Management Science

Appointment Scheduling with Limited Distributional Information

Ho-Yin Mak, Ying Rong, Jiawei Zhang

To cite this article:
Ho-Yin Mak, Ying Rong, Jiawei Zhang (2014) Appointment Scheduling with Limited Distributional Information. Management Science

Published online in Articles in Advance 21 May 2014

. http://dx.doi.org/10.1287/mnsc.2013.1881

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–19
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2013.1881

Appointment Scheduling with Limited Distributional Information

Ho-Yin Mak
Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, hymak@ust.hk

Ying Rong
Antai College of Economics and Management, Shanghai Jiao Tong University, Shanghai 200052, China, yrong@sjtu.edu.cn

Jiawei Zhang
Department of Information, Operations, and Management Sciences, Stern School of Business, New York University, New York, New York 10012, jzhang@stern.nyu.edu

In this paper, we develop distribution-free models that solve the appointment sequencing and scheduling problem by assuming only moments information of job durations. We show that our min–max appointment scheduling models, which minimize the worst-case expected waiting and overtime costs out of all probability distributions with the given marginal moments, can be exactly formulated as tractable conic programs. These formulations are obtained by exploiting hidden convexity of the problem. In the special case where only the first two marginal moments are given, the problem can be reformulated as a second-order cone program. Based on the structural properties of this formulation, under a mild condition, we derive the optimal time allowances in closed form and prove that it is optimal to sequence jobs in increasing order of job duration variance. We also prove similar results regarding the optimal time allowances and sequence for the case where only means and supports of job durations are known.

Keywords: appointment scheduling and sequencing; service operations; robust optimization
History: Received February 8, 2012; accepted October 28, 2013, by Dimitris Bertsimas, optimization. Published online in *Articles in Advance*.

1. Introduction

In the healthcare industry, appointment scheduling problems arise in numerous settings, such as scheduling outpatient appointments in primary care and specialty clinics, and scheduling surgeries for operating rooms. Many appointment systems in healthcare involve the following two-stage scheduling process. First, patients and surgeons enter a preliminary booking stage, in which they select preferred dates and time windows for their appointments or surgeries. Second, given a group of appointments booked within a day or within a schedule block in the first stage, the planner has to assign them to the various resources (e.g., different operating rooms and surgeons) and determine their planned starting times. This latter step is typically performed a few days in advance of the appointment dates. An important feature of appointment scheduling problems is that durations of jobs (e.g., surgeries) are typically not known in advance.

In this paper, we focus on the appointment scheduling problem for a single resource arising from the latter of the two planning stages discussed in the previous paragraph. Given a set of jobs with random durations, we have to determine their planned starting times. Equivalently, we have to determine the sequence in which to perform the jobs, and once the sequence is fixed, we determine the time allowances for the jobs. In the sequel, we refer to the determination of time allowances and job sequence as the *scheduling* and *sequencing* decisions, respectively. Correspondingly, we refer to the problem of determining time allowances for the jobs given a predetermined sequence as the *appointment scheduling* problem, and the problem of jointly determining the sequence and time allowances as the *appointment sequencing* problem.

Because of the uncertainty of job durations, any job can be completed before or after the planned starting time of the subsequent job. Both possibilities incur penalties, because these result in the resources becoming idle or the subsequent job waiting to start. Also, if the last job is completed after the due date (the end of working hours of the day), the resources have to work overtime, which is often quite costly. Therefore, the key performance measures in the appointment scheduling problem are the waiting times of patients, the idle times of resources (e.g., operating rooms and surgeons), and the overtime.

In the literature, it is a common assumption that the probability distribution of job durations is known to the decision maker. This is a valid assumption in many situations where there are sufficient data available, so that fitting a distribution is possible. However, there is evidence that the probability distribution of job durations can be hard to estimate in some circumstances, because of the lack of data. For example, Denton et al. (2007) find that there are, on average, only 21 data points available per surgery type at Fletcher Allen Health Care, a health center serving Vermont and upstate New York. A quote from Macario (2010) further shows that the amount of data broken down by surgery types and surgeons is even more limited:

> [F]or approximately half of the cases scheduled in [operating rooms] in hospitals in the United States on any given weekday, only 5 or fewer cases of the same procedure type and by the same surgeon have been performed during the preceding year.

Fitting distributions for stochastic planning requires a large amount of data (see, e.g., the discussion by Levi et al. 2012 regarding an inventory problem). On the other hand, the data requirement for estimating moments (see Delage and Ye 2010 for an excellent discussion), although significant, is relatively less burdensome. This motivates us to study the appointment sequencing and scheduling problem that uses only the (marginal) moments information of the job durations.

Another motivation of our study is the intractability of the appointment sequencing problem when the job duration distribution is known. As to be discussed in the literature review section, the appointment sequencing problem is very difficult to solve computationally. Although a simple heuristic of sequencing jobs by increasing order of variance (OV) is known to produce good sequences, there has been no proof of its optimality when the problem involves three or more jobs. With a model of uncertainty in which the variabilities of job durations are completely characterized by their variances (i.e., when only the means and variances of job durations are known), we shall prove that the OV indeed produces the optimal sequence under reasonable conditions. An analogous result is also proved for the case where only the means and supports of job durations are given.

In the remainder of this section, we first present the mathematical model, briefly review the related literature, and then summarize our main results. We also note that although we will focus on healthcare applications throughout this paper, appointment scheduling problems have applications in other domains such as scheduling cargo ships at a seaport (Sabria and Daganzo 1989) and parts on a shop floor (Wang 1993).

1.1. The Model

We first formulate the stochastic appointment scheduling problem when the sequence of jobs are given. There are n jobs to be scheduled during a time interval $[0, T]$. The jobs need to be processed in a predetermined sequence $1, 2, \ldots, n$. We shall relax this assumption and study the optimal sequence in §§3.3 and 4. Job i requires a random service duration $\tilde{p}_i$. Throughout this paper, we use boldface notation to denote vectors. For example, we use $\tilde{\mathbf{p}}$ to denote $(\tilde{p}_1, \ldots, \tilde{p}_n)$. The planner determines time allowances s_i for each job i. If job i cannot be started at its planned start time due to a delay of completion of the previous job, a waiting time cost will be incurred per unit time of delay. We normalize the unit waiting time cost for all jobs to 1. Furthermore, if the last job is completed after time T, then an overtime cost of γ is charged per unit time of delay. We require the sum of time allowances to be no larger than T, so that all jobs are *scheduled* to be completed by time T; that is, the feasible region of $\mathbf{s} = (s_1, \ldots, s_n)$ is given by the set

$$\mathcal{S} = \left\{ \mathbf{s} \geq 0, \sum_{i=1}^{n} s_i \leq T \right\}.$$

For notational brevity, we do not consider penalty for idle time, which occurs whenever a job is completed before the next job is scheduled to begin. We note that the case with uniform idle time penalty costs can be easily handled with a change of notation.

Given time allowances $\mathbf{s}$ and a realization of the random service times, denoted by $\mathbf{p}$, the waiting times and the overtime can be computed recursively as follows. Let W_i denote the waiting time of the ith job ($i = 1, \ldots, n$), and let W_{n+1} denote the overtime. Then we have $W_1 = 0$ and

$$W_i = \max\{0, W_{i-1} + p_{i-1} - s_{i-1}\}, \quad i = 2, \ldots, n+1. \tag{1}$$

The total waiting and overtime cost is denoted by

$$f(\mathbf{s}, \mathbf{p}) = \sum_{i=2}^{n} W_i + \gamma W_{n+1}. \tag{2}$$

The random job durations $\tilde{\mathbf{p}}$ follow a joint probability distribution F. For any job i, the support of $\tilde{p}_i$ is denoted by D_i. Let $\mathbb{D} = D_1 \times \cdots \times D_n$. For any positive integer q, the qth moment of $\tilde{p}_i$ is denoted by M_{iq}. We also use μ_i and σ_i to denote the mean and standard deviation of $\tilde{p}_i$, respectively; that is, $M_{i1} = \mu_i$ and $M_{i2} = \mu_i^2 + \sigma_i^2$. The key assumption of our model is that the exact joint distribution F is unknown to the planner. The only information available to the planner is the support $\mathbb{D}$ and some marginal moments M_{iq}, for any job $i = 1, 2, \ldots, n$, and for any $q \in Q$ where Q is a finite set of positive integers. More specifically, the distribution F needs to satisfy the following constraints:

$$\int_{\mathbb{D}} dF(\mathbf{p}) = 1, \tag{3}$$

$$\int_{\mathbb{D}} p_i^q \, dF(\mathbf{p}) = M_{iq} \quad \text{for } 1 \le i \le n,\ q \in Q, \tag{4}$$

$$dF(\mathbf{p}) \ge 0 \quad \text{for } \mathbf{p} \in \mathbb{D}. \tag{5}$$

We denote by $\mathcal{F}(\mathbb{D}, Q)$ the set of distributions that satisfy the constraints (3)–(5).

Throughout this paper, we will make the following rather mild technical assumption. It states that the given marginal moments vector lies in the interior of the set of all feasible moment vectors.

Assumption 1. *For each job i, the vector $(M_{iq}: q \in Q)$ lies in the interior of the set $\{(\int_{D_i} p_i^q \, dF_i(p_i): q \in Q): F_i$ is a probability distribution over support $D_i\}$.*

When $Q = \{1, 2\}$, Assumption 1 holds if $M_{i2} > M_{i1}^2$ or, equivalently, $\sigma_i > 0$. For further discussions on Assumption 1, we refer the interested readers to Bertsimas and Popescu (2005).

We are now ready to formulate our appointment scheduling model as the following min–max problem:

$$\min_{\mathbf{s} \in \mathcal{S}} \max_{F \in \mathcal{F}(\mathbb{D}, Q)} E_F[f(\mathbf{s}, \tilde{\mathbf{p}})]; \tag{6}$$

that is, we choose the time allowances $\mathbf{s}$ to minimize the worst-case expected value of $f(\mathbf{s}, \tilde{\mathbf{p}})$ among all distributions in $\mathcal{F}(\mathbb{D}, Q)$.

Two special cases are of particular interest to us. In the first case, $Q = \{1, 2\}$ and $\mathbb{D} = \mathbb{R}^n$; that is, only the means and variances of individual job durations are known. We refer to this version of problem (6) as the *mean-variance* model throughout our paper. In the second case, $Q = \{1\}$ and $D_i = [\mu_i - \underline{d}_i, \mu_i + \bar{d}_i]$, where $\bar{d}_i$ and $\underline{d}_i$ are strictly positive numbers for all $i = 1, \ldots, n$; that is, only the means and supports of individual job durations are known. In this case, problem (6) is referred to as the *mean-support* model.

1.2. Related Literature

Excellent surveys of appointment scheduling research in health care are provided by Cayirli and Veral (2003) and Gupta and Denton (2008). The classical healthcare appointment scheduling problem assumes that the probability distribution of job durations is given and considers the objective of minimizing the expected costs of waiting and idle times. Kaandorp and Koole (2007) show that the expected cost function is $L^\natural$-convex in the planned start time variables when job durations are exponentially distributed, whereas Begen and Queyranne (2011) prove the $L^\natural$ convexity by assuming general discrete distributions. Ge et al. (2013) extend the result of Begen and Queyranne (2011) to the case where the cost function can be written as piecewise linear convex functions of waiting times and idle times. By utilizing the result from Orlin (2010) and Murota (2003), Begen and Queyranne (2011) show for the first time that the appointment scheduling problem is polynomial-time solvable. More specifically, they propose an algorithm that requires $O(n^7)$ expected-cost evaluations.

Sample average approximation (SAA) is a common method used to solve the appointment scheduling problem. Denton and Gupta (2003) solve the two-stage stochastic linear programming formulation using an L-shaped algorithm. Begen et al. (2012) show that the number of samples required to achieve $(1 + \epsilon)$ multiplicative error bound with probability $(1 - \delta)$ is $O(n^6(1/\epsilon^4)\ln(n/\delta))$.

The existing algorithms based on $L^\natural$ convexity involve high-order, although polynomial, complexities. They also require efficient means to evaluate the expected cost given solution values, which may not be readily available for general distributions. On the other hand, SAA algorithms could also be computationally intensive, and require knowledge of the precise distribution of job durations, or at least access to sufficient numbers of independent samples. Moreover, incorporating sequencing decisions could only make the problem computationally less tractable.

Indeed, appointment sequencing seems to be an extremely difficult problem. Weiss (1990) and Denton et al. (2007) prove that sequencing jobs by OV is optimal when there are two jobs and their durations are independent. Because of the notable difficulty, most existing works (Denton et al. 2007, Mancilla and Storer 2012, Mak et al. 2014) on the appointment sequencing problem focus on the development of efficient heuristics to obtain near-optimal sequences. Numerical studies in these papers show that OV and its variants are effective heuristics even when the number of jobs is more than two. The intuition is that OV reduces the likelihood of delays propagating down the schedule, by arranging jobs with less variable durations in front.

With limited data available, it is often difficult to fit the precise distributional form or obtain a large number of independent samples as inputs to planning models. One approach to address this issue is to formulate robust optimization models, which make no assumptions on probability distributions, but instead specify uncertainty sets in which the uncertain parameters (job durations in our context) lie. Decisions are chosen against the worst-case scenarios among all realizations of the uncertain parameters; see, for example, Soyster (1973), Ben-Tal and Nemirovski (2000), Bertsimas and Sim (2003), and Ben-Tal et al. (2009). For the appointment scheduling problem, Mittal and Stiller (2011) study a model where job durations lie in interval uncertainty sets. They provide a heuristic to generate time allowances that balance the maximal waiting costs and the maximal idle costs of jobs. They show that such a heuristic is optimal (with respect to the robust objective) when unit idle time costs are nondecreasing in the job sequence. Their computational study reveals that the expected cost of their robust optimal schedule is typically within 20% of the expected cost of the stochastic optimal solution given knowledge of the job duration distributions.

Our min–max appointment scheduling model (6) belongs to the class of distributionally robust optimization models in the literature. These models assume partial distributional information of the uncertain parameters such as support and moments (mean, covariance, etc.), and decisions are chosen to optimize the worst-case expected objective value among all possible distributions with the specified support and moments information (Scarf 1958). Such models have been useful for providing upper and lower bounds on expected objective values of stochastic programs (Žáčková 1966, Dupacova 1977, Ermoliev et al. 1985, Birge and Wets 1987, Prékopa 1988). More recently, there have been efforts in formulating distributionally robust optimization models as tractable conic programs (Bertsimas et al. 2010, Goh and Sim 2010, Natarajan et al. 2011, Chen et al. 2011, Zhu et al. 2013). For cases when the moments are not precisely known, Delage and Ye (2010) propose a model to incorporate confidence regions for the moments rather than just to utilize their point estimations.

One of the major tasks in the appointment scheduling problem is to compute the expected completion time of the last job, which shares common characteristics with stochastic project scheduling problems with random activity durations. A number of studies (e.g., Meilijson and Nádas 1979; Klein Haneveld 1986; Birge and Maddox 1995; Bertsimas et al. 2004, 2006) utilize support and marginal moments information to generate useful bounds on the expected completion time and/or tardiness for stochastic project scheduling problems.

The paper most relevant to ours is Kong et al. (2013). Under the assumption of given mean, covariance matrix, and nonnegative support of job durations, they derive a copositive programming formulation for the appointment scheduling problem. Although the formulation is convex, it is not necessarily polynomial-time solvable. The authors propose a semidefinite programming relaxation as a solution approach. It is notable that, by specifying the complete covariance matrix, Kong et al. (2013) consider a cross-moment model (CMM) of uncertainty. In contrast, by assuming knowledge of only marginal moments, we consider a marginal moment model (MMM) of uncertainty.

1.3. Our Results and Discussions of the Model

Compared to the CMM, one limitation of the MMM is that it does not specify any correlation structure of job durations. For example, assume that the job durations are known to be independent. The information of independence can be accurately (but not exactly) captured by the CMM by assuming, as an input to the model, zero correlations between all pairs of job durations. On the other hand, for the MMM, the worst-case distribution corresponding to a given solution could be highly correlated, which does not accurately reflects the independence of job durations.

However, there is evidence that job durations can be correlated in healthcare settings for various reasons, such as the presence of student doctors and different work loads among days (Cayirli and Veral 2003, Gupta and Denton 2008). In such situations, if estimation of the exact dependence structure is difficult without sufficient data, then the MMM might be used as a conservative approximation.

We elect to use the MMM framework to study the appointment sequencing and scheduling problem mainly for the following reasons. First, it is noted in the literature that the MMM often leads to computationally more tractable formulations than the CMM, at the cost of not capturing the dependence structure between uncertain parameters. For example, the extremal expected value for a 0-1 integer program under the MMM is polynomial-time computable as long as its deterministic counterpart is Bertsimas et al. (2004, 2006). However, computing the extremal expected objective value of a general linear program (LP) under the CMM is *NP*-hard (Bertsimas et al. 2010). For the appointment scheduling problem, we show in §2 that the MMM leads to computationally tractable conic programming formulation. In particular, the mean-variance model can be formulated as a second-order conic program (SOCP), whereas the mean-support model can be formulated as a linear program.

Second, as we show in §3, we are able to derive an analytic solution for our mean-variance model. Under a mild assumption, we can determine optimal time allowances by a very simple procedure, which can be implemented on a spreadsheet. More importantly, the analytic solution enables us to solve the min–max appointment sequencing problem. In particular, we show that sequencing jobs by OV is optimal in our model. Our analysis enhances understanding on this commonly used heuristic from a robust optimization perspective. In §4, we obtain a similar result for our mean-support model, for which we show it is optimal to sequence jobs by increasing order of width of support.

Third, in our numerical tests discussed in §3.1, we observed excellent performance of the solutions obtained from the mean-variance appointment scheduling model. The solutions are compared against "optimal" schedules that are obtained using the SAA approach. The latter approach assumes complete information about the probability distributions of job durations, which are used to compute the expected costs for both solutions. In §3.3, we also provide numerical tests to show that OV is optimal or very close to optimal, even for problems where the job durations are independent.

The remainder of this paper is organized as follows. We present the conic programming formulation of the general min–max appointment scheduling problem in §2. Then we focus on the mean-variance model and the mean-support model in §§3 and 4, respectively. For both cases, we analyze structural properties of the optimal sequences and schedules. Finally, §5 concludes our paper. The proofs of analytical results in §§2 and 3 are provided in the appendix, whereas the proofs of analytical results in §4 are provided in the online supplement (available at http://ihome.ust.hk/~hymak/Papers/Schedule_Appendix.pdf).

2. Conic Programming Approach

In this section, we provide a tractable conic programming formulation for the min–max problem (6). Our first step is to analyze the inner maximization problem. More specifically, for any fixed $\mathbf{s} \in \mathcal{S}$, we consider the maximization problem

$$\max_{F \in \mathcal{F}(\mathbb{D}, Q)} E_F[f(\mathbf{s}, \tilde{\mathbf{p}})]. \tag{7}$$

This is a moment problem with the probability measure F being the decision variable. To proceed, we first obtain a characterization of cost function $f(\mathbf{s}, \mathbf{p})$, defined in Equations (1) and (2), as discussed below. We notice that there is a linear programming representation of (1) and (2). Taking the dual of this linear program leads to the following proposition. This result can also be found in Kong et al. (2013).

Proposition 1. *For any given* $\mathbf{s}$ *and* $\mathbf{p}$, *it holds that*

$$f(\mathbf{s}, \mathbf{p}) = \max_{\mathbf{y} \in \Lambda} \sum_{i=1}^{n} (p_i - s_i) y_i, \tag{8}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ *and*

$$\Lambda = \left\{ \begin{array}{ll} y_i - y_{i-1} \geq -1 & \text{for } 2 \leq i \leq n, \\ y_n \leq \gamma & \\ \mathbf{y} \geq 0 & \end{array} \right\}. \tag{9}$$

In view of Equation (8), problem (7) becomes

$$\max_{F \in \mathcal{F}(\mathbb{D}, Q)} E_F\left[\max_{\mathbf{y} \in \Lambda} \sum_{i=1}^{n} (\tilde{p}_i - s_i) y_i \right]. \tag{10}$$

In the literature, several results are known concerning problems similar to (10). In particular, Bertsimas et al. (2004, 2006) show that if the extreme points of Λ were binary, then problem (10) could be formulated as a semidefinite program. However, this result does not apply to our problem because the extreme points of Λ are not binary, but instead integral (when γ is an integer). In this case, it is possible to apply the approach proposed by Natarajan et al. (2009), which essentially maps the extreme points of Λ into a higher-dimensional space using a binary expansion. In particular, by introducing binary variables $\mathbf{Y} = (Y_{ik}: i = 1, \ldots, n; k = 0, \ldots, \infty)$, any nonnegative integer variable y_i can be represented as

$$y_i = \sum_{k=0}^{\infty} k Y_{ik}, \quad \sum_{k=0}^{\infty} Y_{ik} = 1, \quad Y_{ik} \in \{0, 1\}, \quad \text{for } k = 0, 1, \ldots, \infty.$$

Then there is a unique, one-to-one correspondence between the extreme points of Λ and $\mathcal{B}$, which is defined as

$$\mathcal{B} = \left\{ Y: \begin{array}{ll} \displaystyle\sum_{k=0}^{\infty} k Y_{ik} - \sum_{k=0}^{\infty} k Y_{i-1,k} \geq -1 & \text{for } 2 \leq i \leq n, \\ \displaystyle\sum_{k=0}^{\infty} k Y_{nk} \leq \gamma & \\ \displaystyle\sum_{k=0}^{\infty} Y_{ik} = 1 & \text{for } 1 \leq i \leq n, \\ Y_{ik} \in \{0, 1\} & \text{for } i = 1, 2, \ldots, n, k = 0, 1, \ldots \end{array} \right\}.$$

Then problem (10) can be formulated as a concave maximization problem. In particular, if $\mathcal{F}(\mathbb{D}, Q)$ is the set of distributions with given (marginal) means and standard deviations, Natarajan et al. (2009) show that problem (10) can be formulated as

$$\max_{Y \in CH(\mathcal{B})} \sum_{i=1}^{n} \left((\mu_i - s_i) \sum_{k=0}^{\infty} k Y_{ik} + \sigma_i \sqrt{ \left(\sum_{k=0}^{\infty} k Y_{ik} \right)^2 - \sum_{k=0}^{\infty} k Y_{ik}^2 } \right), \tag{11}$$

where $CH(\mathcal{B})$ is the convex hull of $\mathcal{B}$. However, the tractability of (11) relies on an explicit characterization of $CH(\mathcal{B})$. Indeed, how to efficiently compute $CH(\mathcal{B})$ for special classes of problems has been suggested for future research (Natarajan et al. 2009). It is not clear to us how to apply this approach to derive the main results of our models.

In this paper, we take a different approach to analyze the moment problem (10). Rather than analyzing $CH(\mathcal{B})$, our approach relies directly on the special structure of Λ. To begin, we first show the following result by analyzing the dual of the moment problem (10), which is a semi-infinite linear program.

Lemma 1. *For any given* $\mathbf{s} \in \mathcal{S}$, *the optimal objective value of problem* (10) *is equal to*

$$\min_{\boldsymbol{\rho}} \max_{\mathbf{y} \in \Lambda} \sum_{i=1}^{n} h_i(y_i, \rho) + \sum_{i=1}^{n} \sum_{q \in Q} M_{iq} \rho_{iq}, \tag{12}$$

where $\boldsymbol{\rho} = (\rho_{iq}: i = 1, \ldots, n;\ q \in Q)$, *and*

$$h_i(y_i, \boldsymbol{\rho}) = \max_{p_i \in D_i} \left((p_i - s_i) y_i - \sum_{q \in Q} \rho_{iq} p_i^q \right). \tag{13}$$

Lemma 1 reduces the stochastic optimization problem (10) to a deterministic min–max problem (12). In the literature, one approach to tackle such a min–max problem is to take the dual of the inner maximization problem over $\mathbf{y}$, and as a result, reformulate the min–max problem as a min–min problem, which ends up being convex in certain cases. Another potential approach is to interchange the ordering of optimization to max–min, using the saddle point theorem. However, neither approach would directly work in our case, because the (maximization) objective function in problem (12) is not concave in $\mathbf{y}$. (In fact, it is straightforward to see that $h_i(y_i, \boldsymbol{\rho})$ is convex in y_i for each i.)

Therefore, we use an alternative approach that utilizes the special structure of the inner maximization problem of the min–max problem (12). For any given $\boldsymbol{\rho}$, we consider

$$\max_{\mathbf{y} \in \Lambda} \sum_{i=1}^{n} h_i(y_i, \boldsymbol{\rho}). \tag{14}$$

One key step in our analysis is to recognize problem (14) as a convex maximization problem over the polyhedron Λ. It then follows that there must exist an optimal solution that is an extreme point of Λ. By analyzing the structural properties of the extreme points of Λ, we shall be able to reformulate this nonlinear program as a linear program. Then, by strong duality of linear programming, the maximization problem (14) can be equivalently formulated as a minimization linear program. Therefore, we can reformulate the min–max problem (12) as a convex minimization problem. This is formally proved in the following proposition.

Proposition 2. *The optimal objective value of problem* (14) *is equal to*

$$\begin{aligned} \min_{\lambda} \quad & \sum_{i=1}^{n} \lambda_i \\ \text{s.t.} \quad & \sum_{i=k}^{\min\{j,n\}} \max_{p_i \in D_i} \left((p_i - s_i)\pi_{ij} - \sum_{q \in Q} \rho_{iq} p_i^q - \lambda_i \right) \le 0 \\ & \qquad \textit{for } 1 \le k \le n,\ k \le j \le n+1, \end{aligned} \tag{15}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, *and*

$$\pi_{ij} = \begin{cases} j - i, & 1 \le i \le j \le n; \\ n + \gamma - i, & 1 \le i \le n,\ j = n+1. \end{cases} \tag{16}$$

As is shown in the proof of Proposition 2, the definition of π_{ij} is motivated by the characterization of extreme points of Λ. More specifically, for any extreme point $\mathbf{y} \in \Lambda$ and for any i, $y_i = \pi_{ij}$ for some $j = 1, \ldots, n+1$.

Proposition 2 enables us to formulate the min–max appointment scheduling problem (6) as a tractable conic program. Indeed, by Proposition 2 and Lemma 1, problem (6) is equivalent to

$$\begin{aligned} \min_{\mathbf{s}, \boldsymbol{\rho}, \boldsymbol{\lambda}} \quad & \sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \sum_{q \in Q} M_{iq} \rho_{iq} \\ \text{s.t.} \quad & \sum_{i=k}^{\min\{j,n\}} \max_{p_i \in D_i} \left((p_i - s_i)\pi_{ij} - \sum_{q \in Q} \rho_{iq} p_i^q - \lambda_i \right) \le 0 \\ & \qquad \text{for } 1 \le k \le n,\ k \le j \le n+1, \\ & \sum_{i=1}^{n} s_i \le T, \\ & \mathbf{s} \ge 0. \end{aligned} \tag{17}$$

Moreover, if we let

$$a_{ij} = \max_{p_i \in D_i} \left((p_i - s_i)\pi_{ij} - \sum_{q \in Q} \rho_{iq} p_i^q - \lambda_i \right),$$

then constraint (17) can be replaced by

$$\sum_{i=k}^{\min\{j,n\}} a_{ij} \le 0 \quad \text{for } 1 \le k \le n,\ k \le j \le n+1 \tag{18}$$

and

$$(p_i - s_i)\pi_{ij} - \sum_{q \in Q} \rho_{iq} p_i^q - \lambda_i - a_{ij} \le 0 \quad \text{for } 1 \le i \le n,\ i \le j \le n+1. \tag{19}$$

Constraint (18) is clearly linear and thus tractable. Constraint (19) asserts that certain polynomial functions (in variable p_i) are nonpositive. By applying results of Nesterov (1997) and Bertsimas and Popescu (2005) on univariate polynomial optimization, constraint (19) can be represented by semidefinite constraints. Therefore, one can formulate problem (6) as a semidefinite program.

In §§3 and 4, we focus on two special cases of the min–max appointment scheduling problem (6): the mean-variance model and the mean-support model. We show that these two models can be formulated as SOCPs and LPs, respectively.

3. The Mean-Variance Model

In this section, we first derive, for the mean-variance model given a fixed job sequence, the optimal time allowances and the corresponding worst-case distribution under a mild condition. Then, we show that OV is optimal among all feasible sequences of the jobs under the same condition. Finally, we provide numerical test on the performance of our formulation. To begin, we show that the mean-variance model of problem (6) can be formulated as an SOCP.

THEOREM 1. *When $Q=\{1,2\}$ and $\mathbb{D}=\mathbb{R}^n$, problem (6) can be formulated as the following nonlinear program:*

$$\min_{\boldsymbol{\beta}>0,\,\boldsymbol{\alpha},\boldsymbol{\lambda},\mathbf{s}} \sum_{i=1}^{n}(\lambda_i+M_{i1}\alpha_i+M_{i2}\beta_i) \tag{20}$$

$$\text{s.t.} \quad \sum_{i=k}^{\min\{n,j\}}\lambda_i \ge \sum_{i=k}^{\min\{n,j\}}\left(\frac{(\pi_{ij}-\alpha_i)^2}{4\beta_i}-s_i\pi_{ij}\right) \quad \text{for } 1\le k\le n,\ k\le j\le n+1, \tag{21}$$

$$\mathbf{s}\in\mathcal{S}.$$

Moreover, the nonlinear constraints (21) are equivalent to the following second-order conic constraints by introducing a new set of decision variables ζ_{ij} for $1\le i\le n$ and $i\le j\le n+1$:

$$\sum_{i=k}^{\min\{n,j\}}(\zeta_{ij}-\lambda_i-\pi_{ij}s_i)\le 0 \quad \text{for } 1\le k\le n,\ k\le j\le n+1$$

and

$$(\beta_i+\zeta_{ij})\ge\sqrt{(\beta_i-\zeta_{ij})^2+(\pi_{ij}-\alpha_i)^2} \quad \text{for } 1\le i\le n,\ i\le j\le n+1.$$

In the following, we will analyze formulation (20). In particular, under a mild assumption, we shall derive the optimal time allowances $\mathbf{s}$ in closed form. The analysis is based on the following result.

LEMMA 2. *The Lagrangian dual of problem (20) can be formulated as the following problem:*

$$\max_{\boldsymbol{\delta}\in\Delta}\min_{\mathbf{s}\in\mathcal{S}}\sum_{i=1}^{n}\left(\sigma_i\sqrt{\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}^2\delta_{kj}-\left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}\right)^2}+(\mu_i-s_i)\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}\right), \tag{22}$$

where

$$\Delta=\left\{\boldsymbol{\delta}\ge 0\colon \sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}=1 \text{ for } i=1,\ldots,n\right\}.$$

Moreover, the optimal objective values of problem (20) and problem (22) are the same.

In formulation (22), δ_{kj} is the dual variable associated with constraint (21) in formulation (20).

3.1. Optimal Job Allowances

In this subsection, we provide an optimal solution to problem (20) and its Lagrangian dual problem (22). The solution relies on a quantity that is defined in the next lemma.

LEMMA 3. *There exists a unique κ, denoted by κ^*, in the interval $(0,\gamma)$ such that*

$$\sum_{i=1}^{n}\left(\mu_i+\frac{\pi_{i,n+1}/2-\kappa}{\sqrt{\kappa\pi_{i,n+1}-(\kappa)^2}}\sigma_i\right)=T. \tag{23}$$

Moreover, κ^* is the optimal solution to the problem

$$\max_{\kappa\in(0,\gamma)}\sum_{i=1}^{n}\left(\sigma_i\sqrt{\pi_{i,n+1}\kappa-\kappa^2}+\mu_i\kappa\right)-\kappa T. \tag{24}$$

We define, for each $i=1,\ldots,n$,

$$\eta_i^*=\frac{\pi_{i,n+1}/2-\kappa^*}{\sqrt{\kappa^*\pi_{i,n+1}-(\kappa^*)^2}}. \tag{25}$$

We are now ready to present our main result of this section.

THEOREM 2. *Assume that $\mu_i+\eta_i^*\sigma_i\ge 0$ for all $i=1,2,\ldots,n$. Then the following holds.*

- *The solution $\boldsymbol{\delta}^*$ defined by (26) is optimal to the Lagrangian dual problem (22):*

$$\begin{aligned}
\delta^*_{1,n+1}&=\frac{\kappa^*}{\pi_{1,n+1}};\\
\delta^*_{i,n+1}&=\frac{\kappa^*}{\pi_{i,n+1}}-\frac{\kappa^*}{\pi_{i-1,n+1}} \quad \text{for } i=2,\ldots,n;\\
\delta^*_{ii}&=1-\frac{\kappa^*}{\pi_{i,n+1}} \quad \text{for } i=1,\ldots,n;\\
\delta^*_{ij}&=0 \quad \text{for } i<j\le n.
\end{aligned} \tag{26}$$

- *The solution $\mathbf{s}^*$ defined by (27) is optimal to problem (20) (and thus optimal to problem (6)):*

$$s_i^*=\mu_i+\eta_i^*\sigma_i. \tag{27}$$

- *The optimal objective values of problem (22) and problem (20) are both equal to*

$$\sum_{i=1}^{n}(\sigma_i\sqrt{\pi_{i,n+1}\kappa^*-(\kappa^*)^2}+\mu_i\kappa^*)-\kappa^*T. \tag{28}$$

Theorem 2 provides a sufficient condition to obtain optimal time allowances without solving the second-order conic program (20) explicitly. One can use a spreadsheet to search for the value of κ^* that solves Equation (23). Consequently, η_i^* can be easily computed by using formula (25). Then one can check whether the condition $\mu_i+\eta_i^*\sigma_i\ge 0$ holds for all i. If so, $s_i^*=\mu_i+\eta_i^*\sigma_i$ is optimal time allowance for job i. This simple procedure can be easily implemented in practice.

Notice that the key assumption of Theorem 2 is that $\mu_i+\eta_i^*\sigma_i\ge 0$ for all $i=1,\ldots,n$. As can be seen in the proof of Claim 2 for Theorem 2 provided in appendix, this assumption is only used to ensure that the time allowance defined by formula (27) is nonnegative, and thus feasible. Therefore, in the appointment scheduling problem (6), if the only constraint associated with $\mathbf{s}$ is $\sum_{i=1}^{n} s_i \leq T$, i.e., removing nonnegativity constraints of $\mathbf{s}$, then Theorem 2 still holds without assuming $\mu_i + \eta_i^* \sigma_i \geq 0$ for all $i = 1, \ldots, n$. In that case, $s_i^* = \mu_i + \eta_i^* \sigma_i$ is always optimal to the relaxed problem (6) and yields the optimal objective value defined by (28).

Theorem 2 shows that the optimal time allowances follow the intuitive "mean plus safety stock" pattern, where η_i^* can be interpreted as the safety factor of job i. Furthermore, the safety factors are decreasing in i (as $\pi_{i,n+1}$ is decreasing in i), and can be negative, i.e., the time allowance can be less than the mean job duration. This implies that more slack should be allocated to the earlier jobs, even if that leads to allocating allowances that are shorter than the expected durations to jobs later in the sequence. This "decreasing safety factor" pattern is a result of our robust optimization model, whose objective is to guard against the worst-case distribution. Note that any delay of completion of the earlier jobs may propagate downstream and cause further delays in subsequent jobs. In view of this, especially when positive correlations between durations of consecutive jobs are possible, it is preferable to avoid delays of earlier jobs by providing larger allowances, as measured by some safety factor times the standard deviations of job durations.

The decreasing safety factor pattern is not only optimal for our robust model, but also optimal in certain cases where the true distribution is given, but the jobs durations are positively correlated. To illustrate, we consider an example with three jobs and the sequence is fixed as 1, 2, 3. For each job i, the job duration is $\tilde{p}_i = \chi + \varepsilon_i$, where χ and ε_i follow an empirical distribution generated by 2,000 samples from independent lognormal distributions where $E[\chi] = 1$, $\text{std}[\chi] = 1.1$, and $E[\varepsilon_i] = 1$, $\text{std}[\varepsilon_i] = 0.55$. The means, standard deviations, and correlations of the job durations, based on the empirical distribution, are provided in Table 1. The optimal time allowances can be obtained by solving the deterministic equivalent of the two-stage stochastic linear programming formulation of the appointment scheduling problem. The results are presented in Table 1. It is clear that the safety factors are decreasing in i.

However, we note that although the decreasing safety factor pattern can be optimal in certain settings, it does not hold universally. There have been previous studies in the literature trying to gain understanding on the pattern of optimal job allowances when the distribution of job durations is known. One result worth discussing is the observation by Denton and Gupta (2003), who report that, for instances in which job durations are independent and identically distributed (i.i.d.) with known distributions, and waiting time cost is small (relative to overtime cost), it is typically optimal to allocate time allowances following a "dome-shape" pattern (i.e., initially increasing, and then decreasing). They also point out that "if, however, the waiting and idle cost coefficients are not equal for all jobs, and/or job duration distributions are not i.i.d., then the solution does not share the dome-shape property" (Denton and Gupta 2003, p. 1011). Indeed, as discussed in Cayirli and Veral (2003) and Gupta and Denton (2008), job durations can be correlated. Besides, waiting time costs can be large in settings such as operating room (surgery) scheduling. Under these circumstances, the dome-shape pattern is not necessarily optimal. Therefore, to some extent, our finding complements the existing insights by enriching understanding on the patterns of optimal time allowances under different practical settings.

Table 1 **An Example of Decreasing Safety Pattern**

	Job		
	1	2	3
Correlation			
Job 1	1	0.22	0.21
Job 2	0.22	1	0.21
Job 3	0.21	0.21	1
Standard deviation σ_i	1.16	1.36	1.22
Mean μ_i	1.99	2.01	2.00
Time allowance s_i^*	2.28	2.25	1.48
Safety factor η_i^*	0.25	0.17	−0.43

To further justify our model, we investigate the performance of the robust scheduling solution when the distributions of job durations are known and independent. We consider problem instances where the number of jobs is from five to eight. Also, the per-unit overtime cost γ is fixed to be 2 for all instances.

We assume that the job durations follow three types of probability distributions: normal, gamma, and lognormal. These distributions can be specified by their means and standard deviations. For each problem instance, only one particular distribution type will be used, but different jobs may follow different distributions. In particular, for each job i, we set $\mu_i \sim U[30, 60]$ and $\sigma_i = \mu_i \cdot \epsilon$, where $\epsilon \sim U[0, 0.3]$. (Hereafter, we use $U[a, b]$ to denote the uniform distribution over $[a, b]$.)

For each set of generated μ_i and σ_i values for $i = 1, 2, \ldots, n$, we set the length of day as $T = \sum_{i=1}^{n} \mu_i + R \cdot \sqrt{\sum_{i=1}^{n} \sigma_i^2}$, where R is a parameter. We allow R take three possible values, 0.5, 0, and -0.5. A smaller value of R implies that the time constraint is tighter. Then we solve the second-order conic program (20) to obtain an optimal *robust* solution, denoted by $\mathbf{s}_{\text{RO}}$. We also generate 1,000 independent samples from one of the distribution families with the given means and standard deviations. Then we can use SAA and solve the problem as a two-stage stochastic program. The solution is denoted by $\mathbf{s}_{\text{SAA}}$, which serves as an approximation to the true optimal solution for the stochastic optimization model that assumes complete knowledge of the probability distributions. To compare the two solutions $\mathbf{s}_{\mathrm{RO}}$ and $\mathbf{s}_{\mathrm{SAA}}$, we evaluate their corresponding expected total costs as follows. We randomly generate 10,000 samples from the distribution, based on which the solution $\mathbf{s}_{\mathrm{SAA}}$ is obtained. For each sample, we compute the total costs corresponding to $\mathbf{s}_{\mathrm{RO}}$ and $\mathbf{s}_{\mathrm{SAA}}$, respectively. Then, we estimate the following measures of the costs over the 10,000 samples for all instances:

- means, denoted by M_{RO} and M_{SAA}, respectively;
- upper semivariances, denoted by SV_{RO} and SV_{SAA}, respectively; and
- tth percentiles, denoted by PT^t_{RO} and PT^t_{SAA}, for $t = 75$, 85, 95, and 99, respectively.

Recall that for any random variable X, its upper semivariance is defined by

$$E[\max(0, X - E[X])^2],$$

which measures the variability of X above its mean (i.e., upside dispersion). Then we compute the percentage differences of the three measures as follows:

$$\frac{M_{\mathrm{RO}}}{M_{\mathrm{SAA}}} - 1, \quad \frac{SV_{\mathrm{RO}}}{SV_{\mathrm{SAA}}} - 1, \quad \text{and} \quad \frac{PT^t_{\mathrm{RO}}}{PT^t_{\mathrm{SAA}}} - 1.$$

For each fixed number of jobs and type of distribution, we generate 20 instances. The average results (over 20 instances in each case) are reported in Table 2. A negative value indicates that the robust solution performs better than the SAA solution with regard to a particular measure.

From Table 2, we can draw a number of observations. First, the expected cost of the robust solution is higher than that of the SAA solution. The average percentage difference ranges from 3.4% to 10.6%, depending on the number of jobs as well as the value of R. Furthermore, the SAA solution also performs better for the 75th and 85th percentiles. However, these differences are not large considering that the robust solution only uses the first two moments of the distribution, whereas the SAA solution is determined with access to 1,000 samples of the true distribution. Second, the increase in expected cost by implementing the robust solution is partially compensated by improvements in performances in the right tail. In particular, the robust solution outperforms the SAA solution for the 95th and 99th percentiles, as well as upper semivariance. This suggests that the robust solution delivers more reliable performance for extreme cases, producing distribution of cost with smaller upside dispersion and a shorter right tail. One possible explanation is that the decreasing safety factor property of the robust solution guards against costly scenarios in which consecutive jobs have long duration realizations simultaneously, and any delays of

Table 2 Percentage Differences (in %) Between Robust Solutions and SAA Solutions for Three Performance Measures: Mean, Upper Semivariance, and Percentiles

Performance measure	R	Distribution	Number of jobs 5	6	7	8
Mean	−0.5	Normal	3.4	3.6	4.7	5.4
		Gamma	3.7	4.0	5.1	5.7
		Lognormal	3.9	4.3	5.4	6.0
	0	Normal	4.8	5.0	6.3	7.2
		Gamma	5.1	5.5	6.8	7.5
		Lognormal	5.3	5.8	7.2	7.9
	0.5	Normal	7.5	7.8	9.1	10.2
		Gamma	7.5	8.0	9.2	10.3
		Lognormal	7.7	8.3	9.5	10.6
Upper semivariance	−0.5	Normal	−4.5	−7.4	−10.4	−15.7
		Gamma	−7.7	−10.6	−12.3	−16.4
		Lognormal	−10.0	−12.3	−14.8	−17.4
	0	Normal	−6.6	−8.6	−11.3	−16.2
		Gamma	−7.3	−10.1	−12.0	−17.4
		Lognormal	−8.7	−11.3	−13.6	−18.6
	0.5	Normal	−6.5	−8.6	−9.6	−14.6
		Gamma	−5.5	−7.5	−8.3	−13.9
		Lognormal	−5.4	−7.6	−8.5	−14.3
75th percentile	−0.5	Normal	5.2	5.5	6.5	6.1
		Gamma	4.3	4.4	5.7	5.9
		Lognormal	2.9	3.8	5.0	6.0
	0	Normal	6.6	7.3	8.5	8.0
		Gamma	6.6	6.6	7.7	7.5
		Lognormal	5.3	6.3	7.4	7.4
	0.5	Normal	12.5	12.8	12.7	12.3
		Gamma	13.2	12.3	12.6	12.1
		Lognormal	13.3	12.9	13.1	12.3
85th percentile	−0.5	Normal	2.4	1.9	3.0	2.5
		Gamma	0.9	0.9	1.8	2.0
		Lognormal	−0.1	0.2	1.0	1.6
	0	Normal	2.5	2.4	3.8	3.4
		Gamma	1.8	1.9	2.7	2.5
		Lognormal	1.0	1.2	2.1	1.7
	0.5	Normal	5.4	5.3	7.1	6.1
		Gamma	5.7	5.8	6.6	5.6
		Lognormal	5.4	6.0	6.5	5.2
95th percentile	−0.5	Normal	−0.4	−1.2	−1.6	−2.7
		Gamma	−1.6	−2.4	−2.3	−3.4
		Lognormal	−2.5	−3.0	−3.2	−3.8
	0	Normal	−1.1	−1.7	−1.8	−3.0
		Gamma	−1.5	−2.3	−2.3	−3.8
		Lognormal	−2.4	−2.9	−2.9	−4.6
	0.5	Normal	−0.9	−1.5	−0.6	−2.0
		Gamma	−0.7	−1.2	−0.5	−2.4
		Lognormal	−0.9	−1.2	−0.7	−2.9
99th percentile	−0.5	Normal	−1.6	−2.6	−4.1	−5.2
		Gamma	−2.5	−3.1	−3.8	−5.3
		Lognormal	−2.4	−3.5	−4.7	−5.5
	0	Normal	−2.5	−3.2	−4.8	−5.8
		Gamma	−2.7	−3.3	−4.2	−6.0
		Lognormal	−2.4	−3.6	−5.0	−6.3
	0.5	Normal	−3.1	−3.9	−4.8	−6.1
		Gamma	−3.0	−3.0	−3.7	−5.7
		Lognormal	−1.8	−2.7	−4.0	−5.4

job completions propagate down the job sequence. Therefore, the robust solution, although slightly conservative in the average case, is very desirable when planners are risk averse and want to guard against extreme scenarios Furthermore, the robust solution enjoys two more advantages over the SAA solution as discussed previously. In particular, it requires less distributional information (means and variances) and is easy to compute (in SOCP form).

3.2. The Worst-Case Distribution

In our min–max model (6), we choose the time allowances $\mathbf{s}$ to minimize the worst-case expected cost among all distributions in the set $\mathcal{F}(\mathbb{D}, Q)$. In this subsection, we assume that the time allowances are fixed to be $\mathbf{s}^*$, and show how to construct the corresponding worst-case distribution with the given first two marginal moments.

In our construction, there are $n+1$ different scenarios. For each scenario $\omega = 1, \ldots, n+1$, we define a vector $\mathbf{Y}_\omega = (Y_{\omega 1}, \ldots, Y_{\omega n})$ such that

$$Y_{\omega i} = \begin{cases} \pi_{i,n+1} & \omega \le i, \\ 0 & \omega > i. \end{cases}$$

This definition implies that $Y_{1i} = \pi_{i,n+1}$ and $Y_{n+1,i} = 0$ for any $i = 1, \ldots, n$. We also define a vector $\mathbf{P}_\omega = (P_{\omega 1}, \ldots, P_{\omega n})$, for each scenario indexed by ω, such that

$$P_{\omega i} = \mu_i - \frac{\sigma_i}{\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2}}(\kappa^* - Y_{\omega i}).$$

Here, $\pi_{i,n+1}$ and κ^* are defined in (16) and (24), respectively.

Now we define a joint probability distribution of $\tilde{\mathbf{p}}^* = (\tilde{p}_1^*, \ldots, \tilde{p}_n^*)$ such that

$$\tilde{\mathbf{p}}^* = \begin{cases} \mathbf{P}_1 \text{ with probability } \chi_1 = \kappa^*/\pi_{1,n+1}, \\ \mathbf{P}_\omega \text{ with probability } \chi_\omega = \kappa^*/\pi_{\omega,n+1} \\ \quad - \kappa^*/\pi_{\omega-1,n+1} \quad \text{for } \omega = 2, \ldots, n, \\ \mathbf{P}_{n+1} \text{ with probability } \chi_{n+1} = 1 - \kappa^*/\pi_{n,n+1}. \end{cases} \tag{29}$$

The vector $\tilde{\mathbf{p}}^*$ defines a probability distribution because $\chi_\omega > 0$ by the fact that $\pi_{\omega-1,n+1} > \pi_{\omega,n+1} > 0$ for $\omega \ge 2$. Also,

$$\sum_{\omega=1}^{n+1} \chi_\omega = \frac{\kappa^*}{\pi_{1,n+1}} + \sum_{\omega=2}^{n}\left(\frac{\kappa^*}{\pi_{\omega,n+1}} - \frac{\kappa^*}{\pi_{\omega-1,n+1}}\right) + 1 - \frac{\kappa^*}{\pi_{n,n+1}} = 1.$$

The next proposition shows that $\tilde{\mathbf{p}}^*$ is indeed a worst-case distribution.

Proposition 3. *If $\mu_i + \eta_i^*\sigma_i \ge 0$ for all $i = 1, 2, \ldots, n$, where η_i^* is defined in* (25), *then we have $E[\tilde{p}_i^*] = \mu_i$ and $E[(\tilde{p}_i^*)^2] = \mu_i^2 + \sigma_i^2$. Moreover, $\tilde{\mathbf{p}}^*$ is a worst-case distribution when $\mathbf{s} = \mathbf{s}^*$.*

One may note that, as defined above, $\tilde{\mathbf{p}}^*$ may possibly take on negative values, because our analysis is based on the assumption that the support $\mathbb{D} = \mathbb{R}^n$. In practice, however, job durations are naturally nonnegative. We note that it is possible to derive a similar SOCP formulation while imposing the requirement that the job durations are always nonnegative.

Remark 1. If $\mathbb{D} = \mathbb{R}_+^n$, one can follow the same steps of the proof of Theorem 1 to reformulate problem (6) as the following SOCP:

$$\begin{aligned} \min_{\boldsymbol{\beta}>0, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \mathbf{s}, \mathbf{c}} \quad & \sum_{i=1}^{n}(\lambda_i + M_{i1}\alpha_i + M_{i2}\beta_i) \\ \text{s.t.} \quad & \sum_{i=k}^{\min\{n,j\}}(\zeta_{ij} - \lambda_i - \pi_{ij}s_i) \le 0 \\ & \qquad \text{for } 1 \le k \le n,\ k \le j \le n+1; \\ & (\beta_i + \zeta_{ij}) \ge \sqrt{(\beta_i - \zeta_{ij})^2 + c_{ij}^2} \\ & \qquad \text{for } 1 \le i \le n,\ i \le j \le n+1; \\ & c_{ij} \ge \pi_{ij} - \alpha_i \quad \text{for } 1 \le i \le n,\ i \le j \le n+1; \\ & c_{ij} \ge 0 \quad \text{for } 1 \le i \le n,\ i \le j \le n+1; \\ & \mathbf{s} \in \mathcal{S}. \end{aligned}$$

The above SOCP is computationally tractable. However, we find that the nonnegativity assumption has almost no impact on the accuracy of the model practically. We carry out the same test as in Table 2 based on formulation (30) and find that the performances of the two models have negligible differences. As discussed throughout §3, formulation (20) is analytically tractable and allows us to derive structural results on the optimal job allowances and sequence. However, such is not the case for formulation (30). Therefore, we focus on formulation (20) throughout this paper.

3.3. Optimal Job Sequence

In this subsection, we consider the appointment sequencing problem. In the literature, it is a popular heuristic to sequence jobs by OV. We will prove the optimality of OV for the mean-variance model for any number of jobs under a mild condition.

Let ψ be a particular sequence of jobs $1, 2, \ldots, n$, i.e., a permutation of the set of integers $\{1, 2, \ldots, n\}$. We use ψ_i to denote the job index of the ith job performed following sequence ψ.

Furthermore, we use ψ^* to denote the OV sequence. Without loss of generality, we assume that $\sigma_1 \le \sigma_2 \le, \cdots, \le \sigma_n$, i.e., jobs are indexed following the OV sequence; that is $\psi_i^* = i$ for $i = 1, \ldots, n$. For any sequence ψ, its worst-case expected cost can be obtained by solving the problem

$$G(\psi) = \min_{\mathbf{s}\in\mathcal{S}} \max_{\mathbf{F}\in\mathcal{F}(\mathbb{R}^n, \{1,2\})} E_{\mathbf{F}}[f(\mathbf{s}, \tilde{\mathbf{p}}_{\psi})]. \tag{30}$$

By Lemma 3, there exists a unique $\kappa_\psi \in (0, \gamma)$ that maximizes

$$\max_{\kappa\in(0,\gamma)} \sum_{i=1}^{n} (\sigma_{\psi_i}\sqrt{\pi_{i,n+1}\kappa - \kappa^2} + \mu_{\psi_i}\kappa) - \kappa T.$$

Then our main result follows.

Theorem 3. *If*

$$\mu_i + \frac{\pi_{i,n+1}/2 - \kappa_{\psi^*}}{\sqrt{\kappa_{\psi^*}\pi_{i,n+1} - (\kappa_{\psi^*})^2}}\sigma_i \geq 0$$

for all $i = 1, \ldots, n$, *then* $G(\psi^*) \leq G(\psi)$ *for any* ψ; *that is, OV is the optimal sequence.*

Remark 2. Consider the case in which the time allowances are not decision variables, but fixed to be $s_i = \mu_i + \tau\sigma_i$ for each i; that is, all the jobs have the same safety factor. Similar to the proof of Theorem 3, one can show that OV is optimal in this case as well. It has been brought to our attention that this result assuming common safety factors may also be proved by applying Proposition 2 in Natarajan et al. (2009). However, their result does not seem to be directly applicable to our main results in this section, i.e., Theorems 2 and 3, as we discussed at the beginning of §2.

In Theorems 2 and 3, we make the assumption that $\mu_i + ((\pi_{i,n+1}/2 - \kappa_{\psi^*})/\sqrt{\kappa_{\psi^*}\pi_{i,n+1} - (\kappa_{\psi^*})^2})\sigma_i \geq 0$. This assumption may not hold in some settings. In such cases, the time allowances $\mathbf{s}^*$ defined in (27) may not be feasible to problem (20). Consequently, OV may not be the optimal sequence in such scenarios. For example, we consider a very special instance with three jobs such that $\mu_1 = 0.1$, $\sigma_1 = 1.5$, $\mu_2 = \sigma_2 = 2$, and $\mu_3 = \sigma_3 = 3$. We first assume that the jobs are sequenced according OV. If $T = 1$ and $\gamma = 20$, then it is easy to check that $\kappa^* = 15.86$ solves Equation (23). Then by formula (27), we get

$$s_1^* = -0.640, \quad s_2^* = 0.811, \quad s_3^* = 0.828.$$

This is clearly not a feasible time allowance since $s_1^* < 0$. Indeed, if we solve the SOCP (20) directly, we can find that the optimal time allowances of the three jobs under the OV sequence are 0, 0.59, and 0.41, respectively, yielding a total expected cost of 123.67. In contrast, the optimal sequence (identified by enumerating the six possible sequences) is to perform job 2 first, job 1 second, and job 3 the last, yielding a total expected cost of 123.16.

In the counterexample above, the mean duration of job 1 is significantly shorter than the standard deviation. Also, the sum of the mean durations of the three jobs, which is 5.1, is significantly longer than the length of day, which is 1. Such scenario is clearly not realistic. Based on our experience of extensive experiments with more realistic parameters of the problem, the assumption we made in Theorems 2 and 3 holds in almost all cases. It indicates that the OV sequence is typically optimal for our min–max model.

Beyond our min–max model, the optimality or near optimality of OV was observed in several earlier studies (Denton et al. 2007, Mancilla and Storer 2012) that assumed that the distributions of job durations are known and they are independent. We have also performed the following numerical study to provide further evidence. We generate instances with five to eight jobs, and set the unit overtime cost, $\gamma = 1$, and the length of day, $T = \sum_{i=1}^{n} \mu_i$. Similar to the computational study in §3.1, the mean and standard deviation of the job durations are randomly generated as follows: μ_i is generated from $U[30, 60]$, and σ_i is generated from $\mu_i \cdot \epsilon$ with $\epsilon \sim U[0, 0.3]$. Once the means and standard deviations of the jobs are generated, we then randomly generate 1,000 samples of the job durations using three types of distributions: normal, gamma, and lognormal.

The SAA approach is used to solve a two-stage stochastic linear integer program that gives the "optimal" sequence and time allowances, together with the "optimal" expected total cost, denoted by M_{SAA}. We solve all instances using CPLEX version 12.2 running on a Lenovo Thinkpad T400 laptop computer with an Intel Core 2 Duo processor and 2 GB memory. For the OV sequence, we use SAA to find "optimal" time allowances and the corresponding expected total cost denoted by M_{OV}. We evaluate the performance of the OV sequence by computing the percentage gap between M_{SAA} and M_{OV}. The percentage gap is defined by

$$\frac{M_{\text{OV}}}{M_{\text{SAA}}} - 1.$$

For a fixed number of jobs and a particular type of probability distributions of job durations, we compute the average and worst percentage gaps over 20 randomly generated instances. The results of the percentage gap and computation time for SAA are reported in Table 3. We observe that, for all instances tested, the expected cost of OV sequences are very close to optimal (i.e., exhibiting very small percentage gaps). This suggests that, for the purpose of choosing the right sequence to perform the jobs, one can do very well using only the variance of job durations, but not full knowledge of the actual probability distributions. This is a desirable property considering that the computation times for finding the "optimal sequence" increase quite rapidly as the number of jobs increases, as observed from Table 3.

Table 3 Verifying the Performance of OV

		Number of jobs			
Distribution	Measures	5	6	7	8
Normal	Average gap of OV (%)	0.08	0.17	0.32	0.27
	Worst gap of OV (%)	1.13	1.13	2.23	1.44
	Average computation time (second)	6	23	106	498
	Longest computation time (second)	12	44	233	1,483
Gamma	Average gap of OV (%)	0.01	0.21	0.21	0.29
	Worst gap of OV (%)	0.14	0.88	1.14	1.24
	Average computation time (second)	6	21	91	434
	Longest computation time (second)	12	36	220	1,428
Lognormal	Mean gap of OV (%)	0.05	0.08	0.16	0.27
	Worst gap of OV (%)	0.62	0.94	0.81	1.22
	Average computation time (second)	5	20	88	414
	Longest computation time (second)	10	37	201	1,272

According to Table 3, M_{OV} could be strictly higher than M_{SAA}. However, this does not imply that OV is not optimal for those problem instances. The reason is that both M_{OV} and M_{SAA} are only approximations of the true expected total cost for the corresponding job sequences, estimated by using 1,000 samples of the true distribution. The differences between M_{OV} and M_{SAA} are too small to be differentiated from the possible estimation errors. Therefore, the OV sequences may or may not be exactly optimal when the time allowances and the expected cost are computed with respect to the true probability distribution. Nevertheless, the numerical results confirm that the OV sequence provides close-to-optimal performance, which is consistent with the findings in Denton et al. (2007) and Mancilla and Storer (2012).

4. The Mean-Support Model

In this section, we consider the mean-support model, in which only the means and supports of individual job durations are known. We prove in Theorem 4 that the min–max appointment scheduling problem can be formulated as a linear program, when the job sequence is fixed. To be consistent with the results for the mean-variance model presented in §3, we abuse the use of some notation such as $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and κ.

Theorem 4. *When $Q=\{1\}$ and $D_i=[\mu_i-\underline{d}_i,\mu_i+\bar{d}_i]$, problem* (6) *can be formulated as the following LP*:

$$\min_{\mathbf{c},\boldsymbol{\alpha},\boldsymbol{\lambda},\mathbf{s}} \sum_{i=1}^{n}(\lambda_i+\mu_i\alpha_i) \tag{31}$$

$$\text{s.t.} \sum_{i=k}^{\min\{n,j\}}(\lambda_i+\pi_{ij}s_i-\xi_{ij})\ge 0 \quad \textit{for } 1\le k\le n,\ 1\le k\le j\le n+1; \tag{32}$$

$$\xi_{ij}+\alpha_i(\mu_i-\underline{d}_i)\ge\pi_{ij}(\mu_i-\underline{d}_i) \quad \textit{for } 1\le i\le n,\ 1\le i\le j\le n+1; \tag{33}$$

$$\xi_{ij}+\alpha_i(\mu_i+\bar{d}_i)\ge\pi_{ij}(\mu_i+\bar{d}_i) \quad \textit{for } 1\le i\le n,\ 1\le i\le j\le n+1; \tag{34}$$

$$\sum_{i=1}^{n}s_i\le T; \tag{35}$$

$$\mathbf{s}\ge\mathbf{0}. \tag{36}$$

Proof. The proofs of results in this section are provided in the online supplement. □

Next, we provide a closed-form expression for the optimal objective value of the LP as well as the optimal time allowances. For notational brevity, we define

$$u(\kappa)=\left(\sum_{i=1}^{n}\mu_i-T\right)\kappa+\sum_{i=1}^{n}\min(\bar{d}_i\kappa,\underline{d}_i(\pi_{i,n+1}-\kappa)). \tag{37}$$

It is obvious that $u(\kappa)$ is a concave function of κ. Therefore, there must exist an optimal solution to the problem

$$\max_{\kappa\in[0,\gamma]} u(\kappa),$$

and let κ^* be the least optimal solution. In fact, $u(\kappa)$ is a piecewise linear concave function of κ, and thus κ^* must be one of its break points. However, besides 0 and γ, any breakpoint κ of $u(\kappa)$ must satisfy

$$\bar{d}_i\kappa=\underline{d}_i(\pi_{i,n+1}-\kappa);$$

that is,

$$\kappa=\frac{\underline{d}_i}{\underline{d}_i+\bar{d}_i}\pi_{i,n+1}$$

for some $i=1,2,\ldots,n$. It implies that we can search for κ^* among at most $n+2$ points. This also motivates the following partition of the set $\{1,2,\ldots,n\}$. Let

$$\Upsilon_1=\left\{i\colon \kappa^*<\frac{\underline{d}_i}{\underline{d}_i+\bar{d}_i}\pi_{i,n+1}\right\},$$

$$\Upsilon_2=\left\{i\colon \kappa^*>\frac{\underline{d}_i}{\underline{d}_i+\bar{d}_i}\pi_{i,n+1}\right\},\quad\text{and}$$

$$\Upsilon_3=\{1,2,\ldots,n\}\backslash\Upsilon_1\backslash\Upsilon_2.$$

We are now ready to present our main results of this section.

Theorem 5. *Assume that $\kappa^*\in(0,\gamma)$. Then the optimal objective value to problem* (31) *is equal to $u(\kappa^*)$. Moreover, any $\mathbf{s}$ that satisfies $\sum_{i=1}^{n}s_i=T$, and the constraint* (38) *as follows is optimal to problem* (31):

$$\begin{aligned} s_i&=\mu_i+\bar{d}_i \quad \textit{for } i\in\Upsilon_1,\\ s_i&=\mu_i-\underline{d}_i \quad \textit{for } i\in\Upsilon_2,\\ s_i&\in[\mu_i-\underline{d}_i,\mu_i+\bar{d}_i] \quad \textit{for } i\in\Upsilon_3. \end{aligned} \tag{38}$$

Notice that when $\kappa^*\in(0,\gamma)$, there always exists $\mathbf{s}$ that satisfies $\sum_{i=1}^{n}s_i=T$ and constraints (38). We use $\mathbf{s}^*$ to denote one such solution.

Theorem 5 is based on the assumption that $\kappa^* \in (0, \gamma)$. In the following, we show a simple condition that guarantees $\kappa^* > 0$. Notice that

$$u(\epsilon) - u(0) = \left(\sum_{i=1}^{n}(\mu_i + \bar{d}_i) - T\right)\epsilon$$

for sufficiently small $\epsilon > 0$. Therefore, when the length of day T is shorter than the sum of maximum possible durations of all the jobs, i.e., $T < \sum_{i=1}^{n}(\mu_i + \bar{d}_i)$, then we must have $u(\epsilon) - u(0) > 0$, which implies that $\kappa^* > 0$. However, we have not found a simple condition that guarantees $\kappa^* < \gamma$.

Finally, notice that the optimal allowance for any job in Υ_1 (Υ_2, respectively) is equal to the upper bound (lower bound, respectively) of the support of its duration. But the optimal allowances for jobs in Υ_3 may not be uniquely defined. However, when Υ_3 is a singleton, there is a unique solution that satisfies $\sum_{i=1}^{n} s_i = T$ and constraint (38). This is the case when $\underline{d}_i/L_i$ is a constant for all i, where $L_i = \underline{d}_i + \bar{d}_i$ denotes the width of the support of the duration of job i. Under this condition, there is at most one job whose optimal time allowance is in the interior of the support of its duration.

In the mean-support setting, we may measure the duration variability of job i by L_i. Similarly, $\underline{d}_i$ and $\bar{d}_i$ can be used to measure the variability below the mean and above the mean, respectively. Motivated by the optimality of OV in the robust mean-variance model, it is natural to suggest that for the robust mean-support model, it is optimal to sequence jobs in increasing order of L_i. It is indeed the case, under a technical condition related to the ratio $\bar{d}_i/L_i$. This ratio can be interpreted as the relative variability above the mean (compared with the width L_i). The condition assumes that there exists $\varphi \in (0, 1)$ such that $\bar{d}_i/L_i = \varphi$ for all i. Intuitively, this assumption states that the supports of job durations have the same degree of symmetry about the respective means.

The result is formally presented in Theorem 6 below, which is an immediate corollary of Theorem 5. Similar to the notation in §3.3, let ψ denote any sequence of the jobs such that ψ_i is the ith job in the sequence. Define L_{ψ_i}, μ_{ψ_i}, $\bar{d}_{\psi_i}$, and $\underline{d}_{\psi_i}$ accordingly. Let ψ^* be the sequence such that $L_{\psi_1^*} \leq \cdots \leq L_{\psi_n^*}$, that is, jobs are sequenced by increasing width of support. Finally, let κ_ψ be the least optimal solution to the problem

$$G(\psi) = \max_{\kappa \in [0, \gamma]} \left(\sum_{i=1}^{n} \mu_{\psi_i} - T\right)\kappa + L_{\psi_i} \sum_{i=1}^{n} \min(\varphi\kappa, (1-\varphi)(\pi_{i, n+1} - \kappa)).$$

Then we have the following:

Theorem 6. *Assume that $\bar{d}_i/L_i$ is a constant in $(0, 1)$ for all i. If $\kappa_\psi \in (0, \gamma)$, then ψ^* is the optimal sequence.*

5. Conclusion

In this paper, we study a stochastic appointment scheduling problem that is prevalent in the healthcare industry. We develop tractable, distribution-free conic programming formulations that are based on partial distributional information of random job durations, i.e., support and moments. Besides computational tractability, we are able to analytically derive structural results for the special cases where the means and variances, and means and supports, of the job durations are known. Also, based on the analytical results, we derive insights that aid appointment planning. In particular, we prove that the widely used heuristic of ordering jobs by variance (ordering by the width of support, respectively) is optimal under a mild condition. This provides further theoretical evidence supporting the use of this popular heuristic. Furthermore, in many cases, the optimal schedule of our model can be obtained with a simple procedure. This result suggests a spreadsheet-implementable heuristic to be used in practice.

The main technical contribution of this paper is a new approach to solve the problem

$$\max_{F \in \mathcal{F}(\mathbb{D}, Q)} E_F\left[\max_{\mathbf{y} \in \Lambda} \sum_{i=1}^{n} (\tilde{p}_i - s_i) y_i\right].$$

Its dual problem is a semi-infinite linear program. The difficulty in analyzing the dual problem is to efficiently handle its infinitely many constraints. We observe, for any given dual solution, that checking its feasibility can be reformulated as a problem of maximizing a separable convex function over the polyhedron Λ. Although such a reformulation is still intractable in general, we have been able to take advantage of the special structure of Λ to further reformulate it as a finite dimensional linear program. The last step is crucial in solving our semi-infinite linear program. We may consider other marginal moments linear programming models where the feasible polyhedron Λ could be different from ours. Our approach might be applicable if the problem of maximizing separable convex functions over that polyhedron admits tractable formulations.

Our research can be extended in several directions. First, healthcare facilities typically have multiple resources, e.g., operating rooms, that can process jobs in parallel. An interesting direction is to extend our modeling framework to the case with multiple parallel resources handling jobs. This poses another challenge; the multiple-resource problem involves the decision of assigning jobs to resources, in addition to sequencing and scheduling. Second, no-shows are prevalent in many healthcare appointment planning problems. Because no-shows lead to idleness of the resource, it is a common practice to allow overbooking. Then, the joint decisions of sequencing, scheduling, and overbooking of jobs, under limited distributional information on job durations and possibly no-show probabilities, are expected to give rise to challenging new problems.

Acknowledgments
The authors gratefully acknowledge Dimitris Bertsimas, the associate editor, and two anonymous referees for their valuable comments that helped improve this paper considerably. The authors also thank Chung-Piaw Teo for helpful discussion. The research of the second author was partially sponsored by the Shanghai Pujiang Program and the National Science Foundation of China [Grant 71202068]. The research of the third author was partially supported by the National Science Foundation of China [Grant 71331004].

Appendix. Proofs of Analytical Results

Additional Lemmas
In the proofs of our analytical results, we will use Lemmas 4–8. The proofs of Lemmas 4–7 are omitted because they either are directly adopted from references or can be easily derived from the first-order condition of unconstrained optimization. The proof of Lemma 8 is provided in the online supplement.

Lemma 4 (Rearrangement Inequality in Hardy et al. 1952). *Suppose that* $0 \le a_1 \le a_2 \le, \cdots, \le a_n$ *and* $b_1 \ge b_2 \ge, \cdots, \ge b_n \ge 0$. *Let* ψ *be a permutation of* $\{1, 2, \ldots, n\}$. *Then* $\sum_{i=1}^n a_i b_i \le \sum_{i=1}^n a_{\psi_i} b_i$ *for any* ψ.

Lemma 5. *Assume that* a *and* b *are positive real numbers. Then* $\min_{x>0} ax + b/x = 2\sqrt{ab}$.

Lemma 6. *Assume that* a, b, c, *and* d *are all nonnegative real numbers, and* $b > a^2$ *and* $c \ge d^2$. *Then*

$$\min_x ax + b\sqrt{c - 2dx + x^2} = \sqrt{(b^2 - a^2)(c^2 - d^2)} + ad.$$

Lemma 7. *Assume that* a *is positive and* b *is any real numbers. Then the following problem*

$$\max_{x\in[0,a]} \sqrt{ax - x^2} - bx$$

has a unique optimal solution $x = (a/2)[1 - b/\sqrt{1 + b^2}]$.

Lemma 8. *Let* $a_1, \ldots, a_m$ *be* $m \ge 2$ *nonnegative real numbers such that* $0 = a_1 < a_2 < \cdots < a_m$. *Let* b *be any real number. Then*

$$\begin{aligned} \max_{\mathbf{x}} \quad & \sqrt{\sum_{j=1}^m a_j^2 x_j - \left(\sum_{j=1}^m a_j x_j\right)^2} - b\sum_{j=1}^m a_j x_j \\ \text{s.t.} \quad & \sum_{j=1}^m x_j = 1, \\ & x_j \ge 0, \quad \text{for } j = 1, \ldots, m \end{aligned} \tag{39}$$

has an optimal solution $\mathbf{x}^*$ *such that* $x_1^* = 1/2 + b/(2\sqrt{1+b^2})$, $x_m^* = 1/2 - b/(2\sqrt{1+b^2})$, *and* $x_j^* = 0$ *for* $1 < j < m$.

Proof of Proposition 1
Recall that $f(\mathbf{s}, \mathbf{p})$ is defined by (1) and (2). It follows that

$$\begin{aligned} f(\mathbf{s}, \mathbf{p}) = \min \quad & \sum_{i=2}^n W_i + \gamma W_{n+1} \\ \text{s.t.} \quad & W_2 \ge p_1 - s_1, \\ & W_{i+1} \ge W_i + p_i - s_i \quad \text{for } i = 2, \ldots, n, \\ & W_i \ge 0 \quad \text{for } i = 2, \ldots, n+1. \end{aligned} \tag{40}$$

By strong duality of linear programming, $f(\mathbf{s}, \mathbf{p})$ is equal to the optimal objective value of the dual problem of (40). Then the proposition follows because the maximization problem in (8) is exactly the dual problem of (40), and (40) is clearly feasible. □

Proof of Lemma 1
Recall that the feasible set $\mathcal{F}(\mathbb{D}, Q)$ of problem (10) is defined by constraints (3)–(5). Let θ be the dual variables associated with (3), and let ρ_{iq} be the dual variables associated with constraints (4) associated with the qth moment. Then, for fixed $\mathbf{s}$, the dual problem of (10) is

$$\begin{aligned} \min_{\theta, \boldsymbol{\rho}} \quad & \theta + \sum_{i=1}^n \sum_{q\in Q} M_{iq}\rho_{iq} \\ \text{s.t.} \quad & \theta + \sum_{i=1}^n \sum_{q\in Q} \rho_{iq} p_i^q \ge f(\mathbf{s}, \mathbf{p}) \quad \text{for } \mathbf{p} \in \mathbb{D}. \end{aligned} \tag{41}$$

By Assumption 1, the strong duality theorem for the moment problem holds (see, e.g., Theorem 2.2 of Bertsimas and Popescu 2005); that is, the optimal objective value of problem (10) is equal to the optimal objective value of the dual problem (41).

To complete the proof, we notice that the constraints of (41) are equivalent to

$$\theta \ge \max_{\mathbf{p}\in\mathbb{D}} \left\{ f(\mathbf{s}, \mathbf{p}) - \sum_{i=1}^n \sum_{q\in Q} \rho_{iq} p_i^q \right\}.$$

By Proposition 1, the above is equivalent to

$$\begin{aligned} \theta \ &\ge \max_{\mathbf{p}\in\mathbb{D}} \max_{\mathbf{y}\in\Lambda} \left\{ \sum_{i=1}^n (p_i - s_i) y_i - \sum_{i=1}^n \sum_{q\in Q} \rho_{iq} p_i^q \right\} \\ &= \max_{\mathbf{y}\in\Lambda} \max_{\mathbf{p}\in\mathbb{D}} \left\{ \sum_{i=1}^n (p_i - s_i) y_i - \sum_{i=1}^n \sum_{q\in Q} \rho_{iq} p_i^q \right\} \\ &= \max_{\mathbf{y}\in\Lambda} \sum_{i=1}^n \left\{ \max_{p_i \in D_i} \left((p_i - s_i) y_i - \sum_{q\in Q} \rho_{iq} p_i^q \right) \right\} \\ &= \max_{\mathbf{y}\in\Lambda} \sum_{i=1}^n h_i(y_i, \boldsymbol{\rho}). \end{aligned}$$

We must have $\theta = \max_{\mathbf{y}\in\Lambda} \sum_{i=1}^n h_i(y_i, \boldsymbol{\rho})$ in any optimal solution to (41) because there is no other constraint on θ. Then we conclude that the dual problem (41) is equivalent to problem (12), which completes the proof. □

Proof of Proposition 2

First notice that, for any i and for any fixed $\boldsymbol{\rho}$, $h_i(y_i, \boldsymbol{\rho})$ is convex in y_i. Therefore, problem (14) is a convex maximization problem. It follows that there exists an optimal solution to problem (14) that is an extreme point of the feasible set Λ. Recall that Λ is defined by (9). It is easy to see (see Zangwill 1966, 1969 for a proof) that for any extreme point $\mathbf{y}$ of Λ, either $y_n = 0$ or $y_n = \gamma > 0$ should hold. And for $i = 1, \ldots, n-1$,

$$y_i \cdot (y_{i+1} - y_i + 1) = 0; \tag{42}$$

that is, for $i \le n-1$, either $y_i = 0$ or $y_i = y_{i+1} + 1$. In the latter case, the value y_i is uniquely determined given y_{i+1}. Applying this fact recursively, we obtain the following result. For any $i \le n$, if $y_i > 0$ and j is the smallest index such that $j > i$ and $y_j = 0$ (for notational convenience, we let $y_{n+1} = 0$), then $y_i = \pi_{ij}$, where π_{ij} is defined in (16). This holds because, if $j \le n$, then $y_i = j - i$; if $j = n+1$, then $y_n = \gamma$, and thus $y_i = n + \gamma - i$.

Thus, it is natural to consider a partition of the integers $1, 2, \ldots, n+1$ into intervals, where each interval $[k, j]$ has the following property. For any $i \in [k, j]$, $y_i = 0$ if and only if $i = j$. Thus $y_i = \pi_{ij}$ for any $i \in [k, j]$. In fact, this defines a unique one-to-one correspondence between any extreme point $\mathbf{y}$ of Λ and a partition of the integers $1, 2, \ldots, n+1$ into intervals. Thus, the problem of finding an optimal extreme point $\mathbf{y}$ can be transformed into finding an optimal partition of the integers $1, 2, \ldots, n+1$ into intervals. This is presented in the following.

For any $k \le j$, we define a binary indicator variable t_{kj} such that $t_{kj} = 1$ if and only if $[k, j]$ is one of the intervals in the partition of $[1, n+1]$. The binary variables (t_{kj}: $1 \le k \le j \le n+1$) represent a partition of $[1, n+1]$ if and only if

$$\sum_{k=1}^{i} \sum_{j=i}^{n+1} t_{kj} = 1 \quad \text{for } i = 1, \ldots, n+1.$$

Also, for $1 \le k \le j \le n+1$, when $t_{kj} = 1$,

$$\sum_{i=k}^{j} h_i(y_i, \boldsymbol{\rho}) = \sum_{i=k}^{j} h_i(\pi_{ij}, \boldsymbol{\rho}),$$

where we have used the notation $h_{n+1}(0, \boldsymbol{\rho}) = 0$ and $\pi_{n+1,\,n+1} = 0$. Therefore, problem (14) is equivalent to

$$\max_{\mathbf{t}} \ \sum_{k=1}^{n+1} \sum_{j=k}^{n+1} \left(\sum_{i=k}^{j} h_i(\pi_{ij}, \boldsymbol{\rho}) \right) t_{kj} \tag{43}$$

$$\text{s.t.} \ \sum_{k=1}^{n+1} \sum_{j=k}^{n+1} t_{kj} = 1 \quad \text{for } i = 1, \ldots, n+1; \tag{44}$$

$$t_{kj} \in \{0, 1\} \quad \text{for } k, j\text{: } 1 \le k \le j \le n+1.$$

For this linear integer program, the matrix associated with the constraint set (44) has the so-called consecutive-ones property, and thus is totally unimodular (see Faigle and Kern 2000). Therefore, its linear programming relaxation, which is obtained by replacing the binary constraints $t_{kj} \in \{0, 1\}$ by nonnegativity constraints $t_{kj} \ge 0$, has a binary optimal solution. Thus, the integer program has the same optimal objective value as its linear programming relaxation as well as the dual problem of the linear programming relaxation. The dual problem is given by

$$\min_{\boldsymbol{\lambda}} \ \sum_{i=1}^{n+1} \lambda_i$$

$$\text{s.t.} \ \sum_{i=k}^{j} \lambda_i \ge \sum_{i=k}^{j} h_i(\pi_{ij}, \boldsymbol{\rho}) \quad \text{for } 1 \le k \le j \le n+1.$$

Since $h_{n+1}(\pi_{n+1,\,n+1}, \boldsymbol{\rho}) = 0$, the dual problem can be simplified as

$$\min_{\boldsymbol{\lambda}} \ \sum_{i=1}^{n} \lambda_i$$

$$\text{s.t.} \ \sum_{i=k}^{\min\{j,n\}} \lambda_i \ge \sum_{i=k}^{\min\{j,n\}} h_i(\pi_{ij}, \boldsymbol{\rho}) \quad \text{for } 1 \le k \le n,\ k \le j \le n+1.$$

We have now shown that this linear program has the same objective value as problem (14). This completes the proof by the definition of $h_i(\pi_{ij}, \boldsymbol{\rho})$. □

Proof of Theorem 1

By applying Proposition 2, and by letting $\alpha_i = \rho_{i1}$ and $\beta_i = \rho_{i2}$, we see that (6) is equivalent to

$$\min_{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{s}} \ \sum_{i=1}^{n} (\lambda_i + \alpha_i M_{i1} + \beta_i M_{i2})$$

$$\text{s.t.} \ \sum_{i=k}^{\min\{n,j\}} \max_{p_i} \big((p_i - s_i)\pi_{ij} - \alpha_i p_i - \beta_i p_i^2 - \lambda_i \big) \le 0 \tag{45}$$

$$\text{for } 1 \le k \le n,\ k \le j \le n+1,$$

$$\mathbf{s} \in \mathcal{S}.$$

We now show the equivalence between (45) and (20). First, for any $i = 1, \ldots, n$, it must hold that $\beta_i > 0$ in any optimal solution to (45). To see this, first observe that problem (45) is bounded from above if we choose $\beta_i > 0$ for all $i = 1, \ldots, n$. Next, we notice that

$$\max_{p_i} ((p_i - s_i)\pi_{ij} - \alpha_i p_i - \beta_i p_i^2) = \max_{p_i} (-\beta_i p_i^2 + (\pi_{ij} - \alpha_i) p_i - s_i \pi_{ij}) = +\infty$$

when $\beta_i < 0$, or when $\beta_i = 0$ and $\alpha_i \ne \pi_{ij}$. But by definition, $\pi_{ij} \ne \pi_{ik}$ if $j \ne k$, and thus $\alpha_i = \pi_{ij}$ can hold for at most one j. Therefore, if $\beta_i \le 0$, then

$$\max_{p_i} ((p_i - s_i)\pi_{ij} - \alpha_i p_i - \beta_i p_i^2) = +\infty$$

will hold except for at most one $j \ge i$. This implies that in this optimal solution, $\sum_{i=1}^{n} \lambda_i = +\infty$, and thus the optimal objective value is $+\infty$. This contradicts the fact that problem (45) is bounded from above. Thus, $\beta_i > 0$ for all $i = 1, \ldots, n$.

Therefore,

$$\max_{p_i} ((p_i - s_i)\pi_{ij} - \alpha_i p_i - \beta_i p_i^2) = \frac{(\pi_{ij} - \alpha_i)^2}{4\beta_i} - \pi_{ij} s_i.$$

Therefore, the constraints of problem (45) is equivalent to

$$\sum_{i=k}^{\min\{n,j\}} \lambda_i \ge \sum_{i=k}^{\min\{n,j\}} \left(\frac{(\pi_{ij} - \alpha_i)^2}{4\beta_i} - s_i \pi_{ij} \right)$$

$$\text{for } 1 \le k \le n,\ k \le j \le n+1,$$

which in turn is equivalent to, by introducing new variables $\boldsymbol{\zeta}$,

$$\sum_{i=k}^{\min\{n,j\}} \zeta_{ij} \le \sum_{i=k}^{\min\{n,j\}} (\lambda_i + s_i\pi_{ij}) \quad \text{for } 1 \le k \le n,\ k \le j \le n+1,$$

$$\zeta_{ij} \ge \frac{(\pi_{ij}-\alpha_i)^2}{4\beta_i} \quad \text{for } 1 \le i \le n,\ i \le j \le n+1.$$

To complete the proof, one can easily see that, when $\beta_i > 0$, the constraint

$$\zeta_{ij} \ge \frac{(\pi_{ij}-\alpha_i)^2}{4\beta_i},$$

or

$$\beta_i\zeta_{ij} \ge (\pi_{ij}-\alpha_i)^2 4, \tag{46}$$

is equivalent to

$$(\beta_i + \zeta_{ij}) \ge \sqrt{(\beta_i - \zeta_{ij})^2 + (\pi_{ij}-\alpha_i)^2}$$

by rearranging terms. □

Proof of Lemma 2

We first derive the Lagrangian dual of problem (20). More specifically, for any $1 \le k \le n$ and $1 \le k \le j \le n+1$, we associate a dual variable $\delta_{kj} \ge 0$ to the nonlinear inequality (21). Note that this is a convex constraint when $\beta_i > 0$, which can be assumed without loss of generality, as shown in the proof of Theorem 1. The Hessian of the nonlinear term $(\pi_{ij}-\alpha_i)^2/(4\beta_i)$ is given by

$$\begin{bmatrix} \dfrac{1}{2\beta_i} & \dfrac{\pi_{ij}-\alpha_i}{2(\beta_i)^2} \\ \dfrac{\pi_{ij}-\alpha_i}{2(\beta_i)^2} & \dfrac{(\pi_{ij}-\alpha_i)^2}{2(\beta_i)^3} \end{bmatrix},$$

whose principal minors are all nonnegative, and is therefore positive semidefinite when $\beta_i > 0$.

For any fixed $\boldsymbol{\delta}$, the Lagrangian dual function is given by

$$\min_{\boldsymbol{\beta}>0,\,\boldsymbol{\alpha},\,\boldsymbol{\lambda},\,\mathbf{s}\in\mathscr{S}} \sum_{i=1}^{n}(\lambda_i + M_{i1}\alpha_i + M_{i2}\beta_i) - \sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\left(\sum_{i=k}^{\min\{n,j\}}\left(\lambda_i + \pi_{ij}s_i - \frac{(\pi_{ij}-\alpha_i)^2}{4\beta_i}\right)\right)$$
$$= \min_{\boldsymbol{\beta}>0,\,\boldsymbol{\alpha},\,\boldsymbol{\lambda},\,\mathbf{s}\in\mathscr{S}} \sum_{i=1}^{n}\left(M_{i1}\alpha_i + M_{i2}\beta_i + \sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\frac{(\pi_{ij}-\alpha_i)^2}{4\beta_i} - \left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}\right)s_i + \left(1 - \sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\right)\lambda_i\right).$$

Notice that variable $\boldsymbol{\lambda}$ is unconstrained. Thus, if $\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj} = 1$, i.e., $\boldsymbol{\delta} \in \Delta$, then the Lagrangian dual function is equal to

$$\min_{\boldsymbol{\beta}>0,\,\boldsymbol{\alpha},\,\mathbf{s}\in\mathscr{S}} \sum_{i=1}^{n}\left(M_{i1}\alpha_i + M_{i2}\beta_i + \sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\frac{(\pi_{ij}-\alpha_i)^2}{4\beta_i} - s_i\left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}\right)\right); \tag{47}$$

otherwise, the dual function is equal to $-\infty$. This implies that we can add the constraint $\boldsymbol{\delta} \in \Delta$ to the Lagrangian dual problem without loss of generality. This is because in the dual problem, we will choose $\boldsymbol{\delta} \ge 0$ to *maximize* the Lagrangian dual function.

Under the constraint that $\boldsymbol{\delta} \in \Delta$, the Lagrangian dual function can be further simplified. In particular, if we optimize over variable $\boldsymbol{\beta} > 0$ in (47), then by Lemma 5, the dual function becomes

$$\min_{\boldsymbol{\alpha},\,\mathbf{s}\in\mathscr{S}} \sum_{i=1}^{n}\left(M_{i1}\alpha_i + \sqrt{M_{i2}\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}(\pi_{ij}-\alpha_i)^2} - s_i\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}\right)$$
$$= \min_{\boldsymbol{\alpha},\,\mathbf{s}\in\mathscr{S}} \sum_{i=1}^{n}\left(M_{i1}\alpha_i + \sqrt{M_{i2}} \cdot \sqrt{\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}^2 - 2\left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}\right)\alpha_i + \alpha_i^2} - s_i\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}\right),$$

where the equation holds because $\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj} = 1$ for any $i = 1, \ldots, n$. Notice that if $\boldsymbol{\delta} \in \Delta$, then

$$\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}^2 \ge \left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}\right)^2.$$

Also, by assumption, $M_{i2} > M_{i1}^2$. Now, we apply Lemma 6 and optimize the Lagrangian dual function over variable $\boldsymbol{\alpha}$. Then the Lagrangian dual function becomes

$$\min_{\mathbf{s}\in\mathscr{S}} \sum_{i=1}^{n}\sqrt{(M_{i2}-M_{i1}^2)\left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}^2 - \left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}\right)\right)} + (M_{i1}-s_i)\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}\pi_{ij}.$$

Recall that $\mu_i = M_{i1}$ and $\sigma_i = \sqrt{M_{i2}-M_{i1}^2}$. Then the Lagrangian dual of problem (20) is (22). Therefore, the lemma follows if strong duality holds for this pair of primal and dual problems. By Bazaraa et al. (2006), this is indeed the case, because problem (20) is convex, and it is clear that there exists a feasible solution such that all the inequality constraints hold as strict inequalities. □

Proof of Lemma 3

Notice that for any $i = 1, \ldots, n$, the quantity $\mu_i + ((\pi_{i,n+1}/2 - \kappa)/\sqrt{\kappa\pi_{i,n+1} - (\kappa)^2})\sigma_i$ is strictly decreasing in κ in the interval $(0, \pi_{i,n+1})$. Therefore, it must also be strictly decreasing in interval $(0, \gamma)$, because $\pi_{i,n+1} \ge \pi_{n,n+1} = \gamma$.

Moreover, for $i = n$, the function $\mu_n + ((\pi_{n,n+1}/2 - \kappa)/\sqrt{\kappa\pi_{n,n+1} - (\kappa)^2})\sigma_n$ approaches ∞ when κ is close to zero, and $-\infty$ when κ is close to $\pi_{n,n+1} = \gamma$. Therefore, there exists a unique $\kappa \in (0, \gamma)$ such that Equation (23) holds.

Furthermore, it is easy to verify that the objective function of problem (24) is continuous and concave in $(0, \gamma)$. Its first derivative is given by $\sum_{i=1}^{n}(\mu_i + ((\pi_{i,n+1}/2 - \kappa)/\sqrt{\kappa\pi_{i,n+1} - (\kappa)^2})\sigma_i - T)$. Therefore, κ^* given by (23) satisfies the first-order condition of problem (24). Thus, it is also the optimal solution to problem (24) because of its concavity. □

Proof of Theorem 2
For simplicity, we rewrite problem (20) as

$$\max_{\boldsymbol{\delta}\in\Delta}\min_{\mathbf{s}\in\mathcal{S}} g(\mathbf{s},\boldsymbol{\delta}), \tag{48}$$

where the function $g(\mathbf{s},\boldsymbol{\delta})$ is defined as

$$g(\mathbf{s},\boldsymbol{\delta}) = \sum_{i=1}^{n}\left(\sigma_i\sqrt{\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}^2\delta_{kj}-\left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}\right)^2} + (\mu_i - s_i)\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}\right).$$

By the saddle point theorem, Theorem 2 follows from the following claims.

Claim 1. *$\mathbf{s}^*$ is an optimal solution to problem* $\min_{\mathbf{s}\in\mathcal{S}} g(\mathbf{s},\boldsymbol{\delta}^*)$.

Claim 2. *$\boldsymbol{\delta}^*$ is an optimal solution to problem* $\max_{\boldsymbol{\delta}\in\Delta} g(\mathbf{s}^*,\boldsymbol{\delta})$.

Claim 3. *It holds that*

$$g(\mathbf{s}^*,\boldsymbol{\delta}^*) = \sum_{i=1}^{n}\left(\sigma_i\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2}\right) + \kappa^*\sum_{i=1}^{n}\mu_i - \kappa^* T.$$

We first prove Claims 1 and 3. By definition of $\mathbf{s}^*$ and the assumption of Proposition 2, we have $\mathbf{s}^* \geq 0$. Moreover, the definitions of η_i^* and κ^* imply that $\sum_{i=1}^{n} s_i^* = T$. Therefore $\mathbf{s}^* \in \mathcal{S}$; that is, $\mathbf{s}^*$ is feasible to problem $\min_{\mathbf{s}\in\mathcal{S}} g(\mathbf{s},\boldsymbol{\delta}^*)$. We next prove the optimality of $\mathbf{s}^*$. For each $i = 1, 2, \ldots, n$, by definition of $\boldsymbol{\delta}^*$ and the fact $\pi_{ii} = 0$, we have that

$$\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}^2\delta_{kj}^* = \sum_{k=1}^{i}\pi_{i,n+1}^2\delta_{k,n+1}^* = \pi_{i,n+1}\kappa^*$$

and

$$\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}^* = \sum_{k=1}^{i}\pi_{i,n+1}\delta_{k,n+1}^* = \kappa^*.$$

This, together with the definition of the function g, implies

$$\begin{aligned} g(\mathbf{s},\boldsymbol{\delta}^*) &= \sum_{i=1}^{n}\left(\sigma_i\sqrt{\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}^2\delta_{kj}^*-\left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}^*\right)^2} + (\mu_i - s_i)\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}^*\right) \\ &= \sum_{i=1}^{n}\left(\sigma_i\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2} + (\mu_i - s_i)\kappa^*\right) \\ &= \sum_{i=1}^{n}\left(\sigma_i\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2}\right) + \kappa^*\sum_{i=1}^{n}\mu_i - \kappa^*\sum_{i=1}^{n}s_i \\ &\geq \sum_{i=1}^{n}\left(\sigma_i\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2}\right) + \kappa^*\sum_{i=1}^{n}\mu_i - \kappa^* T \quad \text{for } \mathbf{s}\in\mathcal{S}. \end{aligned}$$

However, the inequality above holds as an equality when $\mathbf{s} = \mathbf{s}^*$ because $\sum_{i=1}^{n} s_i^* = T$. Thus, $\mathbf{s}^*$ minimizes the function $g(\mathbf{s},\boldsymbol{\delta}^*)$ in the feasible set $\mathcal{S}$. This completes the proof of Claims 1 and 3.

We now prove Claim 2. First, we show that $\boldsymbol{\delta}^* \in \Delta$. The fact $\boldsymbol{\delta}^* \geq 0$ follows from its definition and $0 < \kappa^* < \gamma \leq \pi_{i,n+1} < \pi_{i-1,n+1}$ for any $i = 2, \ldots, n$. Moreover, for any $i = 1, \ldots, n$,

$$\sum_{k=1}^{i}\sum_{j=i}^{n+1}\delta_{kj}^* = \delta_{ii}^* + \sum_{k=1}^{i}\delta_{k,n+1}^* = 1.$$

Thus, $\boldsymbol{\delta}^*$ is a feasible solution to problem $\max_{\boldsymbol{\delta}\in\Delta} g(\mathbf{s}^*,\boldsymbol{\delta})$.

To prove optimality of $\boldsymbol{\delta}^*$, we notice that by definition of g and $\mathbf{s}^*$,

$$g(\mathbf{s}^*,\boldsymbol{\delta}) = \sum_{i=1}^{n}\sigma_i\left(\sqrt{\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}^2\delta_{kj}-\left(\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}\right)^2} - \eta_i^*\sum_{k=1}^{i}\sum_{j=i}^{n+1}\pi_{ij}\delta_{kj}\right).$$

Now we introduce new variables $\boldsymbol{v}$ such that for any $i \leq j$,

$$v_{ij} = \sum_{k=1}^{i}\delta_{kj}. \tag{49}$$

Equation (49) also implies that

$$\begin{aligned} \delta_{1j} &= v_{1j} \quad \text{for } j = 1, \ldots, n+1; \\ \delta_{ij} &= v_{i,j} - v_{i-1,j} \quad \text{for } i = 2, \ldots, n,\ j = i, \ldots, n+1. \end{aligned} \tag{50}$$

With the one-to-one correspondence between $\boldsymbol{v}$ and $\boldsymbol{\delta}$, it is easy to see that the problem $\max_{\boldsymbol{\delta}\in\Delta} g(\mathbf{s}^*,\boldsymbol{\delta})$ is equivalent to

$$\min \ \sum_{i=1}^{n}\sigma_i\left(\sqrt{\sum_{j=i}^{n+1}\pi_{ij}^2 v_{ij} - \left(\sum_{j=i}^{n+1}\pi_{ij}v_{ij}\right)^2} - \eta_i^*\sum_{j=i}^{n+1}\pi_{ij}v_{ij}\right) \tag{51}$$

$$\text{s.t.} \ \sum_{j=i}^{n+1} v_{ij} = 1 \quad \text{for } i = 1, \ldots, n; \tag{52}$$

$$v_{ij} \geq v_{i-1,j} \quad \text{for } i = 2, \ldots, n, \text{ and } j = i, \ldots, n; \tag{53}$$

$$v_{ij} \geq 0 \quad \text{for } i = 1, \ldots, n, \text{ and } j = i, \ldots, n. \tag{54}$$

Also, we can define $\boldsymbol{v}^*$ according to (49) as

$$\begin{aligned} v_{i,n+1}^* &= \sum_{k=1}^{i}\delta_{k,n+1}^* = \frac{\kappa^*}{\pi_{i,n+1}}, \\ v_{ii}^* &= \sum_{k=1}^{i}\delta_{ki}^* = 1 - \frac{\kappa^*}{\pi_{i,n+1}} \quad \text{for } i = 1, \ldots, n, \\ v_{ij}^* &= 0 \quad \text{for } i < j \leq n. \end{aligned} \tag{55}$$

Therefore, to show $\boldsymbol{\delta}^*$ is optimal to problem $\max_{\boldsymbol{\delta}\in\Delta} g(\mathbf{s}^*,\boldsymbol{\delta})$, it suffices to prove $\boldsymbol{v}^*$ is optimal to problem (51), where the constraints are (52), (53), and (54). It is obvious that $\boldsymbol{v}^*$ satisfies all three constraints.

In what follows, we prove a stronger result that $\boldsymbol{v}^*$ is even optimal to problem (51) with constraints (52) and (54) only, i.e., constraint (53) is removed. However, the relaxed problem can be decomposed into n independent subproblems. Indeed, we need only to show that, for each $i = 1, \ldots, n$, (v_{ij}^*: $j = i, \ldots, n+1$) is optimal to problem

$$\begin{aligned} \min \ & \sigma_i\left(\sqrt{\sum_{j=i}^{n+1}\pi_{ij}^2 v_{ij} - \left(\sum_{j=i}^{n+1}\pi_{ij}v_{ij}\right)^2} - \eta_i^*\sum_{j=i}^{n+1}\pi_{ij}v_{ij}\right) \\ \text{s.t.} \ & \sum_{j=i}^{n+1} v_{ij} = 1, \\ & v_{ij} \geq 0 \quad \text{for } j = i, \ldots, n. \end{aligned} \tag{56}$$

This problem has the form of the problem analyzed in Lemma 8. Therefore, by Lemma 8, an optimal solution to problem (5) is given by

$$v_{ii} = \frac{1}{2} + \frac{\eta_i^*}{2\sqrt{1+(\eta_i^*)^2}}, \quad v_{i,n+1} = \frac{1}{2} - \frac{\eta_i^*}{2\sqrt{1+(\eta_i^*)^2}},$$
$$v_{ij} = 0 \quad \text{for } i < j \le n.$$

Substituting η_i^* with $((\pi_{i,n+1}/2-\kappa^*)/\sqrt{\kappa^*\pi_{i,n+1}-(\kappa^*)^2})$, we obtain

$$v_{ii} = 1-\kappa^*/\pi_{i,n+1}, \quad v_{i,n+1} = \kappa^*/\pi_{i,n+1}, \; v_{ij} = 0 \quad \text{for } i<j\le n.$$

Thus, $(v_{ij}^*: j = i, \ldots, n+1)$ is optimal to problem (5) for any $i = 1, \ldots, n$. This completes the proof. □

Proof of Proposition 3
From the definition of $\tilde{\mathbf{p}}^*$, the marginal distribution of $\tilde{p}_i^*$ is

$$\tilde{p}_i^* = \begin{cases} \mu_i - \dfrac{\sigma_i}{\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2}}(\kappa^* - \pi_{i,n+1}) \\ \qquad \text{with probability } \displaystyle\sum_{\omega=1}^{i} \chi_\omega = \frac{\kappa^*}{\pi_{i,n+1}}, \\ \mu_i - \dfrac{\sigma_i}{\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2}}\kappa^* \\ \qquad \text{with probability } \displaystyle\sum_{\omega=i+1}^{n+1} \chi_\omega = 1 - \frac{\kappa^*}{\pi_{i,n+1}}. \end{cases}$$

It is easy to verify that $E[\tilde{p}_i^*] = \mu_i$ and $E[(\tilde{p}_i^*)^2] = \mu_i^2 + \sigma_i^2$; that is, $\tilde{\mathbf{p}}^*$ is a feasible solution to the moment problem (10), i.e., a feasible distribution in $\mathcal{F}(\mathbb{R}^n, \{1,2\})$ for a given $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. In what follows, we prove that $\tilde{\mathbf{p}}^*$ is indeed a worst-case distribution, i.e., an optimal solution to the moment problem (10). By Theorem 3, it is sufficient to prove $E[f(\mathbf{s}^*, \tilde{\mathbf{p}}^*)] = \sum_{i=1}^n (\sigma_i\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2} + \mu_i\kappa^*) - \kappa^* T$, which implies that $E[f(\mathbf{s}^*, \tilde{\mathbf{p}}^*)]$ is equal to the worst-case cost. Therefore, $\tilde{\mathbf{p}}^*$ is the worst-case distribution under $\mathbf{s}^*$.

To see this, we first observe that $\mathbf{Y}_\omega \in \Lambda$ for any ω, i.e., it is a feasible solution to (8). Then it follows that

$$E[f(\mathbf{s}^*, \tilde{\mathbf{p}}^*)] \ge E_\omega\left[\sum_{i=1}^n (P_{\omega i} - s_i^*) Y_{\omega i}\right].$$

On the other hand,

$$\begin{aligned} E_\omega&\left[\sum_{i=1}^n (P_{\omega i} - s_i^*) Y_{\omega i}\right] \\ &= \sum_{\omega=1}^{n+1}\sum_{i=1}^n ((P_{\omega i} - (\mu_i + \eta_i^*\sigma_i)) Y_{\omega i})\chi_\omega \\ &= \sum_{i=1}^n \sum_{\omega=1}^{n+1} \left(\mu_i Y_{\omega i} - \frac{\sigma_i}{\sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2}}(\kappa^* Y_{\omega i} - Y_{\omega i}^2) \right. \\ &\qquad \left. - (\mu_i + \eta_i^*\sigma_i) Y_{\omega i}\right)\chi_\omega \\ &= \sum_{i=1}^n \left(\sigma_i \sqrt{\pi_{i,n+1}\kappa^* - (\kappa^*)^2} + \mu_i\kappa^*\right) - \kappa^* T. \end{aligned}$$

The last equality follows from $\sum_{\omega=1}^{n+1} Y_{\omega i}\chi_\omega = \kappa^*$ and $\sum_{i=1}^n (\mu_i + \eta_i^*\sigma_i) = T$. This completes the proof. □

Proof of Theorem 3
For any sequence ψ, we have

$$G(\psi) = \min_{\sum_{i=1}^n s_i \le T,\, \mathbf{s}\ge 0} \; \max_{\mathbf{F}\in\mathcal{F}(\mathbb{R}^n, \{1,2\})} E_{\mathbf{F}}[f(\mathbf{s}, \tilde{\mathbf{p}}_\psi)].$$

However, by Theorem 2 and the discussion following it, we get

$$\begin{aligned} &\min_{\sum_{i=1}^n s_i \le T} \; \max_{\mathbf{F}\in\mathcal{F}(\mathbb{R}^n, \{1,2\})} E_{\mathbf{F}}[f(\mathbf{s}, \tilde{\mathbf{p}}_\psi)] \\ &\quad = \sum_{i=1}^n \left(\sigma_{\psi_i}\sqrt{\pi_{i,n+1}\kappa_\psi - (\kappa_\psi)^2} + \mu_{\psi_i}\kappa_\psi\right) - \kappa_\psi T. \end{aligned}$$

The equality holds even without assuming

$$\mu_{\psi_i} + \frac{\pi_{i,n+1}/2 - \kappa_\psi}{\sqrt{\kappa_{\psi^*}\pi_{i,n+1} - (\kappa_{\psi^*})^2}}\sigma_{\psi_i} \ge 0,$$

because we relax the constraint $\mathbf{s} \ge 0$. It follows that

$$G(\psi) \ge \sum_{i=1}^n \left(\sigma_{\psi_i}\sqrt{\pi_{i,n+1}\kappa_\psi - (\kappa_\psi)^2} + \mu_{\psi_i}\kappa_\psi\right) - \kappa_\psi T.$$

On the other hand, if

$$\mu_i + \frac{\pi_{i,n+1}/2 - \kappa_{\psi^*}}{\sqrt{\kappa_{\psi^*}\pi_{i,n+1} - (\kappa_{\psi^*})^2}}\, \sigma_i \ge 0 \text{ for all } i = 1, \ldots, n,$$

Theorem 2 leads to

$$G(\psi^*) = \sum_{i=1}^n \left(\sigma_i\sqrt{\pi_{i,n+1}\kappa_{\psi^*} - (\kappa_{\psi^*})^2} + \mu_i\kappa_{\psi^*}\right) - \kappa_{\psi^*} T.$$

Therefore, to prove Theorem 3, it is sufficient to show

$$\begin{aligned} &\sum_{i=1}^n \left(\sigma_{\psi_i}\sqrt{\pi_{i,n+1}\kappa_\psi - (\kappa_\psi)^2} + \mu_{\psi_i}\kappa_\psi\right) - \kappa_\psi T \\ &\quad \ge \sum_{i=1}^n \left(\sigma_{\psi_i}\sqrt{\pi_{i,n+1}\kappa_{\psi^*} - (\kappa_{\psi^*})^2} + \mu_{\psi_i}\kappa_{\psi^*}\right) - \kappa_{\psi^*} T \\ &\quad \ge \sum_{i=1}^n \left(\sigma_i\sqrt{\pi_{i,n+1}\kappa_{\psi^*} - (\kappa_{\psi^*})^2} + \mu_i\kappa_{\psi^*}\right) - \kappa_{\psi^*} T. \end{aligned}$$

The first inequality follows from Lemma 3, that κ_ψ is the maximizer of $\sum_{i=1}^n (\sigma_{\psi_i}\sqrt{\pi_{i,n+1}\kappa - \kappa^2} + \mu_{\psi_i}\kappa) - \kappa T$. The second inequality follows from Lemma 4 and the assumption that σ_i is increasing in i. This completes the proof. □

References

Bazaraa MS, Sherali HD, Shetty CM (2006) *Nonlinear Programming: Theory and Algorithms* (Wiley-Interscience, Hoboken, NJ).

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) 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.

Ben-Tal A, El Ghaoui L, Nemirovski A (2009) *Robust Optimization*. Princeton Series in Applied Mathematics (Princeton University Press, Princeton, NJ).

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 (2003) Robust discrete optimization and network flows. *Math. Programming* 98(1):49–71.

Bertsimas D, Natarajan K, Teo CP (2004) Probabilistic combinatorial optimization: Moments, semidefinite programming, and asymptotic bounds. *SIAM J. Optim.* 15(1):185–209.

Bertsimas D, Natarajan K, Teo CP (2006) Persistence in discrete optimization under data uncertainty. *Math. Programming* 108(2): 251–274.

Bertsimas D, Doan XV, Natarajan K, Teo CP (2010) Models for minimax stochastic linear optimization problems with risk aversion. *Math. Oper. Res.* 35(3):580–602.

Birge JR, Maddox MJ (1995) Bounds on expected project tardiness. *Oper. Res.* 43(5):838–850.

Birge JR, Wets RJB (1987) Computing bounds for stochastic programming problems by means of a generalized moment problem. *Math. Oper. Res.* 12(1):149–162.

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

Chen L, He S, Zhang S (2011) Tight bounds for some risk measures, with applications to robust portfolio selection. *Oper. Res.* 59(4):847–865.

Delage E, Ye Y (2010) Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Oper. Res.* 58(3):595–612.

Denton B, 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.

Dupacova J (1977) Minimaxová úloha stochastického lineárnıho programovánı a momentový problém. *Ekonomicko-Matematický Obzor* 13:279–307.

Ermoliev Y, Gaivoronski A, Nedeva C (1985) Stochastic optimization problems with incomplete information on distribution functions. *SIAM J. Control Optim.* 23(5):697–716.

Faigle U, Kern W (2000) On the core of ordered submodular cost games. *Math. Programming* 87(3):483–499.

Ge D, Wan G, Wang Z, Zhang J (2013) A note on appointment scheduling with piecewise linear cost function. *Math. Oper. Res.*, ePub ahead of print November 13, http://dx.doi.org/10.1287/moor.2013.0631.

Goh J, Sim M (2010) Distributionally robust optimization and its tractable approximations. *Oper. Res.* 58(4):902–917.

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

Hardy GH, Littlewood JE, Polya G (1952) *Inequalities* (Cambridge University Press, Cambridge, UK).

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

Klein Haneveld WK (1986) Robustness against dependence in PERT: An application of duality and distributions with known marginals. *Stochastic Programming 84, Part I*, Mathematical Programming Study, Vol. 27 (North-Holland, Amsterdam), 153–182.

Kong Q, Lee C, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Oper. Res.* 61(3):526–5463.

Levi R, Perakis G, Uichanco J (2012) The data-driven newsvendor problem: New bounds and insights. Technical report, Working paper, Massachusetts Institute of Technology, Cambridge, MA.

Macario A (2010) Is it possible to predict how long a surgery will last? *Medscape Anesthesiology* (July 14), http://www.medscape.com/viewarticle/724756.

Mak HY, Rong Y, Zhang J (2014) Sequencing appointments for service systems using inventory approximations. *Manufacturing Service Oper. Management* 16(2):251–262.

Mancilla C, Storer R (2012) A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Trans.* 44(8):655–670.

Meilijson I, Nádas A (1979) Convex majorization with an application to the length of critical paths. *J. Appl. Probab.* 16(3):671–677.

Mittal S, Stiller S (2011) Robust appointment scheduling. Working paper, Massachusetts Institute of Technology, Cambridge.

Murota K (2003) *Discrete Convex Analysis*, SIAM Monographs on Discrete Mathematics and Applications, Vol. 10 (Society for Industrial and Applied Mathematics, Philadelphia).

Natarajan K, Song M, Teo CP (2009) Persistency model and its applications in choice modeling. *Management Sci.* 55(3):453–469.

Natarajan K, Teo CP, Zheng Z (2011) Mixed 0-1 linear programs under objective uncertainty: A completely positive representation. *Oper. Res.* 59(3):713–728.

Nesterov Y (1997) Structure of non-negative polynomials and optimization problems. Center for Operations Research and Econometrics, Université Catholique de Louvain, Louvain-la-Neuve Belgium.

Orlin JB (2010) Improved algorithms for computing Fisher's market clearing prices: Computing Fisher's market clearing prices. *Proc. 42nd ACM Symp. Theory Comput.* (Association for Computing Machinery, New York), 291–300.

Prékopa A (1988) Boole-Bonferroni inequalities and linear programming. *Oper. Res.* 36(1):145–162.

Sabria F, Daganzo CF (1989) Approximate expressions for queueing systems with scheduled arrivals and established service order. *Transportation Sci.* 23(3):159–165.

Scarf H (1958) A min–max solution of an inventory problem. *Stud. Math. Theory Inventory Production* 10:201–209.

Soyster AL (1973) Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper. Res.* 21(5):1154–1157.

Wang PP (1993) Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Res. Logist.* 40(3):345–360.

Weiss EN (1990) Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Trans.* 22(2):143–150.

Žáčková J (1966) On minimax solutions of stochastic linear programming problems. *Časopis Pro pěstování Matematiky* 91(4):423–430.

Zangwill WI (1966) A deterministic multi-period production scheduling model with backlogging. *Management Sci.* 13(1):105–119.

Zangwill WI (1969) A backlogging model and a multi-echelon model of a dynamic economic lot size production system–a network approach. *Management Sci.* 15(9):506–527.

Zhu Z, Zhang J, Ye Y (2013) Newsvendor optimization with limited distribution information. *Optim. Methods Software* 28(3):640–667.