MANAGEMENT SCIENCE

Appointment Scheduling and the Effects of Customer Congestion on Service

Journal:	*Management Science*
Manuscript ID	Draft
Manuscript Type:	Operations Management
Keywords:	appointment scheduling, server behavior, optimization

Abstract: This paper addresses an appointment scheduling problem in which the server responds to congestion of the service system. Using waiting time as a proxy for how far behind schedule the server is running, we characterize the congestion induced behavior of the server as a function of customer waiting time. Decision variables are the scheduled arrival times for a specific sequence of customers. The objective of our model is to minimize a weighted cost incurred for customer waiting time, server overtime and server speedup in response to congestion. We provide alternative formulations of this problem as a simulation optimization (SO) model and a stochastic integer programming (SIP) model, respectively. We show the SIP model can solve moderate sized instances exactly under certain assumptions about server response to congestion. We further show that the SO model achieves near optimal solutions for moderate sized problems while also being able to scale up to much larger problem instances. We present theoretical results for both models and we characterize the solutions of special cases of the problem. We carry out a series of experiments to illustrate the characteristics of the optimal schedules and to measure the importance of accounting for server behavior when scheduling appointments. Moreover, we illustrate the importance of congestion effects using a case study for an outpatient clinic at a large medical center. Finally, we summarize the most important managerial insights obtained from this study.

Submitted to *Management Science*
manuscript (Please, provide the manuscript number!)

Appointment Scheduling and the Effects of Customer Congestion on Service

(Authors' names blinded for peer review)

This paper addresses an appointment scheduling problem in which the server responds to congestion of the service system. Using waiting time as a proxy for how far behind schedule the server is running, we characterize the congestion induced behavior of the server as a function of customer waiting time. Decision variables are the scheduled arrival times for a specific sequence of customers. The objective of our model is to minimize a weighted cost incurred for customer waiting time, server overtime and server speedup in response to congestion. We provide alternative formulations of this problem as a simulation optimization (SO) model and a stochastic integer programming (SIP) model, respectively. We show the SIP model can solve moderate sized instances exactly under certain assumptions about server response to congestion. We further show that the SO model achieves near optimal solutions for moderate sized problems while also being able to scale up to much larger problem instances. We present theoretical results for both models and we characterize the solutions of special cases of the problem. We carry out a series of experiments to illustrate the characteristics of the optimal schedules and to measure the importance of accounting for server behavior when scheduling appointments. Moreover, we illustrate the importance of congestion effects using a case study for an outpatient clinic at a large medical center. Finally, we summarize the most important managerial insights obtained from this study.

Subject classifications: appointment scheduling; server behavior; optimization.
Area of review: operations management.

1. Introduction

Appointment systems are intended to reduce service variability by lowering customer waiting and increasing server utilization. In many cases, the system performance is highly affected by the service variability because longer than expected service time for a particular customer results in waiting of the next customer and possibly overtime, whereas shorter than expected service time results in idling of the server. Previous literature on appointment scheduling assumes exogenous service times that are independent of the state of the service system. However, a number of authors have noted that congestion, or crowding of the service system, may affect the service time as a result of

the server's behavior in response to these factors (see Rising et al. (1973), Deveugele et al. (2002) and Cayirli et al. (2008) for examples). In many cases, when the server perceives congestion, the resulting service time is observed to decrease so the server can catch up. In this article we refer to this response by the server as *congestion behavior*.

Congestion has many aspects such as the number of waiting customers or crowding of a waiting area with limited space. In this study on appointment scheduling, we treat customer waiting time as a proxy for congestion since it represents the time by which the daily schedule has been delayed and it is correlated with crowding. The anticipation of congestion behavior complicates the optimization of appointment scheduling because it makes the service time endogenous due to dependency on the waiting time. Thus, appointment time decisions should incorporate service times that are affected by waiting times.

A specific example of scheduling outpatient appointments at Mayo Clinic in Rochester, MN provided additional motivation for our study. We analyzed over 14,037 observations of electronic time-stamp data of patient events during their appointment including scheduled appointment time, when the patient checked in, and the duration of time the patient was with the provider. We limited our analysis to full-time providers who had at least 8 appointments and at least 100 observations in total. We aggregated the data into 11 groups by the combination of provider and appointment type, and further aggregated data by the patient waiting time in 20-minute intervals. We observed that the mean service time was negatively correlated with the waiting time among 9 of the 11 groups. This provides some additional empirical evidence in support of waiting time dependent congestion behavior, in addition to that already reported in the literature, which we discuss in Section 6.

In this article, we extend the previously proposed appointment scheduling problem by Denton and Gupta (2003) to account for server congestion behavior. The resulting service time depends on the amount of waiting time by the customer, which we refer to as *congestion-dependent service time* or service time for short. Thus, the resulting optimization problem is an endogenous stochastic program in which the optimal appointment times depend on the customer waiting time. We formulate this problem as a simulation optimization (SO) model and a stochastic integer programming (SIP) model, respectively. The SIP model can solve moderate sized instances exactly under certain assumptions about server response to congestion, and it can serve as a reference to assess the accuracy of the SO model. The SO model, on the other hand, achieves near optimal solutions for moderate sized problems, while also being able to scale up to much larger problem instances.

We use the models to carry out a series of experiments, including a case study based on an outpatient clinic example at Mayo Clinic, to provide results to answer the following two research questions:

- Is it important to anticipate the server's congestion behavior when optimizing the design of appointment schedules?
- What is the nature of the optimal schedule when congestion effects are present, and how can results from the model inform appointment scheduling in practice?

The remainder of this article is organized as follows: Section 2 reviews the related literature. In Section 3, the appointment scheduling problem with congestion behavior is described and formulated as an SO model. Section 4 reformulates the problem as a two-stage SIP model when the congestion behavior is a piecewise linear function. Section 5 analyzes the characteristic of Karush-Kuhn-Tucker (KKT) solutions on special cases of the problem, and Section 6 presents numerical results. Finally, Section 7 concludes the paper and summarizes some important managerial insights.

2. Literature Review

Appointment scheduling has been extensively studied in the literature (see Cayirli and Veral (2003) and Gupta and Denton (2008) for reviews). Prior optimization models by Weiss (1990), Denton and Gupta (2003) and Robinson and Chen (2003) demonstrated a "dome shape" for customer interarrival times when scheduling identical customers. The rationale is that customers in the middle of the day warrant additional planned service time since they are more likely to experience waiting time, and hence they have the potential to disrupt the remainder of the schedule when waiting occurs. Recently, customer behaviors like no-shows and dynamic arrivals have been addressed in several papers. For instance, analytical models were proposed by Muthuraman and Lawley (2008) and Chakraborty et al. (2010) to sequentially schedule customers who may or may not show up. Other studies used sample approximation to characterize customer no-show, i.e., the distribution of service time was adjusted to incorporate no-show probabilities (Erdogan and Denton (2013), Begen and Queyranne (2011)). An example by Erdogan and Denton (2013) considered a dynamic model that sets appointment times dynamically as customers request appointments (one at a time). The authors reported that optimal interarrival times were also dome shaped; although, the dome shape was less pronounced due to the presence of no-shows. No-shows have also been addressed in appointment scheduling using heuristics. Hassin and Mendel (2008) reported that the optimal equal spacing interarrivals were, for a wide range of relative cost values, almost linearly increasing with the showing-up probability. More recently, Cayirli et al. (2012) proposed a procedure that sets appointment times based on mean and standard deviation of the service time to account for no-shows and walk-in customers.

Customer no-show rates were considered to vary over the time of day by Kong et al. (2016). The authors proposed a distribution-free appointment scheduling model with schedule-dependent no-show rates. They formulated a conic program, and solved the problem iteratively using a descent

algorithm guided by dual prices. They reported that anticipating the schedule-dependent no-show rate can reduce the worse case expected cost by 30%-60%.

Congestion behavior has been observed in (but is not limited to) healthcare systems such as outpatient clinics and emergency departments (see Rising et al. (1973), Deveugele et al. (2002) and Batt and Terwiesch (2012) for examples). There exist several queuing models that have incorporated congestion behavior. Early work by Harris (1967) considered an M/G/1 queuing system and treated each customer's service time as a stochastic process classified by the total number of customers in the queue. Posner (1973) considered an M/M/1 queuing model in which the service time depends on the amount of waiting time by the customer. Recently, Chan et al. (2014) considered a queuing network in which the service time and the probability of customer's returning to service were modeled as step functions depending on the number of waiting customers in the queue.

While the intent of the above articles to model behavioral aspects of service systems, the topic of this article differs in several ways. The most significant differences are as follows:

- In contrast to the literature on no-shows, which focuses on customer behavior, we focus on server behavior.
- In contrast to the queueing literature, which seeks to describe service systems that experience a congestion effect, we focus on optimization of scheduled systems.

We present, to our knowledge, the first appointment scheduling optimization models that account for server congestion behavior, including a stochastic programming model and a simulation optimization model. Second, we show the cost function of these models has valuable properties including continuous differentiability, monotonicity and convexity under certain assumptions about server response to congestion. Third, we compare a series of numerical results from these models to establish the importance of anticipating congestion in practice and to characterize the resulting schedules.

3. Simulation Optimization Model for General Congestion Behavior

In this section, we describe the SO model for appointment scheduling in the presence of congestion behavior. We begin with this model because it is the more general of the two models we propose. In Section 4 we describe the SIP formulation, which assumes a piecewise linear structure to the congestion response.

We consider n pre-sequenced customers, possibly with different waiting costs, denoted by a vector $\alpha \in \mathbb{R}^n$ where element α_i denotes the waiting cost per unit time for customer i. The server has a nominal session length, denoted by d, which is extensible with overtime cost per unit time of 1. Customer waiting time and server overtime costs are commonly considered in the literature on appointment scheduling. In this paper, we also consider an additional cost for service speedup

since speedup may, in some environments, have undesirable consequences such as utilization of additional resources, increased risk of poor service, or server fatigue. We measure service speedup as the amount of service time reduction compared to the *nominal service time* in the absence of customer waiting, and we penalize the service time reduction by a vector of cost penalties, β, where element β_i denotes the speedup cost per unit time for customer i. We assume each customer arrives punctually at the scheduled arrival time.

We consider the following decision variables in our models:

a: vector of scheduled arrival times for n customers (note that $a_1 = 0$ by assumption),

$w(\omega)$: vector of waiting times for n customers in scenario ω with element $w_i(\omega)$ denoting the waiting time for customer i (note that $w_1(\omega) = 0$ by assumption),

$l(\omega)$: overtime for the server in scenario ω,

$\delta(\omega)$: vector of service time reduction for n customers in scenario ω with element $\delta_i(\omega)$ denoting the service time reduction for customer i (note that $\delta_1(\omega) = 0$ by assumption).

Note that a is a vector of first stage decision variables made before scenarios are observed and $w(\omega)$, $l(\omega)$ and $\delta(\omega)$ are second stage decision variables made after scenarios are observed.

The service time for each customer depends on the waiting time due to congestion behavior. We let $\xi_i(\omega)$ denote the nominal service time for customer i in scenario ω, and we define the congestion-dependent service time for customer i in scenario ω as $z_i(\omega, w_i(\omega))$ where $z_i(\omega, w_i(\omega))$ is assumed to be nonincreasing in $w_i(\omega)$ and $z_i(\omega, 0) = \xi_i(\omega)$. Moreover, we make the following assumptions on $z_i(\omega, w_i(\omega))$.

ASSUMPTION 1. *$z_i(\omega, w_i(\omega))$ is Lipschitz-continuous in $w_i(\omega)$; specifically, for any $w_i(\omega)$ and $\widehat{w}_i(\omega)$, there exists a constant $L(\omega)$ such that:*

$$|z_i(\omega, w_i(\omega)) - z_i(\omega, \widehat{w}_i(\omega))| \leq L(\omega)|w_i(\omega) - \widehat{w}_i(\omega)|.$$

ASSUMPTION 2. *$z_i(\omega, w_i(\omega))$ is piecewise differentiable in $w_i(\omega)$ with a finite set of nondifferentiable points.*

As we will show, Assumptions 1-2 are necessary to guarantee continuity and differentiability of the SO model, respectively. They state mild requirements of $z_i(\omega, w_i(\omega))$ as a function of $w_i(\omega)$. There exist many examples that satisfy these assumptions, such as piecewise linear functions, and the logit function which we will use later in our numerical experiments.

Under Assumptions 1-2, $z_i(\omega, w_i(\omega))$ is differentiable in $w_i(\omega)$ everywhere except at a finite set of saddle points of measure 0. The third and final assumption about service times is as follows:

ASSUMPTION 3. *$z_i(\omega, w_i(\omega))$ is a continuous random variable with known distribution of finite density.*

Assumption 3 is necessary to guarantee that the objective function of the SO model is almost surely differentiable. Although $z_i(\omega, w_i(\omega))$ may have positive probability at some values (e.g., when customer i does not show, $z_i(\omega, w_i(\omega)) = 0$), it can be easily adjusted to make Assumption 3 true. Specifically, $z_i(\omega, w_i(\omega))$ can be smoothed by perturbation, i.e., when customer i did not show, $z_i(\omega, w_i(\omega))$ is set to a sufficiently small and positive random variable. As a result, the distribution density of $z_i(\omega, w_i(\omega))$ becomes finite.

3.1. Simulation Optimization Model

Our objective is to minimize the total weighted cost of customer waiting time, server overtime and service speedup. Waiting time is incurred if a customer arrives before the previous service completes, overtime is incurred if the server finishes services later than the session length, and service speedup is incurred if the service time is reduced from its nominal value due to congestion behavior. The congestion anticipated appointment scheduling problem can be formulated as the following SO model:

$$\min_{a \in \mathbb{R}_+^n} g(a) = \mathbb{E}[f(a, \omega)], \tag{1}$$

where $g(a)$ is the expected cost for a fixed schedule, a, and $f(a, \omega)$ is the sample path cost for schedule a and scenario ω, which is determined as follows:

$$f(a, \omega) = \sum_{i=2}^{n} \alpha_i w_i(\omega) + \sum_{i=2}^{n} \beta_i \delta_i(\omega) + l(\omega), \ \forall \omega, \tag{2a}$$

$$\delta_i(\omega) = \xi_i(\omega) - z_i(\omega, w_i(\omega)), \quad 2 \le \forall i \le n, \ \forall \omega, \tag{2b}$$

$$w_i(\omega) = \left(a_{i-1} + w_{i-1}(\omega) + z_{i-1}(\omega, w_{i-1}(\omega)) - a_i\right)^+, \quad 2 \le \forall i \le n, \ \forall \omega, \tag{2c}$$

$$l(\omega) = \left(a_n + w_n(\omega) + z_n(\omega, w_n(\omega)) - d\right)^+, \ \forall \omega, \tag{2d}$$

where $(x)^+ = \max(x, 0)$. Equation (2a) determines the sample path cost incurred by customer waiting, service speedup and server overtime. Equation (2b) determines the amount of service time reduction. Equations (2c) and (2d) determine customer waiting time and server overtime, respectively.

3.2. Continuity and Differentiability

The SO model defined by (1) has the following properties (all proofs of lemmas and theorems are in the Appendix):

LEMMA 1. *Under Assumption 1, $f(a, \omega)$ is Lipschitz-continuous in a.*

LEMMA 2. *Under Assumptions 1-3, $f(a, \omega)$ is almost surely differentiable in a.*

Using Lemmas 1 and 2, we can prove the following theorem which is necessary for the SO approach we propose:

THEOREM 1. *Under Assumptions 1-3, $g(a)$ is continuously differentiable in a.*

Using Theorem 1, we can derive the following lemma.

LEMMA 3. *Under Assumptions 1-3, the gradient of $\nabla g(a)$ exists for all a, and $\nabla g(a) = E[\nabla f(a,\omega)]$ where $\nabla f(a,\omega)$ is the gradient of $f(a,\omega)$ with respect to a.*

Lemma 3 shows that $\nabla g(a)$ can be estimated based on the mean of $\nabla f(a,\omega)$ over a sampled set of scenarios. We let $f_i'(\omega) = \frac{\partial f(a,\omega)}{\partial a_i}$ and $g_i' = \frac{\partial g(a)}{\partial a_i}$ denote the derivatives of $f(a,\omega)$ and $g(a)$ with respect to a_i, respectively. We let $z_i'(\omega) = \frac{\partial z_i(\omega, w_i(\omega))}{\partial w_i(\omega)}$ denote the derivative of $z_i(\omega, w_i(\omega))$ with respect to $w_i(\omega)$. Note that $z_i'(\omega) \le 0$ by assumption. The next theorem establishes how to compute the gradient of the sample path cost.

THEOREM 2. *The derivatives of the sample path cost (2a) can be expressed as follows:*

$$f_i'(\omega) = \begin{cases} -\alpha_i + z_i'(\omega)(\beta_i - \lambda_{i+1}), & if\ w_i(\omega) > 0, \\ \lambda_{i+1}, & if\ w_i(\omega) = 0, \end{cases} \tag{3}$$

where λ_{i+1} defines the incremental cost when the completion time of customer i is delayed by one unit time; specifically, it is shown as follows:

$$\lambda_i = \begin{cases} \alpha_i + (1 + z_i'(\omega))\lambda_{i+1} - z_i'(\omega)\beta_i, & if\ i \le n,\ w_i(\omega) > 0, \\ 0, & if\ i \le n,\ w_i(\omega) = 0, \\ I(l(\omega)), & if\ i = n+1, \end{cases} \tag{4}$$

where $I(x) = 1$ if $x > 0$ and $I(x) = 0$ if $x \le 0$.

Based on Theorem 2 the gradient of the sample path cost can be estimated efficiently for each scenario ω. Therefore, we can use a stochastic approximation method to solve the SO model (Kushner and Yin (2003)).

4. Stochastic Programming Model for Piecewise Linear Congestion Behavior

In this section, we show that when congestion behavior is represented by a piecewise linear function, the congestion anticipated appointment scheduling problem can be formulated as a two-stage SIP model based on the sample average approximation (Kleywegt et al. (2002)). We further reduce the SIP to a (convex) stochastic linear program under certain assumptions about server response to congestion.

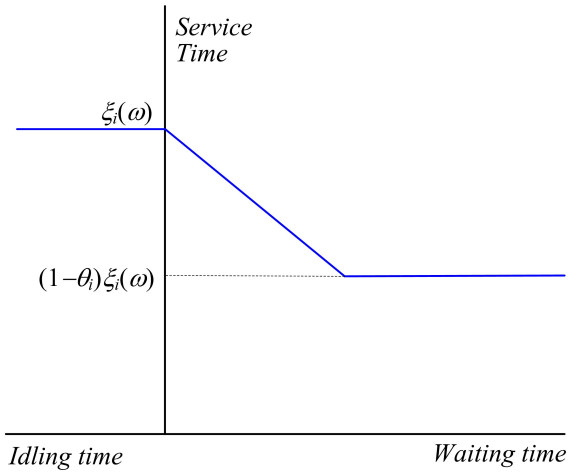

Figure 1 **Change in service time with respect to waiting time when congestion response is a piecewise linear function.**

4.1. Piecewise Linear Congestion Behavior

We assume there exists a maximum proportion of service time reduction for each customer, denoted by θ_i. We let $\gamma_i(\omega)$ denote the service time reduction rate and $\gamma_i(\omega) = -z_i'(\omega)$ when $z_i(\omega, w_i(\omega))$ is decreasing. We characterize the *piecewise linear congestion behavior* as follows (illustrated by Figure 1): $z_i(\omega, w_i(\omega))$ is $\xi_i(\omega)$ when there is no waiting time, then, as $w_i(\omega)$ grows, $z_i(\omega, w_i(\omega))$ decreases linearly at the rate of $\gamma_i(\omega)$, until it reaches $\xi_i(\omega)(1-\theta_i)$. Formally, $z_i(\omega, w_i(\omega))$ can be expressed as follow:

$$z_i(\omega, w_i(\omega)) = \begin{cases} \xi_i(\omega) - \gamma_i(\omega) w_i(\omega), & if\ 0 \le w_i(\omega) \le \frac{\theta_i \xi_i(\omega)}{\gamma_i(\omega)}, \\ (1-\theta_i)\xi_i(\omega), & if\ w_i(\omega) > \frac{\theta_i \xi_i(\omega)}{\gamma_i(\omega)}. \end{cases} \tag{5}$$

4.2. Stochastic Programming Model

We introduce the following additional auxiliary decision variables to formulate the two-stage SIP model:

$Z(\omega)$: vector of service times for n customers where element $Z_i(\omega)$ denotes the service time for customer i in scenario ω (note that $Z_1(\omega) = \xi_1(\omega)$ by assumption),

$s(\omega)$: vector of idling times of the server where element $s_i(\omega)$ denotes the idling time prior to serving customer i in scenario ω (note that $s_1(\omega) = 0$ by assumption),

$v_i(\omega)$: binary decision variable with $v_i(\omega) = 1$ denoting $w_i(\omega) > 0$, and $v_i(\omega) = 0$ otherwise,

$u_i(\omega)$: binary decision variable with $u_i(\omega) = 1$ denoting $w_i(\omega) > \frac{\theta_i \xi_i(\omega)}{\gamma_i(\omega)}$, and $u_i(\omega) = 0$ otherwise.

Using these additional decision variables, the congestion anticipated appointment scheduling problem can be formulated as the following two-stage SIP model:

$$\min_{a \in \mathbb{R}^n} \mathbb{E}\left\{ \sum_{i=2}^{n} \alpha_i w_i(\omega) + \sum_{i=2}^{n} \beta_i \delta_i(\omega) + l(\omega) \right\} \tag{6a}$$

$$s.t.\quad \delta_i(\omega)+Z_i(\omega)=\xi_i(\omega),\quad 2\le\forall i\le n,\ \forall\omega, \tag{6b}$$

$$w_{i-1}(\omega)-w_i(\omega)+s_i(\omega)+Z_{i-1}(\omega)=a_i-a_{i-1},\quad 2\le\forall i\le n,\ \forall\omega, \tag{6c}$$

$$w_n(\omega)+Z_n(\omega)-l(\omega)\le d-a_n,\ \forall\omega, \tag{6d}$$

$$Z_i(\omega)+\gamma_i(\omega)w_i(\omega)\ge\xi_i(\omega),\quad 2\le\forall i\le n,\ \forall\omega, \tag{6e}$$

$$Z_i(\omega)\ge(1-\theta_i)\xi_i(\omega),\quad 2\le\forall i\le n,\ \forall\omega, \tag{6f}$$

$$Z_i(\omega)+\gamma_i(\omega)w_i(\omega)-\gamma_i(\omega)\bar{w}_iu_i(\omega)\le\xi_i(\omega),\quad 2\le\forall i\le n,\ \forall\omega, \tag{6g}$$

$$Z_i(\omega)+\theta_i\xi_i(\omega)u_i(\omega)\le\xi_i(\omega),\quad 2\le\forall i\le n,\ \forall\omega, \tag{6h}$$

$$\gamma_i(\omega)w_i(\omega)-\left(\gamma_i(\omega)\bar{w}_i-\theta_i\xi_i(\omega)\right)u_i(\omega)\le\theta_i\xi_i(\omega),\quad 2\le\forall i\le n,\ \forall\omega, \tag{6i}$$

$$\gamma_i(\omega)w_i(\omega)-\theta_i\xi_i(\omega)u_i(\omega)\ge0,\quad 2\le\forall i\le n,\ \forall\omega, \tag{6j}$$

$$w_i(\omega)-\bar{w}_iv_i(\omega)\le0,\quad 2\le\forall i\le n,\ \forall\omega, \tag{6k}$$

$$s_i(\omega)+\bar{s}v_i(\omega)\le\bar{s},\quad 2\le\forall i\le n,\ \forall\omega, \tag{6l}$$

$$a_i,Z_i(\omega),\delta_i(\omega),w_i(\omega),s_i(\omega),l(\omega)\ge0;v_i(\omega),u_i(\omega)\in\{0,1\},\quad 2\le\forall i\le n,\ \forall\omega, \tag{6m}$$

where $\bar{w}_i$ and $\bar{s}$ are upper bounds of waiting times and idling times, respectively. The objective (6a) is to minimize the total expected cost of customer waiting time, service speedup, and server overtime. Constraint (6b) determines the amount of service time reduction. Constraint (6c) determines waiting time and idle time. Constraint (6d) determines the overtime. Constraints (6e)-(6h) jointly determine the congestion-dependent service time. Constraints (6i)-(6j) jointly determine $u_i(\omega)$, and constraints (6k)-(6l) jointly determine $v_i(\omega)$. Constraint (6m) restricts the decision variables to be nonnegative or binary.

It is always feasible to set the upper bounds as sufficiently large numbers for waiting times and idling times in model (6); however, a tighter bound helps to strengthen the model, and thus reduces the computation time. Mancilla and Storer (2012) presented upper bounds for waiting times and idling times. Based on their formulations, we incorporate the congestion response of the server on service times, and derive $\bar{w}_i$ and $\bar{s}$ as follows:

$$\bar{w}_i\le\sum_{j=1}^{i-1}\left\{\max_\omega\xi_i(\omega)-\min_\omega\xi_i(\omega)\left(1-\theta_i\right)\right\},$$

$$\bar{s}\le\max_i\left\{\max_\omega\xi_i(\omega)-\min_\omega\xi_i(\omega)\left(1-\theta_i\right)\right\}.$$

The formulation (6) can be further strengthened by the following inequalities:

$$u_i(\omega) \le v_i(\omega), \quad 2 \le \forall i \le n, \ \forall \omega, \tag{7a}$$

$$s_i(\omega) + \bar{s} u_i(\omega) \le \bar{s}, \quad 2 \le \forall i \le n, \ \forall \omega. \tag{7b}$$

Inequality (7a) holds because $w_i(\omega) = 0$ leads to $w_i(\omega) \le \frac{\theta_i \xi_i(\omega)}{\gamma_i(\omega)}$. Similarly, inequality (7b) holds because $w_i(\omega) \ge \frac{\theta_i \xi_i(\omega)}{\gamma_i(\omega)}$ leads to $s_i(\omega) = 0$.

Solving model (6) is challenging because it has many binary decision variables in the second stage. These variables are associated to constraints (6g)-(6j) representing the congestion-dependent service time and constraints (6k)-(6l) representing the complementary slackness between $w_i(\omega)$ and $s_i(\omega)$. In the following subsections we establish conditions under which model (6) is more tractable.

4.3. Monotonicity Condition

We now show that the waiting time is monotonic with respect to schedule a under the following assumption:

ASSUMPTION 4. $z_i'(\omega) \ge -1$, $2 \le \forall i \le n$.

Assumption 4 can be interpreted to mean that the service time is reduced no faster than the growth in waiting time, i.e., the reduction in service time does not over compensate for congestion. Therefore, $z_i(\omega, w_i(\omega)) \ge \xi_i(\omega) - w_i(\omega)$. It also implies that, as the waiting time increases for customer i, the completion time for customer i does not decrease.

THEOREM 3. *Under Assumption 4, as a_i increases, the waiting time is nondecreasing for any remaining customer, $j (j > i)$, and the overtime is also nondecreasing.*

Theorem 3 also holds for the SO model when congestion behavior is a more general function considered in Section 3. It can be interpreted to mean that one customer's waiting time cannot be reduced by adding unnecessary waiting time for previous customers. Therefore, Assumption 4 provides an optimality condition that no customers have unnecessary waiting time.

Using Theorem 3, we can derive the next lemma that establishes an important property necessary to reduce the complexity of model (6).

LEMMA 4. *Under Assumption 4, the complementary slackness constraints (6k)-(6l) can be removed from model (6) because they hold automatically.*

Lemma 4 establishes conditions under which the binary variables associated with constraints (6k)-(6l) can be removed; as a result, model (6) becomes much easier to solve.

4.4. Convexity Condition

Under Assumption 4 and the following assumption, it can be shown that model (6) reduces to a (convex) two-stage stochastic linear program:

ASSUMPTION 5. *$\beta_i = 0, \forall i$, which implies there is no cost for service speedup.*

Using Lemma 4, we can prove the following theorem.

THEOREM 4. *Under Assumptions 4 and 5, the SIP model (6) reduces to the following two-stage convex stochastic program:*

$$\min_{a\in\mathbb{R}^n} \mathbb{E}\left\{ \sum_{i=2}^{n} \alpha_i w_i(\omega) + l(\omega) \,\middle|\, (6b)\text{-}(6f),\ a, Z(\omega), \delta(\omega), w(\omega), s(\omega), l(\omega) \geq 0,\ \forall \omega \right\}. \tag{8}$$

This important result establishes conditions under which the appointment scheduling optimization model can be solved very efficiently.

5. Properties of Optimal Appointment Times

In this section, we present some analytical results for special cases of the model when the number of customers is 2, 3 and n. These results help to give some insights into the anticipated structure of numerical results for appointment schedules we present in Section 6.

5.1. Case of Two Customers

We first consider the case of two customers under which there is a single decision for customer 2's arrival. From Theorem 2, the derivative of the sample path cost function is as follows:

$$f_2'(\omega) = \begin{cases} -\alpha_2 + z_2'(\omega)\left(\beta_2 - I(l(\omega))\right), & if\ w_2(\omega) > 0, \\ I(l(\omega)), & if\ w_2(\omega) = 0. \end{cases} \tag{9}$$

We let $\widehat{f}(a,\omega)$ denote the sample path cost when congestion behavior is not anticipated, and $\widehat{f}_i'(\omega)$ denote the derivative of $\widehat{f}(a,\omega)$ with respect to a_i. As $w_2(\omega)$ is independent of $z_2(\omega, w_2(\omega))$, we can rewrite the derivative (9) as follows:

$$f_2'(\omega) = \widehat{f}_2'(\omega) + z_2'(\omega) I(w_2(\omega))\Big(\beta_2 - I(l(\omega))\Big). \tag{10}$$

We let $\widehat{g}(a) = \mathbb{E}[\widehat{f}(a,\omega)]$ be the expected cost for solution a when congestion behavior is not anticipated. It is well-known that $\widehat{g}$ is convex in a, and hence it has a unique optimal solution, which is denoted by $\widehat{a}^*$. Using these properties we can prove the following proposition:

PROPOSITION 1. *When $n = 2$, there exists a KKT point, a^*, such that $a_2^* \leq \widehat{a}_2^*$ if $\beta_2 = 0$, and $a_2^* \geq \widehat{a}_2^*$ if $\beta_2 \geq 1$.*

Proposition 1 can be interpreted to mean that the model we propose that anticipates congestion behavior results in an earlier appointment time being optimal when there is no cost of the service speedup. Conversely, when reducing the service time is more costly than the server overtime, anticipating congestion behavior results in a later appointment time.

5.2. Case of Three Customers

When $n=3$, the decisions become a_2 and a_3, and the problem becomes much more complicated. Deriving the closed form expression for $f_i'(\omega)$ becomes challenging for a general case. Instead, we consider a special case when the service time for customer 3 cannot speedup, i.e., $\theta_3=0$. From Theorem 2, the derivative of the sample path cost function on a_2 is as follows:

$$f_2'(\omega)=\begin{cases}-\alpha_2+z_2'(\omega)\left(\beta_2-\alpha_3-I(l(\omega))\right), & if\ w_2(\omega)>0,\ w_3(\omega)>0,\\ -\alpha_2+z_2'(\omega)\beta_2, & if\ w_2(\omega)>0,\ w_3(\omega)=0,\\ \alpha_3+I(l(\omega)), & if\ w_2(\omega)=0,\ w_3(\omega)>0,\\ 0, & if\ w_2(\omega)=0,\ w_3(\omega)=0.\end{cases} \tag{11}$$

Since $w_2(\omega)$ is independent of $z_2(\omega,w_2(\omega))$ and $z_2(\omega,w_2(\omega))=\xi_2(\omega)$ if $w_2(\omega)=0$, we can rewrite (11) as follows:

$$f_2'(\omega)=\widehat{f}_2'(\omega)+I(w_2(\omega))z_2'(\omega)\Big(\beta_2-I(w_3(\omega))(\alpha_3+I(l(\omega)))\Big). \tag{12}$$

The derivative of the sample path cost function on a_3 is as follows:

$$f_3'(\omega)=\begin{cases}-\alpha_3, & if\ w_3(\omega)>0,\\ I(l(\omega)), & if\ w_3(\omega)=0.\end{cases} \tag{13}$$

Using these properties we can derive the following lemmas:

LEMMA 5. *When $n=3,\ \beta_2=0$ and $\theta_3=0$, $f_2'(\omega)$ is decreasing in a_3, and $f_3'(\omega)$ is decreasing in a_2 and increasing in a_3.*

LEMMA 6. *When $n=3,\ \beta_2=0$ and $\theta_3=0$, if a^* is a KKT point, then $a_2^*<a_3^*$.*

Using Lemma 5 and Lemma 6, we can prove the next proposition:

PROPOSITION 2. *When $n=3,\ \beta_2=0$ and $\theta_3=0$, there exists a KKT point a^* such that $a_i^*\le\widehat{a}_i^*,\ \forall i=2,3$.*

Proposition 2 is an extension of Proposition 1 for the case when $n=3$ and $\theta_3=0$, suggesting the appointment time when congestion is considered is earlier when there is no cost of service speedup. Although this is for the special case in which the 3^{rd} appointment does not admit speedup, in the numerical results section we show this property is generally true when this assumption is relaxed.

5.3. Case of n Customers

When there exist n customers (e.g., $n\ge3$), it becomes difficult to compare a^* and $\widehat{a}^*$. However, we establish a comparison between the optimal costs for models that anticipate and do not anticipate congestion behavior, respectively, with the following proposition:

PROPOSITION 3. *For any solution* a, $g(a) \le \widehat{g}(a)$ *if* $\beta_i = 0, \quad 2 \le \forall i \le n$; $g(a) \ge \widehat{g}(a)$ *if* $\beta_i \ge \sum_{j=i+1}^{n} \alpha_j + 1, \quad 2 \le \forall i \le n$.

Proposition 3 guarantees the existence of a KKT point, a^* such that $g(a^*) \le \widehat{g}(\widehat{a}^*)$ if $\beta_i = 0, \quad 2 \le \forall i \le n$; $g(a^*) \ge \widehat{g}(\widehat{a}^*)$ if $\beta_i \ge \sum_{j=i+1}^{n} \alpha_j + 1, \ 2 \le \forall i \le n$. It can be interpreted to mean that congestion behavior helps to reduce the total cost when there is no cost of service speedup, however, it will eventually increase the total cost if the cost of service speedup becomes large.

6. Numerical Results

In this section we present numerical experiments to show the computational performance of our proposed models, and the impacts of congestion behavior on appointment schedules. We present the results in two parts: 1) results based on hypothetical test instances with a single appointment type and 2) a case study based on an outpatient clinic at Mayo Clinic.

The SO model, which is a continuous optimization model, was solved by a stochastic approximation (SA) method originated in Robbins and Monro (1951). Based on the SA method, Zhang and Xie (2015) proposed a stochastic gradient algorithm for appointment scheduling problems. We applied their algorithm to solve our SO model. Some implementation details of the algorithm are as follows: the initial solution was set as the *mean-value solution* in which customer interarrival times were set to the mean service times; the step size at iteration k was set as $\bar{\mu}/k$ where $\bar{\mu}$ denotes the average over mean service times of all customers; the algorithm was stopped when the number of iterations reached 10^7. The iteration number is sufficiently large since experimental results show that, across all tested instances, the algorithm stopped at a near optimal solution with $||\nabla g(a)|| < 0.25$. The SIP model was solved by branch-and-cut as implemented in CPLEX MIP Solver 12.6 with 500 sampled scenarios approximating the service time distributions for all customers. The maximum running time was 500 seconds. All solutions obtained from the optimization models were evaluated via simulation using an independent set of 10^6 sampled scenarios.

To analyze the impact of congestion behavior, we compared results of the mean-value solution and the following two solutions:

- *Anticipative solution*: obtained from SO and SIP models. It is referred to as the anticipative solution since it anticipates the effect of congestion when it sets customer appointments.
- *Nonanticipative solution*: obtained from a model that did not anticipate congestion behavior (the model from Denton and Gupta (2003)). The model was solved using the CPLEX LP Solver 12.6 with 5,000 samples approximating service times. It is referred to as the nonanticipative solution since it does not anticipate the effect of congestion.

Table 1 The vector norm of the difference in solutions ($||\Delta \mathbf{a}||$) obtained from the SO and SIP models and the CPU seconds (t_{SO}, t_{SIP}) spent to solve the two models where the dash represents "unavailable".

			5 customers			7 customers			9 customers		
α	β	θ	$\|\|\Delta \mathbf{a}\|\|$	t_{SO}	t_{SIP}	$\|\|\Delta \mathbf{a}\|\|$	t_{SO}	t_{SIP}	$\|\|\Delta \mathbf{a}\|\|$	t_{SO}	t_{SIP}
0.33	0	0.05	0.00	8.8	5.6	0.00	14.4	10.3	0.01	15.5	14.7
		0.1	0.00	8.8	5.1	0.00	12.9	15.6	0.01	16.4	16.8
		0.2	0.00	9.2	20.2	0.00	15.0	162.4	—	15.4	>500
3	0	0.05	0.00	8.0	5.1	0.00	11.3	9.1	0.00	13.6	13.0
		0.1	0.00	7.8	4.8	0.00	10.7	8.8	0.00	14.8	12.3
		0.2	0.01	8.0	5.0	0.00	11.0	10.0	0.00	14.5	13.3
3	3	0.05	0.00	8.0	12.3	0.00	11.8	36.2	0.01	14.1	87.3
		0.1	0.00	8.7	44.9	0.00	10.6	68.7	0.01	14.4	238.0
		0.2	0.01	7.6	53.8	0.00	11.2	192.7	—	13.0	>500

6.1. Hypothetical Tests

Unless otherwise specified, we considered $n = 7$; customers were identical with $\alpha_i = 3$ or $\alpha_i = 0.33$ representing cases of high waiting cost and low waiting cost, respectively; the cost for service speedup was set to $\beta_i = 3$ or $\beta_i = 0$ representing cases where the speedup was penalized or not, respectively. The nominal service time was chosen to be uniformly distributed with mean time $\mu = 1.0$ and standard deviation $\sigma = 0.2$. The length of session was set as $d = \sum \mu_i$. Congestion behavior was modeled as a piecewise linear function and the service time reduction rate was set to $\gamma_i(\omega) = \frac{\theta \xi_i(\omega)}{\tau}$ where τ represents the time threshold beyond which the service time is not reduced, and we set $\tau = 0.2$, unless otherwise specified. Therefore, the magnitude of the congestion effect was determined by the maximum proportion of service time reduction, θ.

6.1.1. Computational Performance of Models When congestion behavior is a piecewise linear function, the problem can be solved by either of our two proposed models. The SO model may find a local optimum when the average cost function is nonconvex; the SIP model, on the other hand, finds exact solutions but may require substantially longer computation. We compared the results obtained from the SO model (1) and the SIP model (6), primarily for the purpose of establishing the validity of results from the SO approach.

Table 1 presents the comparison of the two models in terms of their computational performance. Based on the results, we made the following observations:

- The vector norm of the difference in solutions, $||\Delta \mathbf{a}||$, was negligible in all cases; thus the SO and SIP models obtained approximately the same solution in all instances. The SIP model was solved to optimality, which means the SO model also found near optimal solutions.
- The SO model often required less computation time than the SIP model, particularly for the most difficult SIP instances and especially when θ was large, i.e., when service times were highly sensitive to congestion.

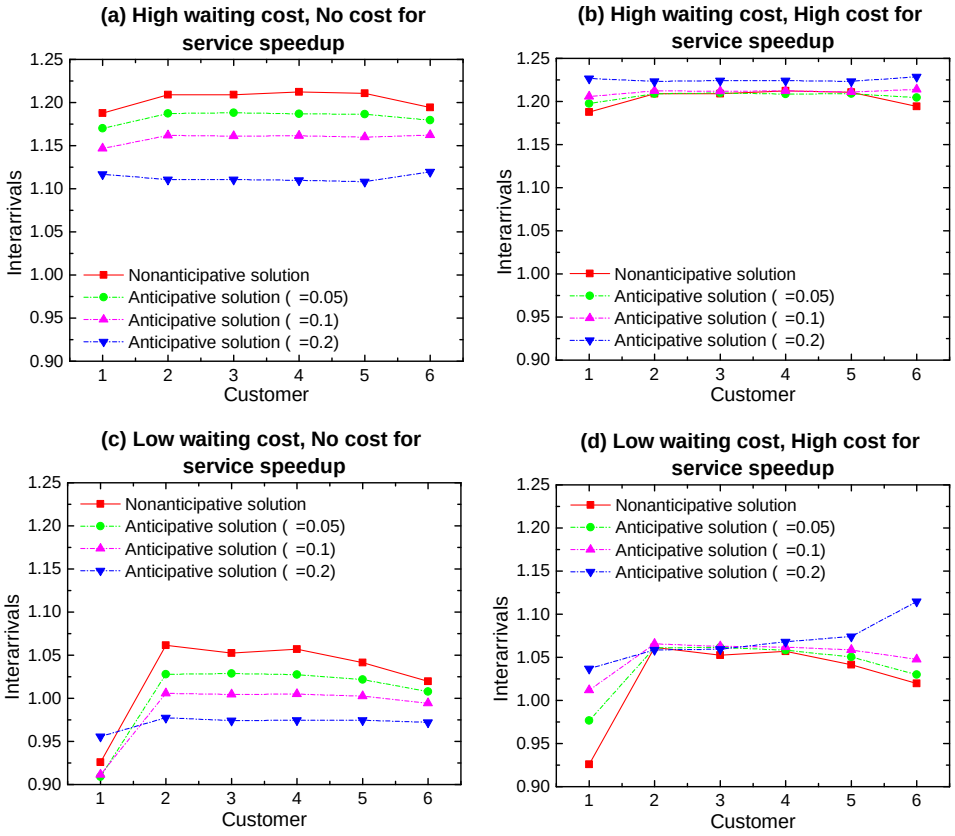

Figure 2 **Interarrival times for scheduling 7 identical customers with mean time of 1.0 and standard deviation of 0.2 when congestion behavior was considered where θ denotes the maximum proportion of service time reduction.**

6.1.2. Impact of Congestion Behaviour Figure 2 illustrates the interarrival times obtained from the anticipative and nonanticipative solutions under different maximum proportions of service time reduction. From the results, we made the following observations:

- When there was no cost for service speedup, the optimal interarrival times were dome shaped, but the shape was less pronounced as the maximum proportion of service time reduction, θ, increased; moreover, when θ was sufficiently large, the shape of interarrival times became flatted. This could be explained by the fact that when θ was sufficiently large, the server can easily catch up delays of the schedule, the propagation of waiting time was mitigated, and thus all customers could be regarded as almost the same. When service speedup was costly, the optimal interarrival times were no longer dome shaped, and the shape became quite sensitive to θ.
- Anticipating congestion behavior resulted in a smaller average time of interarrivals when there is no cost for service speedup, and a longer average time when the speedup is costly. A similar observation was suggested by Proposition 1 for the case of $n = 2$.

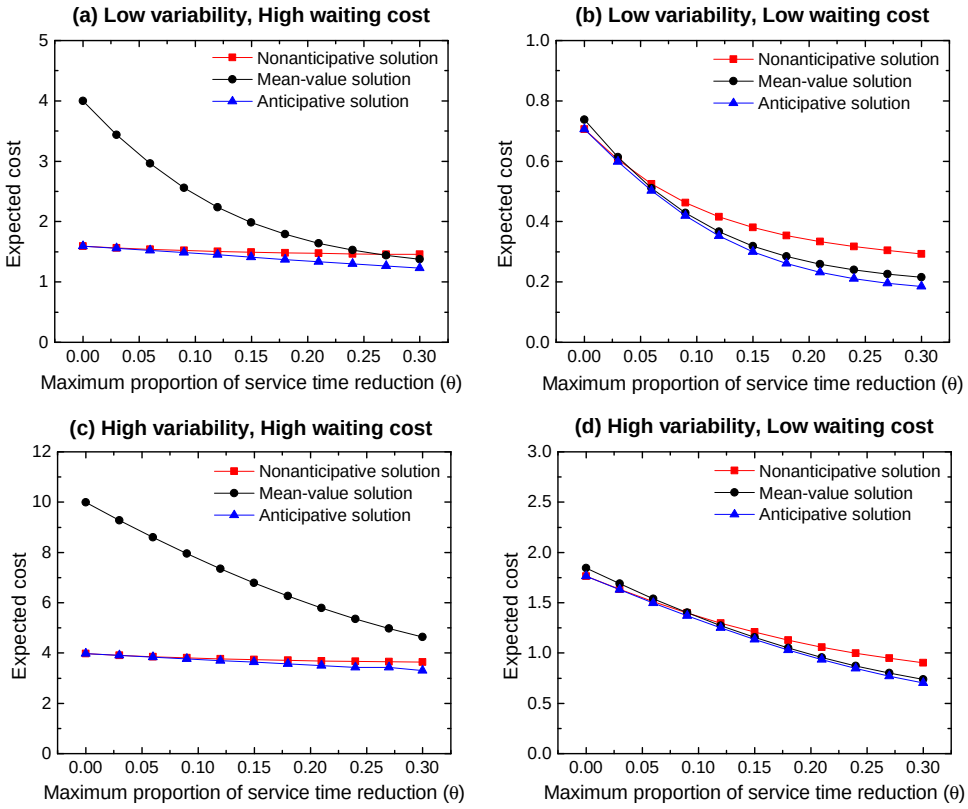

Figure 3 Expected costs for scheduling 7 identical customers with mean time of 1.0 and standard deviation of 0.2 under nonanticipative solutions, mean-value solutions and anticipative solutions.

Figure 3 illustrates the expected cost of the nonanticipative solution, mean-value solution and our anticipative solution under congestion response when there is no cost for service speedup. From the results, we made the following observations:

- Proposition 3 suggests that congestion behavior helps to reduce the total cost when there is no cost of service speedup. From the experiments, we found that the value of anticipating congestion behavior, which is measured by the relative cost reduction, was generally high, ranging from 10% to 58% across the 4 test instances and when $\theta = 0.3$. The value was especially high for cases with low customer waiting costs which may accurately represent practice.
- As the congestion response rate increased, the mean-value solution performed well in all cases. Similar results were also observed for cases with higher variability when the standard deviation is 0.5.

6.1.3. Nonlinear congestion behavior We considered an additional case where congestion behavior was a nonlinear function, and particularly when the congestion-dependent service time

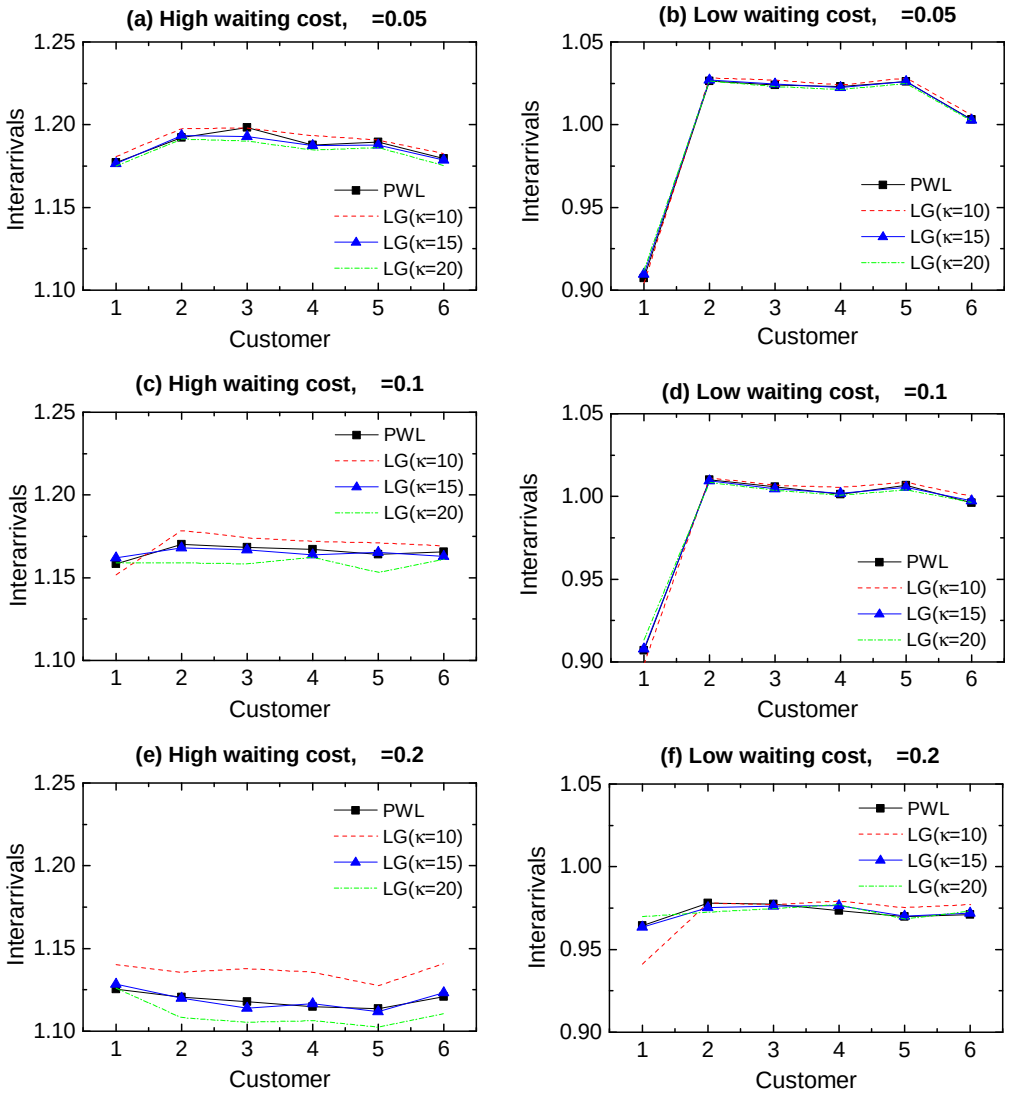

Figure 4 Interarrival time solutions for scheduling 7 identical customers with mean time of 1.0 and standard deviation of 0.2 under piecewise linear (PWL) and logit (LG) congestion response functions of server behavior.

was defined as the following logit function:

$$z_i(\omega, w_i(\omega)) = \xi_i(\omega)\left[1 + \theta_i - \frac{2\theta_i}{1 + e^{-\kappa w_i(\omega)}}\right] \tag{14}$$

We chose this function because it requires only two parameters and it can model smooth variation in congestion response (compared to the piecewise linear response model). The logit function was varied by setting κ to 10, 15 and 20, respectively. From the perspective of $z_i'(\omega)$, the congestion effect on service time for the logit function, compared with the piecewise linear function, was less pronounced when $\kappa = 10$, more pronounced when $\kappa = 20$, and approximately equally pronounced when $\kappa = 15$.

We solved six instances with the logit function with different waiting costs and θ's. Figure 4 illustrates the shape of interarrival times for the six instances. From the results, we observed that the solutions obtained using piecewise linear and logit functions for modeling congestion behavior were similar, especially when $\kappa = 15$; the difference between the two solutions was less than 0.43%, which means the piecewise linear function can be used to approximate a more general function in the optimization of appointment scheduling, and the approximation error is negligible when the two functions have the same magnitude of congestion effect on service time. Using the piecewise linear function may be preferred in some cases since it requires fewer parameters to characterize congestion behavior, and it can be used to solve exact solutions with the SIP model.

6.2. Case Study: Scheduling of an Outpatient Clinic

We considered a specific example of scheduling outpatient appointments at Mayo Clinic, a non-profit organization engaged in clinical practice, education, and research with major campuses in Rochester, MN, Scottsdale and Phoenix, AZ, and Jacksonville, FL. We considered an outpatient clinic at the Rochester campus that served over 24,000 appointments. The outpatient clinic employed physicians, physician assistants, and nurse practitioners for various outpatient consultative appointment types. The data in our analysis included electronic time-stamp data of patient events during their appointment including scheduled appointment time, when the patient checked in, and the duration of time the patient was with the provider. We limited our analysis to patients with a clinical appointment, appointments at a single check in location, appointments with confirmed attendance and confirmed room location, and patients with a single appointment in the clinic on a particular day. This resulted in 14,037 observations for our study.

Mayo Clinic is an academic medical center where many providers have substantial research and education commitments. Based on the clinical workload and frequency of appointments, we selected "full-time" providers who had at least 8 appointments per day and at least 100 observations in total because they represent what is most common in practice. We aggregated the data into groups by combinations of the provider and the appointment type, and we further aggregated the data for each group by patient waiting time into 20-minute intervals, and we truncated data for intervals with less than 30 observations. Finally, we obtained 11 available groups which accounted for 16% of the total data. Out of the 11 groups, we observed the mean service time was negatively correlated with the waiting time in 9 groups. We also conducted the two-sample t-test (Keselman et al. (2004)) to make pairwise comparisons of the mean service times for each waiting time interval. The results showed the mean service times were different in different intervals (significance level was 0.05) in 2 groups, in which the mean service time was also negatively correlated with the waiting time, suggesting they exhibited a significant congestion effect, i.e., the mean service time was reduced

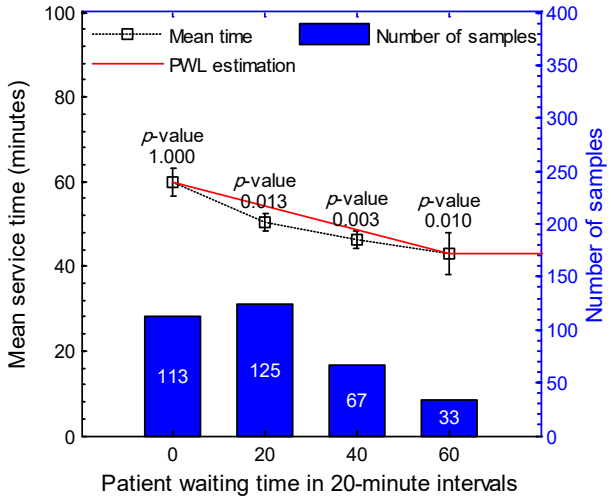

Figure 5 **Piecewise linear function estimation (PWL estimation) of congestion behavior on mean service times for type A patients.**

as the waiting time increased (see Figure 5 for an example, with estimates of the p-values for the pairwise comparisons). For our case study, we selected a particular provider with 3 appointment types (referred to as "A", "B", and "O", respectively) as an example of our numerical experiments. Among the appointment types, the most common, type A, exhibited a congestion effect but the others did not.

Table 2 **Summary of model parameters for the three appointment types for a particular provider considered in the case study.**

Appointment Type	Percent of Appointments Per Day	Mean service Time (minutes)	Standard Deviation (minutes)	Maximum Proportion, θ	Time Threshold, τ (minutes)
A	56%	59.94	34.14	0.28	60
B	20%	37.39	32.11	0	0
O	24%	53.57	26.95	0	0

6.2.1. Estimation of congestion behavior We used the mean of service times in each interval to characterize congestion behavior for appointment type A. We let v denote the interval in which the mean time is the minimum. The nominal mean time and standard deviation were estimated based on service times in interval 0. If the mean time in interval v was smaller than the nominal mean time with significance level less than 0.05, we regard the mean time decreased from interval 0 to v, and kept constant in intervals beyond v. We estimated the congestion behavior as a piecewise linear function as illustrated in Figure 5 for appointment type A. The maximum proportion of service time reduction was estimated as $\theta_i = \frac{\mu_0 - \mu_v}{\mu_0}$, where μ_0 and μ_v are mean times

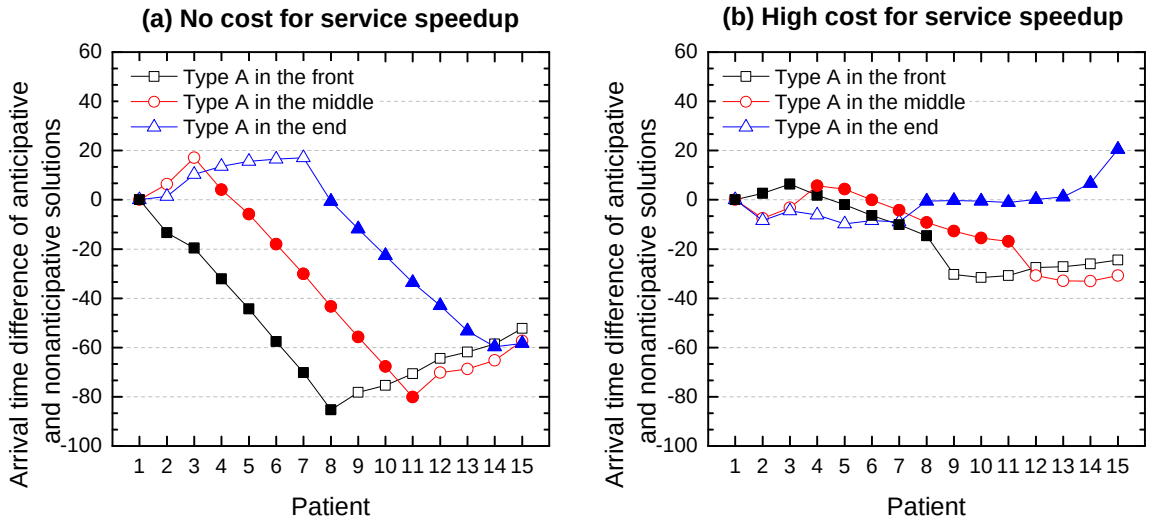

Figure 6 **Arrival time difference of anticipative and nonanticipative solutions when type A patients, which were filled markers, were positioned in the front, middle, and end of the schedule, respectively.**

in intervals 0, and v, respectively. The threshold was estimated as $\tau_i = v$, and the reduction rate was estimated as $\gamma_i(\omega) = \frac{\theta_i \xi_i(\omega)}{\tau_i} \mid \tau_i > 0$ and $\gamma_i(\omega) = 0 \mid \tau_i = 0$, respectively. Table 2 summarizes the estimated parameters for all types which were used to solve the models.

6.2.2. Results Based on the data for the selected provider, we considered $n = 15$ patients of three types. The numbers of patients of each type were allocated based on historical percentages (8 of A, 3 of B, and 4 of O). Parameters in Table 2 were used to generate test instances. We observed service times have long right tails, so they were assumed to be lognormally distributed. Consistent with previous literature (Robinson and Chen (2010)), we regarded the patient waiting time as less valuable than the provider's overtime, so we set $\alpha_i = 0.33$. We considered two cases of service speedup costs, setting $\beta_i = 3$ or $\beta_i = 0$, and the length of session was set to $d = \sum \mu_i$. From Table 2, only type A exhibited speedup. Thus, we considered three possible sequence positions for patients of type A including the front, middle and end of the schedule.

We compared the results obtained from the anticipative and nonanticipative solutions in terms of the arrival time difference as shown in Figure 6. We aggregated patients into three batches: batch *A* included all patients of type A, batch *Pre* included all patients prior to type A patients, and batch *Post* included all patients following type A patients. From the results, we made the following observations:

- When there was no cost for service speedup and congestion behavior was anticipated, we observed the arrival time difference for batch A was usually negative, and decreasing. This means considering congestion induced earlier arrival times than the nonanticipative solution, and less interarrival time for each patient. The intuition behind this is that such a schedule caused more

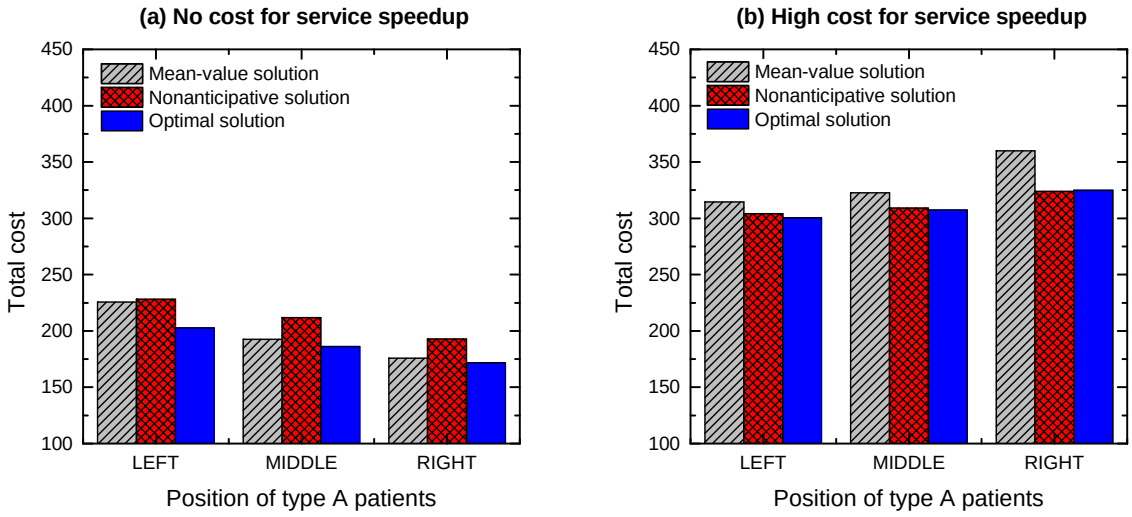

Figure 7 Cost comparison of the mean-value solution, the nonanticipative solution and the anticipative solution when type A patients were positioned in the front, middle, and end of the schedule, respectively.

waiting time, thus reducing the service time due to the service speedup, and ultimately reducing the overtime cost. Batch Pre had later arrival times than the nonanticipative solution, since such a schedule delayed the completion of batch Pre, thus increasing the waiting time of batch A, and causing more of a congestion response. Batch Post had earlier arrival times, since they followed batch A, which completed earlier in the anticipative solution. Both batch Pre and batch Post were allocated more interarrival time for each patient since batch A had a shorter interarrival time.

- As the speedup cost increased from 0 to 3, for the anticipative solution the interarrival time for batch A increased, and the time for other batches decreased. The intuition behind this is that increasing the interarrival time for batch A caused less waiting time, thus reducing the service speedup. Since batch A had a longer interarrival time, the other two batches were allocated less interarrival time for each patient.

It is worth noting that when each of the batches has one patient or less, the problem is similar to the special cases we considered in Section 5. We showed in Proposition 1 that if there are two batches (Pre and A) with one patient in each, batch A will have an earlier arrival time when there is no cost for service speedup, and a later arrival time when there is a sufficiently large speedup cost ($\beta > 1$ for the special case). We also showed in Proposition 2 that if there are three batches with one patient in each, batches A and Post will have earlier arrival times when $\beta = 0$. All of these properties were observed for these larger problems.

Figure 7 presents the cost comparison of the mean-value solution, the nonanticipative solution and the anticipative solution when type A patients were positioned in different places. From the results, we observed that the anticipative solution always resulted in a lower or approximately equal cost compared to the nonanticipative solution. Specifically, the relative cost reduction was

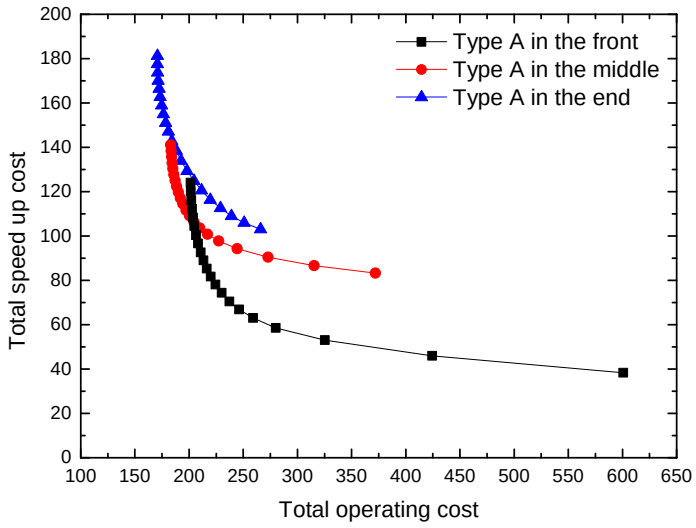

Figure 8 **Pareto solutions of the operating cost versus the speedup cost when type A patients were positioned in the front, middle, and end of the schedule, respectively.**

11% on average when there was no cost for service speedup, and 1% on average when the unit speedup cost was 3. Therefore, it is particularly rewarding to anticipate congestion behavior when there is no cost for speeding up. Another observation was that, when there is no cost for speedup, the mean-value solution generally outperformed the nonanticipative solution, and it had similar performance to the anticipative solution, suggesting the mean-value solution may work well in some cases; perhaps explaining its common use in practice.

6.2.3. Sensitivity analysis In our examples we assumed the service speedup cost was 0 or 3; however, in practice it may be difficult to estimate this cost. Therefore we conducted experiments where we varied the weight of the criteria to estimate the tradeoff. We introduced a parameter, ρ, to adjust the weights between the operating cost (i.e. sum of patient waiting costs and the provider overtime cost) and the service speedup cost as follows:

$$Total\ cost = (1-\rho) \times operating\ cost + \rho \times speed\ up\ cost.$$

We varied ρ from 0 to 1, and thus the larger ρ the more value of the speedup cost. Figure 8 presents Pareto optimal solutions of the operating cost and the speedup cost when patients of type A were positioned in the front, middle, and end of the schedule. From the results, we made the following observations:

- When the operating cost was highly valued, positioning type A in the end is preferred. From the congestion behavior perspective, the intuition behind this is that waiting times propagated

with the number of patients. When type A patients were positioned later in the sequence, they experienced more waiting time and thus more service time speed up, and ultimately the overtime was reduced.

- When the speedup cost was highly valued, positioning type A in the front is preferred. The intuition is that positioning type A patients in the front resulted in less waiting time, so it helped to reduce the service speedup cost.
- Speedup could be limited substantially with very little impact on operating cost. This is important because in health service systems speedup can cause errors and/or fatigue.

7. Conclusion

This paper addressed an appointment scheduling problem when the server responds to congestion. The service time is an endogenous random variable depending on the state of congestion. We characterized congestion behavior as a piecewise differentiable function on the amount of waiting time. Decisions were scheduled appointment times for a sequence of customers to minimize a weighted cost of customer waiting time, server overtime and a penalty for service speedup. We provided alternative formulations of this problem as a simulation optimization (SO) model and a stochastic integer programming (SIP) model, respectively. We presented theoretical results for both models including conditions under which the SO model becomes continuously differentiable and the SIP model reduces to a convex stochastic program. A series of experimental results showed the SO model is likely to find the optimal solution and it is much easier to solve than the SIP model, and can scale up to large problem instances.

We summarize some of the most important insights in terms of answers to the research questions posed in the introduction:

- We found that it could indeed be very important to consider server congestion behavior, particularly when waiting costs were low and/or congestion response was sensitive to waiting time. In the latter case, in particular, the mean-value solution performed very well. This could explain, in part, the popularity of scheduling according to mean service time in practice.
- The nature of the optimal schedule can be summarized as follows: 1) When scheduling identical customers, the optimal interarrival times exhibited a dome shape pattern if service speedup was not costly; in contrast, when the speedup was costly, the shape became quite sensitive to parameters such as waiting cost per unit time and the magnitude of the congestion effect. 2) The optimal schedule was insensitive to the shape of congestion behavior. Therefore, using a piecewise linear function could provide a good approximation to a more general function of congestion behavior. This approximation would be useful for applying the SIP model to solve exact solutions, especially when the model reduces to a convex stochastic program. 3) It was possible to tradeoff the speedup

cost and the operating cost by adjusting arrival times and positions of customers; perhaps most importantly, speedup could be limited substantially with very little impact on operating cost.

To the authors' knowledge, this is the first paper to incorporate congestion response in optimization models for appointment scheduling. The usefulness of the proposed models are limited by the need to better understand congestion factors in practice; however, our example from an outpatient clinic, and many examples cited in the literature suggest this effect is common. Additional empirical studies are needed to better understand the effect in the outpatient setting. One possible direction of future research is to consider other congestion factors like sequence-dependent service time. Another possible direction is to extend our models to consider multiple servers and appointment sequencing problems under congestion behavior. We believe this article will help lay the foundation for some of these future studies.

References

Batt, Robert J, Christian Terwiesch. 2012. Doctors under load: An empirical study of state-dependent service times in emergency care. *Working Paper* .

Begen, Mehmet A., Maurice Queyranne. 2011. Appointment Scheduling with Discrete Random Durations. *Mathematics of Operations Research* **36**(2) 240–257.

Cayirli, Tugba, Emre Veral. 2003. Outpatient scheduling in health care: a review of literature. *Production and Operations Management* **12**(4) 519–549.

Cayirli, Tugba, Emre Veral, Harry Rosen. 2008. Assessment of patient classification in appointment system design. *Production and Operations Management* **17**(3) 338–353.

Cayirli, Tugba, Kum Khiong Yang, Ser Aik Quek. 2012. A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management* **21**(4) 682–697.

Chakraborty, Santanu, Kumar Muthuraman, Mark Lawley. 2010. Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions* **42**(5) 354–366.

Chan, Carri W, Galit Yom-Tov, Gabriel Escobar. 2014. When to use speedup: An examination of service systems with returns. *Operations Research* **62**(2) 462–482.

Denton, Brian, Diwakar Gupta. 2003. A Sequential Bounding Approach for Optimal Appointment Scheduling. *IIE Transactions* **35**(11) 1003–1016.

Deveugele, Myriam, Anselm Derese, Atie van den Brink-Muinen, Jozien Bensing, Jan De Maeseneer. 2002. Consultation length in general practice: cross sectional study in six european countries. *British Medical Journal* **325**(7362) 472.

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

Gupta, Diwakar, Brian Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions* **40**(9) 800–819.

Harris, Carol M. 1967. Queues with state-dependent stochastic service rates. *Operations Research* **15**(1).

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

Keselman, HJ, Abdul R Othman, Rand R Wilcox, Katherine Fradette. 2004. The new and improved two-sample t test. *Psychological Science* **15**(1) 47–51.

Kleywegt, Anton J, Alexander Shapiro, Tito Homem-de Mello. 2002. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization* **12**(2) 479–502.

Kong, Qingxia, Shan Li, Nan Liu, Chung-Piaw Teo, Zhenzhen Yan. 2016. Appointment scheduling under schedule-dependent patient no-show behavior. *Working paper* .

Kushner, Harold, G George Yin. 2003. *Stochastic approximation and recursive algorithms and applications*, vol. 35. Springer Science & Business Media.

Mancilla, Camilo, Robert Storer. 2012. A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions* **44**(8) 655–670.

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

Posner, M. 1973. Single-server queues with service time dependent on waiting time. *Operations Research* **21**(2).

Rising, Edward J, Robert Baron, Barry Averill. 1973. A systems analysis of a university-health-service outpatient clinic. *Operations Research* **21**(5) 1030–1047.

Robbins, Herbert, Sutton Monro. 1951. A stochastic approximation method. *The annals of mathematical statistics* 400–407.

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

Robinson, Lawrence W, Rachel R Chen. 2010. A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management* **12**(2) 330–346.

Weiss, Elliott N. 1990. Models for Determining Estimated Start Times and Case Orderings In Hospital Operating Rooms. *IIE Transactions* **22**(2) 143–150.

Zhang, Zheng, Xiaolan Xie. 2015. Simulation-based optimization for surgery appointment scheduling of multiple operating rooms. *IIE Transactions* **47**(9) 998–1012.

Appendix

Proof of Lemma 1

Proof: Here and throughout the appendix, we let $c_i(\omega)$ denote the completion time for customer i

in scenario ω and $\triangle$ is the difference operator. We have $|\triangle c_1(\omega)| = 0$ as $a_1 \equiv 0$ and $z_1(\omega) \equiv \xi_1(\omega)$ by assumption. Assuming $||\triangle a|| \leq h$ where h is a constant number, we have the following inequalities:

$$|\triangle w_2(\omega)| \leq |\triangle c_1(\omega) - \triangle a_2| \leq |\triangle c_1(\omega)| + |\triangle a_2| \leq h, \tag{15a}$$

$$|\triangle z_2(\omega, w_2(\omega))| \leq L(\omega)|\triangle w_2(\omega)| \leq L(\omega)h, \tag{15b}$$

$$\begin{aligned} |\triangle c_2(\omega)| &= |\triangle a_2 + \triangle w_2(\omega) + \triangle z_2(\omega, w_2(\omega))| \\ &\leq |\triangle a_2| + |\triangle w_2(\omega)| + |\triangle z_2(\omega, w_2(\omega))| \leq (L(\omega)+2)h, \end{aligned} \tag{15c}$$

where inequality (15a) is from the definition of waiting time, inequality (15b) is from Assumption 1 where $L(\omega)$ is a constant defined in Assumption 1, and inequality (15c) is from the definition of completion time. By induction, we have

$$|\triangle w_i(\omega)| \leq |\triangle c_{i-1}(\omega)| + |\triangle a_i| \leq (L(\omega)+3)^{i-2} h, \tag{16a}$$

$$|\triangle z_i(\omega, w_i(\omega))| \leq L(\omega)|\triangle w_i(\omega)| \leq L(\omega)\,(L(\omega)+3)^{i-2} h, \tag{16b}$$

$$|\triangle c_i(\omega)| \leq |\triangle a_2| + |\triangle w_2(\omega)| + |\triangle z_2(\omega, w_2(\omega))| \leq (L(\omega)+1)\,(L(\omega)+3)^{i-2} h. \tag{16c}$$

Thus, $w_i(\omega)$ and $z_i(\omega, w_i(\omega))$ are Lipschitz-continuous in a. Letting $K = (L(\omega)+3)^{n-1}$, we have $\triangle w_i(\omega) \leq Kh$, $\triangle z_i(\omega, w_i(\omega)) \leq Kh$, and $\triangle c_i(\omega) \leq Kh$ for all i and ω and thus $l(\omega) \leq Kh$. As a result, we have the following inequality:

$$|\triangle f(a,\omega)| \leq \sum_{i=2}^{n} \alpha_i |\triangle w_i(\omega)| + \sum_{i=2}^{n} \beta_i |\triangle z_i(\omega, w_i(\omega))| + |\triangle l(\omega)|, \tag{17a}$$

$$\leq \left(\sum_{i=2}^{n} \alpha_i + \sum_{i=2}^{n} \beta_i + 1 \right) Kh. \tag{17b}$$

Therefore, $f(a,\omega)$ defined as (2a) is Lipschitz-continuous in a. □

Proof of Lemma 2

Proof: The sample path cost function, $f(a,\omega)$, is differentiable everywhere except at points with one of the following conditions:

- (i) customer i arrives at exactly when $i-1$ completes, i.e., $a_{i-1} + w_{i-1}(\omega) + z_{i-1}(\omega, w_{i-1}(\omega)) = a_i$,
- (ii) the last customer, n, completes at exactly when the session ends, i.e., $a_n + w_n(\omega) + z_n(\omega, w_n(\omega)) = d$,
- (iii) $z_i(\omega, w_i(\omega))$ is nondifferentiable at $w_i(\omega)$.

Conditions *(i)* and *(ii)* occur with probability zero because $w_i(\omega) + z_i(\omega, w_i(\omega))$ is a continuous random variable with finite density according to Assumption 3, and independent of a_i and d, respectively. Condition *(iii)* also occurs with probability zero because the set of nondifferential points of $z_i(\omega, w_i(\omega))$ is finite according to Assumption 2. As a result, $f(a,\omega)$ is differentiable everywhere except at finite saddle points at measure 0. □

Proofs of Theorem 1 and Lemma 3

We omit to prove Theorem 1 and Lemma 3 as similar proofs can be found in the literature, e.g., proofs of Theorem 1 and Lemma 4 in Zhang and Xie (2015).

Proof of Theorem 2

Proof: We first prove equation (3): we let ϵ denote a sufficiently small positive constant. When $w_i(\omega) = 0$ and a_i is increased by ϵ, $w_i(\omega)$ and $z_i(\omega, w_i(\omega))$ are unchanged, but $c_i(\omega)$ is increased by ϵ, and thus $f(a,\omega)$ is increased by $\lambda_{i+1}\epsilon$. When $w_i(\omega) > 0$ and a_i is increased by ϵ, $w_i(\omega)$ and $z_i(\omega, w_i(\omega))$ are decreased by ϵ and $z_i'(\omega)\epsilon$, respectively, the start time of customer i is unchanged, and thus $c_i(\omega)$ is also decreased by $z_i'(\omega)\epsilon$. As a result, $f(a,\omega)$ is decreased by $[\alpha_i + z_i'(\omega)(\lambda_{i+1} - \beta_i)]\,\epsilon$. We next prove equation (4): when $w_i(\omega) = 0$ and $c_{i-1}(\omega)$ is increased by ϵ, $w_i(\omega)$ is unchanged and thus $f(a,\omega)$ is unchanged. When $w_i(\omega) > 0$ and $c_{i-1}(\omega)$ is increased, $w_i(\omega)$ and $z_i(\omega, w_i(\omega))$ are increased by ϵ and $z_i'(\omega)\epsilon$, respectively, the start time of customer i is increased by ϵ, and thus $c_i(\omega)$ is increased by $(1 + z_i'(\omega))\,\epsilon$. As a result, $f(a,\omega)$ is increased by $[\alpha_i + (1 + z_i'(\omega))\lambda_{i+1} - z_i'(\omega)\beta_i]\,\epsilon$. When $i = n+1$ and $c_{i-1}(\omega)$ is increased by ϵ, the overtime, $l(\omega)$, is increased by ϵ if $l(\omega) > 0$, and unchanged if $l(\omega) = 0$. □

Proof of Theorem 3

Proof: When a_i is increased, $\triangle c_i(\omega) \geq 0$ and thus $\triangle w_{j+1}(\omega) \geq 0$. According to Assumption 4, $\triangle z_{i+1}(\omega) \geq -\triangle w_{i+1}(\omega)$, and thus $\triangle c_{i+1}(\omega) = \triangle w_{i+1}(\omega) + \triangle z_{i+1}(\omega) \geq 0$. By induction, we have $\triangle c_j(\omega) \geq 0,\ \triangle w_j(\omega) \geq 0,\ \forall j > i$, and $\triangle l(\omega) > 0$. □

Proof of Lemma 4

Proof: When both $w_i(\omega)$ and $s_i(\omega)$ are positive, the decrease of both $w_i(\omega)$ and $s_i(\omega)$ results in less waiting time and nonincreasing completion time for customer i according to Assumption 4, and thus it does not increase waiting time for the remaining customers. As a result, without increasing the total cost, $w_i(\omega)$ and $s_i(\omega)$ can be decreased until either of them reaches zero, suggesting the complementary slackness constraints (6k)-(6l) automatically hold when the optimal

solution is achieved. □

Proof of Theorem 4

Proof: Under Assumptions 4-5, increasing the service time for customer i does not reduce waiting time for the remaining customers and thus there is no reduction in the total cost. Therefore, $Z_i(\omega) \leq z_i(\omega, w_i(\omega))$ when the optimal solution is achieved. On the other hand, constraints (6e)-(6f) jointly enforce $Z_i(\omega) \geq z_i(\omega, w_i(\omega))$; as a result, $Z_i(\omega) = z_i(\omega, w_i(\omega))$ automatically holds for each i and ω. Under Assumptions 4, constraints (6k)-(6l) can be removed according to Lemma 4. Therefore, model (6) reduces to the convex program (8). □

Proof of Proposition 1

Proof: The KKT condition holds if *(i)* $g_2'(a_2) = 0$, or *(ii)* $g_2'(0) > 0$ and $a_2 = 0$. When $\beta_2 = 0$, we have $f_2' \geq \widehat{f}_2'$ and $g_2' \geq \widehat{g}_2'$ according to (10). If $g_2'(0) \geq 0$, the proposition can be proved by setting $a_2^* = 0$. If $g_2'(0) < 0$, we have $g_2'(\widehat{a}_2^*) \geq \widehat{g}'(\widehat{a}_2^*)$ where $\widehat{a}_2^*$ is the KKT point to $\widehat{g}(a)$ and thus $\widehat{g}'(\widehat{a}_2^*) = 0$. Since g_2' is continuous in a_2 (Theorem 1), there exists a KKT point, $a_2^* \in (0, \widehat{a}_2^*]$, such that $g_2'(a_2^*) = 0$. Similarly, when $\beta_2 \geq 1$, we have $f_2' \leq \widehat{f}_2'$ and $g_2' \leq \widehat{g}'$. If $\widehat{a}_2^* = 0$, $a_2^* \geq \widehat{a}_2^*$ holds automatically. If $\widehat{a}_2^* > 0$, we have $g_2'(\widehat{a}_2^*) \leq 0$. Since $g_2'(\infty) = E[I(l(\omega))] > 0$ and g_2' is continuous in a_2, there must exists a KKT point, $a_2^* \in [\widehat{a}_2^*, \infty)$, such that $g_2'(a_2^*) = 0$. □

Proof of Lemma 5

Proof: We rewrite $f_2'(\omega)$ and $f_3'(\omega)$ as follows:

$$f_2'(\omega) = \begin{cases} -\alpha_2 + z_2'(\omega)\beta_2 - z_2'(\omega) I(w_3(\omega)) \left(\alpha_3 + I(c_2(\omega) + \xi_3(\omega) - d)\right), & if\ w_2(\omega) > 0, \\ I(w_3(\omega)) \left(\alpha_3 + I(c_2(\omega) + \xi_3(\omega) - d)\right), & if\ w_2(\omega) = 0. \end{cases}$$

$$f_3'(\omega) = \begin{cases} -\alpha_3, & if\ w_3(\omega) > 0, \\ I(c_2(\omega) + \xi_3(\omega) - d), & if\ w_3(\omega) = 0. \end{cases}$$

$f_2'(\omega)$ increases with $I(w_3(\omega))$ and $w_3(\omega)$ is decreasing in a_3 and thus $f_2'(\omega)$ is decreasing in a_3. $f_3'(\omega)$ decreases with $w_3(\omega)$. As $w_3(\omega)$ increases with a_2 and decreases with a_3, $f_3'(\omega)$ decreases with a_2, and increases with a_3. □

Proof of Lemma 6

Proof: Assuming that $a_3 \leq a_2$, we have $w_3(\omega) > 0,\ \forall \omega$ and thus $g_3' < 0$. As increasing a_3 reduces the total cost, any solution with $a_3 \leq a_2$ can not be a KKT point. □

Proof of Proposition 2

Proof: From Lemma 6, when $n=3$, there only exist two cases of KKT points: *i)* $g_2'=g_3'=0$, and *ii)* $g_3'=0,\ g_2'>0$, and $a_2=0$. According to (12)-(13), $g_2'\geq\widehat{g}_2',\ \ g_3'\geq\widehat{g}_3'$ when $\beta_2=0$. Assuming $\widehat{a}^*$ is a KKT point to $\widehat{g}$, we have $g_2'(\widehat{a}^*)\geq 0$ and $g_3'(\widehat{a}^*)\geq 0$. From Lemma 5, decreasing $\widehat{a}_2^*$ reduces g_2', but increases g_3', and decreasing $\widehat{a}_3^*$ reduces g_3', but increases g_2'. Therefore, $\widehat{a}_2^*$ and $\widehat{a}_3^*$ can be decreased alternatively to make either *i)* or *ii)* is satisfied. Thus, there exists a KKT point a^* such that $a_i^*\leq\widehat{a}_i^*,\ \ 2\leq\forall i\leq n$. □

Proof of Proposition 3

Proof: First, $w_i(\omega)$ and $l(\omega)$ are nonincreasing with congestion behavior. When $\beta=0$, we have $g(a)\leq\widehat{g}(a)$. Assuming $z_i(\omega,w_i(\omega))$ is reduced by δ, $w_i(\omega)$ and $l(\omega)$ can be reduced by at most δ. Therefore, $g(a)\geq\widehat{g}(a)$ when $\beta_i\geq\sum_{j=i+1}^{n}\alpha_j+1,\ \ 2\leq\forall i\leq n$. □