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

INFORMS JOURNAL ON COMPUTING
Vol. 34, No. 5, September–October 2022, pp. 2845–2865
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

A Simulation Optimization Approach for the Appointment Scheduling Problem with Decision-Dependent Uncertainties

Tito Homem-de-Mello,[a] Qingxia Kong,[b,*] Rodrigo Godoy-Barba[c]

[a] School of Business, Universidad Adolfo Ibáñez, 7941169 Santiago, Chile; [b] Rotterdam School of Management, Erasmus University Rotterdam, 3000 DR Rotterdam, Netherlands; [c] Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, 7941169 Santiago, Chile
*Corresponding author

Contact: tito.hmello@uai.cl, https://orcid.org/0000-0002-2044-3306 (TH-d-M); q.kong@rsm.nl, https://orcid.org/0000-0002-6812-9535 (QK); rodrigogoba@gmail.com (RG-B)

Received: December 4, 2020
Revised: August 11, 2021; December 3, 2021; March 7, 2022; April 22, 2022; May 13, 2022
Accepted: May 13, 2022
Published Online in Articles in Advance: July 18, 2022

https://doi.org/10.1287/ijoc.2022.1212

Abstract. The appointment scheduling problem (ASP) studies how to manage patient arrivals to a healthcare system to improve system performance. An important challenge occurs when some patients may not show up for an appointment. Although the ASP is well studied in the literature, the vast majority of the existing work does not consider the well-observed phenomenon that patient no-show is influenced by the appointment time, the usual decision variable in the ASP. This paper studies the ASP with random service time (exogenous uncertainty) with known distribution and patient decision-dependent no-show behavior (endogenous uncertainty). This problem belongs to the class of stochastic optimization with decision-dependent uncertainties. Such problems are notoriously difficult as they are typically nonconvex. We propose a stochastic projected gradient path (SPGP) method to solve the problem, which requires the development of a gradient estimator of the objective function—a nontrivial task, as the literature on gradient-based optimization algorithms for problems with decision-dependent uncertainty is very scarce and unsuitable for our model. Our method can solve the ASP problem under arbitrarily smooth show-up probability functions. We present solutions under different patterns of no-show behavior and demonstrate that breaking the assumption of constant show-up probability substantially changes the scheduling solutions. We conduct numerical experiments in a variety of settings to compare our results with those obtained with a distributionally robust optimization method developed in the literature. The cost reduction obtained with our method, which we call the value of distribution information, can be interpreted as how much the system performance can be improved by knowing the distribution of the service times, compared to not knowing it. We observe that the value of distribution information is up to 31% of the baseline cost, and that such value is higher when the system is crowded or/and the waiting time cost is relatively high.

Summary of Contribution: This paper studies an appointment scheduling problem under time-dependent patient no-show behavior, a situation commonly observed in practice. The problem belongs to the class of stochastic optimization problems with decision-dependent uncertainties in the operations research literature. Such problems are notoriously difficult to solve as a result of the lack of convexity, and the present case requires different techniques because of the presence of continuous distributions for the service times. A stochastic projected gradient path method, which includes the development of specialized techniques to estimate the gradient of the objective function, is proposed to solve the problem. For problems with a similar structure, the algorithm can be applied once the gradient estimator of the objective function is obtained. Extensive numerical studies are presented to demonstrate the quality of the solutions, the importance of modeling time-dependent no-shows in appointment scheduling, and the value of using distribution information about the service times.

History: Accepted by J. Paul Brooks, Area Editor for Applications in Biology, Medicine, & Healthcare.
Supplemental Material: The software that supports the findings of this study is available within the paper and its online supplement [https://doi.org/10.1287/ijoc.2022.1212] or is available from the IJOC GitHub software repository (https://github.com/INFORMSJoC) at http://dx.doi.org/10.5281/zenodo.6471628.

Keywords: stochastic optimization • decision-dependent uncertainty • appointment scheduling • no-show • healthcare management

1. Introduction

Deciding appointment times for patients is one of the critical decisions outpatient healthcare providers have to make to better utilize critical resources, reduce overtime work of healthcare providers, and, at the same time, shorten the waiting time of patients. Shorter interarrival times can lead to higher utilization of the resources but longer patient waiting times. Conversely, longer interarrival times can reduce patient waiting but increase idle time and overtime of the service system. Appointment systems are stochastic in nature. For instance, service duration—an important factor to consider in the appointment scheduling problem (ASP)—is usually random. Therefore, deciding optimal schedules for the patients involves trading off competing objectives under a stochastic setting.

In practice, there exists a second source of uncertainty, incurred by the possibility that a patient does not show up for his or her appointment—that is, patient no-show behavior. Every healthcare system bears the risk of patient no-show, which has significant impacts on the healthcare system including increased system costs, underutilized critical resources, and reduced patient outcome (Moore et al. 2001, Lacy et al. 2004, Daggy et al. 2010). Besides individual characteristics such as demographic and social-economic factors, home-clinic distance, and no-show history, current literature reports that patient no-show behavior can actually depend on time (Dantas et al. 2018). Evidence shows that the show-up probability depends not only on the appointment day and day of the week (Gallucci et al. 2005, Samorani and LaGanga 2015) but also on the time of the day at which the patient is scheduled. Moore et al. (2001) analyze a hospital in South Carolina; their results suggest that morning appointments are more likely to be kept than afternoon ones. LaGanga (2011) presents evidence from a mental health center showing that show-up rates vary with the time of the day. Dantas et al. (2019) in a Bariatric clinic study show that patients tend to show up less toward the later hours of the day. Kong et al. (2020) analyze two independent data sets and have interesting results: on weekdays, patients in a U.S. adult healthcare facility are more likely to show up at the beginning of the day or late in the afternoon, whereas in a Chilean pediatric practice, the show-up rates on weekdays are lower in the early morning, and a peak is seen in the middle of the day. It is important to mention that the consideration of time-dependent no-show behavior is not just a matter of increasing modeling accuracy; it can actually lead to much better system performance. For example, LaGanga and Lawrence (2012) show that the system performance (measured by net utility) using time-dependent no-show rates is twice as good as the performance obtained using average no-show rates across 16 slots in a day.

Despite the widely recognized evidence on time-dependent no-show behavior, scarcely any work on appointment scheduling has taken this behavior into consideration. A major difficulty in this problem is the fact that show-up probabilities depend on the scheduled time of appointment; thus, the problem belongs to the class of stochastic optimization models with decision-dependent uncertainty—or, more specifically, decision-dependent probabilities, following the taxonomy of Hellemo et al. (2018). Such problems are notoriously difficult to solve, especially because of their lack of convexity, and solution methods found in the literature often exploit the structure of the problem. A recent survey of such methods can be found in Hellemo et al. (2018). It is worthwhile noticing, however, that the methods discussed in Hellemo et al. (2018, p. 377) always assume a finite (and moderately sized) number of possible scenarios—indeed, the authors of that paper say that "[we] are not familiar with any attempts to model and solve problems with decision-dependent probabilities using continuous probability distributions." In our case, although the decision-dependent random variables do have finite support (as they represent the attendance or not of each patient), the presence of continuous distributions to represent service times requires different techniques to solve the problem.

In this paper, we present a method to solve the ASP with time-dependent show-up probabilities and stochastic service times. Our approach has two main ingredients. The first one, which is key for subsequent developments, is the estimation of gradients of the objective function $\mathbb{E}[f(\mathbf{x},\xi)]$ with respect to the decision variables $\mathbf{x}$, where ξ denotes the random variables in the problem. The topic of gradient estimation of stochastic functions is very well studied; see, for instance, Fu and Hu (1997), Glasserman (1991), and Rubinstein and Shapiro (1993) for comprehensive discussions and also Fu (2015) for a more recent survey of existing techniques. These general approaches typically assume that either (i) the underlying probability distributions are exogenous and do not depend on the decision variables (i.e., ξ does not depend on $\mathbf{x}$) or (ii) the dependence on the decision variables occurs *only* in the probability distributions, as in the case, for example, of queuing systems where the distribution of service times depends on some controllable parameter—that is, $\mathbb{E}[f(\mathbf{x},\xi)] = \mathbb{E}[g(\xi(\mathbf{x}))]$ for some function g. Our case, however, is more involved; indeed, because the decision variables are the arrival times and the show-up probabilities depend on such times, the decision variables appear as parameters both in the integrand of the objective function and in the probability distributions, so we actually have $\mathbb{E}[f(\mathbf{x},\xi(\mathbf{x}))]$. By exploiting the structure of the uncertainty, we are able to derive expressions for the gradient that can be easily estimated with Monte Carlo simulation. A particular advantage of our approach is that it allows for a wide variety of functional forms to represent the dependence of show-up probabilities on time; the only requirement is that such a function be differentiable. Such flexibility contrasts with previous works in the literature, which deal only with piecewise-linear function forms (see Kong et al. (2020)).

The second ingredient of our approach is the optimization algorithm. By making a mild assumption on the decision variables—namely, that no two patients can be scheduled to arrive at the *exact* same moment—we show that the objective function is differentiable. Thus, by making use of the gradient estimation technique discussed in the preceding, we can apply a sampling-based first-order stochastic optimization method. We develop a new stochastic projected gradient path (SPGP) method to solve the constrained stochastic problem with decision-dependent uncertainties, building on the stochastic trust-region response-surface method (STRONG) proposed by Chang et al. (2013) to find stationary points and Cauchy line searches originally designed for deterministic problems (Conn et al. 2000). STRONG combines the advantage of the trust-region methods and the response-surface methodology to solve unconstrained simulation optimization problems. However, because our problem is constrained, we integrate a Cauchy line search with the method to incorporate gradient projections. Along the way, we have made an important enhancement to the line search procedure to avoid the calculation of projections onto tangent cones, which can be computationally expensive. That enhancement may be of independent interest, as it is not related to our specific problem.

Although the resulting algorithm yields locally optimal solutions (which is natural, because the problem is nonconvex), by applying a multistart procedure, we are able to obtain good solutions. As discussed earlier, we are not aware of any other approach that can model arbitrarily smooth show-up probability functions; the only available benchmark to our method is the work of Kong et al. (2020), who model this problem as a distributionally robust optimization and solve for the schedules using a set of approximations and heuristics. The method proposed in that paper uses only the first two moments and thus avoids more complexity brought on by estimating any specific distribution. That method, however, can be very conservative under some extreme circumstances—for example, when the system is overcrowded. Using a set of computational studies, we show that our approach makes significant improvements compared with Kong et al. (2020) in terms of the quality of the solutions. As the latter method only assumes knowledge of the first two moments of the service time distribution, the cost reduction obtained with our method can be interpreted as how much the system performance can be improved by knowing the distribution of the service times compared with not knowing it—a quantity we call the *value of distribution information.* In our experiments, the value of distribution information reached up to 31% of the baseline cost. Moreover, the gains were higher in the cases where the system is more crowded or the waiting time is more costly.

Finally, we present solutions under different show-up functions and demonstrate that breaking the assumption of constant show-up probability substantially changes the solution patterns and significantly reduces system costs. We observe that significant improvements can be obtained by incorporating the time-dependent no-show behavior: under most show-up function forms, significant improvements in performance (22.4%–49.2%) can be observed, whereas relatively milder improvements (3.7%–5.6%) are seen in the increasing show-up function. Such improvements thus demonstrate that if the time-dependent no-show behavior is ignored or one does not have access to a time-dependent no-show model, then a simplistic approach is likely to be adopted and the loss in terms of costs can be significant.

To summarize, the contributions of this paper are the following:

- We study the appointment scheduling problem where patient no-show behavior depends on the appointment time, a phenomenon that often occurs in practice but has very much been neglected in the literature except in a couple of papers. To the best of our knowledge, ours is the very first paper to consider the case where the no-show behavior pattern across time can be very general.
- We formulate the appointment scheduling problem with time-dependent no-shows as a stochastic optimization problem with *decision-dependent uncertainty* and show that, in general, the problem is *nonconvex*.
- We prove that the objective function of the problem is differentiable under mild assumptions on the show-up and service processes. Moreover, we provide unbiased consistent estimators of the gradient of the objective function, a result that has independent interest because the literature on gradient estimators for stochastic optimization problems with decision-dependent uncertainty is very scarce.
- Using the gradient estimators described in the preceding, we adapt a trust-region based stochastic optimization method for unconstrained problems proposed in the literature to our constrained setting by incorporating a projected gradient path algorithm. Along the way, we provide an enhancement to that algorithm that bypasses the calculation of tangent cones without sacrificing its convergence properties—a

result that also has independent interest, as it is valid beyond the setting of this paper. Even though our method may yield local optima (because of the lack of convexity of the problem), the multistart procedure adopted in the paper shows very good performance compared with the approach of Kong et al. (2020), which is the only available benchmark in the literature.

2. Literature Review

The appointment scheduling problem is extensively studied in the literature of operations management. Interested readers can refer to Cayirli and Veral (2003), Ahmadi-Javid et al. (2017), and Gupta and Denton (2008) for excellent literature reviews on ASP. The following literature review focuses on appointment scheduling problems with patient no-show behavior.

Cayirli and Veral (2003) provide an extensive review of related literature, identifying differences in the design of the appointment system, performance measures, sources of uncertainties, and methodologies applied. In relations to source of uncertainties, there are cases where only the randomness of service time has been considered (Denton and Gupta 2003, Robinson and Chen 2003), only patients' no-show behavior (Zacharias and Pinedo 2014) is considered, or both are considered (Hassin and Mendel 2008, Erdogan and Denton 2013, Jiang et al. 2017, Kong et al. 2020). Our paper considers two main sources of uncertainty: stochastic service time and patients' time-dependent no-show.

Erdogan and Denton (2013) formulate two appointment scheduling models using stochastic linear programming. One model extends the sequential bounding approach developed in Denton and Gupta (2003) and incorporates patient no-shows in the ASP. Their computational experiments show that the optimal interarrival times decrease as the no-show probabilities increase. Hassin and Mendel (2008) investigate the ASP using a single-server model with no-shows. They assume exponential service-time distribution and time-invariant no-show; that is, the show-up probabilities remain unchanged across the planning session. They report continuous scheduling decisions (versus template) and show that the optimal schedules in the no-show context still exhibit a dome-shaped pattern. We use that paper as one of the benchmarks for the stochastic projected gradient path algorithm developed in our paper and show that our method can replicate the optimal patterns reported in Hassin and Mendel (2008). Zacharias and Pinedo (2014) study an overbooking model for scheduling arrivals within a day of heterogeneous patients (i.e., having different show-up probabilities and different waiting costs). They divide the working day into a finite number of slots in which a number of patients is assigned to every slot. They present priority rules and structural results for the optimal schedule and find optimal schedules for the case of homogeneous patients. Jiang et al. (2017) consider a distributionally robust model under static patient no-show and random service duration, studying the case's risk-neutral and risk-averse system operator in order to incorporate his or her risk preferences.

Two works are directly related to this paper. Kong et al. (2020) and LaGanga and Lawrence (2012) present some results about the effect of this time dependence. LaGanga and Lawrence (2012) present an appointment scheduling model that considers no-show and deterministic service time. A heuristic is proposed based on a gradient search algorithm and adapted to solve the case with slot dependent show-up probability. Although LaGanga and Lawrence give one of the first results on the slot-dependent case, the solution procedure they propose is based on a heuristic as a result of its computational difficulty. Therefore, some research opportunities are left open for new models that consider stochastic service times and a more systematic optimization approach.

Kong et al. (2020) present empirical evidence using data from two different countries that patient no-show behavior depends on the time of the day. They develop a distributionally robust model that allows for schedule-dependent show-up rates under stochastic service times. Their computational results suggest that significant efficiency gains can be achieved when time-of-day effect on show-up probabilities is incorporated. Kong et al. (2020) is arguably the first work that incorporates the slot-dependent no-show into appointment scheduling problem with random service time. Their methodology, however, has several limitations: First, the solution procedure consists of several layers of approximations (semidefinite programming approximations to copositive and completely positive cones) and heuristics (iterative approach). It is not even clear whether the solution generates upper or lower bounds. Second, because of the cone structure, the current solution proposed in the paper can only deal with piecewise-linear show-up functions with time. Third, because it solves for the worst-case scenarios, the solution generated can be very conservative in some cases—for example, when the system is overcrowded or the show-up probabilities are large toward the end of the session.

Inspired by those limitations, the present work proposes a methodology to find solutions to the *exact* appointment scheduling problem with *known* service time distributions. In the first crucial step, we exploit the structure of the problem and derive the gradient estimator of the objective function. We then develop the SPGP method to solve the constrained stochastic problem with decision-dependent uncertainties, improving on two existing originally designed for unconstrained and deterministic problems. In the following, we briefly review the stochastic optimization literature on decision-dependent uncertainties.

2.1. Decision-Dependent Uncertainties

The stochastic programming problem with decision-dependent uncertainty—where the decision variables influence the underlying stochastic process—is known to be difficult to solve. Pflug (1990) is arguably the first paper that works on this topic, where the optimization decisions can influence the stochastic process. Later on, several papers appear to tackle the problem formulated in different contexts/fields, such as facility location (Basciftci et al. 2021), operations scheduling in power systems (Basciftci et al. 2020), gas field planning (Goel and Grossmann 2004), and project decisions (Jonsbråten et al. 1998). Goel and Grossmann (2006) address a class of stochastic programs with decision-dependent uncertainties. They use mixed-integer disjunctive programs to describe the decision dependency, prove some structural properties of size reduction, and propose a branch-and-bound algorithm. Hellemo et al. (2018) give an excellent review on the recent work of decision-dependent uncertainty. These techniques usually need to assume finite support of the decision variables and represent the problem with scenario trees. Our problem, however, consists of continuous random variables (service times). We tackle this problem by using a simulation optimization approach.

3. The Model

3.1. Problem Description

In this section, we introduce the stochastic model for the appointment schedule problem with patient time-dependent no-show behavior and our model assumptions. Following the literature (E.G., Cayirli and Veral 2003), our paper assumes that a fixed number of patients are to be scheduled to arrive at a fixed-length clinic session with a single service provider. Patients may not show up for their appointments, thus incurring patient no-show; in particular, we assume that patient show-up probabilities depend on the time of arrivals. If they show up, they always arrive at the scheduled time (i.e., nonpunctual arrivals are not considered). Walk-in and emergency arrivals are not considered either. The arrival sequence is predetermined; thus arrival times are the only (continuous) decision variables. Service time is assumed to be stochastic. This stochastic nature of the system together with patient no-show behavior may incur patient waiting time, service provider idle time, and overtime. The objective is to minimize a weighted sum of the three performance measures by deciding the optimal schedule for each patient.

As is customary in the literature, the decision-making process in this paper is assumed to be static (offline); that is, all appointment requests are known in advance to the scheduling decision. Some recent work (Erdogan et al. 2015, Truong 2015, Parizi and Ghate 2016) studies the *online* appointment scheduling problem, in which they dynamically assign arriving patients to the remaining slots. Thus, the present work focuses on drawing insights into the appointment design under time-dependent no-show over a longer planning horizon.

3.2. Problem Formulation

Table 1 presents the notation used in this paper. Among them, $\mathbf{x} = (x_i)$, $i = 0, \dots, n$ are the decision variables and denote the allocated service interval for patient i, with $x_0\,(\geq 0)$ denoting the arrival time of the first patient. Each element in $\mathbf{x}$ is nonnegative, and $\sum_{i=0}^{n-1} x_i \leq T$, which indicate that each patient is allocated with a nonnegative time interval and all patients arrive within the clinic session $[0, T]$. The scheduled arrival time for patient i is represented by $\sum_{j=0}^{i-1} x_j$ for $i \geq 1$. The sources of randomness of the system are represented by $\mathbf{A}$ and $\mathbf{U}$, which represent the attendance status of each patient and the patient's service

Table 1. Notation

Notation	Description
n	Number of patients to be scheduled
T	Length of the clinic session
A_i	Random attendance status of patient i.
U_i	Random service time for patient i
W_i	Waiting time of patient i
S_i	Idle time on time (slot) assigned to patient i
L	Overtime of the session
x_i	Time assigned to patient i (decision variable)
TC	Total cost
c_w	Cost of waiting time per unit of time
c_I	Cost of idle time per unit of time
c_o	Cost of overtime per unit of time
$p(t)$	Probability that a patient shows up when his or her appointment is scheduled at time t

time, respectively. More specifically, $\mathbf{A}$ is a vector of n Bernoulli random variables ($A_i = 1$ if patient i shows up and $A_i = 0$ otherwise), and $\mathbf{U}$ is a vector of n random variables characterizing the service times. Throughout this paper, we assume that each U_i is a nonnegative random variable with $\mathbb{E}[U_i] < \infty$.

Similar to Denton and Gupta (2003), we next present expressions for the total cost of a given schedule. The total cost is a weighted sum of waiting time, idle time, and overtime of the system, and it is expressed as

$$TC(\mathbf{x}, \mathbf{A}, \mathbf{U}) = c_o L + c_I x_0 + \sum_{i=1}^{n} [c_w A_i W_i + c_I S_i], \tag{1}$$

where L represents the overtime, W_i the waiting time of patient i, and S_i the idle time during the allocated service interval (x_i) for patient i. They are expressed as follows:

$$W_1 = 0, \tag{2}$$

$$W_i = \max(0,\ W_{i-1} + A_{i-1}U_{i-1} - x_{i-1}), \quad i = 2, \dots, n, \tag{3}$$

$$x_n = T - \sum_{j=0}^{n-1} x_j. \tag{4}$$

$$S_i = \max(0,\ x_i - A_i U_i - W_i) \quad i = 1, \dots, n, \tag{5}$$

$$L = \max(0,\ W_n + A_n U_n - x_n). \tag{6}$$

Because of the crucial *decision-dependent uncertainty* characteristics of the problem, the distribution of the random vector $\mathbf{A}$ depends on the scheduling decisions $\mathbf{x}$, thereby turning the problem into a difficult stochastic optimization problem. We shall then write $\mathbf{A}(\mathbf{x})$ to emphasize that dependence. Note also that the random variables W, S, and L depend on $\mathbf{x}$ as well, but the dependence is omitted to simplify the notation. The optimization problem aims to find a schedule given by $\mathbf{x}$ that minimizes the expected total cost, subject to nonnegativity constraints, and it is formulated as follows:

$$\min_{\mathbf{x} \in \mathcal{X}} \{f(\mathbf{x}) := \mathbb{E}_{\mathbf{A}(\mathbf{x}),\mathbf{U}}[TC(\mathbf{x}, \mathbf{A}(\mathbf{x}), \mathbf{U})]\}, \tag{7}$$

where $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0,\ \sum_{j=0}^{n-1} x_j \leq T\}$. Note that $\mathbf{x}$ corresponds to the vector $(x_0, \dots, x_{n-1})$; we have dropped the last component (x_n) because it is defined via (4) directly as a function of $x_0, \dots, x_{n-1}$. Notice, moreover, that because $W_i \leq \sum_{j=1}^{i-1} U_j$, it follows from our aforementioned assumption, $\mathbb{E}[U_j] < \infty$, that $\mathbb{E}_{\mathbf{A}(\mathbf{x}),\mathbf{U}}[TC(\mathbf{x}, \mathbf{A}(\mathbf{x}), \mathbf{U})] < \infty$ for all $\mathbf{x} \in \mathcal{X}$; that is, the objective function in (7) is well defined for all $\mathbf{x} \in \mathcal{X}$.

One difficulty is to calculate the objective function $f(\mathbf{x})$, which involves the recursive equations (3) and (5). Assuming that the service times $\mathbf{U}$ do not depend on the time of the day—a reasonable assumption in this context—it follows from the classical tower property in probability theory that we can express the function as a conditional expectation as follows:

$$f(\mathbf{x}) = \mathbb{E}_{\mathbf{U}}[\varphi(\mathbf{U}, \mathbf{x})], \tag{8}$$

where

$$\varphi(\mathbf{U}, \mathbf{x}) := \mathbb{E}_{\mathbf{A}(\mathbf{x})}[TC(\mathbf{x}, \mathbf{A}(\mathbf{x}), \mathbf{U}) \mid \mathbf{U}]. \tag{9}$$

As we shall see shortly, the structure of the problem allows us to find an exact expression for the inner expectation $\varphi(\mathbf{U}, \mathbf{x})$, which represents the expected total cost for a given vector of service times. Suppose that the no-show behavior of patients is not affected by the behavior of other patients; that is, the Bernoulli components of the vector $\mathbf{A}(\mathbf{x})$ are mutually independent. By noticing that each random variable A_i is Bernoulli, we represent a scenario of patients' attendance realizations as a binary vector, denoted by $\omega_{\mathbf{A}}$, where a value of 1 in the component i of the vector indicates that patient i shows up, and 0 indicates otherwise. Thus, the probability of a realization $\omega_{\mathbf{A}} = (\omega_{A_1}, \dots, \omega_{A_n})$, which yields the probability mass function of the random vector $\mathbf{A}(\mathbf{x})$, is given by

$$\mathbb{P}(\mathbf{A}(\mathbf{x}) = \omega_{\mathbf{A}}) = \prod_{i=1}^{n} P_i(\mathbf{x}, \omega_{A_i}), \tag{10}$$

where $P_i(\mathbf{x}, \omega_{A_i})$ represents the probability mass function of the random variable $A_i(\mathbf{x})$ (i.e., $\mathbb{P}(A_i(\mathbf{x}) = \omega_{A_i})$). Recall from Table 1 that $p(t)$ represents a function that returns the probability that the patient shows up when their appointment is at time t. It follows that we can express the function $P_i(\mathbf{x}, \omega_{A_i})$ as

$$P_i(\mathbf{x}, \omega_{A_i}) := \mathbb{P}(A_i(\mathbf{x}) = \omega_{A_i}) = \begin{cases} p\left(\sum_{j=0}^{i-1} x_j\right) & \omega_{A_i} = 1, \\ 1 - p\left(\sum_{j=0}^{i-1} x_j\right) & \omega_{A_i} = 0. \end{cases} \tag{11}$$

From Expressions (10) and (11), we can see explicitly the *endogenous uncertainty*, where the probability of a scenario depends on the decision vector $\mathbf{x}$. Moreover, this dependence is defined by a highly nonlinear function. Also, the same equations imply that the function $\varphi(\mathbf{U}, \mathbf{x})$ defined in (9) can be expressed as

$$\varphi(\mathbf{U}, \mathbf{x}) = \sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} \left(TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) \prod_{i=1}^{n} P_i(\mathbf{x}, \omega_{A_i}) \right), \tag{12}$$

and thus the objective function $f(\mathbf{x})$ can be expressed as

$$f(\mathbf{x}) = \mathbb{E}_{\mathbf{U}}\left[\sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} \left(TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) \prod_{i=1}^{n} P_i(\mathbf{x}, \omega_{A_i}) \right) \right]. \tag{13}$$

The exact evaluation of $f(\mathbf{x})$ has two obstacles. First, calculating $\varphi(\mathbf{U}, \mathbf{x})$ would require enumeration of all no-show scenarios, the number of which grows exponentially with the number of patients (2^n). Second, the stochasticity of service time implies that it is necessary to calculate a multidimensional integral.

Another distinctive feature of the objective function is that it may be *nonconvex*, as a result of the time-dependent nature of the show-up probabilities. To see that, consider a very simplified system with only one patient and no overtime cost, and suppose that the service time is uniformly distributed on $(0, T)$. If the patient's show-up probability is constant over time, the solution is trivial—simply schedule the patient at the beginning of the time interval, as such a solution minimizes the amount of incurred idle time. However, the situation changes when the patient's show-up probability is a function $p(\cdot)$ of time. With a bit of calculation, we see that in this case the expected total cost function $f(x)$ given in (13) can be expressed as

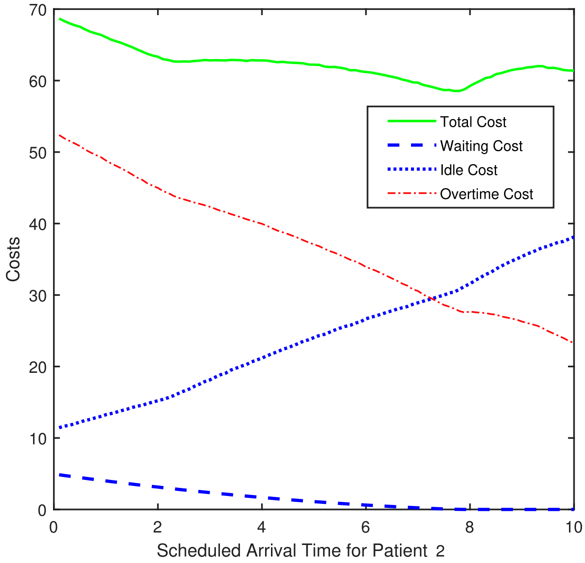
$$f(x) = c_I\left(T - p(x)\,\frac{T^2 - x^2}{2T}\right), \qquad (14)$$

where, as before, c_I is the idle time cost. When $p(\cdot)$ is constant, we see that $f(\cdot)$ is convex, and the value of x that minimizes $f(x)$ is, indeed, $x = 0$. However, it is easy to see that when $p(\cdot)$ is not constant, $f(x)$ in (14) can be nonconvex and can have more than one local minimum over $[0, T]$. This occurs, for example, when $p(\cdot)$ is the quadratic function $p(x) = (x - T/2)^2/(T/2)^2$, which models the situation where the patient's show-up probability is the highest at the beginning and at the end of the time interval but is the lowest in the middle of the day.

A lack of convexity may occur even when the show-up probability function is linear. We use a simple numerical example with two patients as an illustration. In this example, the session length is 10, service time follows a uniform distribution between 7.5 and 8.5, and the show-up probability decreases linearly from 0.8 to 0.2. The cost parameters are $c_w = 1$, $c_I = 10$, and $c_o = 15$. The first patient arrives at time 0 (i.e., $x_0 = 0$). Figure 1 displays the estimated objective function value as a function of x_1, which, in this case, corresponds to the scheduled arrival time for the second patient.

Despite its simplicity (and the choice of artificial parameters), the example in Figure 1 already illustrates the nonconvex feature of the problem, and it helps to provide some intuition on the nonconvexity. First, it easy to see that the system is congested, because both patients have an average service time of 8 units (with a small uniform variation around that value) and the session length is 10 units of time. So, inevitably, there will be considerable overtime cost. Naturally, the first patient should be scheduled to arrive at time 0. If second patient is scheduled to arrive too early, there will be waiting costs in addition to overtime costs, so this is clearly a bad solution. A good option appears to be to schedule the second patient to arrive at around the expected service completion time of patient 1 (i.e., at time $t = 8$). Scheduling the second patient to arrive after that point seems counterproductive, as it incurs unnecessary extra idle and overtime costs. However, *because the show-up probability of patient 2 decreases over time*, if the second patient is scheduled to arrive sufficiently late, the total expected costs decrease again, as there is a higher chance that the patient will not show up, in which case there is no overtime. That explains the "bump" in the function at the end of the session, thereby rendering $x_1 = 10$ a local optimal solution. We see then the nonconvexity is inherently related to the time-dependent nature of the show-up probabilities.

Figure 1. (Color online) Estimated Total, Waiting, Idle, And Overtime Costs for the Case of Two Patients, with Uniform Service Time ($U \sim U(7.5, 8.5)$) and Linearly Decreasing (from 0.8 to 0.2) Show-up Function and Cost Parameters $c_w = 1$, $c_I = 10$, and $c_o = 15$



In Sections 4 and 5 we present a sampling-based gradient algorithm that overcomes the obstacles with the exact evaluation of $f(\mathbf{x})$. To deal with the nonconvexity and avoid local optima, we adopt multiple starting points and select the solution that generates the lowest objective function value.

The following assumption will be made for the remainder of the paper.

Assumption 1 . *The no-show behavior of patients is not affected by the behavior of other patients; that is, the Bernoulli components of the vector* $\mathbf{A}(\mathbf{x})$ *are mutually independent. Moreover, the service times* $\mathbf{U}$ *are mutually independent and have continuous distributions.*

4. Gradient Estimation

Our first task before we describe our gradient-based optimization algorithm is to discuss how to estimate the gradient of the objective function in (7). Even though the objective function can be estimated through simulation of the service time and attendance status of each patient, obtaining an estimator of the gradient is not a trivial task. Fu (2015) summarizes some of the frequently used methodologies for gradient estimation of expected-value functions, such as finite differences, infinitesimal perturbation analysis, likelihood ratio, and measure-valued differentiation. Each of these techniques has its own limitations: finite differences typically yields noisy estimators that are sensitive to the choice of the perturbation parameter(s); likelihood ratio relies strongly on the selection of an appropriate distribution independent of $\mathbf{x}$ to sample from, but such a distribution may not be good uniformly for all $\mathbf{x}$; and infinitesimal perturbation analysis differentiates the function on each sample path but, in principle, cannot be applied when the distribution of the random variables depends on $\mathbf{x}$. Measure-valued differentiation can, in principle, be applied when only the distribution depends on the parameter(s) of interest.

In what follows, we present an unbiased estimator of the gradient of $f(\mathbf{x})$, obtained by exploiting the structure of the objective function. Our approach can be viewed as a form of a conditional Monte Carlo technique (Fu and Hu 1997). To proceed, consider the representation of $f(\mathbf{x})$ given in (13). Then, by defining $P(\mathbf{x}, \omega_{\mathbf{A}}) := \prod_{i=1}^{n} P_i(\mathbf{x}, \omega_{A_i})$ and observing that the term inside the expectation in (13) is a finite sum of a term that depends on $\mathbf{U}$ times a term that does not depend on $\mathbf{U}$, we can write

$$f(\mathbf{x}) = \sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} P(\mathbf{x}, \omega_{\mathbf{A}}) \mathbb{E}_{\mathbf{U}}[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})], \tag{15}$$

and thus, if both $P(\cdot, \omega_{\mathbf{A}})$ and $\mathbb{E}_{\mathbf{U}}[TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})]$ are differentiable, then $f(\cdot)$ is differentiable as well. Theorem 1 shows some properties of the function $TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})$.

Theorem 1. *Let* $\mathbf{x}$ *be an arbitrary point belonging to the interior of the simplex* $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0, \sum_{j=0}^{n-1} x_j \leq T\}$, *and let* $\omega_{\mathbf{A}} \in \{0,1\}^n$ *be an arbitrary attendance scenario. Suppose also that* Assumption 1 *holds. Then, the following conclusions hold:*

(i) *The function* $TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})$ *is differentiable at* $\mathbf{x}$ *with probability 1 (w.p.1) (i.e., for almost every realization of U), and its gradient can be calculated.*

(ii) *There exists a constant* $M > 0$ *such that*

$$|TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) - TC(\mathbf{y}, \omega_{\mathbf{A}}, \mathbf{U})| \leq M \|\mathbf{x} - \mathbf{y}\| \; w.p.1. \tag{16}$$

(iii) *The function* $\mathbb{E}_{\mathbf{U}}[TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})]$ *is differentiable at* $\mathbf{x}$, *and* $\nabla \mathbb{E}_{\mathbf{U}}[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})] = \mathbb{E}_{\mathbf{U}}[\nabla TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})]$.

Proof. We first show statement (i). For each $i = 1, \ldots, n$, let B_i denote the time at which patient i begins being served. Also, let $r_i := x_0 + \cdots + x_{i-1}$ denote the arrival time of the ith patient, and let $\mathcal{R}$ denote the region of feasible values of $r_1, \ldots, r_n$. Let $\omega_{\mathbf{A}} \in \{0,1\}^n$ be an arbitrary attendance scenario. Then, we have that

$$B_1 = r_1, \tag{17}$$

$$B_i = \max(r_i, B_{i-1} + \omega_{A_{i-1}} U_{i-1}), \quad i = 2, \ldots, n. \tag{18}$$

Note that B_i is a function of $r_1, \ldots, r_i$ only. Moreover, it is easy to see that the beginning-of-service time B_i relates to the waiting times W_i and idle times S_i, as follows:

$$W_i = B_i - r_i, \tag{19}$$

$$S_i = B_{i+1} - B_i - \omega_{A_i} U_i, \tag{20}$$

where B_{n+1} is defined as in (18), and $r_{n+1} \equiv T$.

In view of (18), it follows that $W_i \geq 0$ and $S_i \geq 0$. Also, both W_i and S_i are functions of $r_1, \ldots, r_i$ only. Moreover, from (6), we see that the overtime L can be written as the waiting time of the $(n+1)$st "patient," and thus we have

$$L = W_{n+1} = B_{n+1} - r_{n+1}. \tag{21}$$

Our first claim is that the function B_i can be written as the maximum of i affine functions. More specifically,

$$B_i(r_1, \ldots, r_i) = \max\Bigg(r_i, r_{i-1} + \omega_{A_{i-1}} U_{i-1}, r_{i-2} + \omega_{A_{i-1}} U_{i-1} + \omega_{A_{i-2}} U_{i-2}, \ldots, r_1 + \sum_{j=1}^{i-1} \omega_{A_j} U_j\Bigg),$$
$$i = 2, \ldots, n+1. \tag{22}$$

In what follows, we show how Equation (22) is derived. For $i = 1$, the statement is trivially true, because the function $B_1(r_1)$ is just the identity function. Suppose that the statement is true for $1, \ldots, i-1$. Then, from (18) and the induction hypothesis, we have

$$\begin{aligned}
B_i &= \max(r_i, B_{i-1} + \omega_{A_{i-1}} U_{i-1}) \\
&= \max\Bigg(r_i, \max\Bigg(r_{i-1}, r_{i-2} + \omega_{A_{i-2}} U_{i-2}, \ldots, r_1 + \sum_{j=1}^{i-2} \omega_{A_j} U_j\Bigg) + \omega_{A_{i-1}} U_{i-1}\Bigg) \\
&= \max\Bigg(r_i, \max\Bigg(r_{i-1} + \omega_{A_{i-1}} U_{i-1}, r_{i-2} + \omega_{A_{i-2}} U_{i-2} + \omega_{A_{i-1}} U_{i-1}, \ldots, r_1 + \sum_{j=1}^{i-1} \omega_{A_j} U_j\Bigg)\Bigg) \\
&= \max\Bigg(r_i, r_{i-1} + \omega_{A_{i-1}} U_{i-1}, r_{i-2} + \omega_{A_{i-2}} U_{i-2} + \omega_{A_{i-1}} U_{i-1}, \ldots, r_1 + \sum_{j=1}^{i-1} \omega_{A_j} U_j\Bigg),
\end{aligned}$$

and so we see that (22) holds.

Let $\mathbf{x}$ be an arbitrary point belonging to the *interior* of the simplex $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0, \sum_{j=0}^{n-1} x_j \leq T\}$. We show now that the function $TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})$ is differentiable at $\mathbf{x}$ with probability 1 (i.e. for almost every realization of U). Recall that $\mathbf{r}$ denotes the corresponding vector of arrival times. Note that the assumption that $\mathbf{x}$ belongs to the interior of $\mathcal{X}$ implies that $\mathbf{r}$ satisfies

$$0 < r_1 < r_2 < \cdots < r_n < r_{n+1} \equiv T.$$

From Equations (19) and (20), together with the relation $r_i = x_0 + \cdots + x_{i-1}$, we see that it suffices to show that B_i is differentiable at $\mathbf{r}$. From (22), we see that B_i is differentiable except at the "kink" points where at least two of the affine functions on the right-hand side of (22) are equal (note that the kink points depend on the scenario of U). Outside those kink points, we have, for any $i = 1, \dots, n+1$ and $j = 1, \dots, n$,

$$\frac{\partial B_i}{\partial r_j} = \begin{cases} 1 & \text{if } j = k^i, \\ 0 & \text{otherwise,} \end{cases} \tag{23}$$

where k_i is the index of the function attaining the maximum in (22). Furthermore,

$$\begin{aligned} &P(B_i \text{ is not differentiable at } \mathbf{r}) \\ &\leq P\left(r_{k_s^i} + \sum_{j=k_s^i}^{i-1} \omega_{A_j} U_j = r_{k_t^i} + \sum_{j=k_t^i}^{i-1} \omega_{A_j} U_j \text{ for some} \right. \\ &\qquad \left. k_s^i, k_t^i \in \{1, \dots, i\},\ k_s^i < k_t^i \right) \\ &\leq \sum_{k_s^i, k_t^i \in \{1,\dots,i\} : k_s^i < k_t^i} P\left(r_{k_s^i} + \sum_{j=k_s^i}^{i-1} \omega_{A_j} U_j = r_{k_t^i} + \sum_{j=k_t^i}^{i-1} \omega_{A_j} U_j \right) \\ &= \sum_{k_s^i, k_t^i \in \{1,\dots,i\} : k_s^i < k_t^i} P\left(0 < r_{k_t^i} - r_{k_s^i} = \sum_{j=k_s^i}^{k_t^i - 1} \omega_{A_j} U_j \right). \end{aligned} \tag{24}$$

Consider the term inside the sum in (24). If all the $\omega_{A_j}, j = k_s^i, \dots, k_t^i - 1$ are 0, then it is clear that the probability on the right-hand side of that equation is 0. Suppose now that the set $J := \{j \in \{k_s^i, \dots, k_t^i - 1\};, \omega_{A_j} = 1\}$ is nonempty. Then, the term $\sum_{j \in J} U_j$ is a sum of continuous *independent* random variables, and therefore it has continuous distribution. Although the value of $r_{k_t^i} - r_{k_s^i}$ may vary according to the scenario of U, there are only finitely many possible choices for both $r_{k_t^i}$ and $r_{k_s^i}$ (namely, $r_1, \dots, r_n$, defined in terms of the chosen point $\mathbf{x}$). It follows that

$$P(B_i \text{ is not differentiable at } \mathbf{r}) \leq P\left(\sum_{j \in J} U_j = r_{k_t^i} - r_{k_s^i} > 0 \right) = 0,$$

and therefore B_i is differentiable with probability 1. It follows from (19)–(21) that W_i, S_i, and L are also differentiable with probability 1 with respect to $\mathbf{r}$—and hence also with respect to $\mathbf{x}$. By definition of the total cost function TC in (1), we conclude that, for each attendance scenario $\omega_{\mathbf{A}} \in \{0,1\}^n$, the function $TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})$ is differentiable at $\mathbf{x}$ with probability 1 (i.e. for almost every realization of U).

We now prove statement (ii); that is, we show that $TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})$ is a Lipschitz function w.p.1. From the relations (19)–(21), we have for $i, j = 1, \dots, n$ that

$$\begin{aligned} \frac{\partial W_i}{\partial r_j} &= \frac{\partial B_i}{\partial r_j} - \mathbb{I}_{\{j=i\}}, \\ \frac{\partial S_i}{\partial r_j} &= \frac{\partial B_{i+1}}{\partial r_j} - \frac{\partial B_i}{\partial r_j}, \\ \frac{\partial L}{\partial r_j} &= \frac{\partial B_{n+1}}{\partial r_j}, \end{aligned}$$

which can be calculated using (23). Then, from (1), we write

$$\frac{\partial TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})}{\partial r_j} = c_o \frac{\partial L}{\partial r_j} + c_I \mathbb{I}_{\{j=1\}} + \sum_{i=1}^{n} \left[c_w \omega_{A_i} \frac{\partial W_i}{\partial r_j} + c_I \frac{\partial S_i}{\partial r_j} \right], \tag{25}$$

and so it is easy to see that the derivatives of TC with respect to $\mathbf{r}$ are bounded by a constant. By using the chain rule, we can write

$$\begin{aligned} \frac{\partial TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})}{\partial x_j} &= \sum_{i=1}^{n} \frac{\partial TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})}{\partial r_i} \frac{\partial r_i}{\partial x_j} \\ &= \sum_{i=1}^{n} \frac{\partial TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})}{\partial r_i} \mathbb{I}_{\{j<i\}}, \end{aligned} \tag{26}$$

and so we see that the derivatives of TC with respect to $\mathbf{x}$ are also bounded; thus $TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})$ is Lipschitz.

Finally, we prove statement (iii) of the theorem. The results in statements (i) and (ii) ensure that the general conditions for differentiability of expected value functions established in Shapiro et al. (2009, theorem 7.49) are fulfilled; that result, in turn, ensures that (a) $\mathbb{E}_{\mathbf{U}}[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})]$ is differentiable, and (b) $\nabla \mathbb{E}_{\mathbf{U}}[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})] = \mathbb{E}_{\mathbf{U}}[\nabla TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})]$. □

Remark 1. The condition that $\mathbf{x}$ belongs to the interior of the set cannot be relaxed. To see that, consider the term B_2. We have

$$B_2 = \max(r_2, B_1 + A_1 U_1),$$

so B_2 is nondifferentiable if and only if $r_2 = B_1 + A_1 U_1 = r_1 + A_1 U_1$ (i.e., if $A_1 U_1 = r_2 - r_1$). If $r_1 = r_2$, then such an event occurs if $A_1 = 0$, an event of positive probability.

Consider the term B_3. We have

$$B_3 = \max(r_3, B_2 + A_2 U_2),$$

so B_3 is nondifferentiable if and only if $r_3 = B_2 + A_2 U_2$. If $r_3 = r_2$, then a sufficient condition for this to occur is that

$$A_2 = 0 \text{ and } r_2 \geq r_1 + A_1 U_1.$$

The probability of the latter event is

$$P(A_2 = 0) \times [P(U_1 \le r_2 - r_1)P(A_1 = 1) + P(A_1 = 0)] \\ \ge P(A_2 = 0) \times P(A_1 = 0) > 0.$$

We can generalize this argument to show that, given r such that $r_{i+1} = r_i$, the probability that B_{i+1} is not differentiable at r is at least $P(A_1 = 0) \times \cdots \times P(A_i = 0) > 0$. Note that the condition $r_{i+1} = r_i$ is equivalent to $x_i = 0$.

Remark 2. When there are no no-shows—that is, all show-up probabilities are equal to 1—the resulting model is very similar to that studied by Homem-de-Mello et al. (1999), who discuss methods to find the optimal release times of jobs in a manufacturing plant. In that paper it is shown that the objective function is differentiable whenever the service times have continuous distributions, and a gradient-based algorithm is presented to solve the problem. However, the presence of no-shows in our problem, complicated further by the fact that the show-up probabilities depend on the scheduled time of arrivals, makes the present problem significantly harder. Indeed, as mentioned earlier, the problem is nonconvex, unlike the problem in Homem-de-Mello et al. (1999); moreover, the techniques used in that paper to show differentiability do not apply to our case because a crucial assumption in Homem-de-Mello et al. (1999) is that the service times have continuous distributions. In our setting, however, although the service times also have continuous distributions, the corresponding service time is 0 if a patient does not show up. Consequently, when combined with the show-up variables, the service times may have an atom at 0, and thus the aforementioned assumption in Homem-de-Mello et al. (1999) does not hold, which, in turn, has required us to develop a proof of Theorem 1 from first principles.

The results in Theorem 1 allow us make statements about the derivatives of the function $f(\mathbf{x})$ defined in (15). These are summarized in Proposition 2.

Proposition 2. *If the show-up function $p(t)$ is differentiable and* Assumption 1 *holds, then the function $f(\mathbf{x})$ in* (15) *is differentiable on the interior of the simplex $\mathcal{X}$. If, in addition, $p(t)$ is continuously differentiable, then for any $\mathbf{x}$ belonging to the interior of the simplex $\mathcal{X}$, we have that*

$$\nabla f(\mathbf{x}) = \mathbb{E}_{\mathbf{U}}[\nabla \varphi(\mathbf{U}, \mathbf{x})], \tag{27}$$

where $\varphi(\mathbf{U}, \mathbf{x})$ is as defined in (12).

Proof. Under Assumption 1, part (iii) of Theorem 1 ensures differentiability of $\mathbb{E}_{\mathbf{U}}[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})]$ with respect to $\mathbf{x}$ on the interior of the simplex $\mathcal{X}$. Thus, it immediately follows from (11) and (15) that if the show-up function $p(t)$ is differentiable, then $f(\mathbf{x})$ is differentiable at any $\mathbf{x}$ belonging to the interior of the simplex $\mathcal{X}$.

We now show the second statement of the proposition. Note initially that differentiability of $\varphi(\mathbf{U}, \mathbf{x})$ w.p.1 follows from Theorem 1 and the assumption that $p(t)$ is differentiable. We can write the gradient of $f(\mathbf{x})$ in (15) as follows:

$$\begin{aligned} \nabla f(\mathbf{x}) &= \nabla \left(\sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} P(\mathbf{x}, \omega_{\mathbf{A}}) \, \mathbb{E}_{\mathbf{U}}[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})] \right) \\ &= \nabla \mathbb{E}_{\mathbf{U}} \left[\sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} P(\mathbf{x}, \omega_{\mathbf{A}}) \, TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) \right] \\ &= \nabla \mathbb{E}_{\mathbf{U}}[\varphi(\mathbf{U}, \mathbf{x})], \end{aligned} \tag{28}$$

where the interchange of expectation and summation in the second equality is valid because the sum is over a finite number of terms.

Next, we show that we can interchange the gradient and expectation operators in (28). To do so, first observe that the derivatives of $\varphi(\mathbf{U}, \mathbf{x})$ are bounded by an integrable random variable. Indeed, by calculating the derivatives in (12) via the product rule, we see that $\nabla \varphi(\mathbf{U}, \mathbf{x})$ is given by a sum of terms involving $TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})$, $\nabla TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})$, $P(\mathbf{x}, \omega_{\mathbf{A}})$, and $\nabla P(\mathbf{x}, \omega_{\mathbf{A}})$. The latter three terms are bounded by constants—the derivatives of $TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})$ are bounded by virtue of Theorem 1, and the assumption of continuous differentiability of $p(t)$ ensures that its derivative is bounded on $[0, T]$; because $p(\cdot) \in [0, 1]$, it follows that the derivatives of $P(\mathbf{x}, \omega_{\mathbf{A}})$ are bounded on $[0, T]$. Thus, we see that the derivatives of $\varphi(\mathbf{U}, \mathbf{x})$ are bounded by a constant times $TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})$, and the latter is integrable, as argued earlier (see the discussion following (2)).

It follows that the derivatives of $\varphi(\mathbf{U}, \mathbf{x})$ are bounded by an integrable random variable (i.e., $\varphi(\mathbf{U}, \cdot)$ is Lipschitz w.p.1 with an integrable Lipschitz constant). By applying again the result in (Shapiro et al. 2009, theorem 7.49), we conclude that

$$\nabla \mathbb{E}_{\mathbf{U}}[\varphi(\mathbf{U}, \mathbf{x})] = \mathbb{E}_{\mathbf{U}}[\nabla \varphi(\mathbf{U}, \mathbf{x})],$$

thus completing the proof. □

Proposition 2 shows that, in principle, having an expression for $\nabla \varphi(\mathbf{U}, \mathbf{x})$ would allow us to estimate $\nabla f(\mathbf{x})$ by sampling from the distribution of service times and calculating the sample average estimator of the right-hand side of (27). However, obtaining the derivatives of $\varphi(\mathbf{U}, \mathbf{x})$ is not trivial because it explicitly considers all attendance scenarios. Indeed, by using (12) we can compute the derivative of $\varphi(\mathbf{U}, \mathbf{x})$ with respect to x_k, $k = 0, \ldots, n-1$, as

$$\begin{aligned} &\frac{\partial \varphi(\mathbf{U}, \mathbf{x})}{\partial x_k} \\ &= \sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} \left[\frac{\partial TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})}{\partial x_k} \cdot \left(\prod_{i=1}^n P_i(\mathbf{x}, \omega_{A_i}) \right) + TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) \right. \\ &\qquad \left. \cdot \sum_{j=1}^n \left(\frac{\partial P_j(\mathbf{x}, \omega_{A_j})}{\partial x_k} \cdot \left[\prod_{i:\, i \neq j}^n P_i(\mathbf{x}, \omega_{A_i}) \right] \right) \right], \end{aligned} \tag{29}$$

so we see that the expression involves the explicit enumeration of all attendance scenarios. However, by exploiting the structure of this expression, we can derive an alternative approach. For a fixed attendance scenario $\omega_{\mathbf{A}}$ and a realization of service times $\mathbf{U}$, define the function

$$\psi_k(\mathbf{x},\omega_{\mathbf{A}},\mathbf{U}) := \frac{\partial TC(\mathbf{x},\omega_{\mathbf{A}},\mathbf{U})}{\partial x_k} + TC(\mathbf{x},\omega_{\mathbf{A}},\mathbf{U}) \cdot \left(\sum_{j=1}^{n} \frac{\partial P_j(\mathbf{x},\omega_{A_j})}{\partial x_k} \cdot \frac{1}{P_j(\mathbf{x},\omega_{A_j})}\right). \tag{30}$$

It is easy to see that by multiplying the term on the right-hand side in (30) by $\prod_{i=1}^{n} P_i(\mathbf{x},\omega_{A_i})$, we obtain the expression inside the sum over $\omega_{\mathbf{A}}$ in (29). Then, we can write the derivative of φ with respect to x_k in (29) as

$$\begin{aligned}\frac{\partial\varphi(\mathbf{U},\mathbf{x})}{\partial x_k} &= \sum_{\omega_{\mathbf{A}}\in\{0,1\}^n} \psi_k(\mathbf{x},\omega_{\mathbf{A}},\mathbf{U}) \cdot \prod_{i=1}^{n} P_i(\mathbf{x},\omega_{A_i}) \\ &= \mathbb{E}_{\mathbf{A}(\mathbf{x})}[\psi_k(\mathbf{x},\mathbf{A}(\mathbf{x}),\mathbf{U})\,|\,\mathbf{U}],\end{aligned} \tag{31}$$

where the second equality follows from (10).

The reformulation of the derivative in (31) is the fundamental step that allows us to obtain an unbiased sampling estimator of the gradient of the objective function f. Recall from (27) that $\nabla f(\mathbf{x}) = \mathbb{E}_{\mathbf{U}}[\nabla\varphi(\mathbf{U},\mathbf{x})]$. Then, by using (31), we can write

$$\begin{aligned}\frac{\partial f(\mathbf{x})}{\partial x_k} &= \mathbb{E}_{\mathbf{U}}\left[\frac{\partial\varphi(\mathbf{U},\mathbf{x})}{\partial x_k}\right] = \mathbb{E}_{\mathbf{U}}[\mathbb{E}_{\mathbf{A}(\mathbf{x})}[\psi_k(\mathbf{x},\mathbf{A}(\mathbf{x}),\mathbf{U})\,|\,\mathbf{U}]] \\ &= \mathbb{E}_{\mathbf{A}(\mathbf{x}),\mathbf{U}}[\psi_k(\mathbf{x},\mathbf{A}(\mathbf{x}),\mathbf{U})],\end{aligned} \tag{32}$$

where the last equality follows from applying again the tower property of expectations.[1] The gradient of the objective function can now be estimated though simulating the vectors of random service times and no-shows. This is summarized in the following theorem.

Theorem 3. *Let* $\{(\mathbf{A}^s(\mathbf{x}),\mathbf{U}^s)\}$, $s = 1,\dots,S$ *be a sample of size* S *of the bivariate random vector* $(\mathbf{A}(\mathbf{x}),\mathbf{U})$ *(note that the samples of* $\mathbf{A}(\mathbf{x})$ *depend on the particular point* $\mathbf{x}$*). Then, for* $k = 0,\dots,n-1$, *a consistent and unbiased estimator* $G_k(\mathbf{x})$ *of* $\frac{\partial f(\mathbf{x})}{\partial x_k}$ *can be obtained as*

$$G_k(\mathbf{x}) := \frac{1}{S}\sum_{s=1}^{S} \psi_k(\mathbf{x},\mathbf{A}^s(\mathbf{x}),\mathbf{U}^s), \tag{33}$$

where ψ_k *is defined in* (30).

The result in Theorem 3, which follows directly from (32), is key for estimating the gradient of f. As discussed at the beginning of this section, the issue of estimating gradients in stochastic systems where *both* the performance function and the underlying distributions depend on the parameters of interest is very challenging. By exploiting the structure of the problem, Theorem 3 shows that we can bypass these difficulties and estimate that gradient by the sample average of appropriately defined functions. The strong law of large numbers ensures that $G_k(\mathbf{x})$ defined in (33) converges with probability 1 to the expression in (32). Moreover, when the service times U have finite variance, by defining $\mathbf{G}(\mathbf{x})$ as the vector $(G_0(\mathbf{x}),\dots,G_{n-1}(\mathbf{x}))$ we can apply a multivariate version of the central limit theorem to conclude that

$$\sqrt{S}(\mathbf{G}(\mathbf{x}) - \nabla f(\mathbf{x})) \xrightarrow{d} \text{Normal}(0,\Sigma(\mathbf{x})), \tag{34}$$

where "$\xrightarrow{d}$" denotes convergence in distribution, and $\Sigma(\mathbf{x})$ is the covariance matrix of $[\psi_k(\mathbf{x},\mathbf{A}(\mathbf{x}),\mathbf{U})]_{k=0,\dots,n-1}$. In particular, (34) shows that the gradient estimator $G(\mathbf{x})$ defined via (33) converges to $\nabla f(\mathbf{x})$ at rate $S^{-1/2}$.

This results allow us to use a sampling-based optimization approach to solve the original problem, as discussed in the next section.

5. Algorithms

The analysis of the problem has unveiled some of its features: a nonlinear nonconvex stochastic objective function and a compact and convex feasible set. Using the techniques described in Section 4, a sampling-based estimator of the gradient is obtained in order to apply a gradient-based method for nonlinear stochastic optimization. In Section 5.1 we discuss one such method available in the literature. However, as our problem is constrained, we use a modified projected gradient algorithm described in the literature, as discussed in Section 5.2. Finally, in Section 5.3 we describe how we combine these methods to solve the proposed ASP.

5.1. Description of the STRONG Method

The algorithm we propose to solve this problem is built on the STRONG algorithm proposed by Chang et al. (2013). STRONG, or the stochastic trust-region response-surface method, is an algorithm designed for *unconstrained* stochastic optimization that is proved to converge to critical points. The strategy of the algorithm is to explore the behavior of the objective function within small subregions, called *trust regions*, to find new candidate solutions. As the objective function is estimated via sampling, the improvement of a candidate solution is evaluated under statistical tests.

STRONG consists of an outer loop in which either Stage I or Stage II is selected in each iteration, depending on whether the current trust region is considered large or small. When the algorithm is in Stage I (large trust region), a linear approximation of the objective function is built with the gradient estimator.

Then, a subproblem is solved with this approximation in which the feasible region is delimited by the trust region, and simulation runs are required to estimate the expected total cost of the candidate solution. The improvement is assessed with two tests: a *ratio-comparison* (*RC*) *test* and a *sufficient-reduction test*. The RC test intends to test whether the trust region is trustworthy, whereas the second test intends to tell whether there is sufficient reduction. After the evaluation of the candidate solution, a new trust region size is defined. Stage I is similar to Stage I, but it differs from it in two ways: first, a quadratic approximation is used instead of a linear one; second, if no satisfactory solution is found in the subproblem, the algorithm goes into an Inner Loop, where the sample size is increased to ensure that a better solution is found or the stopping criterion is satisfied.

A brief outline of the STRONG algorithm, as described in Chang et al. (2013), is as follows.:

Step 0. Set the iteration counter $k = 0$. Select an initial solution x_0 and an initial sample size n_0 for the current solution to estimate the objective function. Select a trust region size Δ^0, a switch threshold size $\tilde{\Delta}$, and the other algorithm parameters.

Step 1. Let $k = k + 1$. Denote Δ_k as the size of the new trust region. If $\Delta_k > \tilde{\Delta}$, go to Stage I. Otherwise, go to Stage II.

Step 2. If the termination criterion is satisfied, stop and return the solution. Otherwise, go to Step 1.

The complete algorithm, convergence properties, and further details of STRONG are described in Chang et al. (2013).

5.2. Projected Gradient Path Algorithm

As was previously discussed, STRONG is an algorithm designed for an unconstrained setting. However, the appointment scheduling problem is constrained.

Generally speaking, the nature of the feasible set strongly determines the methods that can be used to find solutions to optimization problems. As Conn et al. (2000) point out, the geometry of the feasible region is relevant to the ease of projecting unfeasible solutions onto the feasible set. In Conn et al. (2000, chap. 12), an algorithm under the trust-region methodology for deterministic problems with convex constraints is presented. The purpose of that algorithm is to find, according to a given approximation of the objective function, a step from the current solution that gives a feasible solution and leads to a decrease in the objective function. The technique is based on a line search along the projected-gradient path to compute a generalized Cauchy point, so we shall refer to that algorithm as the *projected gradient path* (*PGP*) *algorithm*.

We first give a brief description of the PGP algorithm; further details can be found in Conn et al. (2000). The projected gradient path for any $\mathbf{x}$ onto the feasible set $\mathcal{X}$ is given by the following expression, for all $t \geq 0$:

$$\mathbf{q}(t, \mathbf{x}) := \Pi_{\mathcal{X}}(\mathbf{x} - t\nabla f(\mathbf{x})), \tag{35}$$

where $\Pi_{\mathcal{X}}(\mathbf{x})$ is the Euclidean projection of $\mathbf{x}$ onto $\mathcal{X}$. At each iteration ν, given the current feasible solution $\mathbf{x}^\nu$, let $m_\nu(\mathbf{x}^\nu)$ be an approximation of the objective function f at $\mathbf{x}^\nu$ (in our case, it is an affine approximation), let $\mathbf{g}_\nu$ be an estimator of the gradient $\nabla f(\mathbf{x}^\nu)$, and let Δ_ν be the trust region size. The algorithm can be outlined as follows.

Step 0. Define the required parameters. Set $t_{\min} = 0, \ldots\ t_{\max} = \infty$, $t_0 = \frac{\Delta_\nu}{\|\mathbf{g}_\nu\|}$, and $j = 0$.

Step 1. Compute the candidate point $\mathbf{q}(t_j, \mathbf{x}^\nu) = \Pi_{\mathcal{X}}(\mathbf{x}^\nu - t_j \mathbf{g}_\nu)$ and evaluate $m_\nu(\mathbf{q}(t_j, \mathbf{x}^\nu))$.

Step 2. Set $j := j + 1$, update the value of t_j, $t_{\min} = 0$ and/or $t_{max} = \infty$, and go back to Step 1 until (i) the candidate point lies within the trust region; (ii) there is sufficient decrease in the approximation m_ν (from the current point to the candidate point); and (iii) at least one of the following three conditions holds: (a) the candidate point is not too close to the current point, (b) the decrease in the approximation m_ν is not too large, or (c) the norm of the projection of $\mathbf{g}_\nu$ onto the tangent cone at $\mathbf{q}(t_j, \mathbf{x}^\nu)$ with respect to the feasible set is sufficiently small.

We actually implemented a variation of the preceding algorithm that does not require the calculation of the projection onto the tangent cone. To describe that change, we need to define some notation. Let $\mathbf{s}_\nu(t_j)$ denote the vector $\mathbf{q}(t_j, \mathbf{x}^\nu) - \mathbf{x}^\nu$, and let $\Pi_{\mathcal{T}(\mathbf{x})}$ denote the projection operator onto the tangent cone at a point $\mathbf{x}$ with respect to $\mathcal{X}$. In Conn et al. (2000), the condition representing criterion (c) in Step 2 is expressed as

$$\|\Pi_{\mathcal{T}(\mathbf{q}(t_j, \mathbf{x}^\nu))}[-\mathbf{g}_\nu]\| \leq \kappa_{\text{epp}} \frac{|\langle \mathbf{g}_\nu, \mathbf{s}_\nu(t_j)\rangle|}{\Delta_\nu}, \tag{36}$$

where $\kappa_{epp} \in (0, \frac{1}{2})$ is a constant. Given the current solution $\mathbf{x}^\nu \in \mathcal{X}$ and $\theta \geq 0$, let $\mathcal{M}(\mathbf{x}^\nu, \theta)$ denote the *criticality measure* defined in Conn et al. (2000) as

$$\mathcal{M}(\mathbf{x}^\nu, \theta) := |\min\{\langle \mathbf{g}_\nu, \mathbf{d}\rangle : \mathbf{x}^\nu + \mathbf{d} \in \mathcal{X}, \|\mathbf{d}\| \leq \theta\}|. \tag{37}$$

In our implementation, we replaced Condition (36) with

$$\frac{\mathcal{M}(\mathbf{x}^\nu, \theta_j) - |\langle \mathbf{g}_\nu, \mathbf{s}_\nu(t_j)\rangle|}{2\theta_j} \leq \kappa_{epp} \frac{|\langle \mathbf{g}_\nu, \mathbf{s}_\nu(t_j)\rangle|}{\Delta_\nu}, \tag{38}$$

where $\theta_j := \|\mathbf{s}_\nu(t_j)\| + 1$. We show now that the replacement of (36) with (38) does not affect the proof of convergence of Conn et al. (2000), provided some extra condition is imposed.

Proposition 4. *Suppose that* (38) *is used in place of* (36) *in algorithm* 12.2.2 *of Conn et al.* (2000)*. Then, the arguments in the proof of theorems* 12.2.1 *and* 12.2.2 *of Conn et al.* (2000) *remain valid, provided that the constants* κ_{frd} *and* κ_{epp} *that appear in the algorithm satisfy the condition*

$$\kappa_{\text{frd}} \leq \frac{2}{4\kappa_{epp}+1}. \tag{39}$$

Proof. By theorem 12.1.5 in Conn et al. (2000) and the fact that $\theta_j > \|\mathbf{s}_\nu(t_j)\|$, we have the inequality

$$\frac{\mathcal{M}(\mathbf{x}^\nu,\theta_j) - |\langle \mathbf{g}_\nu, \mathbf{s}_\nu(t_j)\rangle|}{2\theta_j} \leq \|\Pi_{\mathcal{T}(\mathbf{q}(t_j,\mathbf{x}^\nu))}[-\mathbf{g}_\nu]\|. \tag{40}$$

As argued in the proof of theorem 12.2.1 in Conn et al. (2000), Condition (36) holds for j sufficiently large when $t_{max} = \infty$ for all j. Thus, in that case it follows from (40) that (38) holds as well, and so the arguments in aforementioned proof remain valid.

Consider now theorem 12.2.2 in Conn et al. (2000). Notice initially that we have $\|\mathbf{s}_\nu(t_j)\| \leq \Delta_\nu$, as this is assumed by that theorem. Thus, (38) implies that

$$\mathcal{M}(\mathbf{x}^\nu,\theta_j) \leq \frac{2(\Delta_\nu+1)\kappa_{epp}|\langle \mathbf{g}_\nu, \mathbf{s}_\nu(t_j)\rangle|}{\Delta_\nu} + |\langle \mathbf{g}_\nu, \mathbf{s}_\nu(t_j)\rangle|,$$

and thus we have

$$\begin{aligned} |\langle \mathbf{g}_\nu, \mathbf{s}_\nu(t_j)\rangle| &\geq \frac{\Delta_\nu}{2(\Delta_\nu+1)\kappa_{epp}+\Delta_\nu}\mathcal{M}(\mathbf{x}^\nu,\theta_j) \\ &\geq \frac{\Delta_\nu}{2(\Delta_\nu+1)\kappa_{epp}+\Delta_\nu}\mathcal{M}(\mathbf{x}^\nu,1), \end{aligned} \tag{41}$$

where the last inequality follows the fact that $\mathcal{M}(\mathbf{x}^\nu,\cdot)$ is a nondecreasing function (Conn et al. 2000, theorem 12.1.5) and $\theta_j \geq 1$. It suffices now to show that (39) and (41) together imply that

$$|\langle \mathbf{g}_\nu, \mathbf{s}_\nu(t_j)\rangle| \geq \frac{1}{2}\kappa_{frd}\min(\Delta_\nu,1)\,\mathcal{M}(\mathbf{x}^\nu,1), \tag{42}$$

where κ_{frd} is another constant used in the algorithm. Note that (42) is exactly the inequality used in the proof of theorem 12.2.2 of Conn et al. (2000), which is the basis for convergence of the trust region method. Indeed, suppose that $\Delta_\nu > 0$ and $\mathcal{M}(\mathbf{x}^\nu,1) > 0$ (otherwise, (42) already holds trivially). When (41) holds, a sufficient condition for (42) to hold is that

$$\frac{\Delta_\nu}{2(\Delta_\nu+1)\kappa_{epp}+\Delta_\nu}\mathcal{M}(\mathbf{x}^\nu,1) \geq \frac{1}{2}\kappa_{frd}\min(\Delta_\nu,1)\,\mathcal{M}(\mathbf{x}^\nu,1);$$

that is,

$$\kappa_{frd} \leq \frac{2\Delta_\nu}{2(\Delta_\nu+1)\kappa_{epp}+\Delta_\nu}\,\frac{1}{\min(\Delta_\nu,1)}. \tag{43}$$

It is easy to see that the function on the right-hand side of (43) is continuous for $\Delta_\nu > 0$, strictly decreasing for $\Delta_\nu \in (0,1]$, and strictly increasing for $\Delta_\nu \geq 1$; thus it attains its minimum at $\Delta_\nu = 1$. Therefore, (43) holds for all $\Delta_\nu > 0$, provided that (39) holds. It follows that Conditions (39) and (41) together imply (42), and so the arguments in the proof of theorem 12.2.2 of Conn et al. (2000) can be applied, and convergence is ensured. □

Remark 3. The contribution of Proposition 4 is general and goes beyond the particular problem discussed in this paper. The value of the contribution is that it completely avoids the projection onto tangent cones, which can be very difficult in some problems. As shown in Proposition 4, the only price to pay for not dealing with tangent cones is a mild restriction on the constants used in the algorithm. Note also that the specification of the constants in Conn et al. (2000) dictates that $\kappa_{frd} \in (0,1)$ and $\kappa_{epp} \in (0,\frac{1}{2})$; thus, by choosing $\kappa_{epp} \leq \frac{1}{4}$, we ensure that (39) holds. In our implementation, we used $\kappa_{epp} = 0.125$.

5.3. Combining STRONG and PGP

As discussed earlier, our algorithm to solve the appointment scheduling problem enhances STRONG by incorporating the PGP algorithm of Conn et al. (2000), which is designed to find improved and feasible solutions in a constrained deterministic context.

Recall from the description of STRONG in Section 5.1 that both in Stage I and in Stage II a subproblem is solved with an approximation of the objective function in which the feasible region is delimited by the trust region, a step that has the purpose of finding new candidate solutions that yield expected cost reduction. In the original description of STRONG in Chang et al. (2013), a linear approximation is used in Stage I, whereas a quadratic approximation is used in Stage II. Note, however, that the quadratic approximations allowed in Chang et al. (2013) are fairly general; that is, they are not necessarily based on second derivatives. In particular, the setting of Chang et al. (2013) allows for the case where the quadratic term is 0. In our algorithm, we use linear functions in Stage II by setting the quadratic term to be 0, as the second-order derivative estimators are too noisy.

Our algorithm to solve the appointment scheduling problem, which we shall denote by SPGP (short for stochastic PGP), can then be described as follows:

- Apply the STRONG method to the problem.
- To solve the subproblems defined in Stage I and in Stage II, use the PGP method described in Section 5.2, enhanced by our modification discussed in Proposition 4.

Note that because the original problem may be nonconvex, as discussed in Section 3, convergence is ensured only to stationary points, so the SPGP algorithm could end up at a local minimum. As there is no guarantee of global optimality, a multistart strategy is adopted. We discuss the details in Section 6.

6. Computational Studies

In this section, we apply our SPGP approach to explore the insights into scheduling solutions under time-dependent show-up probabilities. In Section 6.1 we describe the experimental design and computational setup. In Section 6.2 we explore scheduling patterns under different show-up functions, as opposed to the static case, and perform an out-of-sample simulation study to demonstrate the value of incorporating patient schedule-dependent no-show behavior. Finally, in Section 6.3 we compare our scheduling solutions with those obtained with the distributionally robust optimization (DRO) model developed in Kong et al. (2020) to show the value of distribution information. All algorithms and procedures are programmed in Matlab 2018a; the code is available in the journal's GitHub repository at https://github.com/INFORMSJoC/2020.0311.

6.1. Experimental Design and Computational Setup

In the basic experiment setup, we solve for the scheduling solutions for 12 patients in a clinic session that spans six time units ($T = 6$). Three types of costs occur in the system, which are the costs of the patient's waiting time, the physician's idle time, and overtime. Following Zacharias and Pinedo (2014) and Kong et al. (2020), we set the three costs to be 0.1,1, and 1.5, respectively.[2]

Both uncertainties in service time and attendance status are incorporated in the computational study. We solve for scheduling solutions under three types of service time distributions and six types of attendance functions. The three types of service distributions are exponential, log-normal, and beta. Following the setting in Hassin and Mendel (2008), both the mean and the standard deviation of the service duration are set to be 1. As mentioned before, our method has the advantage of dealing with any smooth show-up functions. As an illustration, we consider six different show-up patterns: constant, increasing, decreasing, quadratic concave, quadratic convex, and cosine. These functions represent different patterns of no-shows. Figure 2 depicts the show-up probabilities during a session of six time units. Following the experimental design in Kong et al. (2020), we set 0.1 as the lowest show-up probability and 0.9 the highest. Note that the average show-up rate is 0.5 in the increasing, decreasing, and cosine cases; 0.67 in the concave case; and 0.33 in the convex case. We remark that such a big range may not represent the reality; however, it allows us to examine the behavior of the solution in extreme cases. In Section 6.3, we will use more realistic show-up rates to compare our method with the DRO approach.

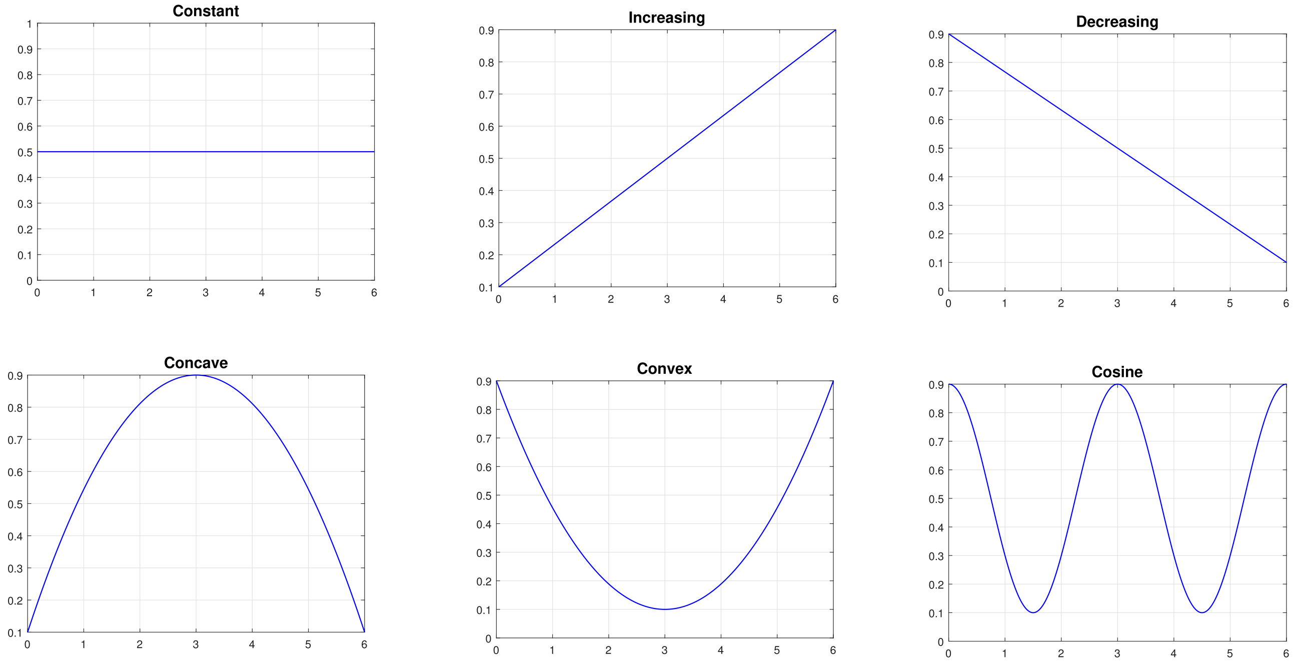
The parameters used in the algorithms are presented in Appendix A in the online supplement. Note that the value of the parameters are selected based on the recommendation in the literature. We do not vary or try to optimize them in the algorithm. We randomly select 20 different starting points, run the SPGP algorithm to generate scheduling solutions, and then generate new samples for both service duration and no-show status (sample size of 1 million) to estimate the total expected costs of all the scheduling solutions and their corresponding 95% confidence intervals. We then select the schedule that generates the least expected total cost.

6.2. Scheduling Patterns Under Different Show-up Functions

We proceed to solve for the scheduling solutions under six different show-up functions and three different service time distributions by running the SPGP algorithm described in Section 5. The schedule solutions indicate the (inter)arrival time of each patients. From those and the show-up function used, we can calculate the expected show-up rate of each patient. Figure 3 shows the arrival times of the 12 patients in a clinical session, together with the corresponding show-up rate at the time of arrival. Each point in this figure represents one patient arrival, whereas the x axis shows the arrival time, and y axis corresponds to the show-up probability. Note that the results shown are solved under exponential service distribution; similar results under two other distributions are reported in Appendix B in the online supplement. As we can see from Figure 3, a front-loading pattern is observed when there is a nondecreasing show-up function at the beginning, as in the constant, increasing, and quadratic concave cases. Interestingly, a load of patients also appear at the first lowest point of the cosine pattern. Patients are allocated to the front or to the low show-up time slots, possibly to reduce the workload and thus overtime. It is worthwhile noting that few patients are scheduled to arrive in the latter half of the session, which is probably due to the high probability of show-up and the uncertainty in service time. This strategy will result in, on average, less overtime and a balanced idle time and waiting time.

The front-loading pattern, however, disappears when a decreasing function is involved at the beginning, including patterns decreasing, quadratic convex, and cosine. Under those cases, most patients are more equally allocated across the decreasing time frame, exhibiting a *spreading-out* pattern. Moreover, when a decreasing function is involved in the second half of the session (decreasing and concave), patients are, in general, scheduled to arrive later. The observations

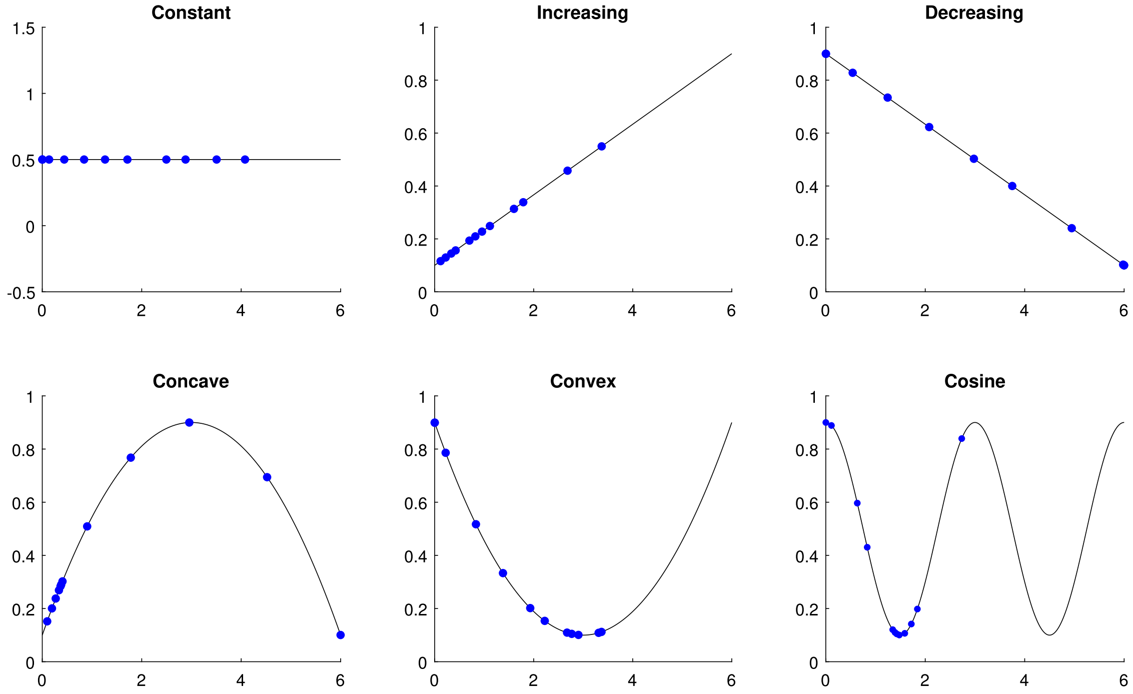
Figure 2. (Color online) Show-up Probabilities as a Function of Time (y Axis: Show-up Probability; x Axis: Time (Same Unit as the Session Length's))



could probably be explained by two reasons: first, a higher show-up rate at the beginning of the session may probably result in a patient waiting queue, the burden of which can be mitigated by spreading patients across the session; and second, assigning patients to a lower show-up time (toward the end of the session) can reduce the workload of the system.

Next, we explore the patterns of the interarrival times under different show-up functions, similar as some appointment scheduling literature, such as Denton and

Figure 3. (Color online) Scheduling Decisions Under Different Show-up Patterns and Exponential Service Distribution (y Axis: Show-up Probability; x Axis: Time)



Gupta (2003) and Kong et al. (2013). To understand Figure 4, we first assume two hypothetical patients: patient 0, who arrives at the beginning of the session, and patient 13, who arrives at the end of the session. The x axis of Figure 4 represents patient i ($i = 0, 1, \ldots, 12$), and the y axis demonstrates the interarrival times between patients i and $i+1$, $i = 0, 1, \ldots, 12$. Figure 4 shows patient interarrivals under exponential service time; results under two other distributions are reported in Appendix B in the online supplement.

The dome-shaped pattern in the interarrival times has been well reported in the previous literature on standard appointment scheduling model where no-shows are not considered. Denton and Gupta (2003); Robinson and Chen (2003), and Hassin and Mendel (2008) observe similar patterns when patient no-show is constant. The left upper chart in Figure 4 depicts the interarrival times for the constant, increasing, and decreasing cases. Whereas the interarrival times show, in general, an increasing trend in both the constant and increasing cases, a dome shape exists in the decreasing case. Moreover, the decreasing case allocates much larger interarrival times to patients arriving in the middle of the session compared with the two other cases.

Compared with the constant case, the schedule under the concave case shows a delayed dome shape (see the right upper chart in Figure 4). It assigns more patients to arrive at the beginning of the session and then spread the patients across the session. The schedule under the convex case, however, allocates fewer patients to the beginning of the session, at which time the show-up rate is high (see the left lower chart). Note that although the schedules under the constant and convex cases seem similar, their performance under the convex show-up environment may differ greatly. We will show the performance comparison later. In the cosine case, the interarrival times for the first few patients are higher than in the constant case and then go almost to 0; the times increase for the last patients (see the right lower chart). In addition, a big gap is left from the last arrival to the end of the session.

6.2.1. The Value of Incorporating Patient Schedule-Dependent No-Show Behavior.

In the preceding discussion, we examined the scheduling patterns under different show-up functions. We now investigate the value of incorporating the patient schedule-dependent no-show behavior—that is, the performance improvement under various show-up probabilities compared with constant show-up rates. The purpose of such a comparison is to quantify the improvement obtained by using a more elaborate model in which the show-up rates are allowed to depend on the scheduled time compared with a model whereby the show-up rates are incorrectly assumed to be constant.

We proceeded as follows. For each type of service distribution and each show-up pattern discussed earlier in this section, we generated best "static schedules" under constant show-up rates, and we ran a Monte Carlo simulation to estimate the total cost of those static schedules, but under the time-dependent no-show environment. The 95% confidence intervals

Figure 4. (Color online) Interarrival Time of Patients Under Different Show-up Patterns and Exponential Service Distribution (y Axis: Interarrival Time; x Axis: Patients)

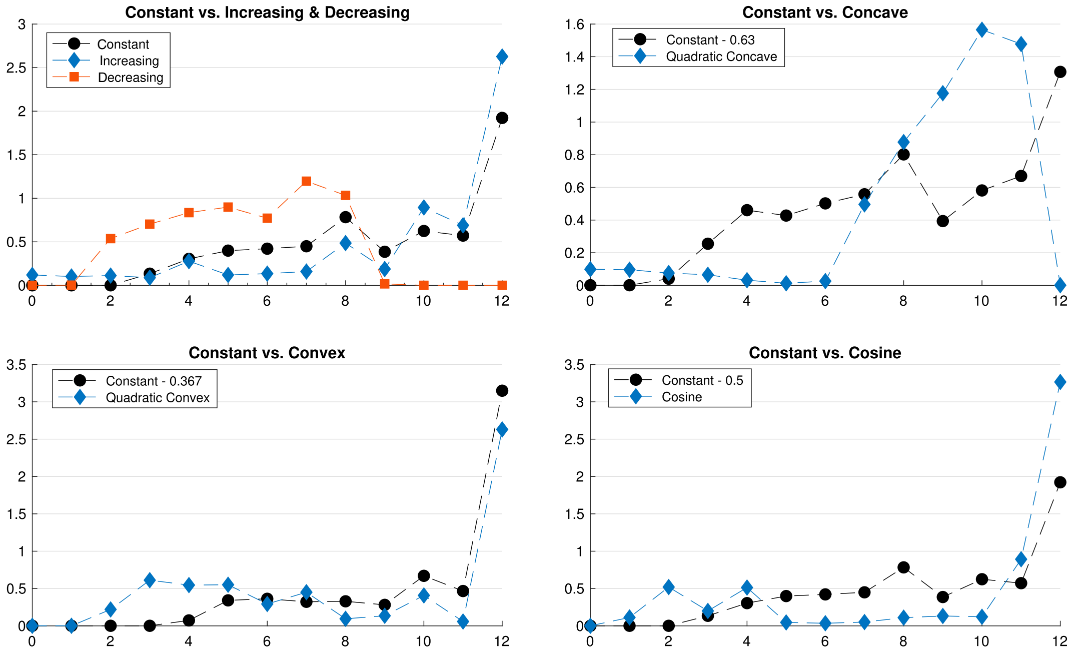


Table 2. Value of Time-Dependent Schedule

		95% confidence interval of simulated total cost		
Service distribution	Show-up pattern	Static schedule	Time dependent	Cost reduction (average cost) (%)
Exponential	Increasing	[4.039, 4.046]	[3.870, 3.876]	4.20
	Decreasing	[6.925, 6.947]	[3.931, 3.941]	43.25
	Concave	[5.263, 5.278]	[3.781, 3.790]	28.18
	Convex	[5.014, 5.029]	[3.259, 3.267]	35.02
	Cosine	[5.402, 5.420]	[3.212, 3.219]	40.58
Log-normal	Increasing	[4.053, 4.062]	[3.827, 3.835]	5.58
	Decreasing	[6.246, 6.267]	[3.784, 3.796]	39.42
	Concave	[4.808, 4.823]	[3.734, 3.742]	22.36
	Convex	[4.646, 4.661]	[3.242, 3.250]	30.24
	Cosine	[4.873, 4.891]	[3.236, 3.246]	33.62
Beta	Increasing	[3.981, 3.987]	[3.832, 3.840]	3.71
	Decreasing	[7.627, 7.648]	[3.875, 3.886]	49.19
	Concave	[5.142, 5.153]	[3.592, 3.598]	30.15
	Convex	[5.550, 5.566]	[3.205, 3.212]	42.28
	Cosine	[5.876, 5.893]	[3.212, 3.218]	45.37

of the simulated total costs are reported in the third column of Table 2. As a comparison, the fourth column presents the corresponding 95% confidence intervals under time-dependent schedules. Each simulation is run with one million samples. We note that as the sample size is quite large, all the 95% confidence intervals are small, and all the comparisons are statistically significant. These results show nontrivial cost reductions ranging from 3.71% to 49.19% when the assumption of constant show-up probability is broken.

6.3. The Value of Distribution Information

We now compare our SPGP method with the DRO model developed in Kong et al. (2020) under the time-dependent show-up probabilities. We emphasize that the goal of the comparison is not to show the superiority of one method over the other; such a conclusion would be unfair, as these two approaches make very different assumptions about the system being modeled. Indeed, in our model we assume that the probability distribution of the service times is known (perhaps from data), whereas in Kong et al. (2020) it is assumed that only the first two moments of service distribution are known, and therefore a DRO approach is warranted. In that sense, we view the approaches as complementary to each other rather than competitors. In light of this discussion, we use the experiments in this section to show the *value of distribution information*—that is, how much we can improve the performance (i.e., total expected cost) by knowing the distribution of the service times compared with not knowing it. Armed with that information, the decision maker might decide, for example, that it is worthwhile investing more resources into collecting (better-quality) data if the extra cost of such resources is offset by the value of distribution information.

We proceeded as follows. We generated schedules under the assumption that the service time distribution is log-normal, a situation often encountered in practice. More specifically, we assumed a log-normal distribution with mean 1 and standard deviation 1. We then ran both methods under a variety of settings, described in Table 3.

We explain in brief the choice of some critical parameters. First, we select two session lengths, $T = 6$ and $T = 12$. The number of patients was chosen to allow for some level of *overbooking*. In Zacharias and Pinedo (2014), the authors propose a model to predict optimal overbooking levels in a clinic. However, the prediction model in that paper was established for problems with *static* no-show, whereas the no-show rates vary in our problem. We therefore roughly based our choice on the optimal overbooking levels when using the average, minimum, and maximum show-up probabilities.

Regarding the cost parameters, as discussed in the previous section, we followed the literature and fixed the overtime cost and idle cost to be 1.5 and 1.0, respectively. To investigate how waiting time costs influence the performance comparison, we ran experiments where

Table 3. Experimental Design

Parameters	Value or range
Session length	6 and 12
Number of patients	8, 10, 12 and 14, 16
Waiting time cost	0.1, 0.5, 0.9
Show-up pattern	Increasing and decreasing
Show-up probability range	[0.4 0.8] and [0.8 0.4]
Service distribution	Log-normal
Computational time	Time used in SPGP

the waiting time cost was set to 0.1 (low), 0.5 (medium), and 0.9 (high).

For the show-up pattern, we only use the increasing and decreasing show-up functions in this set of experiments because the DRO model cannot deal with smooth show-up functions such as the concave, convex, and cosine patterns presented in the previous section. Moreover, on the basis of evidence collected from literature (Geraghty et al. 2008, LaGanga and Lawrence 2012, Huang et al. 2017; Kong et al. 2020), we set the show-up probability to be between 0.4 and 0.8 in both the increasing and decreasing cases.

Finally, we remark that the DRO method implemented in Kong et al. (2020) has no stopping criteria. Thus, we choose the number of iterations in such a way that the two methods (SPGP and DRO) have the same computational time. Note also that, as remarked earlier, the SPGP method uses multiple starting points in order to avoid local minima; for this set of experiments, we used 5 and 10 starting points. The time allocated to DRO then corresponds to the *sum* of the computational times for all starting points in each case.

The results are reported in Tables 4 and 5.[3] We summarize our observations as follows:

- We observe that distribution information provides savings of up to 31.12% of the baseline cost. The average savings is 12.63% for the case where $T = 6$ and 8.01% for the case where $T = 12$.
- The value of distribution information is higher for the case of more congested systems ($T = 6$ and a higher number of patients). This observation could have either one of the possible explanations: (i) in a congested system, the DRO method tends to perform more conservatively (i.e., by focusing on avoiding the worst-case performance), or (ii) the information from the distributions can be better exploited in a more congested system, thereby achieving more significant improvements in the schedule.
- When the waiting time cost is higher, our SPGP model tends to perform better; that is, the distribution

Table 4. Performance Comparison Between the SPGP and DRO Methods Under $T = 6$

	Increasing show-up rate [0.4, 0.8]				Decreasing show-up rate [0.8, 0.4]			
		Total costs				Total costs		
	Run time (sec)	SPGP	DRO	Impv (%)	Run time (sec)	SPGP	DRO	Impv (%)
				Waiting time cost $c_w = 0.1$				
$n = 8$								
5 starts	335.2	3.398	3.572	4.87	8,724.1	3.311	3.48	4.87
10 starts	571.3	3.398	3.574	4.94	8,870.5	3.304	3.483	5.14
$n = 10$								
5 starts	505.8	3.377	3.583	5.76	759.1	4.127	5.651	26.98
10 starts	2,001.8	3.377	3.612	6.52	2,403.3	4.126	5.105	19.17
$n = 12$								
5 starts	2,230.1	3.7	3.855	4.03	1,417.6	5.332	7.741	31.12
10 starts	13,119	3.698	3.859	4.19	2,526.9	5.305	7.294	27.28
				Waiting time cost $c_w = 0.5$				
$n = 8$								
5 starts	7,726.4	4.426	4.788	7.57	452.4	4.885	5.824	16.12
10 starts	12,871	4.42	4.896	9.74	828.5	4.885	5.724	14.64
$n = 10$								
5 starts	4,887.2	5.364	5.743	6.61	11,573	6.453	6.785	4.89
10 starts	5,281.7	5.347	5.683	5.91	13,316	6.453	6.768	4.65
$n = 12$								
5 starts	834.9	6.996	8.772	20.24	331.2	8.385	10.719	21.77
10 starts	1,348.7	6.979	8.684	19.64	1,928.2	8.361	9.991	16.31
				Waiting time cost $c_w = 0.9$				
$n = 8$								
5 starts	1,525.5	5.395	6.27	13.95	585.3	6.055	6.731	10.04
10 starts	2,493.9	5.395	6.018	10.35	1,381	6.042	6.625	8.8
$n = 10$								
5 starts	542.4	7.136	8.925	20.05	2,608.5	8.062	8.899	9.41
10 starts	2,319.3	7.127	8.708	18.15	3,184.3	8.038	8.77	8.34
$n = 12$								
5 starts	11,186	9.959	11.231	11.32	522	10.654	13.469	20.9
10 starts	11,704	9.959	11.157	10.74	825.2	10.654	13.267	19.7

Table 5. Performance Comparison Between the SPGP and DRO Methods Under T = 12

	Increasing show-up rate [0.4, 0.8]				Decreasing show-up rate [0.8, 0.4]			
		Total costs				Total costs		
	Run time (sec)	SPGP	DRO	Impv (%)	Run time (sec)	SPGP	DRO	Impv (%)
Waiting time cost $c_w = 0.1$								
n = 14								
5 starts	14,622	6.159	6.360	3.16	14,957	5.446	5.791	5.97
10 starts	30,402	6.148	6.318	2.68	22,2204	5.446	5.691	4.31
n = 16								
5 starts	604.1	5.879	6.008	2.14	854.6	5.679	7.054	19.49
10 starts	1,289	5.850	6.007	2.61	1,636	5.679	6.990	18.76
Waiting time cost $c_w = 0.5$								
n = 14								
5 starts	617.9	8.211	8.255	0.53	734.9	8.994	9.477	5.1
10 starts	1,300.3	8.204	8.232	0.36	2,127	8.994	9.361	3.92
n = 16								
5 starts	734.9	8.994	9.477	5.10	662.4	9.882	11.869	16.74
10 starts	2,127	8.994	9.361	3.92	1,924.4	9.779	11.493	14.92
Waiting time cost $c_w = 0.9$								
n = 14								
5 starts	448.1	10.169	11.357	10.46	1,214.8	10.662	11.983	11.03
10 starts	1,601.4	10.11	11.26	10.21	1,791.3	10.662	11.917	10.53
n = 16								
5 starts	977.5	11.863	14.071	15.69	18,204	12.637	13.406	5.73
10 starts	1,783	11.863	13.977	15.13	45,423	12.637	13.127	3.73

information provides a higher value. Such a finding may result from the fact that the total cost is more sensitive to the choice of schedule when waiting cost is higher; thus, more is gained by exploiting the knowledge of the service distribution to obtain a better schedule.

- We further observe that the value of distribution information is relatively low in cases with increasing show-up rates, $T = 12$, and lower waiting time costs. The increasing show-up rates usually imply a more front-loaded schedule (see Figure 3), which coincides with the conservative nature of the DRO model to avoid high overtime cost. This, combined with a less congested system ($T = 12$) and lower waiting time costs, diminishes the disadvantage of the DRO model.
- The difference between using 5 or 10 starting points in the SPGP method is practically negligible, suggesting that 5 starting points are enough to avoid local minima.
- Running more iterations in the DRO does not necessarily generate better solutions with regard to the total expected cost. In some cases, fewer iterations may generate a much better solution than more. This is because the DRO approach, by design, optimizes against the worst-case distribution; thus, if the worst-case distribution is too far from the distribution being simulated, the solution obtained with more iterations may have a higher expected cost than an earlier solution.

7. Conclusions

This paper studies the appointment scheduling problem with random service time and patient time-dependent no-show behavior. This problem is difficult because of its nonconvex nature, and little work had been done on this topic. Limited previous literature relies on either heuristics or approximations to solve this problem, thus having obvious limitations. In this paper, we address this problem using a simulation optimization approach. As an important first step, we derive the gradient estimator of the objective function and then develop a projected gradient path method to solve the constrained stochastic optimization problem under decision-dependent uncertainties. The SPGP method builds on two existing methods, STRONG and PGP, originally designed for unconstrained and deterministic optimization problems, respectively. When adapting the Cauchy line search within the PGP algorithm, we have made an important modification that significantly simplifies the computations. To the best of our knowledge, this is the first work that uses a simulation optimization approach method to solve a class of stochastic optimization problems with decision-dependent uncertainties.

Solutions obtained from six patterns of no-show behavior yield considerable differences from the schedules under the assumption of constant show-up probabilities, as well as significant cost reductions when breaking from that assumption. Moreover, in our experiments we have observed that knowledge of the distribution of the underlying random variables can lead to significant cost improvements compared with the case where only the first two moments of that distribution are known, particularly in the case of congested systems. These findings provide important managerial insight, which can help decision makers to decide whether it is worth investing more resources into collecting more or better-quality data. For example, this insight can be particularly relevant for clinics without a "checkout" system (i.e., they do not record when their patients leave at the end of the service). Thus, these clinics do not have service time data and will have to use the DRO model. In such a situation, the value of distribution information can help the decision makers to evaluate whether they should introduce a checkout step, which has a cost, in order to collect service time data—which can then be used to estimate the service time distribution, thereby allowing for the use of SPGP and, consequently, an improvement of their scheduling performance.

We have considered an off-line scheduling problem, assuming that the system already knows the number of arriving patients and does not consider patient nonpunctual arrivals and walk-in patients. We focus on providing a general guideline to appointment scheduling design in the long run. In the objective function, we do not explicitly consider the revenues generated from seeing patients but use the idle time to implicitly capture it. Theoretically, if patients are assigned to those periods during which the no-show rates are higher, fewer patients would show up, leading to higher idle cost, which is equivalent to a revenue loss.

Future work is needed on incorporating *endogenous uncertainty* into different issues regarding patient scheduling, such as *online* appointment scheduling, in which the decisions have to be taken sequentially according to the demand of patients. Furthermore, other sources that might affect the show-up probability need attention. One important extension is heterogeneous patient characteristics. Likewise, other sources of uncertainty, such as patients' lack of punctuality and cancellations, can be analyzed.

Acknowledgments

The authors are grateful to the area editor, the anonymous referees, and the associate editor for their valuable comments and suggestions.

Endnotes

[1] Note that the tower property is valid whenever the expectation is well defined. In the case of (32), we see that the function ψ_k defined in (30) is a sum of a bounded term (as shown in the proof of Theorem 1) and a nonnegative term. Therefore, the expectation of ψ_k is well defined.

[2] In Section 6.3, we use the change in the value of the waiting time costs.

[3] Please note that we include the computational times just to allow for an assessment of the difficulty of solving one instance *relative to* another; the absolute times are not as relevant, as our focus was on the methodology rather than on efficient implementations.

References

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

Basciftci B, Ahmed S, Gebraeel N (2020) Data-driven maintenance and operations scheduling in power systems under decision-dependent uncertainty. *IISE Trans.* 52(6):589–602.

Basciftci B, Ahmed S, Shen S (2021) Distributionally robust facility location problem under decision-dependent stochastic demand. *Eur. J. Oper. Res.* 292(2):548–561.

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

Chang K-H, Hong LJ, Wan H (2013) Stochastic trust-region response-surface method (STRONG)—A new response-surface framework for simulation optimization. *INFORMS J. Comput.* 25(2):230–243.

Conn AR, Gould NI, Toint PL (2000) *Trust Region Methods* (Society for Industrial and Applied Mathematics, Philadelphia).

Daggy J, Lawley M, Willis D, Thayer D, Suelzer C, DeLaurentis P-C, Turkcan A, Chakraborty S, Sands L (2010) Using no-show modeling to improve clinic performance. *Health Informatics J.* 16(4):246–259.

Dantas L, Fleck J, Cyrino Oliveira F, Hamacher S (2018) No-shows in appointment scheduling—A systematic literature review. *Health Policy* 122(4):412–421.

Dantas L, Hamacher S, Cyrino Oliveira F, Barbosa S, Viegas F (2019) Predicting patient no-show behavior: A study in a bariatric clinic. *Obesity Surgery* 29(2):40–47.

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

Erdogan SA, Denton B (2013) Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS J. Comput.* 25(1):116–132.

Erdogan SA, Gose A, Denton BT (2015) On-line appointment sequencing and scheduling. *IIE Trans.* 47(11):1267–1286.

Fu M (2015) Stochastic gradient estimation. Fu MC, ed. *Handbook of Simulation Optimization, International Series in Operations Research & Management Science*, vol. 216 (Springer, New York), 105–147.

Fu M, Hu J-Q (1997) *Conditional Monte Carlo: Gradient Estimation and Optimization Applications* (Springer, New York).

Gallucci G, Swartz W, Hackerman F (2005) Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psych. Services* 56(3):344–346.

Geraghty M, Glynn F, Amin M, Kinsella J (2008) Patient mobile telephone "text" reminder: A novel way to reduce non-attendance at the ENT out-patient clinic. *J. Laryngology Otology* 122(3):296–298.

Glasserman P (1991) *Gradient Estimation via Perturbation Analysis* (Kluwer Academic Publishers, Norwell, MA).

Goel V, Grossmann IE (2004) A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Comput. Chemical Engrg.* 28(8):1409–1429.

Goel V, Grossmann IE (2006) A class of stochastic programs with decision dependent uncertainty. *Math. Programming* 108(2–3):355–394.

Gupta D, Denton B (2008) Appointment scheduling in healthcare: Challenges and opportunities. *IIE Trans.* 40(9):1800–1809.

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

Hellemo L, Barton PI, Tomasgard A (2018) Decision-dependent probabilities in stochastic programs with recourse. *Comput. Management Sci.* 15(3–4):369–395.

Homem-de-Mello T, Shapiro A, Spearman ML (1999) Finding optimal material release times using simulation based optimization. *Management Sci.* 45(1):86–102.

Huang Z, Ashraf M, Gordish-Dressman H, Mudd P (2017) The financial impact of clinic no-show rates in an academic pediatric otolaryngology practice. *Amer. J. Otolaryngology* 38(2):127–129.

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

Jonsbråten TW, Wets RJ-B, Woodruff DL (1998) A class of stochastic programs with decision dependent random elements. *Ann. Oper. Res.* 82:83–106.

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

Kong Q, Li S, Liu N, Teo C-P, Yan Z (2020) Appointment scheduling under schedule-dependent patient no-show behavior. *Management Sci.* 66(8):3480–3500.

Lacy NL, Paulman A, Reuter MD, Lovejoy B (2004) Why we don't come: Patient perceptions on no-shows. *Ann. Family Medicine* 2(6):541–545.

LaGanga LR (2011) Lean service operations: Reflections and new directions for capacity expansion in outpatient clinics. *J. Oper. Management* 29(5):422–433.

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

Moore CG, Wilson-Witherspoon P, Probst JC (2001) Time and money: Effects of no-shows at a family practice residency clinic. *Family Medicine* 33(7):522–527.

Parizi MS, Ghate A (2016) Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. *Comput. Oper. Res.* 67(March):90–101.

Pflug G (1990) On-line optimization of simulated Markovian processes. *Math. Oper. Res.* 15(3):381–395.

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

Rubinstein RY, Shapiro A (1993) *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method* (John Wiley & Sons, Chichester, UK).

Samorani M, LaGanga LR (2015) Outpatient appointment scheduling given individual day-dependent no-show predictions. *Eur. J. Oper. Res.* 240(1):245–257.

Shapiro A, Dentcheva D, Ruszczynski A (2009) *Lectures on Stochastic Programming: Modeling and Theory* (Society for Industrial and Applied Mathematics, Philadelphia).

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

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