



Integer Programming Approaches for Appointment Scheduling with Random No-Shows and Service Durations

Ruiwei Jiang,^a Siqian Shen,^a Yiling Zhang^a

^aDepartment of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109

Contact: ruiwei@umich.edu,  <http://orcid.org/0000-0002-3941-5057> (RJ); siqian@umich.edu,  <http://orcid.org/0000-0002-2854-163X> (SS); zyiling@umich.edu (YZ)

Received: December 4, 2015

Revised: September 10, 2016; March 15, 2017

Accepted: May 3, 2017

Published Online in Articles in Advance:
October 9, 2017

Subject Classifications: programming:
stochastic; programming: integer; algorithms:
cutting plane/facet

Area of Review: Optimization

<https://doi.org/10.1287/opre.2017.1656>

Copyright: © 2017 INFORMS

Abstract. We consider a single-server scheduling problem given a fixed sequence of appointment arrivals with random no-shows and service durations. The probability distribution of the uncertain parameters is assumed to be ambiguous, and only the support and first moments are known. We formulate a class of distributionally robust (DR) optimization models that incorporate the worst-case expectation/conditional value-at-risk penalty cost of appointment waiting, server idleness, and overtime into the objective or constraints. Our models flexibly adapt to different prior beliefs of no-show uncertainty. We obtain exact mixed-integer nonlinear programming reformulations and derive valid inequalities to strengthen the reformulations that are solved by decomposition algorithms. In particular, we derive convex hulls for special cases of no-show beliefs, yielding polynomial-sized linear programming models for the least and the most conservative supports of no-shows. We test various instances to demonstrate the computational efficacy of our approaches and to compare the results of various DR models given perfect or ambiguous distributional information.

Funding: The first author is supported in part by the National Science Foundation [Grant CMMI-1555983] and the second and third author are supported in part by the National Science Foundation [Grant CMMI-1433066].

Supplemental Material: The e-companion is available at <https://doi.org/10.1287/opre.2017.1656>.

Keywords: appointment scheduling • no-show uncertainty • distributionally robust optimization • mixed-integer programming • valid inequalities • totally unimodularity • convex hulls

1. Introduction

We consider an appointment scheduling problem that involves a single server and a set of appointments following a fixed order of arrivals. A system operator needs to schedule an arrival time for each appointment with random no-shows and service duration. This problem is fundamental for establishing service quality and operational efficiency in many service systems, and has been studied in the context of surgery planning in hospitals (see, e.g., Denton and Gupta 2003), call-center staffing (see, e.g., Gurvich et al. 2010), and cloud computing server operations (see, e.g., Shen and Wang 2014). Random no-shows are often observed in outpatient appointment scheduling (e.g., Lee et al. 2005, Berg et al. 2014), which may cause equipment and personnel idleness and losses of opportunities of serving other appointments. As observed by Ho and Lau (1992), random no-shows affect the performance of an appointment system more than random service durations. Additionally, even though many administrative mechanisms exist to reduce the likelihood of no-shows, it is not entirely possible to eliminate no-shows and their negative impacts (see, e.g., Barron 1980, Moore et al. 2001). In view of these challenges, Cayirli and

Veral (2003) point out that a better approach is to adapt the appointment schedule to the anticipated no-shows.

A common goal is to minimize the expected cost associated with appointment waiting time, server idle time, and overtime if the distributional information is fully accessible. In Section 1.1, we provide an extensive review of the literature on variants of stochastic appointment scheduling under specific objectives, metrics, and applications. In reality, it is challenging to accurately estimate the probability distribution of no-shows and service durations. The data of no-shows could be limited because of low probability of occurrence and the heterogeneity of appointments. In view of a wide range of plausible substitutes (e.g., log-normal, normal, and uniform) for modeling the service-time uncertainty, one could easily misspecify its distribution. Then, with ambiguous estimates of no-show and service-duration distributions, we could schedule unnecessarily long (respectively, short) time in between appointments, resulting in significant server idleness (respectively, appointment waiting or server overtime). To address the distributional ambiguity issue, Kong et al. (2013) propose a distributionally robust (DR) model using a cross-moment

ambiguity set that consists of all distributions with common mean and covariance of the random service durations. They obtain a copositive cone programming reformulation and solve a semidefinite program to approximate the optimal results. The most relevant to this paper, Mak et al. (2015) study a DR model using a marginal-moment ambiguity set of the random service durations. They obtain tractable reformulations by successfully solving a nonconvex optimization problem based on a binary encoding of its feasible region. More importantly, both Kong et al. (2013) and Mak et al. (2015) point out that DR models can yield appointment schedules that (1) perform stably under various probability distributions of service durations and (2) improve the appointment system performance under extreme scenarios. These observations motivate us to study DR models when faced with both no-show and service-duration uncertainty.

In this paper, we generalize the DR appointment scheduling model in Mak et al. (2015) by incorporating heterogeneous no-shows and their distributional ambiguity. We aim to produce appointment schedules with good out-of-sample performance, even only given a few historical data. To the best of our knowledge, this paper is the first to consider both discrete (no-shows) and continuous (service durations) randomness for DR appointment scheduling. This generalization results in a challenging mixed-integer nonlinear program (MINLP), which the approach by Mak et al. (2015) fails to solve. The main contribution of the paper is to derive effective integer programming approaches for solving the generalized DR model, including valid inequalities that effectively accelerate the computation of the MINLP (see our computational studies in Section 5). We also show that these valid inequalities recover the convex hulls for two important special cases, leading to polynomial-sized linear programming (LP) reformulations that are computationally tractable and can be implemented in desktop solvers to benefit practitioners.

1.1. Literature Review

The studies of stochastic appointment scheduling (see, e.g., Gupta and Denton 2008, Erdogan and Denton 2013, Berg et al. 2014) often assume uncertain service durations following fully known distributions. Denton and Gupta (2003) formulate a two-stage stochastic LP model for appointment scheduling and demonstrate that the optimal time intervals allocated in-between appointments form a “dome shape” if the unit idleness costs are high relative to the unit waiting costs. Mittal et al. (2014), Begen and Queyranne (2011), Begen et al. (2012), and Ge et al. (2013) develop approximation algorithms for deriving near-optimal solutions to various stochastic or robust appointment scheduling problems. Pinedo (2012) conducts a comprehensive survey

of various scheduling problems including their models, theories, and applications.

Ho and Lau (1992) are among the first to take into account no-show uncertainty in scheduling problems. They propose a heuristic approach to double-book the first two arrivals and subsequently schedule the remaining appointments. Erdogan and Denton (2013) incorporate no-shows into a stochastic LP model by Denton and Gupta (2003), and also discuss a stochastic dynamic programming variant of the problem. Cayirli and Veral (2003), Hassin and Mendel (2008), Liu et al. (2010), and Robinson and Chen (2010) demonstrate the impact of no-shows on static and dynamic appointment scheduling, and discuss general policies to mitigate negative effects such as system idleness. A number of heuristic policies and approximation algorithms have been proposed to schedule appointments under uncertain no-shows (see, e.g., Muthuraman and Lawley 2008, Zeng et al. 2010, Cayirli et al. 2012, Lin et al. 2011, Luo et al. 2012, LaGanga and Lawrence 2012, Zacharias and Pinedo 2014, Parizi and Ghathe 2016, Kong et al. 2016). To the best of our knowledge, no papers have handled no-shows in a DR framework, which could lead to intractable binary integer programming models due to the discrete nature of 0-1 no-shows (see our computational studies later in Section 5).

In this paper, we assume a fixed sequence of appointment arrivals. We refer to Denton et al. (2007), Gupta and Denton (2008), Mak et al. (2015), Mancilla (2009), Mak et al. (2014), and He et al. (2015) for studies that also involve sequencing decisions, and Denton et al. (2010), Gurvich et al. (2010), and Shylo et al. (2012) for studies that optimize server allocation under random service durations. Deng et al. (2016) and Deng and Shen (2016) analyze integrated models for optimizing server allocation, appointment sequencing, scheduling decisions under service time uncertainty, and formulate chance constraints for restricting server overtime use. For DR appointment scheduling, we refer to Kong et al. (2013), Mak et al. (2015), Kong et al. (2016), and Zhang et al. (2017). For generic DR optimization using moment-based ambiguity sets, we refer to Scarf et al. (1958), Bertsimas and Popescu (2005), Bertsimas et al. (2010), and Delage and Ye (2010).

1.2. Contributions of the Paper

We summarize the main contributions of this paper as follows.

1. Depending on system operators' risk preferences, we formulate DR models that incorporate the worst-case expected/conditional value-at-risk (CVaR) of waiting, idleness, and overtime costs as objective or constraints. Meanwhile, the DR models can flexibly adapt to different prior beliefs of the maximum number of consecutive no-shows, covering from the least conservative case (i.e., no consecutive no-shows) to the most conservative case (i.e., arbitrary no-shows).

2. We develop effective solution approaches for each DR model. The exact reformulations of the DR models result in mixed-integer trilinear programs. We linearize and derive valid inequalities to strengthen the reformulations, which can significantly reduce computational time of solving various instances by using decomposition algorithms. For special no-show beliefs, our derivation leads to polynomial-sized LP reformulations that can readily be implemented in LP solvers.

3. We test diverse instances to show the computational efficacy and demonstrate the performance of DR models under various uncertainties and levels of conservativeness. We provide guidelines for choosing appropriate DR models and no-show beliefs, depending on historical no-show rates, computation budget, and targeted tradeoffs between quality of service and operational cost.

1.3. Structure of the Paper

The remainder of the paper is organized as follows. In Section 2, we formulate the DR expectation/CVaR models and their variants based on different risk preferences. In Section 3, we derive an MINLP of the DR expectation model, as well as valid inequalities for accelerating a generic cutting-plane algorithm. In Section 4, we derive polynomial-sized LP reformulations for special cases of no-show beliefs. In Section 5, we test various instances to demonstrate the computational efficacy and solution performance of different DR models. In Section 6, we summarize the paper and provide future research directions. In the e-companion (EC), we describe models and approaches for problems under a general waiting-time cost and a DR CVaR setting, respectively. We also present all the proofs there.

Notation: The convex hull of a set X is denoted by $\text{conv}(X)$. The abbreviation “w.l.o.g.” represents “without loss of generality.” We follow the convention that $\sum_{k=i}^j a_k = 0$ if $i > j$.

2. Formulations of DR Appointment Scheduling

We consider n appointments arriving at a single server following a fixed order of arrivals given as $1, \dots, n$. Each appointment i has a random service duration s_i . We interpret the possibility of random no-show for appointment i by a 0-1 Bernoulli random variable q_i such that $q_i = 1$ if appointee i shows up, and $q_i = 0$ otherwise. The goal is to optimize the decision of scheduling an arrival time for each appointment, or equivalently, assigning time intervals between appointments i and $i + 1$ for all $i = 1, \dots, n - 1$.

2.1. Modeling Waiting, Idleness, and Overtime Under Uncertainty

Let variable x_i represent the scheduled time interval between appointments i and $i + 1$, $\forall i = 1, \dots, n - 1$.

Under random no-shows and service durations, one or multiple of the following three scenarios can happen: (i) an appointment cannot start on time due to overtime operations of previous appointments, (ii) the server is idle and waiting for the next appointment due to an early finish or no-shows of previous appointments, and (iii) the server cannot finish serving all appointments within a given time limit, denoted by T . For all $i = 1, \dots, n$, let variable w_i represent the waiting time of appointment i , and variable u_i represent the server idle time after finishing appointment i . Also, let variable W represent the server's overtime beyond the fixed time limit T to finish all n appointments. The feasible region of decision x is defined as

$$X = \left\{ x: x_i \geq 0, \forall i = 1, \dots, n, \sum_{i=1}^n x_i = T \right\}, \quad (1)$$

to ensure that we assign nonnegative time in between all consecutive appointments, and appointment n is scheduled to arrive before the end of the service horizon T . The dummy variable $x_n \geq 0$ represents $T - \sum_{i=1}^{n-1} x_i$, i.e., the time from the scheduled start of the last appointment to the server time limit.

Given decision $x \in X$ and a realization of uncertain parameters (q, s) , the appointment waiting time $w = [w_1, \dots, w_n]^T$ and server idleness $u = [u_1, \dots, u_n]^T$ are given by

$$\begin{aligned} w_i &= \max\{0, q_{i-1}s_{i-1} + w_{i-1} - x_{i-1}\}, \quad \text{and} \\ u_{i-1} &= \max\{0, x_{i-1} - q_{i-1}s_{i-1} - w_{i-1}\}, \quad \forall i = 2, \dots, n. \end{aligned} \quad (2)$$

We denote nonnegative parameters c_i^w , c_i^u , and c^o as unit penalty costs of waiting, idleness of appointment i , and server overtime, respectively, which satisfy $c_{i+1}^u - c_i^u \leq c_{i+1}^w$ for all $i = 1, \dots, n - 1$. This assumption is standard (see Denton and Gupta 2003, Ge et al. 2013, Kong et al. 2013, Mak et al. 2015). In fact, if this assumption does not hold and $c_{i+1}^u > c_i^u + c_{i+1}^w$ for some i , then the system operator would rather enforce idleness even if appointment $i + 1$ has arrived and keep it waiting, which is not realistic due to practical concerns. Under this assumption, we can formulate a linear program to compute the total cost of waiting, idleness, and overtime for given x, q, s :

$$Q(x, q, s) := \min_{w, u, W} \left\{ \sum_{i=1}^n (c_i^w w_i + c_i^u u_i) + c^o W \right\} \quad (3a)$$

$$\text{s.t. } w_i - u_{i-1} = q_{i-1}s_{i-1} + w_{i-1} - x_{i-1}, \quad \forall i = 2, \dots, n, \quad (3b)$$

$$W - u_n = q_n s_n + w_n - x_n, \quad (3c)$$

$$\begin{aligned} w_i &\geq 0, w_1 = 0, u_i \geq 0, W \geq 0, \\ \forall i &= 1, \dots, n. \end{aligned} \quad (3d)$$

The objective function (3a) minimizes a linear cost function of the waiting, idleness, and overtime.

Constraints (3b) yield either the waiting time of appointment i or the server's idle time after finishing appointment $i - 1$, both of which will have the same solutions values as given by (2) (see Proposition 1 in Ge et al. 2013). Similarly, constraint (3c) yields either the overtime W or the idle time u_n . Since appointment 1 always arrives at time 0, we have $w_1 = 0$ and all the waiting, idleness, and overtime variables are nonnegative according to constraints (3d).

In (3), note that the waiting-time costs $c_i^w w_i$ are modeled from the perspective of servers (e.g., operating rooms). In particular, we assume that appointment no-shows take place after the server has been set up for serving the appointments. Hence, the waiting-time costs stem from equipment and personnel idleness, as well as from the losses of opportunities of serving other appointments, and they are incurred regardless whether the appointments show up. From the perspective of appointments, the waiting-time costs should be modeled as $c_i^w w_i q_i$, i.e., the waiting-time costs are waived if an appointment does not show up. In this paper, we focus on the DR models and solution methods for the former case, i.e., server-based waiting-time costs. In EC.1, we elaborate how our DR approaches can adapt for a more general setting that incorporates both server-based and appointment-based waiting-time costs.

2.2. Supports and Ambiguity Set

The classical stochastic appointment scheduling approaches seek an optimal $x \in X$ to minimize the expectation of random cost $Q(x, q, s)$ subject to uncertainty (q, s) with a known joint probability distribution denoted as $\mathbb{P}_{q,s}$. We assume that $\mathbb{P}_{q,s}$ is only known belonging to an ambiguity set $\mathcal{F}(D, \mu, \nu)$ that is determined by the support D of (q, s) and the mean values $\mu = [\mu_1, \dots, \mu_n]^T$ and $\nu = [\nu_1, \dots, \nu_n]^T$, where μ_i represents the mean $\mathbb{E}[s_i]$, and ν_i represents the average show-up probability $\mathbb{E}[q_i]$ of appointment i for each $i = 1, \dots, n$. We consider support $D = D_q \times D_s$ where D_q is the support of random no-show parameter q , and D_s is the support of random service duration parameter s . We assume upper and lower bounds of the duration of each appointment i and accordingly

$$D_s := \{s \geq 0: s_i^l \leq s_i \leq s_i^u, \forall i = 1, \dots, n\}.$$

The full support $D_q = \{q: q \in \{0, 1\}^n\}$ contains all no-show scenarios. However, it often leads to overly conservative schedules. In this paper, we parameterize the no-show support by an integer $K \in \{2, \dots, n + 1\}$ such that $D_q = D_q^{(K)}$ rules out consecutive no-shows in any K consecutive appointments. Accordingly,

$$D_q^{(K)} := \left\{ q \in \{0, 1\}^n: \sum_{j=i}^{i+K-1} q_j \geq 1, \forall i = 1, \dots, n - K + 1 \right\}.$$

Note that (i) when $K = 2$, $D_q^{(2)}$ rules out all consecutive no-shows, and (ii) when $K = n + 1$, we have $D_q^{(n+1)} = \{0, 1\}^n = D_q$ as the full support. Also, the parameterized supports $D_q^{(2)} \subset D_q^{(3)} \subset \dots \subset D_q^{(n+1)}$ form a spectrum of conservativeness levels, with $D_q^{(2)}$ being the least conservative and $D_q^{(n+1)}$ being the most general/conservative. In practice, the system operator has the flexibility to select parameter K according to her targeted conservativeness, regardless of whether the ruled-out realizations may still occur. The conservativeness refers to the trade-off between optimality and robustness (see, e.g., Ben-Tal and Nemirovski 2000, Bertsimas and Sim 2004). If we select $K = n + 1$, then support $D_q^{(n+1)} \equiv \{0, 1\}^n$ contains all possible no-show scenarios and so is most robust. Meanwhile, $D_q = D_q^{(n+1)}$ leads to the largest value of $\sup_{\mathbb{P}_{q,s} \in \mathcal{F}(D, \mu, \nu)} \mathbb{E}_{\mathbb{P}_{q,s}}[Q(x, q, s)]$ among all $K \in \{2, \dots, n + 1\}$. In this sense, $D_q^{(n+1)}$ is the most conservative. On the contrary, $D_q^{(2)}$ is the least conservative because it leads to the smallest values of $\mathbb{P}\{q \in D_q^{(K)}\}$ and $\sup_{\mathbb{P}_{q,s} \in \mathcal{F}(D, \mu, \nu)} \mathbb{E}_{\mathbb{P}_{q,s}}[Q(x, q, s)]$ among all $K \in \{2, \dots, n + 1\}$.

Although $D_q^{(K)}$ with $K \neq n + 1$ does not contain all possible no-shows, we can select a value of K such that $\mathbb{P}\{q \in D_q^{(K)}\}$ exceeds a sufficiently high probability. In Section 2.4, we provide a practical and rigorous guideline for how to select the value of K . More importantly, $D_q^{(K)}$ can lead to better out-of-sample performance. For example, the DR schedules given by $D_q^{(2)}$ outperform those using $D_q^{(n+1)}$ in the out-of-sample simulations, in which arbitrary consecutive no-shows may still happen (see Section 5.4).

We specify the ambiguity set $\mathcal{F}(D, \mu, \nu)$ as

$$\mathcal{F}(D, \mu, \nu) := \left\{ \mathbb{P}_{q,s} \geq 0: \begin{cases} \int_{D_q \times D_s} d\mathbb{P}_{q,s} = 1 \\ \int_{D_q \times D_s} s_i d\mathbb{P}_{q,s} = \mu_i, \quad \forall i = 1, \dots, n \\ \int_{D_q \times D_s} q_i d\mathbb{P}_{q,s} = \nu_i, \quad \forall i = 1, \dots, n \end{cases} \right\}, \quad (4)$$

where $\mathbb{P}_{q,s}$ matches the mean values of service durations and no-shows. The ambiguity set $\mathcal{F}(D, \mu, \nu)$ does not incorporate higher moments (e.g., variance and correlations) of service time and no-shows for several reasons. First, with a small amount of data, it is often unclear whether/how the service time and no-shows are correlated. Second, the introduction of higher moments undermines the computational tractability of the DR models, which can be achieved by using $\mathcal{F}(D, \mu, \nu)$ and the solution algorithm derived later. Finally, as we find in the computational study (see Section 5), the DR models based on $\mathcal{F}(D, \mu, \nu)$ already provide near-optimal results, and the benefit of incorporating higher moments is not significant in our case.

2.3. DR Models with Different Risk Measures

We consider DR appointment scheduling models that impose a min–max DR objective and/or DR constraints. Specifically, given $x \in X$, we consider a risk measure ρ of $Q(x, q, s)$ where

(i) a *risk-neutral* system operator sets $\rho(Q(x, q, s)) = \mathbb{E}_{\mathbb{P}_{q,s}}[Q(x, q, s)]$ (i.e., the expected total cost of waiting, idleness, and overtime);

(ii) a *risk-averse* system operator sets $\rho(Q(x, q, s)) = \text{CVaR}_{1-\epsilon}(Q(x, q, s))$ (i.e., the CVaR of the total cost with $1 - \epsilon \in (0, 1)$ confidence).

Then, the DR models impose a generic min–max DR objective in the form

$$\min_{x \in X} \sup_{\mathbb{P}_{q,s} \in \mathcal{F}(D, \mu, \nu)} \rho(Q(x, q, s)) \quad (5a)$$

and/or generic DR constraints in the form

$$\sup_{\mathbb{P}_{q,s} \in \mathcal{F}(D, \mu, \nu)} \rho(Q(x, q, s)) \leq \bar{Q}, \quad (5b)$$

where $\bar{Q} \in \mathbb{R}$ represents a bounding threshold for the risk measure from above. Both DR objective (5a) and constraints (5b) protect the risk measure by hedging against all probability distributions in $\mathcal{F}(D, \mu, \nu)$. A DR model can impose either or both of DR objective (5a) and constraints (5b), and can use either expectation or CVaR as risk measures in (5a)–(5b), i.e., $\rho(Q(x, q, s)) = \mathbb{E}_{\mathbb{P}_{q,s}}[Q(x, q, s)]$ or $\rho(Q(x, q, s)) = \text{CVaR}_{1-\epsilon}(Q(x, q, s))$. Furthermore, the system operator can tune the cost parameters c_i^w , c_i^u , and c^o to let $Q(x, q, s)$ represent different consequences (e.g., performance metric, quality of service, resource opportunity cost, etc.) associated with waiting, idleness, and overtime in (5a)–(5b). For example, by setting $c^o = 1$ and $c_i^w = c_i^u = 0$ for all i , we use

$$\sup_{\mathbb{P}_{q,s} \in \mathcal{F}(D, \mu, \nu)} \text{CVaR}_{1-\epsilon}(Q(x, q, s)) \leq \bar{Q} \quad (6)$$

to constrain the CVaR of overtime W below threshold \bar{Q} . The CVaR constraints provide a safe guarantee on the performance metrics with high probabilities. For this particular cost parameter setting, constraint (6) guarantees that $\inf_{\mathbb{P}_{q,s} \in \mathcal{F}(D, \mu, \nu)} \mathbb{P}_{q,s}\{W \leq \bar{Q}\} \geq 1 - \epsilon$, i.e., the overtime W is controlled under threshold \bar{Q} with the smallest possible probability being no less than $1 - \epsilon$. This provides an appropriate “end-of-the-day” guarantee (see, e.g., Shylo et al. 2012, Wang et al. 2014, Zhang et al. 2015). For presentation brevity, we have relegated the discussions on the CVaR-based model to EC.2.

2.4. Guideline of Selecting Parameter K

In practice, a system operator may evaluate the probability of the random variables $q = (q_1, \dots, q_n)$ belonging to $D_q^{(K)}$, i.e., $\mathbb{P}(q \in D_q^{(K)})$. Then, she can select a value

of K such that $\mathbb{P}(q \in D_q^{(K)})$ exceeds a given threshold such as 90%. To this end, she can gradually increase K from 2 until $\mathbb{P}(q \in D_q^{(K)})$ exceeds the threshold for the first time.

Observation 2.1. $\mathbb{P}(q \in D_q^{(K)}) = 1$ if $K > n$. If $2 \leq K \leq n$ and the components of q are jointly independent, then $\mathbb{P}(q \in D_q^{(K)}) = 1 - [Q^n]_{1, m(K)+1}$, where $m(i) := \frac{1}{2}i(2n - i + 3)$ for $i = 0, \dots, K$ and Q represents a $(m(K)+1) \times (m(K)+1)$ matrix such that (1) $Q_{m(i)+j, i+j+1} = v_{i+j}$ for all $i = 0, \dots, K-1$ and $j = 1, \dots, n-i$; (2) $Q_{m(i)+j, m(i+1)+j} = 1 - v_{i+j}$ for all $i = 0, \dots, K-2$ and $j = 1, \dots, n-i$; (3) $Q_{m(K-1)+j, m(K)+1} = 1 - v_{K-1+j}$ for all $j = 1, \dots, n-K+1$; and (4) $Q_{m(K)+1, m(K)+1} = Q_{m(i), m(i)} = 1$ for all $i = 1, \dots, K$.

Proof of Observation 2.1. If $K > n$, it is clear that we cannot have K consecutive no-shows and so $\mathbb{P}(q \in D_q^{(K)}) = 1$. If $2 \leq K \leq n$, we construct a Markov chain with $m(K)+1$ states, where state $m(i)+j$ represents i consecutive no-shows in the first $i+j-1$ appointments for all $i = 0, \dots, K-1$ and $j = 1, \dots, n-i+1$, and state $m(K)+1$ represents that K consecutive no-shows happen. By construction, matrix Q is the one-step transition matrix of this Markov chain, where states $m(K)+1$ and $m(i)$ for all $i = 1, \dots, K$ are absorbing. Thus, $[Q^n]_{1, m(K)+1}$, representing component $(1, m(K)+1)$ of matrix Q^n , is equal to the probability of having K consecutive no-shows in the n appointments. \square

Using Observation 2.1, the selection of K can be conveniently done in a spreadsheet. In Figure 1, we display an example of $\mathbb{P}(q \in D_q^{(K)})$ with $n = 10$, $K = 1, \dots, 11$, and $v_i = 0.1, \dots$, or 0.9 for all $i = 1, \dots, 10$. We observe that $K = 2$ is sufficient for $\mathbb{P}(q \in D_q^{(K)}) \geq 90\%$ when $v_i \leq 0.1$; i.e., when the no-show probability for each appointment is no greater than 0.1. This observation motivates us to select $D_q = D_q^{(2)}$ for scheduling appointments with low no-show probabilities.

Next, we develop reformulations and solution methods. For presentation brevity, we only analyze DR expectation models in Sections 3 and 4. We present the results of DR CVaR models in EC.2. All the proofs are organized in EC.3 (for DR expectation models) and EC.4 (for DR CVaR models).

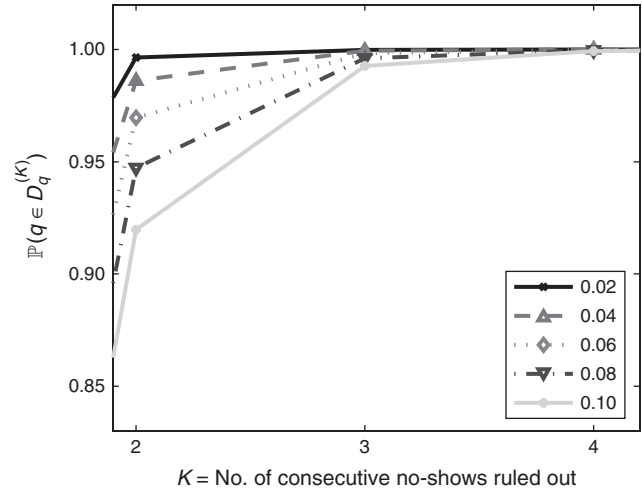
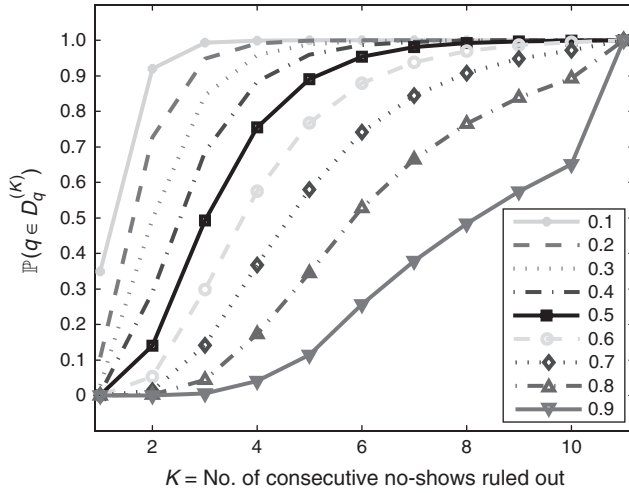
3. Cutting-Plane Approach and Valid Inequalities for DR Expectation Models

We analyze the DR expectation models by specifying a generic objective form (5a) as

$$\min_{x \in X} \sup_{\mathbb{P}_{q,s} \in \mathcal{F}(D, \mu, \nu)} \mathbb{E}_{\mathbb{P}_{q,s}}[Q(x, q, s)], \quad (7)$$

which minimizes the worst-case expected cost of waiting, idleness, and overtime. We first consider the inner maximization problem $\sup_{\mathbb{P}_{q,s} \in \mathcal{F}(D, \mu, \nu)} \mathbb{E}_{\mathbb{P}_{q,s}}[Q(x, q, s)]$ for a fixed $x \in X$, where $\mathbb{P}_{q,s}$ is the decision variable.

Figure 1. An Example of $\mathbb{P}(q \in D_q^{(K)})$ for $n = 10$ Appointments



It can be detailed as a linear functional optimization problem

$$\max_{\mathbb{P}_{q,s} \geq 0} \int_{D_q \times D_s} Q(x, q, s) d\mathbb{P}_{q,s} \quad (8a)$$

$$\text{s.t.} \int_{D_q \times D_s} s_i d\mathbb{P}_{q,s} = \mu_i, \quad \forall i = 1, \dots, n, \quad (8b)$$

$$\int_{D_q \times D_s} q_i d\mathbb{P}_{q,s} = v_i, \quad \forall i = 1, \dots, n, \quad (8c)$$

$$\int_{D_q \times D_s} d\mathbb{P}_{q,s} = 1, \quad (8d)$$

where $D_q = D_q^{(K)}$ for some $K \in \{2, \dots, n+1\}$. Letting ρ_i , γ_i , and θ be dual variables associated with constraints (8b), (8c), and (8d), respectively, we present problem (8) in its dual form as

$$\min_{\rho \in \mathbb{R}^n, \gamma \in \mathbb{R}^n, \theta \in \mathbb{R}} \left\{ \sum_{i=1}^n \mu_i \rho_i + \sum_{i=1}^n v_i \gamma_i + \theta \right\} \quad (9a)$$

$$\text{s.t.} \sum_{i=1}^n s_i \rho_i + \sum_{i=1}^n q_i \gamma_i + \theta \geq Q(x, q, s), \quad \forall (q, s) \in D_q \times D_s. \quad (9b)$$

Here, $\rho = [\rho_1, \dots, \rho_n]^T$, $\gamma = [\gamma_1, \dots, \gamma_n]^T$, and θ are unrestricted, and (9b) are associated with primal variables $\mathbb{P}_{q,s}, \forall (q, s) \in D_q \times D_s$. Under the standard assumptions that μ_i belongs to the interior of set $\{\int_{D_q \times D_s} s_i d\mathbb{Q}_{q,s} : \mathbb{Q}_{q,s} \text{ is a probability distribution over } D_q \times D_s\}$, and that v_i belongs to the interior of set $\{\int_{D_q \times D_s} q_i d\mathbb{Q}_{q,s} : \mathbb{Q}_{q,s} \text{ is a probability distribution over } D_q \times D_s\}$ for each appointment i , strong duality holds between (8) and (9). Furthermore, for a fixed (ρ, γ, θ) , constraints (9b) are equivalent to $\theta \geq \max_{(q,s) \in D_q \times D_s} \{Q(x, q, s) - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i)\}$.

Thus, due to the objective of minimizing θ , the dual formulation (9) is equivalent to

$$\min_{\rho \in \mathbb{R}^n, \gamma \in \mathbb{R}^n} \left\{ \sum_{i=1}^n \mu_i \rho_i + \sum_{i=1}^n v_i \gamma_i + \max_{(q,s) \in D_q \times D_s} \left\{ Q(x, q, s) - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\} \right\}. \quad (10)$$

3.1. MINLP Reformulation and a Generic Cutting-Plane Approach

Note that $Q(x, q, s)$ is a minimization problem, and thus in (10) we have an inner max-min problem. We next analyze the structure of $Q(x, q, s)$ for given solution x and realized value of (q, s) . We formulate $Q(x, q, s)$ in (3) in its dual form as

$$Q(x, q, s) = \max_y \left\{ \sum_{i=1}^n (q_i s_i - x_i) y_i \right\} \quad (11a)$$

$$\text{s.t.} \quad y_{i-1} - y_i \leq c_i^w, \quad \forall i = 2, \dots, n, \quad (11b)$$

$$-y_i \leq c_i^u, \quad \forall i = 1, \dots, n, \quad (11c)$$

$$y_n \leq c^o, \quad (11d)$$

where variable y_{i-1} represents the dual associated with each constraint i in (3b) for all $i = 2, \dots, n$, and variable y_n represents the dual of constraint (3c). Constraints (11b), (11c), and (11d) are related to primal variables w_i , $i = 2, \dots, n$, u_i , $i = 1, \dots, n$, and W in (3), respectively. Therefore, formulation (10) is equivalent to

$$\min_{\rho, \gamma} \left\{ \sum_{i=1}^n \mu_i \rho_i + \sum_{i=1}^n v_i \gamma_i + \max_{(q,s) \in D_q \times D_s} \left\{ Q(x, q, s) - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\} \right\} \quad (12a)$$

$$= \min_{\rho, \gamma} \left\{ \sum_{i=1}^n \mu_i \rho_i + \sum_{i=1}^n v_i \gamma_i + \max_{y \in Y} h(x, y, \rho, \gamma) \right\}, \quad (12b)$$

where Y represents the feasible region of variable y in (11) given by (11b)–(11d), and

$$h(x, y, \rho, \gamma) := \max_{(q, s) \in D_q \times D_s} \left\{ \sum_{i=1}^n (q_i s_i - x_i) y_i - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\}. \quad (12c)$$

The derivation of $h(x, y, \rho, \gamma)$ follows that we can interchange the order of $\max_{(q, s) \in D_q \times D_s}$ and $\max_{y \in Y}$ in (12a). Combining the inner problem in the form of (12b) with the outer minimization problem in (7), we derive a reformulation of the DR expectation model (7) as

$$\min_{x \in X, \rho, \gamma, \delta} \sum_{i=1}^n \mu_i \rho_i + \sum_{i=1}^n \nu_i \gamma_i + \delta \quad (13a)$$

$$\begin{aligned} \text{s.t. } \delta &\geq \max_{y \in Y} h(x, y, \rho, \gamma), \\ &\equiv \max_{y \in Y, (q, s) \in D_q \times D_s} \left\{ \sum_{i=1}^n (q_i s_i - x_i) y_i - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\}. \end{aligned} \quad (13b)$$

Next, we analyze structural properties of $\max_{y \in Y} h(x, y, \rho, \gamma)$ as a function of variables x, ρ , and γ .

Lemma 1. For any fixed values of variables x, ρ , and γ , $\max_{y \in Y} h(x, y, \rho, \gamma) < +\infty$. Furthermore, function $(x, \rho, \gamma) \mapsto \max_{y \in Y} h(x, y, \rho, \gamma)$ is convex and piecewise linear in x, ρ , and γ with a finite number of pieces.

We refer to EC.3.1 for a detailed proof. Lemma 1 indicates that constraint (13b) essentially describes the epigraph of a convex and piecewise linear function of decision variables in model (13). This observation facilitates us applying a separation-based decomposition algorithm to solve model (13) (or equivalently, the DR expectation model (7)), presented in Algorithm 1. This algorithm is finite because we identify a new piece of the function $\max_{y \in Y} h(x, y, \rho, \gamma)$ each time when the set $\{L(x, \rho, \gamma, \delta) \geq 0\}$ is augmented in Step 7, and the function has a finite number of pieces according to Lemma 1.

Algorithm 1 (A decomposition algorithm for solving DR expectation model (7))

- 1: Input: feasible regions X, Y , and $D_q \times D_s$; set of cuts $\{L(x, \rho, \gamma, \delta) \geq 0\} = \emptyset$.
- 2: Solve the master problem

$$\begin{aligned} \min_{x \in X, \rho, \gamma, \delta} \quad & \left\{ \sum_{i=1}^n \mu_i \rho_i + \sum_{i=1}^n \nu_i \gamma_i + \delta \right\} \\ \text{s.t. } \quad & L(x, \rho, \gamma, \delta) \geq 0 \end{aligned}$$

and record an optimal solution $(x^*, \rho^*, \gamma^*, \delta^*)$.

- 3: With (x, ρ, γ) fixed to be (x^*, ρ^*, γ^*) , solve the separation problem

$$\begin{aligned} \max_{y \in Y} h(x, y, \rho, \gamma) \\ \equiv \max_{y \in Y, (q, s) \in D_q \times D_s} \left\{ \sum_{i=1}^n (q_i s_i - x_i) y_i - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\} \end{aligned} \quad (14)$$

and record an optimal solution (y^*, q^*, s^*) .

- 4: if $\delta^* \geq \sum_{i=1}^n (q_i^* s_i^* - x_i^*) y_i^* - \sum_{i=1}^n (\rho_i^* s_i^* + \gamma_i^* q_i^*)$ then
- 5: stop and return x^* as an optimal solution to formulation (7).
- 6: else
- 7: add the cut $\delta \geq \sum_{i=1}^n (q_i^* s_i^* - x_i^*) y_i^* - \sum_{i=1}^n (s_i^* \rho_i + q_i^* \gamma_i)$ to the set of cuts $\{L(x, \rho, \gamma, \delta) \geq 0\}$ and go to Step 2.
- 8: end if.

The main difficulty of the above decomposition algorithm lies in solving the separation problem (14). In general, this problem is a mixed-integer trilinear program because of the integrality restrictions on variables q_i and the trilinear terms $q_i s_i y_i$ in the objective function. This creates obstacles for optimally solving the separation problem if presented in its current form. In Section 3.2, we linearize and reformulate the separation problem (14) as a mixed-integer linear program (MILP) that can readily be solved by optimization solvers. Moreover, we will derive valid inequalities to strengthen this MILP, and test their computational efficiency later.

3.2. MILP Reformulation of the Separation Problem and Valid Inequalities

Our approach is inspired by Mak et al. (2015), where the authors point out that an optimal solution y^* to a similar separation problem but not involving no-shows exists at an extreme point of polyhedron Y . They then successfully decompose the separation problem by appointment for each $i = 1, \dots, n$ and reformulate it by using the extreme points of Y . Different in this paper, for fixed x, ρ , and γ , the separation problem is a mixed-integer trilinear program involving binary variables $q_i, i = 1, \dots, n$. Moreover, except for the case $D_q = D_q^{(n+1)} = \{0, 1\}^n$, $h(x, y, \rho, \gamma)$ is not decomposable by appointment in view of the cross-appointment nature of D_q . Therefore, the approach in Mak et al. (2015) is no longer applicable, and $\max_{y \in Y} h(x, y, \rho, \gamma)$ becomes much more challenging.

Our analysis consists of the following steps. We start by showing the convexity of $h(x, y, \rho, \gamma)$ in variable y . Then, it follows from fundamental convex analysis that maximizing convex function $h(x, y, \rho, \gamma)$ on polyhedron Y will yield an optimal solution at one of the extreme points of Y . Also considering the cost of idleness, we extend the result of extreme-point representation in Mak et al. (2015) and reformulate the separation problem (14) using a polynomial number of binary variables to replace the continuous variables $y_i, i = 1, \dots, n$.

Lemma 2. For fixed x, ρ , and γ , function $h(x, y, \rho, \gamma)$ is convex in variable y .

We refer to EC.3.2 for a proof. According to Lemma 2, an optimal solution y^* to the separation problem (14)

exists at one of the extreme points of Y having linear constraints (11b)–(11d). Consider

$$Y = \{y: c^o \geq y_n \geq -c_n^u, y_n + c_n^w \geq y_{n-1} \geq -c_{n-1}^u, \dots, y_2 + c_2^w \geq y_1 \geq -c_1^u\}. \quad (15)$$

It can be observed that any extreme point \hat{y} of Y satisfy (i) either $\hat{y}_n = -c_n^u$ or $\hat{y}_n = c^o$, and (ii) for all $i = 1, \dots, n-1$, dual constraint $\hat{y}_{i+1} + c_{i+1}^w \geq \hat{y}_i \geq -c_i^u$ is binding at either the lower bound or the upper bound.

This observation motivates us to establish an alternative formulation of (14) using new binary variables. For notation convenience, we define a dummy variable y_{n+1} , which always takes the lower-bound value $-c_{n+1}^u := 0$. There is a one-to-one correspondence between an extreme point of Y and a partition of the integers $1, \dots, n+1$ into intervals. For each interval $\{k, \dots, j\} \subseteq \{1, \dots, n+1\}$ in the partition, y_j takes on the lower bound value $-c_j^u$ and other y_i equal to their upper bounds, i.e., $y_i = y_{i+1} + c_{i+1}^w, \forall i = k, \dots, j-1$. As a result, for each interval $\{k, \dots, j\}$ in the partition and $i \in \{k, \dots, j\}$, the value of y_i is given by

$$y_i = \pi_{ij} := \begin{cases} -c_j^u + \sum_{\ell=i+1}^j c_\ell^w & 1 \leq i \leq j \leq n, \\ c^o + \sum_{\ell=i+1}^n c_\ell^w & 1 \leq i \leq n, j = n+1, \end{cases} \quad (16)$$

and $y_{n+1} = \pi_{n+1, n+1} := 0$. Define binary variables t_{kj} for all $1 \leq k \leq j \leq n+1$, such that $t_{kj} = 1$ if interval $\{k, \dots, j\}$ belongs to the partition (i.e., $t_{kj} = 1$ if $y_i = \pi_{ij}$) and $t_{kj} = 0$ otherwise. For a valid partition, we require each index i belonging to exactly one interval, and thus $\sum_{k=1}^i \sum_{j=i}^{n+1} t_{kj} = 1, \forall i = 1, \dots, n+1$. For notation convenience, we define $x_{n+1} = q_{n+1} = s_{n+1} := 0$. Using binary variables t_{kj} , we reformulate the separation problem (14) as

$$\max_t \max_{(q,s) \in D_q \times D_s} \left\{ \sum_{k=1}^{n+1} \sum_{j=k}^{n+1} \left(\sum_{i=k}^j (q_i s_i - x_i) \pi_{ij} \right) t_{kj} - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\} \quad (17a)$$

$$\text{s.t. } \sum_{k=1}^i \sum_{j=i}^{n+1} t_{kj} = 1, \quad \forall i = 1, \dots, n+1, \quad (17b)$$

$$t_{kj} \in \{0, 1\}, \quad \forall 1 \leq k \leq j \leq n+1. \quad (17c)$$

Note that the objective function (17a) contains trilinear terms $q_i s_i t_{kj}$ with binary variables q_i, t_{kj} and continuous variables s_i . To linearize formulation (17), we define $p_{ikj} \equiv q_i t_{kj}$ and $o_{ikj} \equiv q_i s_i t_{kj}$ for all $1 \leq k \leq j \leq n+1$ and $k \leq i \leq j$. Also, we introduce the following McCormick inequalities (18a)–(18b) and (18c)–(18d) for variables p_{ikj} and o_{ikj} , respectively.

$$p_{ikj} - t_{kj} \leq 0, \quad (18a)$$

$$p_{ikj} - q_i \leq 0, \quad p_{ikj} - q_i - t_{kj} \geq -1, \quad p_{ikj} \geq 0, \quad (18b)$$

$$o_{ikj} - s_i^l p_{ikj} \geq 0, \quad o_{ikj} - s_i^u p_{ikj} \leq 0, \quad (18c)$$

$$o_{ikj} - s_i + s_i^l (1 - p_{ikj}) \leq 0, \quad o_{ikj} - s_i + s_i^u (1 - p_{ikj}) \geq 0. \quad (18d)$$

Thus, the separation problem (14) is equivalent to an MILP as

$$\max_{t,q,s,p,o} \left\{ \sum_{k=1}^{n+1} \sum_{j=k}^{n+1} \sum_{i=k}^j (\pi_{ij} o_{ikj} - x_i \pi_{ij} t_{kj}) - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\} \quad (19a)$$

$$\text{s.t. (17b)–(17c), (18a)–(18d),} \quad (19b)$$

$$s_i \in [s_i^l, s_i^u], q \in D_q \subseteq \{0, 1\}^n. \quad (19c)$$

We can replace Steps 3–8 of Algorithm 1 proposed in Section 3 based on this MILP reformulation:

- 3: With (x, ρ, γ) fixed to be (x^*, ρ^*, γ^*) , solve formulation (19) and record an optimal solution $(t^*, q^*, s^*, p^*, o^*)$.
- 4: if $\delta^* \geq \sum_{k=1}^{n+1} \sum_{j=k}^{n+1} \sum_{i=k}^j (\pi_{ij} o_{ikj}^* - x_i^* \pi_{ij} t_{kj}^*) - \sum_{i=1}^n (\rho_i^* s_i^* + \gamma_i^* q_i^*)$ then
- 5: stop and return x^* as an optimal solution to formulation (7).
- 6: else
- 7: add the cut $\delta \geq \sum_{k=1}^{n+1} \sum_{j=k}^{n+1} \sum_{i=k}^j (\pi_{ij} o_{ikj}^* - \pi_{ij} t_{kj}^* x_i) - \sum_{i=1}^n (s_i^* \rho_i + q_i^* \gamma_i)$ to the set of cuts $\{L(x, \rho, \gamma, \delta) \geq 0\}$ and go to Step 2.
- 8: end if.

Remark 1. We note that Algorithm 1 applies to various types of no-show support D_q . For example, we can specify $D_q = \{q \in \{0, 1\}^n: \sum_{i=1}^n (1 - q_i) \leq Q_{\max}\}$, where Q_{\max} represents the maximum number of no-shows. In this case, we only need to replace the definition of D_q in (19c) when applying Algorithm 1. In fact, Algorithm 1 is general regardless of the specific form of set D_q , to select which we take into account the operator's beliefs and/or preferences, and the computational tractability. In this paper, we specify $D_q = D_q^{(K)}$ due to its flexibility (see Section 2.4) and computational tractability (see Proposition 1 and Theorem 2).

We further identify a set of valid inequalities to strengthen formulation (19). We summarize the valid inequalities in the following proposition and delegate its proof in EC.3.3. The inequalities (20a)–(20f) can be added to the MILP (19) solved in Step 3, to strengthen the reformulation.

Proposition 1. The following inequalities are valid for set $F = \{(t, q, s, p, o): (19b)–(19c)\}$:

$$\sum_{k=1}^i \sum_{j=i}^{n+1} p_{ikj} = q_i, \quad \forall i = 1, \dots, n+1, \quad (20a)$$

$$s_i - \sum_{k=1}^i \sum_{j=i}^{n+1} (o_{ikj} - s_i^l p_{ikj}) \geq s_i^l, \quad \forall 1 \leq i \leq n+1, \quad (20b)$$

$$s_i - \sum_{k=1}^i \sum_{j=i}^{n+1} (o_{ikj} - s_i^u p_{ikj}) \leq s_i^u, \quad \forall 1 \leq i \leq n+1, \quad (20c)$$

$$\sum_{\ell=i}^{i+K-1} p_{\ell kj} \geq t_{kj}, \quad \forall 1 \leq k < j \leq n+1, \forall k \leq i \leq j-K+1, \quad (20d)$$

$$\sum_{k=1}^{i-K+2} \sum_{\ell=i-K+2}^i p_{\ell ki} + \sum_{j=i+1}^{n+1} p_{(i+1)(i+1)j} \geq \sum_{k=1}^{i-K+2} t_{ki}, \quad \forall i=K-1, \dots, n, \quad (20e)$$

$$\sum_{k=1}^i p_{iki} + \sum_{\ell=i+1}^{i+K-1} \sum_{j=i+K-1}^{n+1} p_{\ell(i+1)j} \geq \sum_{j=i+K-1}^{n+1} t_{(i+1)j}, \quad \forall i=1, \dots, n-K+2. \quad (20f)$$

Remark 2. Note that the above inequalities hold valid for all $K = 2, \dots, n+1$. We also note two features that (i) valid inequalities (20a)–(20f) are *polynomially* many, and (ii) all coefficients of these inequalities are in *closed-form*. Features (i) and (ii) can significantly accelerate Algorithm 1, because Feature (i) ensures that model (14) (i.e., (19) after reformulation) remains small by incorporating these inequalities, and Feature (ii) implies that we do not need to separate these inequalities.

4. LP Reformulations of the DR Expectation Model

In this section, we present tractable reformulations of the DR expectation model (7) as we derive the convex hull of separation problem (14) for $D_q = D_q^{(2)}$ (i.e., no conservative no-shows) and $D_q = D_q^{(n+1)}$ (i.e., arbitrary no-shows). This leads to polynomial-sized LP reformulations of model (7). We note that these LP reformulations are derived for these the special cases and do not simplify the general model. For the general cases (i.e., $D_q = D_q^{(K)}$ with $3 \leq K \leq n$), we can apply Algorithm 1 to solve model (7) and obtain a globally optimal appointment schedule.

Case 1. (No Consecutive No-Shows) Recall that F represents the mixed-integer feasible region of formulation (19), i.e., $F = \{(t, q, s, p, o) : (19b)–(19c)\}$. We show that the valid inequalities identified in Proposition 1 are sufficient to describe $\text{conv}(F)$. We first notice that when $K=2$: (i) inequalities (20d) are equivalent to $p_{ikj} + p_{(i+1)kj} \geq t_{kj}$ for all $1 \leq k < j \leq n+1$ and $k \leq i \leq j-1$, and (ii) inequalities (20e) and (20f) are identical and equivalent to

$$\sum_{k=1}^i p_{iki} + \sum_{j=i+1}^{n+1} p_{(i+1)(i+1)j} \geq \sum_{k=1}^i t_{ki}, \quad \forall i=1, \dots, n. \quad (21)$$

This leads to the following theorem, of which a proof is relegated to EC.3.4.

Theorem 1. Polyhedron $CF := \{(t, q, s, p, o) : (17b), (18a), (18c), (20a)–(20d), (21)\}$ is the convex hull of set F , i.e., $CF = \text{conv}(F)$.

Therefore, we can reformulate the separation problem (14) as an LP model:

$$\begin{aligned} \max_{t, q, s, p, o} \quad & \left\{ \sum_{k=1}^{n+1} \sum_{j=k}^{n+1} \sum_{i=k}^j (\pi_{ij} o_{ikj} - x_i \pi_{ij} t_{kj}) - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\} \\ \text{s.t.} \quad & (t, q, s, p, o) \in CF. \end{aligned}$$

To combine the separation problem with the outer minimization problem in (13), we present the above reformulation in its dual form:

$$\min \sum_{i=1}^{n+1} (\alpha_i + s_i^U \tau_i^U - s_i^L \tau_i^L) \quad (22a)$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{i=k}^j (\alpha_i - \sigma_{ikj}) + \sum_{i=k}^{j-1} \lambda_{ikj} + \sum_{i=j}^{\min\{j, n\}} \phi_i \geq - \sum_{i=k}^j \pi_{ij} x_i, \\ & \forall 1 \leq k \leq j \leq n+1, \quad (22b) \end{aligned}$$

$$\zeta_i \leq \gamma_i, \quad \forall 1 \leq i \leq n, \quad (22c)$$

$$\tau_i^L - \tau_i^U \leq \rho_i, \quad \forall 1 \leq i \leq n, \quad (22d)$$

$$\begin{aligned} & \sigma_{ikj} + s_i^L \varphi_{ikj}^L - s_i^U \varphi_{ikj}^U + \zeta_i - s_i^L \tau_i^L + s_i^U \tau_i^U \\ & - \sum_{\ell=\max\{k, i-1\}}^{\min\{j-1, i\}} \lambda_{\ell kj} - \sum_{\ell=\min\{2i-k-1, i\} \vee 1}^{\max\{2i-j, i-1\} \wedge n} \phi_\ell \geq 0, \end{aligned}$$

$$\forall 1 \leq k \leq j \leq n+1, \forall k \leq i \leq j, \quad (22e)$$

$$-\varphi_{ikj}^L + \varphi_{ikj}^U + \tau_i^L - \tau_i^U \geq \pi_{ij},$$

$$\forall 1 \leq k \leq j \leq n+1, \forall k \leq i \leq j, \quad (22f)$$

$$\varphi_{ikj}^L, \varphi_{ikj}^U, \tau_i^L, \tau_i^U, \lambda_{ikj}, \phi_i, \sigma_{ikj} \geq 0,$$

$$\forall 1 \leq k \leq j \leq n+1, \forall k \leq i \leq j, \quad (22g)$$

where we denote $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$ for $a, b \in \mathbb{R}$ for notation convenience. Here, the dual variables α_i , σ_{ikj} , $\varphi_{ikj}^{L/U}$, ζ_i , $\tau_i^{L/U}$, λ_{ikj} , and ϕ_i are associated with constraints (17b), (18a), (18c), (20a)–(20d), and (21), respectively (after transforming all “ \geq ” inequalities into the “ \leq ” form), and constraints (22b)–(22f) are associated with primal variables t_{kj} , q_i , s_i , p_{ikj} , and o_{ikj} , respectively. In (22b), the term $\sum_{i=j}^{\min\{j, n\}} \phi_i$ becomes ϕ_j for all $1 \leq j \leq n$, and will disappear for $j = n+1$. In (22e), when $k \leq i < j$, the term $-\sum_{\ell=\max\{k, i-1\}}^{\min\{j-1, i\}} \lambda_{\ell kj}$ becomes $-\lambda_{ikj} - \lambda_{(i-1)kj}$; when $k < i = j$, it becomes a singleton $-\lambda_{(i-1)kj}$; and when $k = i = j$, it does not appear. Similarly, when $2 \leq k = i = j \leq n$, the term $-\sum_{\ell=\min\{2i-k-1, i\} \vee 1}^{\max\{2i-j, i-1\} \wedge n} \phi_\ell$ becomes $-\phi_i - \phi_{i-1}$; when $j > i = k$ or $k = i = j = n+1$, the term only contains $-\phi_{i-1}$; when $k < i = j$ or $1 = k = i = j$, the term only contains $-\phi_i$; and in all other cases (i.e., when $1 \leq k < i < j \leq n+1$), the term does not appear. We can then reformulate the DR expectation model in an LP form as follows.

Theorem 2. Under no-consecutive no-show assumption (i.e., $D_q = D_q^{(2)}$), the DR expectation model (7) is equivalent to the following linear program:

$$\min \left\{ \sum_{i=1}^n \mu_i \rho_i + \sum_{i=1}^n \nu_i \gamma_i + \sum_{i=1}^{n+1} (\alpha_i + s_i^U \tau_i^U - s_i^L \tau_i^L) \right\}$$

$$\text{s.t. (22b)–(22g), } \sum_{i=1}^n x_i = T, x_{n+1} = 0, x_i \geq 0, \\ \forall i = 1, \dots, n.$$

Case 2. (Arbitrary No-Shows): Given $D_q = \{0, 1\}^n$ and $D_s = \prod_{i=1}^n [s_i^L, s_i^U]$, the optimization problem defining function $h(x, y, \rho, \gamma)$ (see (12c)) is separable by each appointment—i.e.,

$$\begin{aligned} h(x, y, \rho, \gamma) &= \max_{(q, s) \in D_q \times D_s} \left\{ \sum_{i=1}^n (q_i s_i - x_i) y_i - \sum_{i=1}^n (\rho_i s_i + \gamma_i q_i) \right\} \\ &= \sum_{i=1}^n \max_{q_i \in \{0, 1\}, s_i \in [s_i^L, s_i^U]} \{ (q_i s_i - x_i) y_i - (\rho_i s_i + \gamma_i q_i) \}. \end{aligned}$$

To reformulate separation problem (14), recall the observations on polyhedron Y in Section 3.2, and again we represent the extreme points of Y based on variables t_{kj} . It follows that

$$\begin{aligned} \max_{y \in Y} h(x, y, \rho, \gamma) &= \max_{t \geq 0} \sum_{k=1}^{n+1} \sum_{j=k}^{n+1} \left(\sum_{i=k}^j \max_{q_i \in \{0, 1\}, s_i \in [s_i^L, s_i^U]} \{ (q_i s_i - x_i) \pi_{ij} \right. \\ &\quad \left. - (\rho_i s_i + \gamma_i q_i) \} \right) t_{kj} \quad (23a) \end{aligned}$$

$$\text{s.t. } \sum_{k=1}^i \sum_{j=i}^{n+1} t_{kj} = 1, \quad \forall i = 1, \dots, n+1, \quad (23b)$$

$$t_{kj} \in \{0, 1\}, \quad \forall 1 \leq k \leq j \leq n+1. \quad (23c)$$

Because the constraint matrix formed by (23b)–(23c) is totally unimodular (TU), we can relax the integrality constraints (23c) without loss of optimality. Hence, formulation (23a)–(23c) is an LP model in variables t_{kj} and we can take its dual as

$$\min_{\alpha, \beta} \sum_{i=1}^{n+1} \alpha_i \quad (24a)$$

$$\text{s.t. } \sum_{i=k}^j \alpha_i \geq \sum_{i=k}^j \beta_{ij}, \quad \forall 1 \leq k \leq j \leq n+1, \quad (24b)$$

$$\beta_{ij} \geq \max_{q_i \in \{0, 1\}, s_i \in [s_i^L, s_i^U]} \{ (q_i s_i - x_i) \pi_{ij} - (\rho_i s_i + \gamma_i q_i) \}, \\ \forall i = 1, \dots, n, \forall j = i, \dots, n+1, \quad (24c)$$

$$\beta_{n+1, n+1} = 0, \quad (24d)$$

where dual variables α_i , $i = 1, \dots, n+1$ are associated with constraints (23b), constraints (24b) are associated with variables y_{kj} , each variable β_{ij} represents the value of $\max_{q_i \in \{0, 1\}, s_i \in [s_i^L, s_i^U]} \{ (q_i s_i - x_i) \pi_{ij} - (\rho_i s_i + \gamma_i q_i) \}$, and $\beta_{n+1, n+1} = 0$ because $q_{n+1} = s_{n+1} = \pi_{n+1, n+1} = 0$. Finally, for each $i = 1, \dots, n$, the related objective function

$(q_i s_i - x_i) \pi_{ij} - (\rho_i s_i + \gamma_i q_i)$ is linear in variables q_i and s_i , and thus each of constraints (24c) is equivalent to

$$\beta_{ij} \geq -\pi_{ij} x_i - s_i^L \rho_i, \quad (25a)$$

$$\beta_{ij} \geq -\pi_{ij} x_i - s_i^U \rho_i, \quad (25b)$$

$$\beta_{ij} \geq -\pi_{ij} x_i - s_i^L \rho_i - \gamma_i + s_i^L \pi_{ij}, \quad (25c)$$

$$\beta_{ij} \geq -\pi_{ij} x_i - s_i^U \rho_i - \gamma_i + s_i^U \pi_{ij}, \quad (25d)$$

because $q_i \in \{0, 1\}$ and $s_i \in [s_i^L, s_i^U]$ at optimality. It follows that model (13) (i.e., the DR expectation model (7)), is equivalent to the following LP model when $D_q = D_q^{(n+1)}$:

$$\begin{aligned} \min_{x, \rho, \gamma, \alpha, \beta} \quad & \left\{ \sum_{i=1}^n \mu_i \rho_i + \sum_{i=1}^n v_i \gamma_i + \sum_{i=1}^{n+1} \alpha_i \right\} \\ \text{s.t. (24b), (24d), (25a)–(25d), } & \sum_{i=1}^n x_i = T, x_{n+1} = 0, \\ & x_i \geq 0, \quad \forall i = 1, \dots, n. \end{aligned}$$

5. Computational Results

We conduct numerical experiments on the three variants of the DR expectation model (7) (namely, E- $D_q^{(2)}$, E- $D_q^{(n+1)}$, and E- $D_q^{(K)}$ ($K = 3, \dots, n$)) yielded by $D_q = D_q^{(2)}$, $D_q^{(n+1)}$, and $D_q^{(K)}$, respectively. For benchmark, we also solve a stochastic linear program (SLP) that minimizes the expected total cost of waiting, server idleness, and overtime via the sample average approximation (SAA) approach (Kleywegt et al. 2002). We briefly describe the key computational procedures as follows. First, we follow a distribution belief to generate N i.i.d. samples, of which we randomly pick a small subset of data to compute the empirical mean and support information, and use them to compute the (in-sample) optimal solutions and optimal objective values to the DR models.

For the SLP, we solve an LP model:

$$[\text{SLP}] \quad \min_{x, w, u, W} \quad \frac{1}{N} \sum_{m=1}^N \sum_{i=1}^n (c_i^w w_i^m + c_i^u u_i^m) + c^o W^m \quad (26a)$$

$$\text{s.t. } w_i^m - u_{i-1}^m = q_{i-1}^m s_{i-1}^m + w_{i-1}^m - x_{i-1}, \\ \forall i = 2, \dots, n, m = 1, \dots, N, \quad (26b)$$

$$W^m - u_n^m = q_n^m s_n^m + w_n^m + \sum_{i=1}^{n-1} x_i - T, \\ \forall m = 1, \dots, N, \quad (26c)$$

$$\sum_{i=1}^{n-1} x_i \leq T, \quad (26d)$$

$$x_i \geq 0, \quad \forall i = 1, \dots, n-1, \quad (26e)$$

$$w_i^m \geq 0, w_1^m = 0, u_i^m \geq 0, W^m \geq 0, \\ \forall i = 1, \dots, n, m = 1, \dots, N, \quad (26f)$$

where q_i^m and s_i^m are realizations of parameter q_i and s_i of appointment i in scenario m , respectively, for all $i = 1, \dots, n$ and $m = 1, \dots, N$. Variables w_i^m, u_i^m , and W^m represent recourse waiting time of appointment i , server idle time after serving appointment i , and server overtime in scenario m , respectively, for all $m = 1, \dots, N$. Constraints (26b) and (26c) obtain the waiting time/idle time/overtime values for each appointment dependent on values of x_i and (q_i^m, s_i^m) .

Section 5.1 describes how to set the parameter for the above models; Section 5.2 compares the CPU time and details of solving $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, $E-D_q^{(K)}$, and SLP. In Section 5.3, we illustrate optimal objective values given by $E-D_q^{(K)}$ with $K = 2, \dots, n$ for different settings of time limit T and no-show probabilities. In Section 5.4, we compare the performance of optimal schedules of $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP via out-of-sample simulation tests. Specifically, we follow a certain distribution to generate N' data samples, which represent realizations of random service durations and no-shows. The distributions used for generating the in-sample and out-of-sample data could be different, and when they are the same and N is sufficiently large, the SLP is considered being optimized under the “perfect information” (Birge and Louveaux 2011). In reality, it is hard to know the exact true distribution, and thus we also test the case where the distribution is “misspecified.”

5.1. Experiment Setup

We follow procedures in the appointment scheduling literature (e.g., Denton and Gupta 2003, Mak et al. 2015) to generate random instances as follows. For most instances, we consider $n = 10$ appointments, each having a random service duration s_i with the mean $\mu_i \sim U[36, 44]$ (i.e., uniformly sampled in between the values below and above 10% of 40 minutes) and the standard deviation $\sigma_i = 0.5\mu_i$. We set $T = \sum_{i=1}^n \mu_i + R \cdot \sqrt{\sum_{i=1}^n \sigma_i^2}$, where scalar R adjusts the length of time limit T . Note that in this setting, T does not take into consideration the no-show probabilities. In Section 5.5, we report the results when T does depend on the no-show probabilities. Each appointment i has a probability v_i of showing up, and we test $v_1 = \dots = v_n = 0.8$ or $v_1 = \dots = v_n = 0.6$. To approximate the upper and lower bounds s_i^U and s_i^L of each service duration s_i , we respectively use the 80%- and 20%-quantile values of the N in-sample data. We set the ratio $c_i^w:c_i^u:c_i^o = 1:0.5:10$ in all of the DR models and in SLP.

We sample $N=1,000$ realizations $(q_1^m, s_1^m), \dots, (q_n^m, s_n^m)$, $m = 1, \dots, N$ by following log-normal distributions with the given means and standard deviations of s_i and probability $1 - v_i$ of no-shows for each $i = 1, \dots, n$. (The log-normal distribution possesses the long-tail property and has been shown accurately describing the shape of service-time distributions in many service systems (see, e.g., Gul et al. 2011, for a study of five-year

outpatient surgical data in Mayo Clinic).) We optimize the SLP model by using all of the N data points, and only use 20 randomly picked data samples from the N -data set to calculate the first moments of service durations and no-shows, used in all the DR models. Given the optimal schedules produced by different models, we generate $N' = 10,000$ i.i.d. data samples from certain distributions with details given in Section 5.4, to evaluate the performance of each schedule.

We increase the size of instances with $n = 10, 15, \dots, 50$ appointments in Section 5.2 to compare the CPU time of different models and approaches. For each instance, we generate $N = 1,000$ i.i.d. data samples of service durations and no-shows by following the same procedures as above. All LP (i.e., $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP) and MILP (i.e., $E-D_q^{(K)}$ with $K = 3, \dots, n$) models are computed in Python 2.7.10 using Gurobi 5.6.3. The computations are performed on a Windows 7 machine with Intel Core i7-2600M CPU 3.40 GHz and 8 GB of memory. The CPU time limit is set as three hours for solving each instance.

5.2. CPU Time and Computational Details

In this section, we increase the problem size from $n = 10$ to $n = 50$ appointments, and compare CPU time of optimizing different DR models and the SLP. We first vary $R = 0, 0.25, 0.5, 0.75, 1$, and find that the CPU time of all the models are similar for different R values and no-show probabilities. Thus, we fix $R = 0$ and use $T = \sum_{i=1}^n \mu_i = 380.24$ minutes. We consider $v_i = 0.6$ for all appointments $i = 1, \dots, n$, and test 10 replications for each instance. Table 1 reports the average CPU time (in second) of solving models $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, SLP, and $E-D_q^{(K)}$ with $K = 0.3n$ and $K = 0.7n$. Note that the first three are LP models, and specifically, there are $\mathcal{O}(n^3)$ variables and $\mathcal{O}(n^3)$ constraints in the two LP models $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$, but $\mathcal{O}(nN)$ variables and constraints in SLP. The $E-D_q^{(0.3n)}$ and $E-D_q^{(0.7n)}$ models are solved via Algorithm 1, and we present the average time for solving the MILP models with (see columns “Ineq.”) and without (see columns “W/O”) the valid inequalities in Proposition 1. For instances that take longer than three hours to solve, we instead report the optimality gap values (in %) achieved at the end of the computation process.

In Table 1, the CPU time of both $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ are shorter than the one of SLP, especially after $n \geq 40$. The time for solving $E-D_q^{(2)}$ is longer than solving $E-D_q^{(n+1)}$, due to the many more constraints involved in the former. Note that the time presented in Table 1 is only for solving all the models but does not include the time spent on reading in data and constructing the constraints, which is negligible for all the DR models but grows quickly for SLP, ranging from 30 seconds to 60 seconds when $n \geq 35$.

Table 1. Average CPU Time (in Second) of Solving DR Models and SLP with $R = 0$ and $1 - v_i = 0.4$

n	$E-D_q^{(2)}$	$E-D_q^{(n+1)}$	SLP	$E-D_q^{(0.3n)}$		$E-D_q^{(0.7n)}$	
				W/O	Ineq.	W/O	Ineq.
10	0.03	0.00	3.10	6.70	10.68	3.44	1.95
15	0.16	0.01	7.24	52.74	47.56	20.28	4.23
20	0.38	0.01	10.44	158.50	106.67	72.96	15.12
25	2.85	0.02	16.97	409.88	270.60	266.34	47.18
30	5.12	0.05	20.91	1,000.38	187.81	823.14	101.68
35	11.50	0.05	28.18	10,658.06	401.91	7,994.79	340.76
40	28.87	0.15	32.32	(5.76%)	808.38	(10.98%)	614.75
45	26.55	0.20	39.07	(12.49%)	1,739.49	(10.24%)	1,491.17
50	24.64	0.45	44.49	(31.77%)	3,271.83	(43.63%)	3,393.62

All the DR LP models and SLP are efficiently solved for $n = 10, \dots, 50$, while the MILP models $E-D_q^{(K)}$ with either small or large K -values are computationally intractable, reflected by the long CPU seconds taken by instances with $n = 35, 40, 45, 50$, especially when no valid inequalities were added. The addition of valid inequalities in Proposition 1 drastically speeds up the decomposition algorithm, and the effect is much more significant when $n \geq 35$. For instance, none of the $n = 40, 45, 50$ cases were solved within three hours without the valid inequalities, and the average optimality gaps could be as large as 30%~45% when $n = 50$. In contrast, after adding the valid inequalities, the decomposition algorithm quickly converges, and on average, it only takes no more than 15, 30, and 60 minutes to optimize the MILPs over instances with $n = 40, 45$, and 50, respectively.

Next, we present more details of solving the $E-D_q^{(K)}$ MILPs. Table 2 illustrates the number of constraints (“# of Cons.”), the total number of branching nodes (“# of Nodes”), the average CPU seconds taken by the master problem and the subproblem in each iteration, and the number of iterations in the decomposition algorithm before it converges to the optimum or reaches the time limit (“# of Cuts”) for solving both $E-D_q^{(0.3n)}$ and $E-D_q^{(0.7n)}$, with or without the valid inequalities.

In Table 2, we observe that the valid inequalities in Proposition 1 slightly increase the number of constraints but significantly tighten the MILPs, directly reflected by the significantly reduced branching-and-bound nodes in all the instances. In particular, the decomposition algorithm obtains integer solutions at the root node in each iteration for solving $E-D_q^{(0.3n)}$ when $n \geq 35$, and for $E-D_q^{(0.7n)}$ given any values of n we test. Moreover, the valid inequalities significantly reduce the CPU seconds of computing the separation subproblem in each iteration, especially for instances with $n = 35, 40, 45, 50$. Lastly, by adding the valid inequalities to the MILP models, the decomposition algorithm takes almost constant number of iterations

to converge (i.e., being around 100 iterations when $n \geq 20$). However, if no valid inequalities were added, the number of iterations first increases as we increase n from 10 to 35, and drastically decreases as we continue increasing n to 50.

5.3. Optimal Objective Values and Scheduling Solution Patterns

We compare the optimal objective values of the $E-D_q^{(K)}$ models with $K = 2, \dots, n + 1$, and plot their value changes for instances with $R = 0, 0.25, 0.5, 0.75, 1$ and no-show probabilities $1 - v_i = 0.2, 0.4$ for all $i = 1, \dots, n$ when $n = 10$. Recall that parameter K represents the minimum number of consecutive appointments in which consecutive no-shows are ruled out. Therefore, as K increases, the support $D_q^{(K)}$ becomes larger, which leads to smaller feasible region for the scheduling decision vector x , and thus the optimal objective value is nondecreasing for $K = 2, \dots, n + 1$. Figure 2 illustrates the optimal objective values, in which Figure 2(a) corresponds to $1 - v_i = 0.2, \forall i = 1, \dots, n$, and Figure 2(b) corresponds to larger no-show probabilities $1 - v_i = 0.4, \forall i = 1, \dots, n$.

Table 3 presents the detailed optimal objective values in Figure 2. Note that the optimal objective values of $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$, respectively, provide valid lower and upper bounds for the optimal objective value of any $E-D_q^{(K)}$ models with $K = 3, \dots, n$. For each combination of R and $1 - v_i$, we mark the first K -value when the optimal objective value of $E-D_q^{(K)}$ equals to the upper bound, i.e., the optimal objective value of $E-D_q^{(n+1)}$.

In Table 3, when the no-show probability is smaller (i.e., $1 - v_i = 0.2, \forall i = 1, \dots, n$), the differences between the upper and lower bounds are very small, indicating that the two LP models $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ can already provide tight approximations for the MILPs of $E-D_q^{(K)}$ with $K = 3, \dots, n$. Considering the long CPU time of solving the MILPs in Table 1, one can avoid directly solving $E-D_q^{(K)}$ ($K = 3, \dots, n$), and instead use $K = 2$ or

Table 2. Computational Details of Solving the MILP Models $E-D_q^{(0.3n)}$ and $E-D_q^{(0.7n)}$

Model	n	With ineq.					W/O ineq.					# of cuts
		# of cons.	# of nodes	Avg. time (s)		# of cuts	# of cons.	# of nodes	Avg. time (s)			
				Master	Sub				Master	Sub		
$E-D_q^{(0.3n)}$	10	2,239	790	0.00	0.22	48	2,023	1,810	0.00	0.13	49	
	15	6,271	1,363	0.01	0.61	77	5,742	8,510	0.01	0.62	83	
	20	13,346	2,572	0.03	1.04	99	12,435	19,112	0.03	1.52	102	
	25	24,637	2,865	0.08	2.38	110	22,979	31,602	0.11	3.83	104	
	30	40,686	32	0.23	2.00	84	38,247	54,697	0.22	9.04	108	
	35	62,932	0	0.51	3.95	90	59,116	1,245,441	0.53	86.12	123	
	40	91,602	0	1.05	7.28	97	86,459	677,713	2.63	632.67	17	
	45	128,499	0	1.60	16.52	96	121,153	387,066	7.30	1,342.76	8	
	50	173,437	0	3.09	27.48	107	164,071	324,952	17.57	3,582.45	3	
$E-D_q^{(0.7n)}$	10	2,097	0	0.01	0.06	29	2,019	525	0.00	0.13	25	
	15	5,882	0	0.01	0.11	35	5,736	1,975	0.02	0.56	35	
	20	12,626	0	0.04	0.28	48	12,427	6,674	0.05	1.41	50	
	25	23,287	0	0.11	0.71	58	22,969	19,704	0.13	3.79	68	
	30	38,636	0	0.23	1.50	59	38,235	52,000	0.27	10.02	80	
	35	59,691	0	0.48	3.63	83	59,102	848,133	0.42	82.86	96	
	40	87,154	0	0.93	6.06	88	86,443	880,188	3.90	1,538.97	7	
	45	122,121	0	2.07	15.07	87	121,135	385,058	7.89	1,192.12	9	
	50	165,207	0	3.06	28.36	108	164,051	268,133	18.05	10,786.87	1	

$K = n + 1$. On the other hand, when the no-show probability is larger (i.e., $1 - v_i = 0.4, \forall i = 1, \dots, n$), the differences between the results of $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ become larger when R is smaller. In such a case, the choices of different K values will lead to significantly different objective costs. A decision maker can choose either $E-D_q^{(2)}$ or $E-D_q^{(n+1)}$ to optimize the schedule x based on his/her risk preference. Alternatively, he/she can firstly use the two LP models to quickly compute the bounds of the optimal objective value for a general $E-D_q^{(K)}$ for any $K = 3, \dots, n$, and then optimize the $E-D_q^{(K)}$

model for some K by employing the valid inequalities in Proposition 1 and the decomposition algorithm.

We demonstrate in Figure 3 the optimal schedules of instances with $n = 10$ appointments, produced by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP for $R = 0, 1$ and $1 - v_i = 0.2, 0.4, \forall i = 1, \dots, n$. The points (i, x_i) of every model in each subfigure correspond to the time interval (in minutes) assigned in-between the arrivals of appointments i and $i + 1$, for all $i = 1, \dots, 9$.

As shown in Figure 3, SLP almost equally distributes the time in-between each arrival and schedules a long

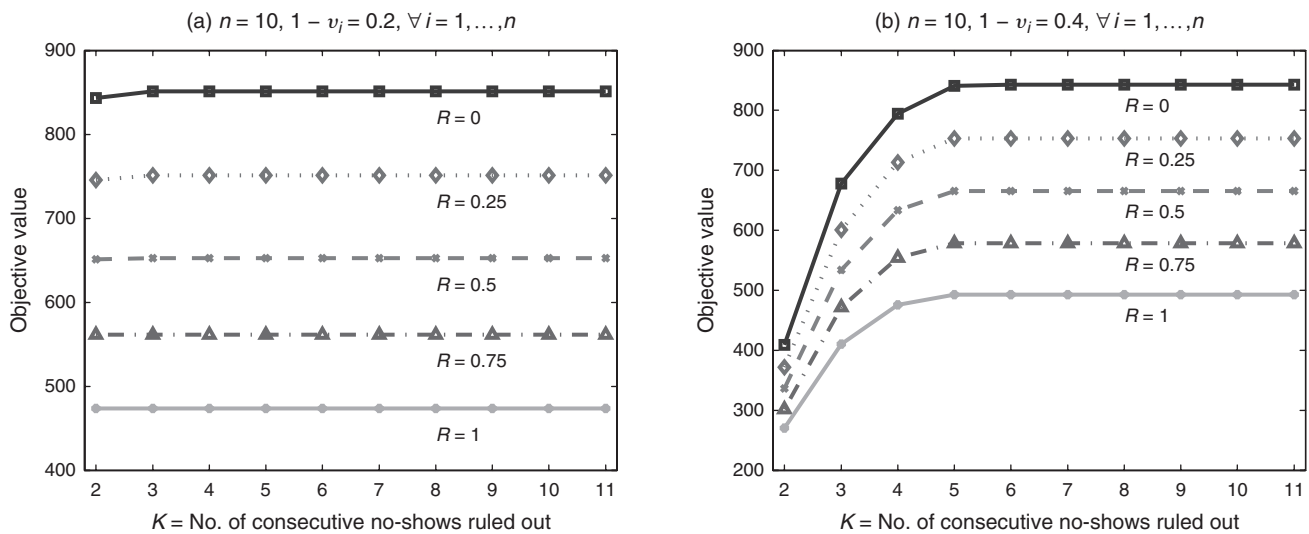
Figure 2. Optimal Objective Values of $E-D_q^{(K)}$ for Different Settings of Parameter R (Time Limit) and $1 - v_i$ (No-Show Probability)

Table 3. Optimal Objective Value Changes According to the Value of K and No-Show Probabilities

No-show	R	$E-D_q^{(2)}$	$E-D_q^{(K)}$, with $K =$								$E-D_q^{(n+1)}$
			3	4	5	6	7	8	9	10	
$1 - v_i = 0.2$	0	843.38	851.08	851.08	851.08	851.08	851.08	851.08	851.08	851.08	851.08
	0.25	745.74	751.32	751.32	751.32	751.32	751.32	751.32	751.32	751.32	751.32
	0.5	651.18	652.54	652.54	652.54	652.54	652.54	652.54	652.54	652.54	652.54
	0.75	561.32	561.32	561.32	561.32	561.32	561.32	561.32	561.32	561.32	561.32
	1	473.69	473.69	473.69	473.69	473.69	473.69	473.69	473.69	473.69	473.69
$1 - v_i = 0.4$	0	409.11	677.66	794.41	840.70	842.91	842.91	842.91	842.91	842.91	842.91
	0.25	371.47	600.99	713.55	752.87	752.87	752.87	752.87	752.87	752.87	752.87
	0.5	336.21	534.22	633.12	665.24	665.24	665.24	665.24	665.24	665.24	665.24
	0.75	301.89	472.07	553.93	578.61	578.61	578.61	578.61	578.61	578.61	578.61
	1	270.55	410.02	475.67	492.69	492.69	492.69	492.69	492.69	492.69	492.69

interval after the last appointment, for all combinations of $1 - v_i$ and R values. As compared to SLP, both $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ schedule longer interarrival time for the early appointments. Intuitively, $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ intend to mitigate the waiting time that may accumulate due to long service durations (also reflected by the

shorter waiting time for the DR models in Tables 4 and 5, reported later in Section 5.4). Additionally, to mitigate the random no-shows, $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ intend to double-book later appointments (reflected by the small x_8 or x_9 value in all cases). When the no-show probability is relative small (i.e., $1 - v_i = 0.2$, $\forall i = 1, \dots, n$)

Figure 3. Appointment Schedules Produced by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP for Different Settings of Parameter R (Time Limit) and $1 - v_i$ (No-Show Probability)

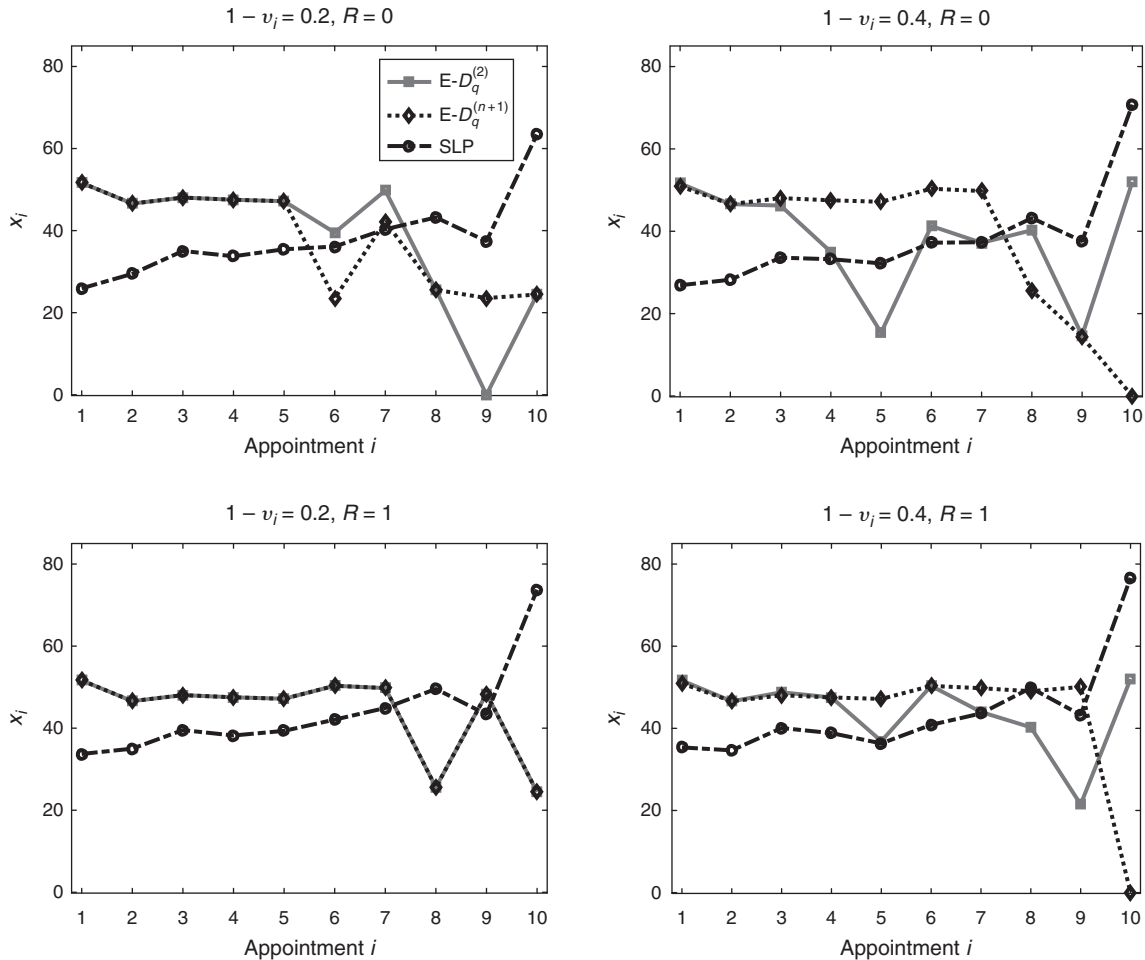


Table 4. Out-of-Sample Performance of Optimal Schedules Given by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP Under Perfect Information with No-Show Probabilities $1 - v_i = 0.4, \forall i = 1, \dots, n$

Metric	Model	R = 0 (in minutes)			R = 0.5 (in minutes)			R = 1 (in minutes)		
		WaitT	OverT	IdleT	WaitT	OverT	IdleT	WaitT	OverT	IdleT
Mean	$E-D_q^{(2)}$	10.06	13.97	16.62	7.33	11.26	19.31	5.50	9.18	22.06
	$E-D_q^{(n+1)}$	6.62	49.27	20.15	4.87	33.59	21.54	3.33	28.11	23.96
	SLP	11.20	3.50	15.57	8.70	2.29	18.41	6.70	1.64	21.31
Median	$E-D_q^{(2)}$	0.00	0.00	12.27	0.00	0.00	14.68	0.00	0.00	20.47
	$E-D_q^{(n+1)}$	0.00	44.92	15.47	0.00	29.22	19.29	0.00	24.93	21.79
	SLP	0.00	0.00	8.19	0.00	0.00	12.95	0.00	0.00	17.45
75%-QT	$E-D_q^{(2)}$	8.40	19.94	33.32	1.79	14.95	34.99	0.00	9.12	40.25
	$E-D_q^{(n+1)}$	0.00	72.90	45.40	0.00	52.22	45.40	0.00	45.27	47.16
	SLP	11.94	0.00	28.83	6.49	0.00	34.38	1.33	0.00	37.08
95%-QT	$E-D_q^{(2)}$	50.88	35.42	46.62	40.08	26.06	48.64	33.52	19.27	50.14
	$E-D_q^{(n+1)}$	40.51	117.78	50.33	30.68	90.84	50.33	22.03	76.66	50.33
	SLP	54.03	23.75	43.20	45.89	12.61	46.05	38.41	7.54	49.85

and the time limit T is sufficiently long (i.e., $R = 1$), both $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ yield the same optimal schedule (also reflected by the same optimal objective value of the two models in Table 3).

5.4. Results of Out-of-Sample Performance

We compare the out-of-sample simulation performance of the optimal schedules of $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP (see Figure 3) under (i) “perfect information” and (ii) misspecified distributional information. Note that $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ models produce solutions that differ the most under large no-show probabilities, and thus we focus on the case when $1 - v_i = 0.4, \forall i = 1, \dots, n$. We examine three cases of the time limit T , by using $R = 0, 0.5, 1$.

We generate two sets of $N' = 10,000$ i.i.d. out-of-sample data $(q_1^m, s_1^m), \dots, (q_n^m, s_n^m), m = 1, \dots, N'$ of the random vector (q, s) following the procedures as follows.

- *Perfect Information:* We use the same distribution (i.e., log-normal) and parameter settings as the ones for generating the N in-sample data to sample the N' data points.

- *Misspecified Distribution:* We keep the same mean values μ_i of the random s_i , v_i of the random q_i , and standard deviation σ_i of the random s_i for each appointment $i = 1, \dots, n$. Therefore, the moment information used in all the DR models and in the SLP remain the same, but we vary the distribution type, as well as correlations among the random service durations and no-shows. Specifically, we follow positively correlated truncated normal distributions with supports $[0, s_i^U], \forall i = 1, \dots, n$ to generate realizations s_1^m, \dots, s_n^m , and follow positively correlated Bernoulli distributions to generate realizations q_1^m, \dots, q_n^m for $m = 1, \dots, N'$. The parameters of the truncated normal distributions and the Bernoulli distributions are designed by following standard statistical methods,¹ to yield

Table 5. Out-of-Sample Performance of Optimal Schedules Given by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP Under Misspecified Distribution with No-Show Probabilities $1 - v_i = 0.4, \forall i = 1, \dots, n$

Metric	Model	R = 0 (in minutes)			R = 0.5 (in minutes)			R = 1 (in minutes)		
		WaitT	OverT	IdleT	WaitT	OverT	IdleT	WaitT	OverT	IdleT
Mean	$E-D_q^{(2)}$	19.89	42.30	13.93	15.75	34.24	16.08	11.51	27.39	18.36
	$E-D_q^{(n+1)}$	12.14	69.59	16.65	10.20	52.16	17.87	8.26	44.56	20.07
	SLP	29.32	38.36	13.53	24.30	31.04	15.76	19.61	24.54	18.07
Median	$E-D_q^{(2)}$	0.00	0.00	5.52	0.00	0.00	9.62	0.00	0.00	14.15
	$E-D_q^{(n+1)}$	0.00	46.25	8.61	0.00	30.62	11.83	0.00	30.56	15.64
	SLP	0.00	0.00	5.09	0.00	0.00	9.16	0.00	0.00	13.00
75%-QT	$E-D_q^{(2)}$	15.34	41.98	27.16	10.70	27.98	30.95	4.91	19.47	35.42
	$E-D_q^{(n+1)}$	6.95	93.23	33.03	2.44	65.31	34.86	0.00	53.56	39.49
	SLP	27.64	11.57	26.36	21.18	0.00	30.33	15.06	0.00	34.45
95%-QT	$E-D_q^{(2)}$	109.75	241.69	46.59	89.11	212.08	49.77	65.02	182.47	48.78
	$E-D_q^{(n+1)}$	62.06	241.69	49.77	58.20	212.08	49.77	52.26	182.47	50.08
	SLP	147.84	241.63	43.20	130.09	212.02	46.05	110.73	182.41	49.84

Table 6. Out-of-Sample Performance of Optimal Schedules Given by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP Under Perfect Information with No-Show Probabilities $1 - v_i = 0.4$, $\forall i = 1, \dots, n$ and No-Show-Dependent Time Limit

Metric	Model	R = 0 (in minutes)			R = 0.5 (in minutes)			R = 1 (in minutes)		
		WaitT	OverT	IdleT	WaitT	OverT	IdleT	WaitT	OverT	IdleT
Mean	$E-D_q^{(2)}$	29.75	52.56	5.06	26.94	28.26	6.37	18.47	21.00	9.37
	$E-D_q^{(n+1)}$	30.74	124.71	12.28	22.62	102.41	13.78	14.77	79.93	15.27
	SLP	34.51	39.40	3.75	28.43	22.56	5.80	21.54	12.76	8.55
Median	$E-D_q^{(2)}$	14.16	42.62	0.00	10.80	8.63	0.00	0.00	0.00	0.00
	$E-D_q^{(n+1)}$	0.00	119.87	0.00	0.00	96.78	0.00	0.00	75.74	4.96
	SLP	19.02	23.80	0.00	11.99	0.00	0.00	4.30	0.00	0.00
75%-QT	$E-D_q^{(2)}$	42.48	77.75	2.48	37.84	44.71	6.73	24.86	33.50	14.68
	$E-D_q^{(n+1)}$	43.24	161.90	24.03	30.07	136.99	25.83	14.59	111.36	25.83
	SLP	48.38	63.39	0.00	39.89	34.70	5.39	29.54	13.68	13.06
95%-QT	$E-D_q^{(2)}$	98.19	145.32	34.88	92.85	109.04	40.23	75.06	90.94	37.75
	$E-D_q^{(n+1)}$	126.95	226.24	48.05	100.50	200.02	48.05	73.78	166.14	48.05
	SLP	108.56	135.82	23.36	95.45	99.25	28.36	81.38	71.52	33.42

positive data correlations and also to keep the first two moments of the N' out-of-sample data the same as the ones of the N in-sample data.

To measure the out-of-sample performance of each solution given by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP, we fix x as an interested solution in Model (26) but use parameters $(q_1^m, s_1^m), \dots, (q_n^m, s_n^m)$, $m = 1, \dots, N'$. We then compute w_i^m , u_i^m , W^m as the waiting time (WaitT), idle time (IdleT), and overtime (OverT) in each scenario m , for $m = 1, \dots, N'$. Table 4 displays means and quantiles of WaitT, IdleT, and OverT, yielded by the optimal solution of each model under perfect distributional information.

Based on Table 4, both $E-D_q^{(2)}$ and $E-D_q^{(n+1)}$ yield slightly better waiting time on average and at different quantiles than the SLP. The $E-D_q^{(2)}$ model results in overtime that is close to the one of SLP (which has

perfect distributional information), but the optimal schedule of $E-D_q^{(n+1)}$ performs badly on average and at all quantiles. For example, when $R = 0$, the schedule by $E-D_q^{(2)}$ lasts 10 minutes longer than the three-minute average overtime given by the SLP optimal schedule, while the optimal schedule of $E-D_q^{(n+1)}$ lasts about 46 minutes longer on average. This is due to the overly conservative no-show support assumption used by $E-D_q^{(n+1)}$. When the distributional information is accurate, $E-D_q^{(n+1)}$ results in overly conservative schedules that perform badly, especially in the overtime metric.

Table 5 illustrates the means and quantiles of WaitT, IdleT, and OverT, yielded by optimal schedules of the three models under misspecified distributional information.

From Table 5, we observe that both DR models yield much better (i.e., 30%–70% shorter) waiting time per

Table 7. Out-of-Sample Performance of Optimal Schedules Given by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP Under Misspecified Distribution with No-Show Probabilities $1 - v_i = 0.4$, $\forall i = 1, \dots, n$ and No-Show-Dependent Time Limit

Metric	Model	R = 0 (in minutes)			R = 0.5 (in minutes)			R = 1 (in minutes)		
		WaitT	OverT	IdleT	WaitT	OverT	IdleT	WaitT	OverT	IdleT
Mean	$E-D_q^{(2)}$	52.66	103.65	4.64	49.78	80.99	6.11	37.38	65.68	8.31
	$E-D_q^{(n+1)}$	41.02	158.23	10.10	31.34	132.34	11.25	22.33	106.81	12.43
	SLP	62.29	100.52	4.33	55.23	80.26	6.04	46.12	63.99	8.14
Median	$E-D_q^{(2)}$	14.74	36.78	0.00	11.91	2.20	0.00	2.93	0.00	0.00
	$E-D_q^{(n+1)}$	4.25	138.88	0.00	0.00	110.64	0.00	0.00	80.42	0.00
	SLP	22.36	33.41	0.00	15.86	3.80	0.00	7.58	0.00	0.00
75%-QT	$E-D_q^{(2)}$	62.00	165.04	2.48	56.99	127.71	6.72	37.89	91.01	13.90
	$E-D_q^{(n+1)}$	50.66	211.26	17.45	35.39	185.49	20.42	19.92	148.20	23.49
	SLP	78.78	164.73	3.43	68.43	127.40	9.10	53.96	90.38	13.88
95%-QT	$E-D_q^{(2)}$	217.42	395.80	32.63	211.46	358.47	36.04	177.68	321.14	37.39
	$E-D_q^{(n+1)}$	169.15	395.85	47.52	140.08	358.53	47.52	108.03	321.20	47.52
	SLP	236.10	395.80	23.36	219.89	358.47	28.44	198.36	321.14	31.64

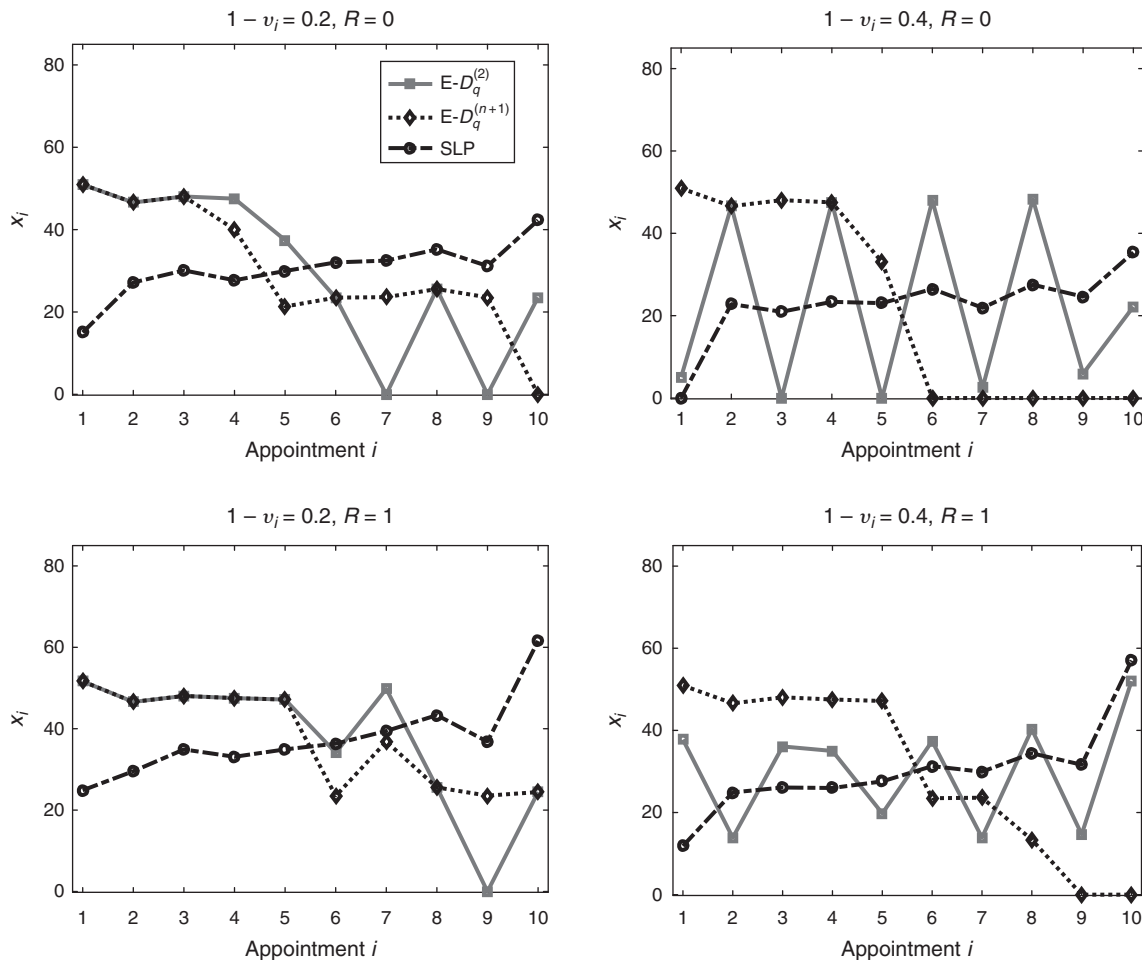
appointment than the optimal schedule given by the SLP, when the distribution type becomes different but the first two moments remain unchanged from the assumed case. The time reduction is reflected in all the metrics, including the mean and 50%–95% quantiles of the random WaitT, for both $R = 0$ and $R = 1$. On the other hand, the three models yield similar IdleT, and the optimal schedule given by $E-D_q^{(n+1)}$ yields slightly longer idle time per appointment but much longer OverT than both $E-D_q^{(2)}$ and SLP. This observation indicates that the optimal schedules given by SLP can become suboptimal when the probability distributions are misspecified, while $E-D_q^{(2)}$ can produce schedules that are less sensitive to misspecification of distribution types.

5.5. No-Show–Dependent Time Limit

In this section, we make the time limit T depend on the no-show probabilities, and report the scheduling solution patterns of SLP, $E-D_q^{(2)}$, and $E-D_q^{(n+1)}$, together with their out-of-sample performance under

perfect information and misspecified distributional information. More specifically, we keep all experiment settings the same as in Sections 5.3–5.4, except for setting $T = \sum_{i=1}^n v_i \mu_i + R \cdot \sqrt{\sum_{i=1}^n [(v_i - v_i^2) \mu_i^2 + v_i \sigma_i^2]}$ to take into account the no-show probabilities. Note that $\mathbb{E}_{\mathbb{P}_{q,s}}[q_i s_i] = v_i \mu_i$ and $\text{Var}(q_i s_i) = (v_i - v_i^2) \mu_i^2 + v_i \sigma_i^2$ if q_i and s_i are independent. Additionally, $v_i < 1$ and so in this setting the time limit is shorter than that in Sections 5.3–5.4. First, we report the out-of-sample performance under perfect information in Table 6 and that under misspecified distributional information in Table 7. From these two tables, we make similar observations as in Section 5.4. Under perfect information, as compared to SLP, $E-D_q^{(2)}$ yields shorter waiting time, longer overtime, and similar idle time, while $E-D_q^{(n+1)}$ yields a much longer overtime in all settings. Under misspecified distributional information, both DR models result in significant reduction in waiting time, and $E-D_q^{(2)}$ yields similar overtime and idle time as SLP does, while $E-D_q^{(n+1)}$ still performs poorly in overtime.

Figure 4. Appointment Schedules Produced by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP for Different Settings of Parameter R (Time Limit) and $1 - v_i$ (No-Show Probability) Under No-Show–Dependent Time Limit



This confirms our conclusion that $E-D_q^{(2)}$ can produce near-optimal schedules that are less sensitive to misspecification of distribution types.

Additionally, we demonstrate in Figure 4 the optimal schedules produced by $E-D_q^{(2)}$, $E-D_q^{(n+1)}$, and SLP for $R = 0, 1$ and $1 - v_i = 0.2, 0.4$, $\forall i = 1, \dots, n$. By comparing this figure with Figure 3, we observe that SLP now intends to double-book the first appointment (reflected by the small x_1 value in three cases) and otherwise behaves similarly. Additionally, $E-D_q^{(n+1)}$ yields similar schedules when the no-show probability is low, but it starts overbooking toward the end when no-shows become more likely (reflected by the small x_i values, $i \geq 6$, when $1 - v_i = 0.4$). Similar pattern changes also happen in $E-D_q^{(2)}$ schedules: $E-D_q^{(2)}$ intends to double-book later appointments when the no-show probability is low; but when no-shows become more likely, it double-books earlier appointments as well. An intuitive explanation is that as the time limit T becomes shorter and no-shows become more likely, double-booking can simultaneously mitigate long waiting time and long idle time (also reflected in Tables 6–7).

6. Conclusions

In this paper, we studied moment-based DR models of the stochastic appointment scheduling problem under uncertainty arising from no-shows and service durations. Our approaches are suitable for an appointment scheduler who has a limited amount of data and considers ambiguous distributions of the two coexisting uncertainties. We derived the following insights on DR appointment scheduling models: (i) one can improve the DR models' ability of utilizing distributional information by using reasonably conservative supports, and (ii) the DR model with the least conservative support of no-shows obtains near-optimal schedules under perfect information and outperforms other DR models and the stochastic program if the distributional type is misspecified.

Based on the computational results, we derived the following recommendations for the practitioner: (i) if the appointment scheduler has accurate information on the no-show probabilities and service duration distributions, then she can double-book the first appointment to better reduce the disruptive effects of no-shows; (ii) if the distributional information is ambiguous, then she can double-book the later appointments to yield more stable performance; and (iii) under ambiguous distributions, she should double-book earlier appointments as well, as the no-show probabilities increase or as the time limit becomes shorter (relative to the number of appointments).

Future research directions include optimizing the length of time limit (i.e., T) and the incorporation of appointment sequencing decisions while considering uncertain no-shows. It is also interesting to investigate

the power of integer programming approaches in other related stochastic and robust optimization models.

Acknowledgments

The authors are grateful to the four referees and the Associate Editor for their constructive comments and helpful suggestions.

Endnote

¹ See https://en.wikipedia.org/wiki/Truncated_normal_distribution.

References

- Barron WM (1980) Failed appointments: Who misses them, why they are missed, and what can be done? *Primary Care* 7(4):563–574.
- Begen MA, Queyranne M (2011) Appointment scheduling with discrete random durations. *Math. Oper. Res.* 36(2):240–257.
- Begen MA, Levi R, Queyranne M (2012) Technical note—A sampling-based approach to appointment scheduling. *Oper. Res.* 60(3):675–681.
- Ben-Tal A, Nemirovski A (2000) Robust solutions of linear programming problems contaminated with uncertain data. *Math. Programming* 88(3):411–424.
- Berg BP, Denton BT, Erdogan SA, Rohleder T, Huschka TR (2014) Optimal booking and scheduling in outpatient procedure centers. *Comput. Oper. Res.* 50:24–37.
- Bertsimas D, Popescu I (2005) Optimal inequalities in probability theory: A convex optimization approach. *SIAM J. Optim.* 15(3):780–804.
- Bertsimas D, Sim M (2004) The price of robustness. *Oper. Res.* 52(1):35–53.
- Bertsimas D, Doan XV, Natarajan K, Teo CP (2010) Models for min-max stochastic linear optimization problems with risk aversion. *Math. Oper. Res.* 35(3):580–602.
- Birge JR, Louveaux FV (2011) *Introduction to Stochastic Programming* (Springer, New York).
- Cayirli T, Veral E (2003) Outpatient scheduling in health care: A review of literature. *Production Oper. Management* 12(4):519–549.
- Cayirli T, Yang KK, Quek SA (2012) A universal appointment rule in the presence of no-shows and walk-ins. *Production Oper. Management* 21(4):682–697.
- Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.
- Deng Y, Shen S (2016) Decomposition algorithms for optimizing multi-server appointment scheduling with chance constraints. *Math. Programming* 157(1):245–276.
- Deng Y, Shen S, Denton BT (2016) Chance-constrained surgery planning under conditions of limited and ambiguous data. SSRN: <http://dx.doi.org/10.2139/ssrn.2432375>.
- Denton BT, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35(11):1003–1016.
- Denton BT, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Sci.* 10(1):13–24.
- Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. *Oper. Res.* 58(4):802–816.
- Erdogan SA, Denton BT (2013) Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS J. Comput.* 25(1):116–132.
- Ge D, Wan G, Wang Z, Zhang J (2013) A note on appointment scheduling with piecewise linear cost functions. *Math. Oper. Res.* 39(4):1244–1251.
- Gul S, Denton BT, Fowler JW, Huschka TR (2011) Bi-criteria scheduling of surgical services for an outpatient procedure center. *Production Oper. Management* 20(3):406–417.
- Gupta D, Denton BT (2008) Appointment scheduling in health care: Challenges and opportunities. *IIE Trans.* 40(9):800–819.

- Gurvich I, Luedtke J, Tezcan T (2010) Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Sci.* 56(7):1093–1115.
- Hassin R, Mendel S (2008) Scheduling arrivals to queues: A single-server model with no-shows. *Management Sci.* 54(3):565–572.
- He S, Sim M, Zhang M (2015) Data-driven patient scheduling in emergency departments: A hybrid robust-stochastic approach. Optimization Online. http://www.optimization-online.org/DB_HTML/2015/11/5213.html.
- Ho CJ, Lau HS (1992) Minimizing total cost in scheduling outpatient appointments. *Management Sci.* 38(12):1750–1764.
- Kleywegt AJ, Shapiro A, Homem-de-Mello T (2002) The sample average approximation method for stochastic discrete optimization. *SIAM J. Optim.* 12(2):479–502.
- Kong Q, Lee CY, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Oper. Res.* 61(3):711–726.
- Kong Q, Li S, Liu N, Teo CP, Yan Z (2016) Appointment scheduling under schedule-dependent patient no-show behavior. <http://www.columbia.edu/~nl2320/doc/Noshow-MS-1030c.pdf>.
- LaGanga LR, Lawrence SR (2012) Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production Oper. Management* 21(5):874–888.
- Lee VJ, Earnest A, Chen MI, Krishnan B (2005) Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. *BMC Health Services Res.* 5(1):51.
- Lin J, Muthuraman K, Lawley M (2011) Optimal and approximate algorithms for sequential clinical scheduling with no-shows. *IIE Trans. Healthcare Systems Engrg.* 1(1):20–36.
- Liu N, Ziya S, Kulkarni VG (2010) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing Service Oper. Management* 12(2):347–364.
- Luo J, Kulkarni VG, Ziya S (2012) Appointment scheduling under patient no-shows and service interruptions. *Manufacturing Service Oper. Management* 14(4):670–684.
- Mak HY, Rong Y, Zhang J (2014) Sequencing appointments for service systems using inventory approximations. *Manufacturing Service Oper. Management* 16(2):251–262.
- Mak HY, Rong Y, Zhang J (2015) Appointment scheduling with limited distributional information. *Management Sci.* 61(2):316–334.
- Mancilla C (2009) Stochastic sequencing and scheduling of an operating room. Ph.D. thesis, Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, Pennsylvania.
- Mittal S, Schulz AS, Sebastian S (2014) Robust appointment scheduling. Jansen K, Rolim J, Devanur N, Moore C, eds. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014)*, Vol. 28 (Schloss Dagstuhl—Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany), 356–37.
- Moore CG, Wilson-Witherspoon P, Probst JC (2001) Time and money: Effects of no-shows at a family practice residency clinic. *Family Medicine* 33(7):522–527.
- Muthuraman K, Lawley M (2008) A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Trans.* 40(9):820–837.
- Parizi MS, Ghate A (2016) Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. *Comput. Oper. Res.* 67:90–101.
- Pinedo M (2012) *Scheduling: Theory, Algorithms, and Systems* (Springer, New York).
- Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing Service Oper. Management* 12(2):330–346.
- Scarf H, Arrow K, Karlin S (1958) A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, Vol. 10 (Stanford University Press, Stanford, CA), 201–209.
- Shen S, Wang J (2014) Stochastic modeling and approaches for managing energy footprints in cloud computing services. *Service Sci.* 6(1):15–33.
- Shylo OV, Prokopyev OA, Schaefer AJ (2012) Stochastic operating room scheduling for high-volume specialties under block booking. *INFORMS J. Comput.* 25(4):682–692.
- Wang Y, Tang J, Fung RYK (2014) A column-generation-based heuristic algorithm for solving operating theater planning problem under stochastic demand and surgery cancellation risk. *Internat. J. Production Econom.* 158:28–36.
- Zacharias C, Pinedo M (2014) Appointment scheduling with no-shows and overbooking. *Production Oper. Management* 23(5):788–801.
- Zeng B, Turkcan A, Lin J, Lawley M (2010) Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Ann. Oper. Res.* 178(1):121–144.
- Zhang Y, Shen S, Erdogan SA (2017) Distributionally robust appointment scheduling with moment-based ambiguity set. *Oper. Res. Lett.* 45(2):139–144.
- Zhang Z, Denton B, Xie X (2015) Branch and price for chance constrained bin packing. Optimization Online. http://www.optimization-online.org/DB_HTML/2015/11/5217.html.

Ruiwei Jiang is an assistant professor in the Department of Industrial and Operations Engineering at the University of Michigan. His research aims to develop data-enabled stochastic optimization (DESO) models and solution methodology that bring together data analytic, integer programming, stochastic programming, and robust optimization. He applies DESO approaches to various engineering problems, including power and water system operations, renewable energy integration, and healthcare resource scheduling.

Siqian Shen is an assistant professor in the Department of Industrial and Operations Engineering at the University of Michigan and also serves as an associate director for the Michigan Institute for Computational Discovery and Engineering (MICDE). Her research focuses on stochastic programming, network optimization, and integer programming. Applications of her work include operations management in healthcare, transportation, and energy.

Yiling Zhang is pursuing the PhD degree in industrial and operations engineering at the University of Michigan. Her research interests are stochastic programming and nonlinear programming with applications in complex service systems.