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

Tito Homem-De-Mello[1], Qingxia Kong[2], and Rodrigo Godoy Barba [3]

[1]*School of Business, Universidad Adolfo Ibáñez, Chile, tito.hmello@uai.cl*
[2]*Rotterdam School of Management, Erasmus University, the Netherlands, q.kong@rsm.nl*
[3]*Black & Veatch, Chile, rodrigogoba@gmail.com*

Abstract

The appointment scheduling problem (ASP) is well studied in the literature. The vast majority of the existing work, however, does not consider the well-established fact that whether a patient shows up for an appointment or not is influenced by the appointment time, the usual decision variable in the ASP. This paper studies the ASP with random service time (exogenous uncertainty) and patient's decision-dependent no-show behavior (endogenous uncertainty). This problem belongs to the class of stochastic optimization models with decision-dependent uncertainties. These problems are notoriously difficult to solve due to the lack of convexity—and also in our case—the presence of continuous distributions (service time), which requires different techniques to solve this problem. We develop a projected gradient path (PGP) method to solve the non-convex constrained stochastic optimization problem. This method improves upon the Stochastic Trust-Region Response-Surface (STRONG) Method by Chang, Hong, and Wan (2013) for unconstrained simulation optimization by integrating a Cauchy line search originally designed for deterministic problems (see Conn, Gould, and Toint (2000)). We derive the gradient estimator by exploiting the structure of the objective function of the problem and then apply the PGP method to solve for the schedules accordingly. To the best of our knowledge, this is the first paper to solve the problem with decision-dependent uncertainties using simulation optimization approach. This method can solve the exact ASP problem under arbitrary smooth show-up probability functions and the performance is shown to bypass the distributionally robust optimization method developed by Kong, Li, Liu, Teo, and Yan (2019) in most cases. We present solutions under different no-show behaviors and demonstrate that breaking the assumption of constant show-up probability substantially changes the scheduling solutions. Finally, we show that significant cost reduction can be achieved when incorporating the time-dependent no-show behavior: While the amount of cost reduction varies according to several factors, most importantly the shape of the show-up probability functions, in our numerical experiments we observed significant cost reductions in all settings, with values of up to 54% in one case.

1 Introduction

Deciding appointment times for patients is one of the critical decisions outpatient healthcare providers have to make to better utilize the critical resources, reduce overtime work of healthcare providers, and in the meantime, to shorten the waiting time of patients. Shorter inter-arrival times can lead to higher utilization of the resources but longer patient waiting time. On the contrary, longer inter-arrival 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 involve 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/her appointment, i.e., patient no-show behavior. Every healthcare system bears the risk of patient no-show, which has significnatly negative impacts on the healthcare system including increased system costs, under-utilized critical resources, and reduced patient outcome (Daggy et al., 2010; Lacy, Paulman, Reuter, & Lovejoy, 2004; Moore, Wilson-Witherspoon, & Probst, 2001). Besides individual characteristics such as demographic and social-economics factors, home-clinic distance, and no-show history (see Dantas, Fleck, Cyrino Oliveira, and Hamacher (2018) for a systematic review), current literature reports that patient no-show behavior can be time dependent. Evidence shows that the show-up probability does not only depend on the appointment day and day of the week (Gallucci, Swartz, & Hackerman, 2005; Samorani & LaGanga, 2015), but also on time of the day in which the patient is scheduled. Moore et al. (2001) analyze a hospital in South Carolina and the 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, Hamacher, Cyrino Oliveira, Barbosa, and Viegas (2019) in a Bariatric clinic study show that patients tend to show up less towards the later hours of the day. Kong et al. (2019) analyze two independent data sets and have interesting results: On weekdays patients in a US adult health care facility are more likely to show up at the beginning of the day or late in the afternoon while in a Chilean pediatric practice, the show-up rates on weekdays are lower in early morning and a peak is seen at the middle of the day.

Despite the widely recognized evidence on time-dependent no-show behavior, scarcely any work on appointment scheduling has taken this phenomenon 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, Barton, and Tomasgard (2018). Such problems are notoriously hard to solve, especially because of their lack of convexity, and solutions 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) 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 find solutions to 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, M. C. Fu and Hu (1997); Glasserman (1991); Rubinstein and Shapiro (1993) for comprehensive discussions and also M. C. 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 depend 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 as well as 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 any functional form 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 can deal only with piecewise linear function form (cf. Kong et al. (2019)).

The second ingredient of our approach is the optimization algorithm. As we show in the paper, 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 and thus, by making use of the gradient estimation technique discussed above, we can apply a sampling-based first-order stochastic optimization method. We develop a gradient projection path method (PGP) to solve the constrained stochastic problem with decision-dependent uncertainties, building upon the STRONG approach proposed by Chang et al. (2013) to find stationary points and Cauchy line searches originally designed for deterministic problems (see Conn et al. (2000)). The STRONG method combines the advantage of the trust-region methods and the response surface methodology to solve unconstrained simulation optimization problems. However, since 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 simplification to the line search procedure—which may be of independent interest as it is not related to our specific problem—that avoids the calculation of projections onto tangent cones, which can be computationally expensive.

While the resulting algorithm yields locally optimal solutions (which is natural since the problem is non-convex), by applying a multi-start procedure we are able to obtain very good solutions. As discussed earlier, we are not aware of any other approach that can model arbitrary smooth show-up probability functions; the only available benchmark to our method is the work of Kong et al. (2019), 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 only uses the first two moments and thus avoids more complexity brought by estimating any specific distribution. That method, however, can be very conservative under some extreme circumstances, e.g., when the system is very crowded. Via a set of computational studies, we show that our approach can replicate the optimal schedules reported in Hassin and Mendel (2008) and makes significant improvements as compared to Kong et al. (2019), not only in terms of the quality of the solutions but also because in Kong et al. (2019) only piece-wise linear show-up functions are allowed. Moreover, 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 (13.1% to 54.0%) can be observed, while relatively milder improvements (4.2% to 12.8%) are seen in the increasing show-up function.

2 Literature Review

The appointment scheduling problem is extensively studied in the literature of Operations Management. Interested readers can refer to Ahmadi-Javid and Klassen (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.

In the literature review presented in Cayirli and Veral (2003) several sources of uncertainty, assumptions and methodologies that have been used are described. Most commonly used methodologies include queuing theory, mathematical programming and simulation studies. There are cases where only the randomness of service time has been considered (Denton & Gupta, 2003; Robinson & Chen, 2003), only patients no-show behavior (Zacharias & Pinedo, 2014) is considered, or both (Erdogan & Denton, 2013; Hassin & Mendel, 2008; Jiang, Shen, & Zhang, 2017; Kong et al., 2019). 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 incorporate patient no-shows in the ASP. Their computational experiments show that the optimal inter-arrival times reduces 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, i.e., the show-up probabilities remain unchanged across the planning session. They report continuous scheduling decisions (vs. 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 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 homogeneous patients case. Jiang et al. (2017) consider a distributionally robust model under static patient no-show and random service duration, studying the cases risk-neutral and risk-averse system operator in order to incorporate his/her risk preferences.

Two works are directly related to this paper. Kong et al. (2019) 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 proposed is based on a heuristic due to the computational difficulty. Therefore, some research opportunities are left open for new models that consider stochastic service times and more systematic optimization approach.

Kong et al. (2019) 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 schedule-dependent show-up rates under stochastic service time. Their computational results suggest that a significant efficiency gain can be achieved when time-of-day effect on show-up probabilities is incorporated. Kong et al. (2019) 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 (SDP approximations to co-positive and completely positive cones) and heuristics (iterative approach). It is not even clear whether the solution generates upper or lower bounds. Second, due to the cone structure, the current solution proposed in the paper can only deal with piece-wise linear show-up functions with time. Third, since 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/and the show-up probabilities are large towards 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 a new

PGP method to solve the constrained stochastic problem with decision-dependent uncertainties, improving upon two existing originally designed for unconstrained and deterministic problems. In the following, we briefly review the stochastic optimization literature on decision-dependent uncertainties.

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 network design problem (Ahmed (2000)), gas field planning (Goel and Grossmann (2004)) and project decisions (Jonsbråten, Wets, and Woodruff (1998)). Goel and Grossmann (2006) address a class of stochastic program 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.

Cayirli and Veral (2003) provide an extensive review of related literature, identifying differences in the definition and formulation of the problem. The differences include the nature of the decision-making (static or dynamic), number of servers, number of doctors, punctuality of patients, presence of no-shows, presence of regular and emergency walk-ins, presence of companions and service times (empirical or theoretical distribution). There are also differences on the performance measure, the design of the appointment system and the methodology applied.

Under this framework, 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, incurring patient no-show; in particular, we assume that patient show-up probabilities depend on time of arrivals. If they show up, they always arrive at the scheduled time, i.e., non-punctual arrivals are not

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 their appointment is scheduled at time t

Table 1: Notation

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's idle time and overtime. The objective is to minimize the weighted sum of the three performance measures by deciding the optimal schedule for each patient.

As customary in the literature, the decision making process in this paper is assumed to be static (offline), i.e., all appointment requests are known in advance to the scheduling decision. Some recent work (Erdogan, Gose, & Denton, 2015; Parizi & Ghate, 2016; Truong, 2015) studies on the *online* appointment scheduling problem, in which they dynamically assign arriving patients to the 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, $\mathbf{x} = (x_i)\ , i = 0, \ldots, n$ are the decision variables and denote the inter-arrival time between patient i and patient $i+1$, with x_0 (≥ 0) denoting the arrival time of the first patient and x_n the time between the arrival of the last patient and the end of the session (time T). Each element in $\mathbf{x}$ is non-negative and $\sum_{i=0}^{n-1} x_i \leq T$, which indicate that each patient is allocated with a non-negative time interval and all patients arrive within the clinic session $[0, T]$. . The scheduled arrival time for patient i is 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 respectively the attendance status of each patient and their service time. More specifically, $\mathbf{A}$ is a vector of n Bernoulli random variables ($A_i = 1$ if patient i

shows-up and $A_i = 0$ if not) and $\mathbf{U}$ is a vector of n random variables characterizing the service times.

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

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

, where the waiting time and idle time for the patient i are given by the following recursive expressions:

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

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

$$S_i = \max\left(0,\, x_i - A_i U_i - W_i\right) \quad i = 1, \ldots, n \tag{4}$$

$$L = \max\left(0,\, W_n + A_n U_n - x_n\right) \tag{5}$$

$$x_n = T - \sum_{j=0}^{n-1} x_j. \tag{6}$$

Due to 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 an schedule given by $\mathbf{x}$ that minimizes the expected total cost, subject to non-negativity constraints. It formulated as follows:

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

where $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} \geq 0,\; \sum_{j=0}^{n-1} x_j \leq T\}$. One difficulty is to calculate the objective function $f(\mathbf{x})$ which involves the recursive equations (3) and (4). Assuming that the service times $\mathbf{U}$ do not depend on the time of the day — a reasonable assumption in this context — we can express the function as a conditional expectation as follows:

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

where

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

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, i.e., 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 ω_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 $\omega_{\mathbf{A}} = (\omega_{A_1}, ..., \omega_{A_n})$, 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., $\mathbb{P}\left(A_i(\mathbf{x}) = \omega_{A_i}\right)$. 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:

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

From expressions (10)-(11) we can see explicitly the *endogenous uncertainty*, where the probability of a scenario depends on the decision variables 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.

Through simulation and interpolation a graph that shows the estimated objective function surface is presented on Figure 1. The axes in the figures correspond to the values of x_1 and x_2 (x_0 is set to 0). This

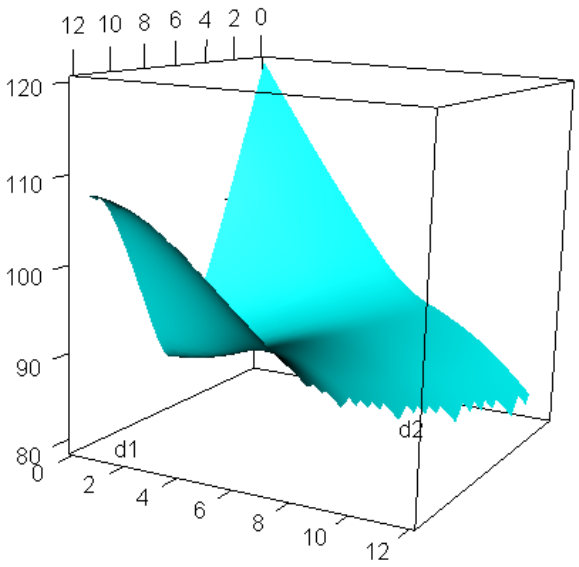


Figure 1: Estimated Objective Function Value under the Case of Three Patients, Uniform Service Time ($U \sim U(2.5, 7.5)$), Linearly Decreasing (from 0.8 to 0.2) Show-up Function and Cost Parameters $c_w = 1$; $c_I = 10$; $c_o = 15$.

is an example of a system with three patients whose show-up probability decreases linearly during the day. Despite the simplicity of this instance, it is already possible to observe the non-linear non-convex feature of the problem. In Sections 4 and 5 we present a gradient-based algorithm to solve the problem.

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

Assumption A: The no-show behavior of patients is not affected by the behavior of other patients, i.e., 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. M. C. 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 very 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 $\mathbf{x}$; 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 only be applied when 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 Conditional Monte Carlo techniques (see M. C. Fu and Hu (1997)). To proceed, consider the representation of $f(\mathbf{x})$ given in (13). Then, by defining $P(\mathbf{x}, \omega_{\mathbf{A}}) := \prod_{i=1}^{n} P_i(\mathbf{x}, \omega_{A_i})$, we can write

$$f(\mathbf{x}) = \sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} P(\mathbf{x}, \omega_{\mathbf{A}})\, \mathbb{E}_{\mathbf{U}} \left[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) \right] \tag{14}$$

and thus, if $P(\cdot, \omega_{\mathbf{A}})$ is differentiable, so is $f(\cdot)$ as long as $\mathbb{E}_{\mathbf{U}} \left[TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U}) \right]$ is differentiable as well. Theorem 1 below 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 A holds. Then, the function* $TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})$ *is differentiable at* $\mathbf{x}$ *with probability one (i.e. for almost every realization of* U*) and its gradient can be calculated. Moreover, there exists a constant* $M > 0$ *such that*

$$|TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) - TC(\mathbf{y}, \omega_{\mathbf{A}}, \mathbf{U})| \leq M \left\| \mathbf{x} - \mathbf{y} \right\| \quad \textit{w.p.1.} \tag{15}$$

Proof. For each $i = 1, \ldots, n$, let B_i denote the time at which patient i begins being served. Also, let $r_i := x_0 + \ldots + x_{i-1}$ 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

$$B_1 = r_1 \tag{16}$$

$$B_i = \max\left(r_i,\, B_{i-1} + \omega_{A_{i-1}} U_{i-1}\right), \quad i = 2, \ldots, n. \tag{17}$$

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:

$$W_i = B_i - r_i, \tag{18}$$

$$S_i = B_{i+1} - B_i - \omega_{A_i} U_i, \tag{19}$$

where B_{n+1} is defined as in (17) and $r_{n+1} \equiv T$.

In view of (17), 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 (5) we see that the overtime L can be written as the waiting time of the "$(n+1)$st" patient and thus we have

$$L = W_{n+1} = B_{n+1} - r_{n+1}. \tag{20}$$

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

$$B_i(r_1, \ldots, r_i) = \max\left(r_i,\, r_{i-1} + \omega_{A_{i-1}} U_{i-1},\, r_{i-2} + \omega_{A_{i-1}} U_{i-1} + \omega_{A_{i-2}} U_{i-2}, \ldots, r_1 + \sum_{j=1}^{i-1} \omega_{A_j} U_j \right), \quad i = 2, \ldots, n+1. \tag{21}$$

For $i = 1$ the statement is trivially true, since the function $B_1(r_1)$ is just the identity function. Suppose that the statement is true for $1, \ldots, i-1$. Then, from (17) and the induction hypothesis we have

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

and so we see that (21) 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 one (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

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

From equations (18) and (19), together with the relation $r_i = x_0 + \ldots + x_{i-1}$, we see that it suffices to show that B_i is differentiable at $\mathbf{r}$. From (21), 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 (21) are equal (note that the kink points depend

on the scenario of U). Outside those kink points we have, for any $i = 1, \ldots, n+1$ and $j = 1, \ldots, n$,

$$\frac{\partial B_i}{\partial r_j} = \begin{cases} 1 & \text{if } j = k^i \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

, where k^i is the index of the function attaining the maximum in (21). Furthermore,

$$\begin{aligned} P\left(B_i \text{ is not differentiable at } \mathbf{r}\right) &\leq P\left(r_{k_s^i} + \sum_{j=k_s^i}^{i-1} \omega_{A_j} U_j = r_{k_t^i} + \sum_{j=k_t^i}^{i-1} \omega_{A_j} U_j \text{ for some } k_s^i, k_t^i \in \{1, \ldots, i\},\ k_s^i < k_t^i\right) \\ &\leq \sum_{k_s^i, k_t^i \in \{1,\ldots,i\} \,:\, k_s^i < k_t^i} P\left(r_{k_s^i} + \sum_{j=k_s^i}^{i-1} \omega_{A_j} U_j = r_{k_t^i} + \sum_{j=k_t^i}^{i-1} \omega_{A_j} U_j\right) \\ &= \sum_{k_s^i, k_t^i \in \{1,\ldots,i\} \,:\, k_s^i < k_t^i} P\left(0 < r_{k_t^i} - r_{k_s^i} = \sum_{j=k_s^i}^{k_t^i - 1} \omega_{A_j} U_j\right). \end{aligned} \tag{23}$$

Consider the term inside the sum in (23). If all the ω_{A_j}, $j = k_s^i, \ldots, k_t^i - 1$ are zero, then it is clear that the probability on the right-hand side of that equation is zero. Suppose now that the set $J := \{j \in \{k_s^i, \ldots, k_t^i - 1\}\,;\, \omega_{A_j} = 1\}$ is non-empty. 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, $r_1, \ldots, r_n$ defined in terms of the chosen point $\mathbf{x}$). It follows that

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

and therefore B_i is differentiable with probability one. It follows from (18)-(20) that W_i, S_i, and L are also differentiable with probability one with respect to $\mathbf{r}$, and hence also with respect to $\mathbf{x}$. By definition (1) of the total cost function TC, 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 one (i.e. for almost every realization of U).

Finally, we show that $TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})$ is a Lipschitz function. From the relations (18), (19) and (20) we have for $i, j = 1, \ldots, n$ 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 (22). Then, from (1) we write

$$\frac{\partial TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})}{\partial r_j} = c_o \frac{\partial L}{\partial r_j} + c_I \mathbb{I}_{\{j=1\}} + \sum_{i=1}^{n} \left[c_w \omega_{A_i} \frac{\partial W_i}{\partial r_j} + c_I \frac{\partial S_i}{\partial r_j} \right], \tag{24}$$

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

$$\frac{\partial TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})}{\partial x_j} = \sum_{i=1}^{n} \frac{\partial TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})}{\partial r_i} \frac{\partial r_i}{\partial x_j} = \sum_{i=1}^{n} \frac{\partial TC(\cdot, \omega_{\mathbf{A}}, \mathbf{U})}{\partial r_i} \mathbb{I}_{\{j<i\}}, \tag{25}$$

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. □

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 non-differentiable iff $r_2 = B_1 + A_1 U_1 = r_1 + A_1 U_1$, 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 non-differentiable iff $r_3 = B_2 + A_2 U_2$. If $r_3 = r_2$, 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

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

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 $P(A_1 = 0) \times \ldots \times P(A_i = 0) > 0$. Note that the condition $r_{i+1} = r_i$ is equivalent to $x_i = 0$.

Corollary 2. *If the show-up function $p(t)$ is differentiable and Assumption A holds, then the function $f(\mathbf{x})$ in* (14) *is differentiable on the interior of the simplex $\mathcal{X}$.*

Proof. It immediately follows from (11) and (14) that, if the show-up function $p(t)$ is differentiable, then $f(\mathbf{x})$ is differentiable so is $\mathbb{E}_{\mathbf{U}}[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})]$. By virtue of Theorem 1, we see that the conditions es-

tablished in Shapiro, Dentcheva, and Ruszczynski (2009, Theorem 7.49) are fulfilled, which ensures that $\mathbb{E}_{\mathbf{U}}\left[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})\right]$ is differentiable and, moreover, $\nabla \mathbb{E}_{\mathbf{U}}\left[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})\right] = \mathbb{E}_{\mathbf{U}}\left[\nabla TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})\right]$. □

Remark 2. When there are no no-shows—i.e., all show-up probabilities are equal to one—the resulting model is very similar to that studied by Homem-de-Mello, Shapiro, and Spearman (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 non-convex, 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 zero if a patient does not show up. Consequently, when combined with the show-up variables, the service times may have an atom at zero and thus the aforementioned assumption in Homem-de-Mello et al. (1999) does not hold.

With the above results at hand, we can now write the derivatives of $f(\mathbf{x})$ in (14) as follows:

$$\begin{aligned}
\nabla f(\mathbf{x}) &= \nabla \left(\sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} P(\mathbf{x},\, \omega_{\mathbf{A}})\, \mathbb{E}_{\mathbf{U}}\left[TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U})\right] \right) \\
&= \nabla \mathbb{E}_{\mathbf{U}} \left[\sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} P(\mathbf{x},\, \omega_{\mathbf{A}})\, TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) \right] \\
&= \mathbb{E}_{\mathbf{U}} \left[\nabla \left(\sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} P(\mathbf{x},\, \omega_{\mathbf{A}})\, TC(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}) \right) \right] \\
&= \mathbb{E}_{\mathbf{U}} \left[\nabla \varphi(\mathbf{U}, \mathbf{x}) \right], \qquad (26)
\end{aligned}$$

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

Note that, in principle, having an expression for $\nabla \varphi\left(\mathbf{U}, \mathbf{x}\right)$ 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 (26). However, obtaining the derivatives of $\varphi\left(\mathbf{U}, \mathbf{x}\right)$ 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 as

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

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

$$\psi_k\left(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}\right) := \frac{\partial TC\left(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}\right)}{\partial\, x_k} + TC\left(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}\right) \cdot \left(\sum_{j=1}^{n} \frac{\partial P_j(\mathbf{x}, \omega_{A_j})}{\partial\, x_k} \cdot \frac{1}{P_j(\mathbf{x}, \omega_{A_j})} \right). \tag{27}$$

Then, we can write the derivative of φ with respect to x_k as

$$\frac{\partial \varphi\left(\mathbf{U}, \mathbf{x}\right)}{\partial\, x_k} \;=\; \sum_{\omega_{\mathbf{A}} \in \{0,1\}^n} \psi_k\left(\mathbf{x}, \omega_{\mathbf{A}}, \mathbf{U}\right) \cdot \prod_{i=1}^{n} P_i(\mathbf{x}, \omega_{A_i}) \;=\; \mathbb{E}_{\mathbf{A}(\mathbf{x})}\left[\psi_k\left(\mathbf{x}, \mathbf{A}(\mathbf{x}), \mathbf{U}\right) \mid \mathbf{U}\right]. \tag{28}$$

The reformulation of the derivative in (28) is the fundamental step that allows us to obtain an unbiased sampling estimator on the gradient of the objective function f. Recall that from (26) that $\nabla f(\mathbf{x}) = \mathbb{E}_{\mathbf{U}}\left[\nabla \varphi(\mathbf{U}, \mathbf{x})\right]$. Then, by using (28) we can write

$$\frac{\partial f(\mathbf{x})}{\partial\, x_k} \;=\; \mathbb{E}_{\mathbf{U}}\left[\frac{\partial \varphi\left(\mathbf{U}, \mathbf{x}\right)}{\partial\, x_k}\right] \;=\; \mathbb{E}_{\mathbf{U}}\left[\mathbb{E}_{\mathbf{A}(\mathbf{x})}\left[\psi_k\left(\mathbf{x}, \mathbf{A}(\mathbf{x}), \mathbf{U}\right) \mid \mathbf{U}\right]\right] \;=\; \mathbb{E}_{\mathbf{A}(\mathbf{x}), \mathbf{U}}\left[\psi_k\left(\mathbf{x}, \mathbf{A}(\mathbf{x}), \mathbf{U}\right)\right], \tag{29}$$

where the last equality follows from the tower property of expectations. 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 theorem below.

Theorem 3. *Let $\{(\mathbf{A}^s(\mathbf{x}), \mathbf{U}^s)\}$, $s = 1, ..., S$ be a sample of size S of the bi-variate 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, a consistent and unbiased estimator of $\frac{\partial f(\mathbf{x})}{\partial\, x_k}$ can be obtained as*

$$\frac{\partial f(\mathbf{x})}{\partial\, x_k} \;\approx\; \frac{1}{S} \sum_{s=1}^{S} \psi_k\left(\mathbf{x}, \mathbf{A}^s(\mathbf{x}), \mathbf{U}^s\right) \tag{30}$$

where ψ_k is defined in (27).

The result in Theorem 3 allows 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 non-linear non-convex 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 non-linear stochastic optimization. In Section 5.1 we discuss one such method available in the literature. However, we need to adapt the method since our problem is constrained, the discussion of which is presented in Section 5.2.

5.1 Description of the STRONG Method

The algorithm we propose to solve this problem is built upon the STRONG algorithm proposed by Chang et al. (2013). STRONG, or 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 small subregions, i.e., the trust region, to find new candidate solutions. The improvement of a candidate solution is evaluated under statistical tests. It contains two stages in the outer loop and an inner loop. When the algorithm is on Stage I 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. Then, simulations runs are required to estimate the expected total cost of the candidate solution. The improvement is assessed with two tests: *Ratio-Comparison (RC) Test* and *Sufficient-Reduction Test*. The RC test intends to test whether the trust region is trustworthy while 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, in particular, the size of the trust region diminishes when no satisfactory solution is found [1]. If no satisfactory solution is found on this stage, the algorithm goes to the Inner Loop where the sample size is increased to ensure that a better solution is found or the stopping criterion is satisfied.

A brief description of the steps of the algorithm is outlined below (see Chang et al. (2013)):

Step 0: Set the iteration counter $k = 0$. Select an initial solution x_0, 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 the 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.

[1] In the original description of STRONG in Chang et al. (2013), a quadratic approximation is also used. However, in our implementation, the quadratic approximation is not used since the estimators of the Hessian matrix are too noisy.

Convergence properties and further details of STRONG are described in Chang et al. (2013).

5.2 Projected Gradient Path Algorithm

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

Generally, 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 to the feasible set. In Conn et al. (2000, Chapter 12) an algorithm under the trust-region methodology for deterministic problems with convex constraints is presented. The technique is based on a line search along the projected-gradient path to compute a generalized Cauchy point to find improved and feasible solutions. The purpose of this 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.

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

$$\mathbf{q}(t, \mathbf{x}) \;:=\; \Pi_{\mathcal{X}}\left(\mathbf{x} - t\nabla f(\mathbf{x})\right), \tag{31}$$

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), $\mathbf{g}_\nu$ an estimator of the gradient $\nabla f(\mathbf{x}^\nu)$, and Δ_ν the trust region size. The algorithm can be outlined as follows:

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

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

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 hold: (a) the candidate point is not too close to the current point, (b) the decrease in the approximation m_ν is not too large, (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 above 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) above is expressed as

$$\left\|\Pi_{\mathcal{T}(\mathbf{q}(t_j,\mathbf{x}^\nu))}\left[-\mathbf{g}_\nu\right]\right\| \;\leq\; \kappa_{epp}\frac{|\langle \mathbf{g}_\nu,\ \mathbf{s}_\nu(t_j)\rangle|}{\Delta_\nu}, \tag{32}$$

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

$$\mathcal{M}(\mathbf{x}^\nu, \theta) \;:=\; \left|\ \min\left\{\langle \mathbf{g}_\nu,\ \mathbf{d}\rangle \;:\; \mathbf{x}^\nu + \mathbf{d} \in \mathcal{X},\ \|\mathbf{d}\| \leq \theta\right\}\ \right|. \tag{33}$$

In our implementation we replaced condition (32) with

$$\frac{\mathcal{M}(\mathbf{x}^\nu, \theta_j) - |\langle \mathbf{g}_\nu,\ \mathbf{s}_\nu(t_j)\rangle|}{2\theta_j} \;\leq\; \kappa_{epp}\frac{|\langle \mathbf{g}_\nu,\ \mathbf{s}_\nu(t_j)\rangle|}{\Delta_\nu}, \tag{34}$$

where $\theta_j := \|\mathbf{s}_\nu(t_j)\| + 1$. We show now that the replacement of (32) with (34) does not affect the proof of convergence of Conn et al. (2000), provided some extra condition is imposed.

Proposition 4. *Suppose that* (34) *is used in place of* (32) *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*

$$\kappa_{frd} \;\leq\; \frac{2}{4\kappa_{epp} + 1}. \tag{35}$$

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

$$\frac{\mathcal{M}(\mathbf{x}^\nu, \theta_j) - |\langle \mathbf{g}_\nu,\ \mathbf{s}_\nu(t_j)\rangle|}{2\theta_j} \;\leq\; \left\|\Pi_{\mathcal{T}(\mathbf{q}(t_j,\mathbf{x}^\nu))}\left[-\mathbf{g}_\nu\right]\right\|. \tag{36}$$

As argued in the proof of Theorem 12.2.1 in Conn et al. (2000), condition (32) holds for j sufficiently large when $t_{max} = \infty$ for all j. Thus, in that case it follows from (36) that (34) 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, (34) 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

$$|\langle \mathbf{g}_\nu,\ \mathbf{s}_\nu(t_j)\rangle| \ \geq\ \frac{\Delta_\nu}{2(\Delta_\nu+1)\kappa_{epp}+\Delta_\nu}\,\mathcal{M}(\mathbf{x}^\nu,\theta_j) \ \geq\ \frac{\Delta_\nu}{2(\Delta_\nu+1)\kappa_{epp}+\Delta_\nu}\,\mathcal{M}(\mathbf{x}^\nu,1), \tag{37}$$

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 (35) and (37) together imply that

$$|\langle \mathbf{g}_\nu,\ \mathbf{s}_\nu(t_j)\rangle| \ \geq\ \frac{1}{2}\kappa_{frd}\min(\Delta_\nu,1)\,\mathcal{M}(\mathbf{x}^\nu,1), \tag{38}$$

where κ_{frd} is another constant used in the algorithm. Note that (38) 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, (38) already holds trivially). When (37) holds, a sufficient condition for (38) to hold is that

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

i.e.,

$$\kappa_{frd} \ \leq\ \frac{2\Delta_\nu}{2(\Delta_\nu+1)\kappa_{epp}+\Delta_\nu}\,\frac{1}{\min(\Delta_\nu,1)}. \tag{39}$$

It is easy to see that the function on the right-hand side of (39) 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, (39) holds for all $\Delta_\nu > 0$ provided that (35) holds. It follows that conditions (35) and (37) together imply (38), 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) dictate 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 (35) holds. In our implementation we used $\kappa_{epp} = 0.125$.

Through the incorporation of the proposed methodology by Conn et al. (2000), which is designed to find improved and feasible solutions in a constrained context, we can then adapt STRONG to our non-linear non-convex constrained stochastic optimization problem. However, convergence is ensured only to stationary

points, so the algorithm could end up at a local minimum. Since there is no guarantee of global optimality, a multi-start strategy is adopted. Then, simulation is used to compare the obtained solutions and select the best one.

6 Computational Studies

In this section, we apply the projected gradient path (PGP) method 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. We then compare our scheduling solutions with those reported in Hassin and Mendel (2008) and its performance with that of the distributionally robust optimization (DRO) model developed in Kong et al. (2019). In Section 6.3, we explore scheduling patterns under different show-up functions, as compared to the static case, and perform an out-of-sample simulation study to demonstrate the value of incorporating patient schedule-dependent no-show behavior.

6.1 Experimental Design and Computational Setup

In the basic experiment set-up, we solve for the scheduling solutions for 12 patients in a clinic session that spans six time units. Three types of costs occur in the system, which are costs of patient waiting time, physician's idle time and overtime. Following Zacharias and Pinedo (2014) and Kong et al. (2019), we set the three costs to be $0.1, 1$ and 1.5, respectively.

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 uniform. 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. Figure 2 depicts the show-up rates during a session of six time units. 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.

The parameters used in the algorithms are presented in Appendix A. We randomly select 20 different starting points, run the Gradient Projection algorithms to generate scheduling solutions and then use different sample sets (sample size 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. If the confidence interval of the selected schedule overlaps with that of the "second-best" schedule

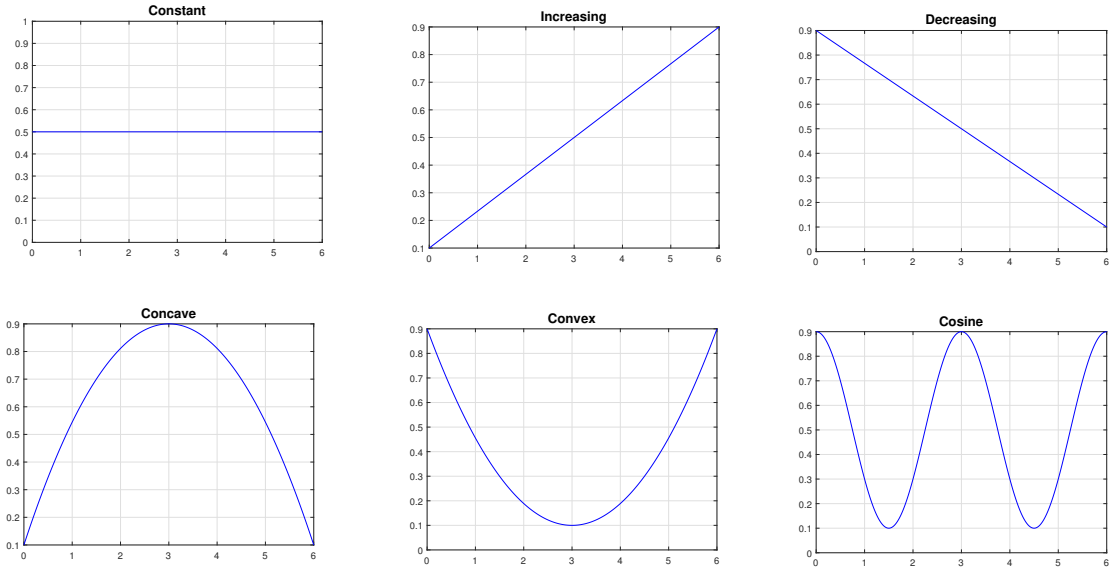


Figure 2: Show-up Probabilities as a Function of Time (Y-axis: Show-up Probability; X-axis: Time)

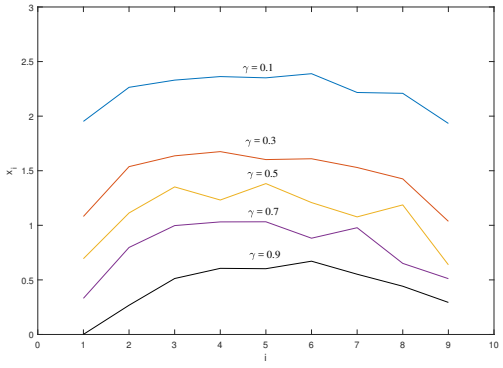
and they exhibit significantly different patterns, more starting points will be added and the same procedure follows.

6.2 Performance Evaluation

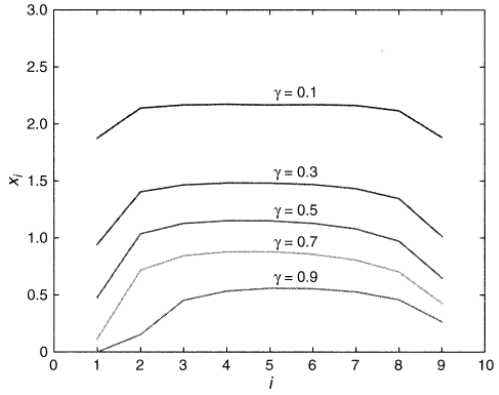
We first apply our algorithm to the setting of Hassin and Mendel (2008) and compare our solutions to those reported in their paper. Hassin and Mendel study the ASP with static no-shows, i.e., the show-up probability remains the same across the scheduling horizon. Its objective function minimizes the weighted sum of patients' waiting time and server's available time. Figures 1 and 2 in that paper report the optimal schedules with 10 patients and show-up rates 0.7 and 0.3 under different values of γ. $\gamma(=\frac{c_s}{c_s+c_w})$ is defined as the relative cost parameter, where c_w denotes the unit waiting cost and c_s the unit server's availability cost.

Note that server's availability counts from time zero to service starting time of the last patient plus possible service time (if the last patient shows up). In our paper, it is equivalent to: 1) set the unit overtime cost (c_o) to be zero and the unit idle time cost c_I the same as the unit server's availability cost c_s ; 2) ignore the idle time (if any) between the service ending time of the last patient and the end of the session; and 3) ask the first patient to arrive at time zero. Since the session length does not play a role in the objective function here, we set the session length $T = 1000$ to make sure that it does not constrain the choice of the inter-arrival times. Figures 3a and 3c demonstrate the scheduling solutions under our approach and Figures 3b and 3d show the optimal schedules reported in Hassin and Mendel (2008). Comparing two sets of figures,

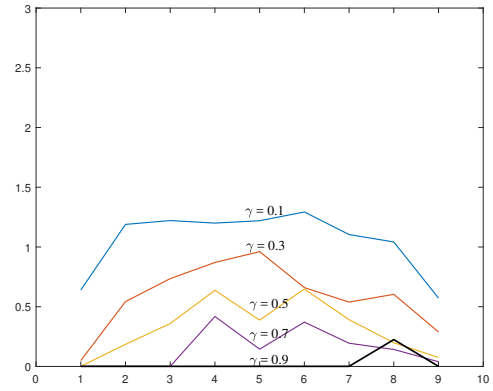
we can see that the schedule solutions derived by our approach are very close (in terms of both pattern and absolute values) to the optimal schedules in Hassin and Mendel (2008). This shows that our method can generate near-optimal solutions when the show-up probability is constant over the scheduling horizon.



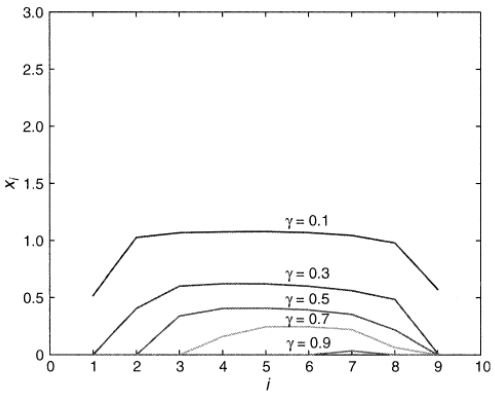
(a) Schedules by PGP ($p = 0.7$)



(b) Schedules in Hassin and Mendel (2008) ($p = 0.7$)



(c) Schedules under PGP ($p = 0.3$)



(d) Schedules in Hassin and Mendel (2008) ($p = 0.3$)

Figure 3: Comparison of Schedules under the PGP Algorithm and in Hassin and Mendel (2008) with 10 Patients, Different Show-up Probabilities, and Different Relative Cost Parameters

Next, we compare our PGP method with the DRO model reported by Kong et al. (2019) under the time-dependent show-up probabilities. To increase the validity of comparison, we set the number of patients to be 8, 10, 12, and 14 in a clinic session of six time units in this particular subsection. We consider two cases: increasing and decreasing show-up probabilities. In the increasing case, the show-up probability increases from 0.1 to 0.9 while the probability drops from 0.9 to 0.1 in the decreasing case. Both cases are solved using the PGP method under exponential, uniform and log-normal service distribution with mean 1 and standard deviation 1, whereas the robust method solves the problem using the same first and second moments. We then run out-of-sample simulations to estimate the expected total costs of the obtained scheduling solutions, assuming that the service times follow a log-normal distribution.

Table 2 shows the expected total costs and performance improvement achieved by using the PGP method. Interestingly, significant performance improvements are observed in the decreasing case (33.89% to 49.84%)

Service Distribution	Show-up Function		(Out-of-Sample) Log-Normal			
			8	10	12	14
(In Sample) Exponential	Increasing	PGP	4.1441	3.9632	3.8445	3.7389
		Robust	4.2361	4.1358	3.9954	3.9490
		Improvement	**2.17%**	**4.17%**	**3.78%**	**5.32%**
	Decreasing	PGP	3.3721	3.7107	3.8151	4.1296
		Robust	3.2404	3.5894	5.8215	8.1927
		Improvement	**-4.06%**	**-3.38%**	**34.47%**	**49.59%**
(In Sample) Log-Normal	Increasing	PGP	4.1153	3.951	3.8279	3.7352
		Robust	4.2361	4.1358	3.9954	3.9490
		Improvement	**2.87%**	**4.47%**	**4.19%**	**5.41%**
	Decreasing	PGP	3.1578	3.4637	3.7918	4.1099
		Robust	3.2404	3.5894	5.8215	8.1927
		Improvement	**2.57%**	**3.5%**	**34.87%**	**49.84%**
(In Sample) Uniform	Increasing	PGP	4.4191	4.2138	4.0138	3.9612
		Robust	4.2361	4.1358	3.9954	3.9490
		Improvement	**-4.14%**	**-1.85%**	**-0.46%**	**0.31%**
	Decreasing	PGP	3.1629	3.5393	3.8487	4.1553
		Robust	3.2404	3.5894	5.8215	8.1927
		Improvement	**2.39%**	**1.40%**	**33.89%**	**49.28%**

Table 2: Total Expected Costs under Scheduling Solutions Generated by Robust Optimization and Gradient Projection Methods: Exponential, uniform, log-normal Distribution with mean 1 and standard deviation 1

when the number of patients is larger; for smaller number of patients, there is a slight decrease (-4.14% to -1.85%) for the exponential and uniform service distributions. The reason is that the robust approach solves for the best solution under the worst-case probability distribution within the ambiguity set, so that it tends to be conservative under some cases. For example, when the system is very crowded, the robust solution assigns almost all patients to time slot where the show-up rate is the lowest. In the increasing case, performance improvement is moderate (2.17% to 5.41%) but consistent for exponential and log-normal distributions. In the case of uniform distributions, the DRO model appears to do a better job since, by design, it protects again misspecification of the probability distribution; indeed, since the out-of-sample distribution is log-normal, when the in-sample distribution is assumed to be uniform there is a high degree of misspecification. Still, it is interesting to notice that in the decreasing case the protection against misspecification does not compensate for the excess of conservatism, and consequently the solutions produced by our approach beat the DRO solutions by a large margin for the cases with higher number of patients.

6.3 Scheduling Patterns under Different Show-up Functions

We proceed to solve for the scheduling solutions under the six different show-up functions and three different service time distributions using our method. Figure 4 shows the arrival times of the 12 patients in a clinical session, together with the corresponding show-up rate at the time of arrival. Note that the results shown are solved under exponential service distribution and similar results are observed under two other distributions and are reported in Appendix B. As we can see from Figure 4, a front loading pattern is observed when

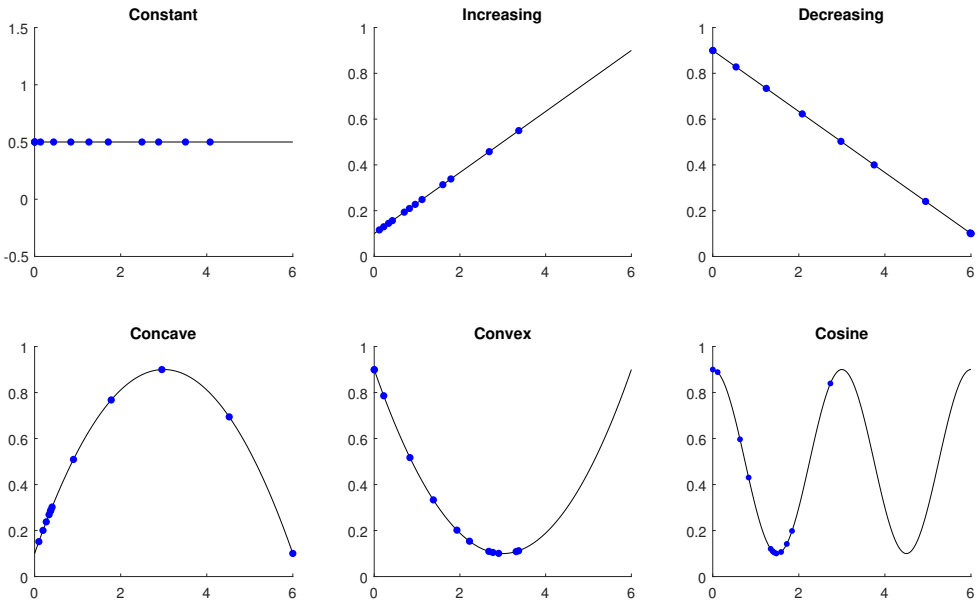


Figure 4: Scheduling Decisions under Different Show-up Patterns and Exponential Service Distribution (Y-axis: show-up probability; X-axis: time of arrivals)

there is a non-decreasing show-up function at the beginning, including Constant, Increasing, and Quadratic Concave. Interestingly, a load of patients also appear at the beginning of the first increasing section 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 to reduce physician overtime. It is worthwhile noting that few patients are scheduled to arrive in the latter half session, which is probably due to 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-load 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. Also, 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 could probably be explained by two reasons: First, a higher show-up rate at the beginning of the session may probably results in patient waiting queue, the burden of which can be mitigated by spreading patients across the session; second, assigning patients to lower show-up time (towards the end of the session) can reduce the workload of the system.

Next, we explore the patterns of the inter-arrival times under different show-up functions. Figure 5 shows patient inter-arrivals under exponential service time and results under two other distributions are reported in Appendix B. Under each show-up pattern, 13 points are presented: The first one shows the arrival time

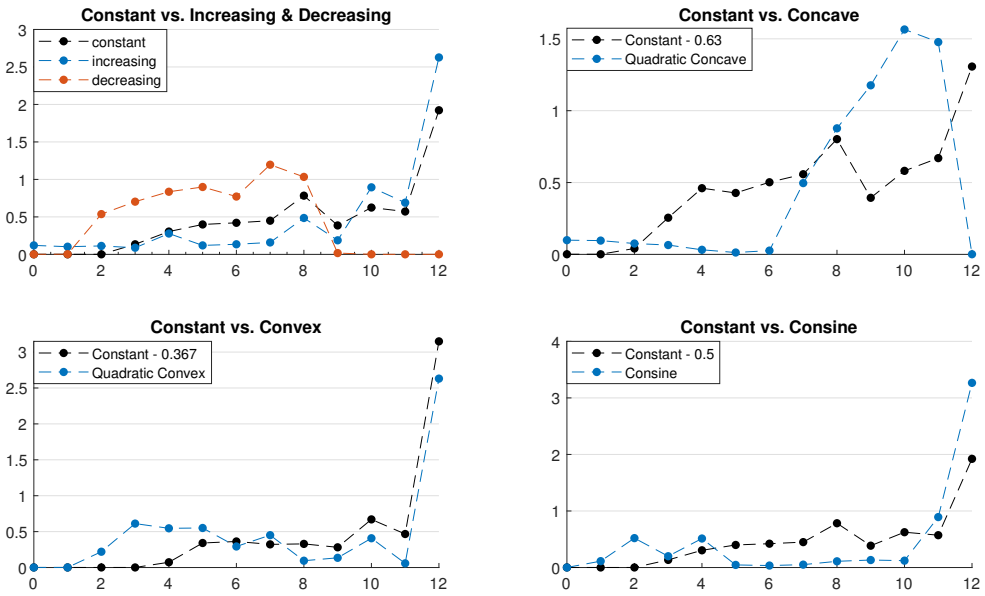


Figure 5: Inter-Arrival Time of Patients under Different Show-up Patterns and Exponential Service Distribution (Y-axis: inter-arrival time ; X-axis: patients that arrives in a sequence).

of the first patient and subsequent point i ($i > 1$) indicates the inter-arrival time between patient i and $i+1$ and the last point means the service time allocated to the last patient.

The dome-shaped pattern in the inter-arrival times has been well reported in the previous literature on standard appointment scheduling model where no-show is 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 5 depicts the inter-arrival times for the Constant, Increasing and Decreasing cases. While the inter-arrivals times show in general an increasing trend in both the constant and the increasing case, a dome-shape exists in the decreasing case. Moreover, the decreasing case allocates much larger inter-arrivals times to patient arriving in the middle of the session, compared to the two other cases.

Compared to the constant case, the schedule under the concave case shows a delayed dome-shape. It assigns more patients to come at the beginning of the session and then spread the patients across the session. The schedule under the convex case, however, allocates less patients to the beginning of the session, at which time the show-up rate is high. Note that although the schedules under 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 inter-arrival times for the first few patients are nigher and then almost zeros except the last patients, as compared to the constant case. Besides, a big gap is left from the last arrival to the end of the session.

The Value of Incorporating Patient Schedule-Dependent No-Show Behavior In the previous subsections, we have validated the performance of the algorithm and examined the scheduling patterns under different show-up functions. Finally we investigate the value of incorporating the patient schedule-dependent no-show behavior, i.e., the performance improvement under variant show-up probabilities compared to constant show-up rates. To do so, we use out-of-sample Monte-Carlo simulation to estimate the total expected system costs under the solutions generated from our methods. Table 3 presents the expected total cost of the obtained scheduled for each case. This value was approximated with one million simulations for each schedule. Additionally, Table 3 presents the expected total cost when applying the solution given by the constant show-up probability to each case.

Service Distribution	Show-up Pattern	Total Expected Cost		**Cost Reduction**
		Static Schedule	Time-Dependent	
Exponential	Increasing	4.0412	3.87	4.24%
	Decreasing	6.9319	3.9320	43.28%
	Concave	5.2676	3.7858	28.13%
	Convex	5.0302	3.2623	35.15%
	Cosine	5.4027	3.2211	40.38%
Log-Normal	Increasing	4.0522	3.8299	5.49%
	Decreasing	6.2572	3.7883	39.42%
	Concave	4.8141	3.7419	22.39%
	Convex	4.6497	3.2400	30.32%
	Cosine	4.8770	3.2405	33.56%
Uniform	Increasing	2.9967	2.6086	12.82%
	Decreasing	4.5529	2.1409	52.98%
	Concave	4.1829	1.9239	54.01%
	Convex	2.2689	1.9718	13.09%
	Cosine	3.3093	1.6796	49.27%

Table 3: Value of Time-Dependent Schedule

These results show an average cost reduction of 30.97% when assumption of constant show-up probability is broken, having the highest percentage of reduction (54.01%) under the concave case and the lowest under the increasing case with a reduction of 4.24%.

7 Conclusions

This paper studies the appointment scheduling problem with random service time and patient time-dependent no-show behavior. This problem is difficult due to its non-convex nature and little work has done on this topic. Limited previous literature either rely on heuristics or approximations to solve this problem and have obvious limitations. In this paper, we address this problem using 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 PGP method is built upon two existing methods (STRONG and Cauchy line search)

originally designed for unconstrained and deterministic optimization problems. When adapting the Cauchy line search, we made an important modification that significantly simplifies the computation. 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 cases show a considerable difference between the solutions of each particular no-show behavior scenario analyzed. The effect of assuming a constant show-up probability was studied, obtaining significant cost reductions when breaking this assumption. It is important to note that since the objective function is non-linear and non-convex, a multi-start strategy was used with the purpose of finding a *near local minimum*. Due to the complexity of the problem, it is not possible to ensure a global optimality for the solutions. Future work is needed on including this *endogenous uncertainty* to different issues on patients scheduling like *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. Heterogeneous patients characteristics is an important extension. Likewise, other sources of uncertainty as patients unpunctuality and cancellations can be analyzed. We consider an off-line scheduling problem, assuming that the system already knows which patients are coming and do not consider patient non-punctual arrivals and walk-in patients. Instead, we focus on providing a general guideline to appointment schedule 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, less patients would show up, leading to higher idle cost, which is equivalent to a revenue loss.

References

Ahmadi-Javid, J. Z., A., & Klassen, K. J. (2017). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, *258(1)*(4), 3-34.

Ahmed, S. (2000). *Strategic planning under uncertainty: Stochastic integer programming approaches* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Banks, J. (1998). *Handbook of simulation: principles, methodology, advances, applications, and practice*. John Wiley & Sons.

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

Chang, K.-H., Hong, L. J., & Wan, H. (2013). Stochastic trust-region response-surface method (strong)-a new response-surface framework for simulation optimization. *INFORMS Journal on Computing*, *25*(2), 230–243.

Conn, A. R., Gould, N. I., & Toint, P. L. (2000). *Trust region methods* (Vol. 1). Siam.

Daggy, J., Lawley, M., Willis, D., Thayer, D., Suelzer, C., DeLaurentis, P.-C., ... Sands, L. (2010). Using no-show modeling to improve clinic performance. *Health Informatics Journal*, *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*, 40–47.

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

Dupacová, J. (2006). Optimization under exogenous and endogenous uncertainty. *University of West Bohemia in Pilsen*.

Erdogan, S. A., & Denton, B. (2013). Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing*, *25*(1), 116–132.

Erdogan, S. A., Gose, A., & Denton, B. T. (2015). On-line appointment sequencing and scheduling. *IIE Transactions*(just-accepted), 00–00.

Fu, M., Lele, S., & Vossen, T. W. M. (2009). Conditional monte carlo gradient estimation in economic design of control limits. *Production and Operations Management*, *18*(1), 60–77.

Fu, M., et al. (2015). *Handbook of simulation optimization* (Vol. 216). Springer.

Fu, M. C. (2015). Stochastic gradient estimation. In M. C. Fu (Ed.), *Handbook of simulation optimization* (pp. 105–147). Springer.

Fu, M. C., & Hu, J.-Q. (1997). *Conditional monte carlo: Gradient estimation and optimization applications*. Springer.

Gallucci, G., Swartz, W., & Hackerman, F. (2005). Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services*.

Glasserman, P. (1991). *Gradient estimation via perturbation analysis*. Norwell, MA: Kluwer Academic Publishers.

Goel, V., & Grossmann, I. E. (2004). A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Computers and Chemical Engineering*, *28 (8)*, 1409–1429.

Goel, V., & Grossmann, I. E. (2006). A class of stochastic programs with decision dependent uncertainty. *Mathematical programming*, *108*(2-3), 355–394.

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, *40 (9)*(1), 800-809.

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

Hellemo, L., Barton, P. I., & Tomasgard, A. (2018). Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, *15*(3-4), 369–395.

Homem-de-Mello, T., Shapiro, A., & Spearman, M. L. (1999). Finding optimal material release times using simulation based optimization. *Manage Sci*, *45*, 86–102.

Homem-de Mello, T. (1998). *Simulation-based methods for stochastic optimization* (Unpublished doctoral dissertation). Georgia Institute of Technology.

Jiang, R., Shen, S., & Zhang, Y. (2017). Distributionally robust appointment scheduling with random no-shows and service durations. *Operations Research*, *65*, 1429–1731.

Jonsbråten, T. W., Wets, R. J., & Woodruff, D. L. (1998). A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, *82*, 83–106.

Kong, Q., Li, S., Liu, N., Teo, C.-P., & Yan, Z. (2019). Appointment scheduling under schedule-dependent patient no-show behavior. *Management Science*.

Lacy, N. L., Paulman, A., Reuter, M. D., & Lovejoy, B. (2004). Why we don't come: patient perceptions on no-shows. *The Annals of Family Medicine*, *2*(6), 541–545.

LaGanga, L. R. (2011). Lean service operations: reflections and new directions for capacity expansion in outpatient clinics. *Journal of Operations Management*, *29*(5), 422–433.

LaGanga, L. R., & Lawrence, S. R. (2012). Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management*, *21*(5), 874–888.

Moore, C. G., Wilson-Witherspoon, P., & Probst, J. C. (2001). Time and money: effects of no-shows at a family practice residency clinic. *FAMILY MEDICINE-KANSAS CITY-*, *33*(7), 522–527.

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

Parizi, M. S., & Ghate, A. (2016). Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. *Computers & Operations Research*, *67*, 90–101.

Pflug, G. (1990). On-line optimization of simulated markovian processes. *Mathematics of Operations Research*, *15*(3), 381—395.

Robinson, L. W., & Chen, R. R. (2003). Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *IIE Transactions*, *35*(3), 295–307.

Rubinstein, R. Y., & Shapiro, A. (1993). *Discrete event systems: Sensitivity analysis and stochastic optimization by the score function method*. Chichester, England: John Wiley & Sons.

Samorani, M., & LaGanga, L. R. (2015). Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research*, *240*(1), 245–257.

Shapiro, A., Dentcheva, D., & Ruszczynski, A. (2009). *Lectures on stochastic programming: Modeling and theory*. SIAM.

Truong, V.-A. (2015). Optimal advance scheduling. *Management Science*.

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

Appendices

A Parameters in the Algorithms

The parameters used in the algorithms are presented in Table 4. The first eight parameters are required for the STRONG algorithm and the last four are required for the PGP method. A detailed explanation of the definition, purpose and conditions of these parameters can be found in Chang et al. (2013) and Conn et al. (2000).

STRONG Algorithm		**PGP Method**	
Parameters	Value	Parameters	Value
η_0	0.01	κ_{epp}	0.125
η_1	0.30	κ_{lbs}	0.95
γ_1	0.95	κ_{ubs}	0.55
γ_2	1.10	κ_{frd}	0.75
Δ^0	2.15		
$\tilde{\Delta}$	1.15		
n_0	5,000		
m_0	15,000		

Table 4: Algorithm Parameters

B Scheduling Solutions under Log-normal and Uniform Distributions

We present in Figures 7-10 the scheduling solutions (both arrival and inter-arrival times) under lognormal and uniform distributions under the six show-up functions: constant, increasing, decreasing, quadratic concave, quadratic convex, and cosine. We can see that the scheduling solutions have similar patterns as those under exponential service time distribution.

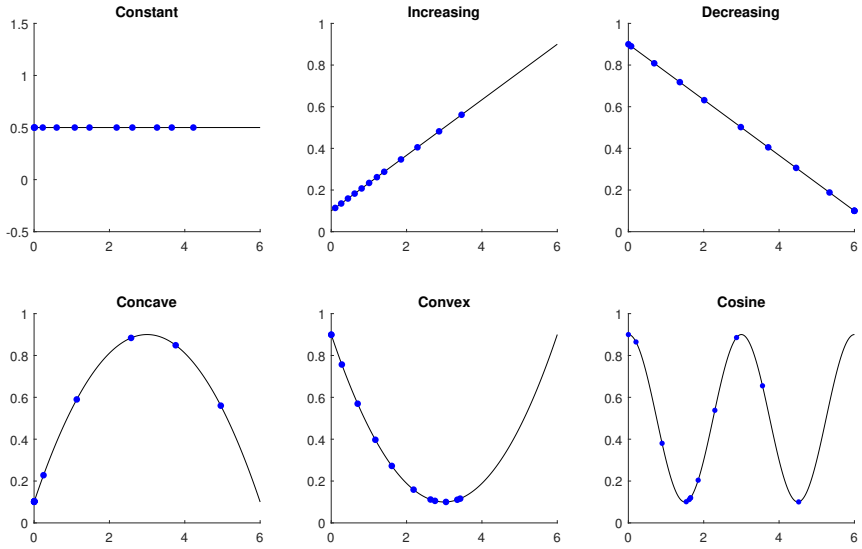


Figure 6: Scheduling Decisions under Different Show-up Patterns and Log-normal Service Distribution (Y-axis: show-up probability; X-axis: time of arrivals)

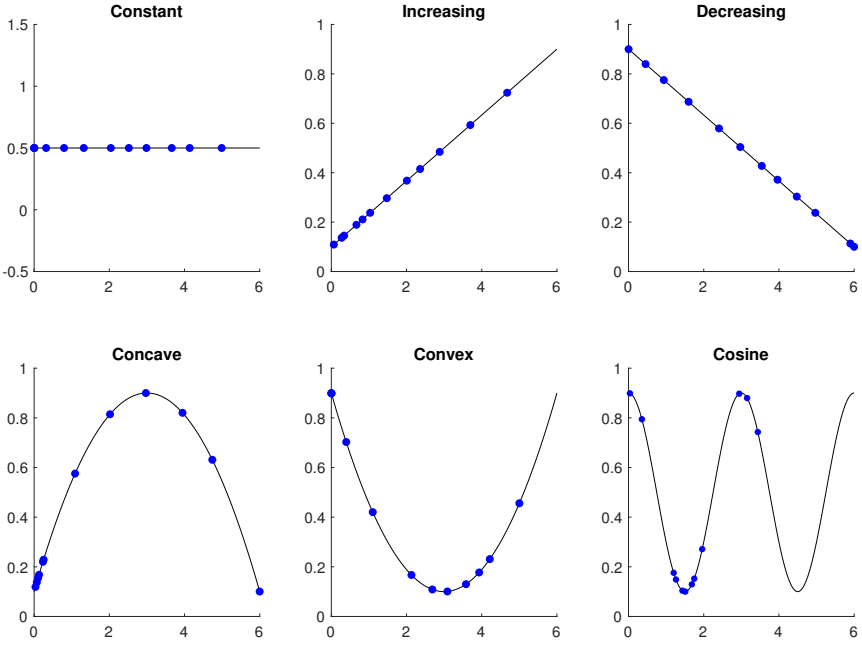


Figure 7: Scheduling Decisions under Different Show-up Patterns and Uniform Service Distribution (Y-axis: show-up probability; X-axis: time of arrivals)

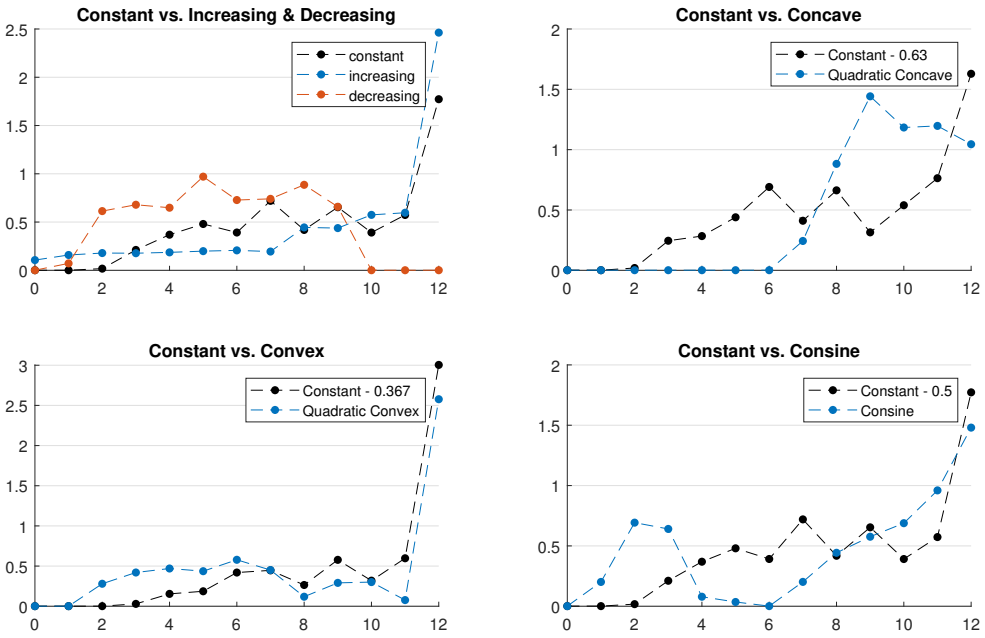


Figure 8: Inter-Arrival Time of Patients under Different Show-up Patterns and Log-Normal Service Distribution (Y-axis: inter-arrival time ; X-axis: patients that arrives in a sequence).

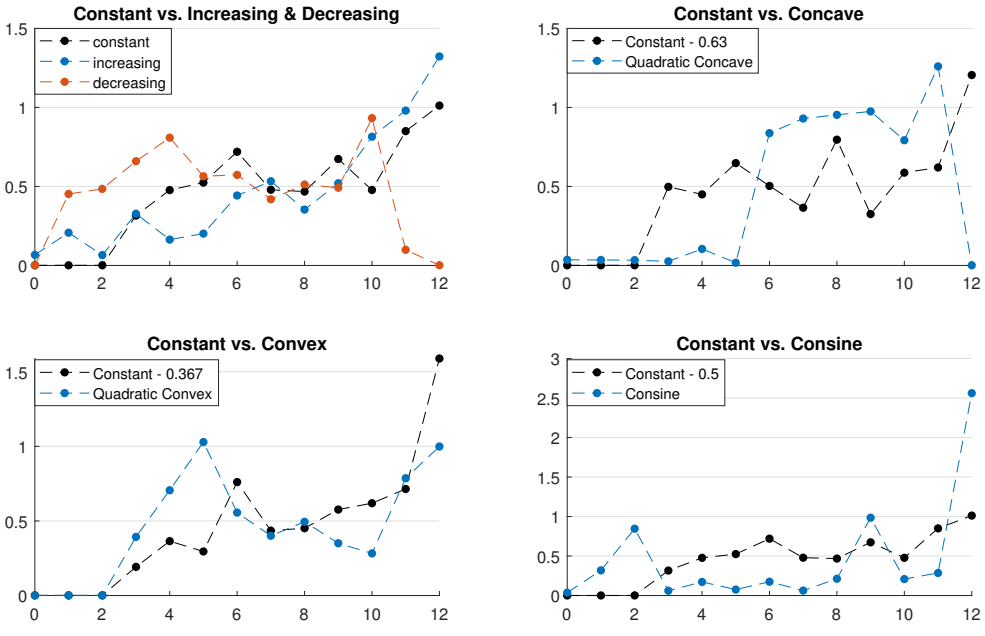


Figure 9: Inter-Arrival Time of Patients under Different Show-up Patterns and Uniform Service Distribution (Y-axis: inter-arrival time ; X-axis: patients that arrives in a sequence).