k\geq K$.

We also know that $\delta+|x|-\sum_{i=1}^{K-2}z_i^{t\pi}(\delta,x)\geq 0$. Thus,

$$\begin{aligned}
z_{K-1}^{t\pi}(\delta,x)&=q^{*t+K-2}\Bigg(\delta+|x|-\sum_{i=1}^{K-2}z_i^{t\pi}(\delta,x)\Bigg)\\
&\leq\delta+|x|-\sum_{i=1}^{K-2}z_i^{t\pi}(\delta,x)
\end{aligned}$$

by Theorem 3. Hence, $\delta+|x|-\sum_{i=1}^{K-1}z_i^{t\pi}(\delta,x)\geq 0$. It follows that $\delta+|x|-\sum_{i=1}^{K-1}z_i^{t\pi}(\delta,x)=0$, or $|z^{t\pi}(\delta,x)|=\delta+|x|$. □

Proof of Proposition 3. By definition,

$$\begin{aligned} z_k^{t\pi}(\delta, x) &= q^{*t+k-1}\left(\delta + |x| - \sum_{i=1}^{k-1} z_i^{t\pi}(\delta, x)\right) \\ &= q^{*t+k-1}\left(|z^{t\pi}(\delta, x)| - \sum_{i=1}^{k-1} z_i^{t\pi}(\delta, x)\right) \\ &= q^{*t+k-1}\left(\sum_{i=k}^{\infty} z_i^{t\pi}(\delta, x)\right) \\ &= q^{*t+k-1}(|\mathbf{s}^{k-1}(z^{t\pi}(\delta, x))|). \end{aligned}$$

The second equality follows from Proposition 2. □

Proof of Theorem 4. In period 1, since $z^0 = x^1 = 0$, the theorem is satisfied for any schedule $z^1$.

Assume that the theorem holds up to period $t-1$. By definition and by Proposition 3, for $k = 2, \ldots, \infty$,

$$x_k^{t\pi} = q^{t*}(|\mathbf{s}^{k-1}(x^{t\pi})|).$$

By Proposition 2,

$$|z^{t\pi}(\delta^t, x^{t\pi})| = |x^{t\pi}| + \delta^t \geq |x^{t\pi}|, \tag{10}$$

and by Proposition 3,

$$z_1^{t\pi}(\delta^t, x^{t\pi}) = q^{t*}(|z^{t\pi}(\delta^t, x^{t\pi})|). \tag{11}$$

Therefore, by Theorem 3,

$$\begin{aligned} x_1^{t\pi} &= q^{t*}(|x^{t\pi}|) \leq z_1^{t\pi}(\delta^t, x^{t\pi}) \leq q^{t*}(|x^{t\pi}|) + \delta^t \\ &= x_1^{t\pi} + \delta^t. \end{aligned} \tag{12}$$

In particular, $x_1^{t\pi} \leq z_1^{t\pi}(\delta^t, x^{t\pi}) \leq x_1^{t\pi} + \delta^t$.

From (12), it also follows that

$$\begin{aligned} |\mathbf{s}(z^{t\pi}(\delta^t, x^{t\pi}))| &= |x^{t\pi}| + \delta^t - z_1^{t\pi}(\delta^t, x^{t\pi}) \\ &\geq |\mathbf{s}(x^{t\pi})|, \end{aligned} \tag{13}$$

which implies that

$$\begin{aligned} z_2^{t\pi}(\delta^t, x^{t\pi}) &= q^{t+1*}(|\mathbf{s}(z^{t\pi}(\delta^t, x^{t\pi}))|) \geq q^{t+1*}(|\mathbf{s}(x^{t\pi})|) \\ &= x_2^{t\pi}, \end{aligned} \tag{14}$$

and

$$\begin{aligned} z_2^{t\pi}(\delta^t, x^{t\pi}) &= q^{t*}(|\mathbf{s}(z^{t\pi}(\delta^t, x^{t\pi}))|) \\ &= q^{t*}(|z^{t\pi}(\delta^t, x^{t\pi})| - z_1^{t\pi}(\delta^t, x^{t\pi})) \\ &= q^{t*}(|x^{t\pi}| + \delta^t - z_1^{t\pi}(\delta^t, x^{t\pi})) \\ &= q^{t*}(|\mathbf{s}(x^{t\pi})| + \underbrace{x_1^{t\pi} + \delta^t - z_1^{t\pi}(\delta^t, x^{t\pi})}_{\geq 0}) \\ &\leq q^{t*}(|\mathbf{s}(x^{t\pi})|) + \underbrace{x_1^{t\pi} + \delta^t - z_1^{t\pi}(\delta^t, x^{t\pi})}_{\geq 0} \\ &= x_2^{t\pi} + \underbrace{x_1^{t\pi} + \delta^t - z_1^{t\pi}(\delta^t, x^{t\pi})}_{\geq 0}. \end{aligned}$$

In other words,

$$x_2^{t\pi} \leq z_2^{t\pi}(\delta^t, x^{t\pi}) \leq x_2^{t\pi} + \underbrace{x_1^{t\pi} + \delta^t - z_1^{t\pi}(\delta^t, x^{t\pi})}_{\geq 0}. \tag{15}$$

Continue to deduce that for every $k = 1, 2, \ldots,$

$$\begin{aligned} x_k^{t\pi} &\leq z_k^{t\pi}(\delta^t, x^{t\pi}) \\ &\leq x_k^{t\pi} + \underbrace{\sum_{i=1}^{k-1} x_i^{t\pi} + \delta^t - \sum_{i=1}^{k-1} z_i^{t\pi}(\delta^t, x^{t\pi})}_{\geq 0}. \end{aligned} \tag{16}$$

Therefore, $\mathbf{s}(z^{t-1\pi}) = x^{t\pi} \leq z^{t\pi}(\delta^t, x^{t\pi})$. □

## References

Ayvaz N, Huh WT (2010) Allocation of hospital capacity to multiple types of patients. *J. Revenue Pricing Management* 9(5):386–398.

Berger T (1971) *Rate-Distortion Theory: A Mathematical Basis for Data Compression* (Prentice Hall, Upper Saddle River, NJ).

Carr S, Duenyas I (2000) Optimal admission control and sequencing in a make-to-stock/make-to-order production system. *Oper. Res.* 48(5):709–720.

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

Chakraborty S, Muthuraman K, Lawley M (2010) Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Trans.* 42(5):354–366.

Denton B, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35(11):1003–1016.

Dexter F, Macario A, Traub RD, Hopwood M, Lubarsky DA (1999) An operating room scheduling strategy to maximize the use of operating room block time: Computer simulation of patient scheduling and survey of patients? Preferences for surgical waiting time. *Anesthesia Analgesia* 89(1):7–20.

Ding Q, Kouvelis P, Milner JM (2006) Dynamic pricing through discounts for optimizing multiple-class demand fulfillment. *Oper. Res.* 54(1):169–183.

Dobson G, Hasija S, Pinker EJ (2011) Reserving capacity for urgent patients in primary care. *Production Oper. Management* 20(3):456–473.

Duran S, Liu T, Simchi-Levi D, Swann JL (2008) Policies utilizing tactical inventory for service-differentiated customers. *Oper. Res. Lett.* 36(2):259–264.

Equitz WHR, Cover TM (1991) Successive refinement of information. *IEEE Trans. Inform. Theory* 37(2):269–275.

Erdelyi A, Topaloglu H (2009) Computing protection level policies for dynamic capacity allocation problems by using stochastic approximation methods. *IIE Trans.* 41(6):498–510.

Esogbue AO, Singh AJ (1976) A stochastic model for an optimal priority bed distribution problem in a hospital ward. *Oper. Res.* 24(5):884–898.

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

George JA, Fox DR, Canvin RW (1983) A hospital throughput model in the context of long waiting lists. *J. Oper. Res. Soc.* 34(1):27–35.

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

Gocgun Y, Ghate A (2012) Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Comput. Oper. Res.* 39(10):2323–2336.

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

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

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

Gupta D, Wang L (2008) Revenue management for a primary-care clinic in the presence of patient choice. *Oper. Res.* 56(3):576–592.

Ha AY (1997a) Inventory rationing in a make-to-stock production system with several demand classes and lost sales. *Management Sci.* 43(8):1093–1103.

Ha AY (1997b) Optimal dynamic scheduling policy for a make-to-stock production system. *Oper. Res.* 45(1):42–53.

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

Huang XM (1995) A planning model for requirement of emergency beds. *Math. Medicine Biol.* 12(3–4):345–353.

Huang Y, Zuniga P (2011) Dynamic overbooking scheduling system to improve patient access. *J. Oper. Res. Soc.* 63(6):810–820.

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

Jouini O, Benjaafar S (2010) Queueing systems with appointment-driven arrivals, non-punctual customers, and no-shows. Working paper, University of Minnesota, Minneapolis.

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

Kim S, Giachetti RE (2006) A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Trans. Systems Man Cybernetics, Part A* 36(6):1211–1219.

Klassen KJ, Rohleder TR (2004) Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. *Internat. J. Service Indust. Management* 15(2):167–186.

LaGanga LR, Lawrence SR (2007) Clinic overbooking to improve patient access and increase provider productivity. *Decision Sci.* 38(2):251–276.

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

Leung JYT (2004) *Handbook of Scheduling: Algorithms, Models, and Performance Analysis* (Chapman and Hall/CRC, New York).

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

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

Lowery JC, Martin JB (1989) Evaluation of an advance surgical scheduling system. *J. Medical Systems* 13(1):11–23.

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

Maglaras C, Zeevi A (2005) Pricing and design of differentiated services: Approximate analysis and structural insights. *Oper. Res.* 53(2):242–262.

McManus ML, Long MC, Cooper A, Mandell J, Berwick DM, Pagano M, Litvak E (2003) Variability in surgical caseload and access to intensive care services. *Anesthesiology* 98(6):1491–1496.

Min D, Yih Y (2010) An elective surgery scheduling problem considering patient priority. *Comput. Oper. Res.* 37(6):1091–1099.

Min D, Yih Y (2014) Managing a patient waiting list with time-dependent priority and adverse events. *RAIRO Oper. Res.* 48(1):53–74.

Moore IC, Strum DP, Vargas LG, Thomson DJ (2008) Observations on surgical demand time series: Detection and resolution of holiday variance. *Anesthesiology* 109(3):408–416.

Mullen PM (2003) Prioritising waiting lists: How and why? *Eur. J. Oper. Res.* 150(1):32–45.

Murray M, Berwick DM (2003) Advanced access. *J. Amer. Medical Assoc.* 289(8):1035–1040.

Muthuraman K, Lawley M (2008) A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Trans.* 40(9):820–837.

Patrick J (2012) A Markov decision model for determining optimal outpatient scheduling. *Health Care Management Sci.* 15(2):91–102.

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

Qu X, Rardin RL, Williams JAS, Willis DR (2007) Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *Eur. J. Oper. Res.* 183(2):812–826.

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

Robinson LW, Chen RR (2003) Scheduling doctors' appointments: Optimal and empirically-based heuristic policies. *IIE Trans.* 35(3):295–307.

Rohleder TR, Klassen KJ (2000) Using client-variance information to improve dynamic appointment scheduling performance. *Omega* 28(3):293–302.

Shaked M, Shanthikumar JG (1988) Stochastic convexity and its applications. *Adv. Appl. Probability* 20(2):427–446.

Sobel MJ, Zhang RQ (2001) Inventory policies for systems with stochastic and deterministic demand. *Oper. Res.* 49(1):157–162.

Swain RW, Kilpatrick KE, Marsh JJ III (1977) Implementation of a model for census prediction and control. *Health Services Res.* 12(4):380–395.

Talluri KT, Van Ryzin G (2005) *The Theory and Practice of Revenue Management*, Vol. 68 (Springer, New York).

Topkis DM (1968) Optimal ordering and rationing policies in a non-stationary dynamic inventory model with $n$ demand classes. *Management Sci.* 15(3):160–176.

Truong VA, Ruzal-Shapiro C (2013) Optimal advanced scheduling, with expediting. Working paper, Columbia University, New York.

Wang PP (1999) Sequencing and scheduling $N$ customers for a stochastic server. *Eur. J. Oper. Res.* 119(3):729–738.

Young JP (1963) A queuing theory approach to the control of hospital inpatient census. Unpublished thesis, Johns Hopkins University, Baltimore.

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.

Zhou Y, Zhao X (2010) A two-demand-class inventory system with lost-sales and backorders. *Oper. Res. Lett.* 38(4):261–266.