Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Scheduling Arrivals to a Stochastic Service Delivery System Using Copositive Cones

Qingxia Kong, Chung-Yee Lee, Chung-Piaw Teo, Zhichao Zheng,

To cite this article:
Qingxia Kong, Chung-Yee Lee, Chung-Piaw Teo, Zhichao Zheng, (2013) Scheduling Arrivals to a Stochastic Service Delivery System Using Copositive Cones. Operations Research 61(3):711-726. https://doi.org/10.1287/opre.2013.1158

OPERATIONS RESEARCH
Vol. 61, No. 3, May–June 2013, pp. 711–726
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2013.1158

Scheduling Arrivals to a Stochastic Service Delivery System Using Copositive Cones

Qingxia Kong
School of Business, Universidad Adolfo Ibáñez, Santiago, Chile, q.kong@uai.cl

Chung-Yee Lee
Department of Industrial Engineering and Logistics Management, Hong Kong University of Science and Technology, Hong Kong, cylee@ust.hk

Chung-Piaw Teo, Zhichao Zheng
Department of Decision Sciences, National University of Singapore, Singapore, Republic of Singapore
{bizteocp@nus.edu.sg, zhichao@nus.edu.sg}

In this paper we investigate a stochastic appointment-scheduling problem in an outpatient clinic with a single doctor. The number of patients and their sequence of arrivals are fixed, and the scheduling problem is to determine an appointment time for each patient. The service durations of the patients are stochastic, and only the mean and covariance estimates are known. We do not assume any exact distributional form of the service durations, and we solve for distributionally robust schedules that minimize the expectation of the weighted sum of patients' waiting time and the doctor's overtime. We formulate this scheduling problem as a convex conic optimization problem with a tractable semidefinite relaxation. Our model can be extended to handle additional support constraints of the service durations. Using the primal–dual optimality conditions, we prove several interesting structural properties of the optimal schedules. We develop an efficient semidefinite relaxation of the conic program and show that we can still obtain near-optimal solutions on benchmark instances in the existing literature. We apply our approach to develop a practical appointment schedule at an eye clinic that can significantly improve the efficiency of the appointment system in the clinic, compared to an existing schedule.

Subject classifications: appointment scheduling; copositive programming; semidefinite programming; network flow.
Area of review: Optimization.
History: Received April 2011; revisions received January 2012, May 2012; accepted September 2012. Published online in *Articles in Advance* May 24, 2013.

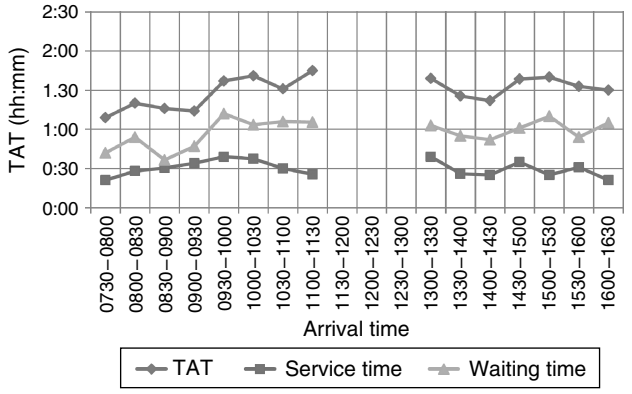
1. Introduction

In many service delivery systems, the core operational activities are largely planned around the arrival times of the customers. The ability to regulate the arrival of customers through a suitable appointment system is thus central to the performance of these systems. The FastPass service of Disney is a well-known example. Customers in the park can obtain a pass to ensure fast service at certain rides if they return at the stipulated time. The temple of Tirumala in India has also used an online appointment system to convert its long waiting line into a virtual queue. This has helped improve service delivery and generated spillover economic benefits to businesses in the vicinity of the temple.[1]

The appointment design problem is also a core problem for healthcare facilities such as outpatient clinics and operating rooms. The appointment system is thus used to regulate the usage of the costly equipment and precious resources in the system. In an eye-care facility that we have visited, there are two consultation sessions per day, each lasting four hours, and the number of doctors available per session is around two to seven. Each doctor has to handle 20 to 30 patients per session. The patients can be classified into "New" (20%) and "Repeat" (80%) patient types. The mean and variance of the consultation times of the new patients are noticeably higher than those of repeat patients, because the conditions of the new patients are hitherto unknown prior to the visit. There are also various operational details that complicate the situation. For instance, patients often have to go for a dilation test prior to seeing the doctor. This process adds to the complexity of finding an optimal appointment strategy for the system.

One key performance indicator in this system is the "turnaround time" (TAT), defined to be *the time from the moment the patient walks into the clinic, to the moment the patient leaves the clinic*. Figure 1 shows the overall median TAT, service time, and waiting time (WT) of patients arriving in different time slots for two different sessions in the clinic, where TAT is the sum of service and waiting times. Clearly, the patients are experiencing long turnaround time, with waiting time far exceeding the actual service time.

Figure 1. Median time from registration to payment.



We note that there are several pertinent features in this system: (i) New patients often have to undergo a series of checks (such as visual acuity, and/or other advanced tests) after the consultation, some of which can take as much as 2.5 hours. To make sure that all the tests and consultations can be performed within the same day, the doctors prefer to see the new patients in the early portion of the morning session. Consequently, early morning slots are reserved primarily for new patients. (ii) The current appointment strategy is to allocate five minutes per patient slot for one hour, followed by a half-hour break. Then, in a four-hour session (e.g., 8 A.M.–12 NOON), there are three contiguous one-hour consultation periods with two half-hour breaks in between. This allows each doctor to see 36 patients.

This leads to the central questions of this paper: is there any (near) optimal strategy to schedule the arrival of patients (based on the patient's classification) such that the waiting time of the patients and overtime work of the doctor are minimized? Furthermore, are there any "distributionally robust" solutions that perform well for a wide range of service-time distributions?

The research on appointment system design over the past few decades has been driven largely by these issues. However, these problems are notoriously difficult. Standard queueing theory does not apply, because we are interested in the transient performance measures of the system. It is technically challenging to calculate the expected waiting time of the nth patient in the sequence, because of the difficulty of propagating the impact of earlier events on this patient. Recently, Begen and Queyranne (2011) showed that the scheduling problem is solvable in polynomial time (in the size of the representation of the discrete distributions). However, this method works well only for discrete distributions with a small number of distinct values. To the best of our knowledge, simulation and stochastic programming methods are still the preferred approaches for the appointment design problem. Unfortunately, the solutions obtained are often sensitive to the samples used to develop the schedules, and hence, very little is known about the structure of the optimal policies, even in the simplest environment with one doctor and when patients arrive punctually according to the appointment schedule.

1.1. Contributions

In this paper, we develop a convex conic programming approach to solve the appointment-scheduling problem. We show that this problem can be suitably reformulated as a two-stage stochastic optimization problem. In the second stage, we construct a network flow model to capture the waiting time of each patient, under a given scheduling policy (from the first stage problem). Our novelty comes in the solution to the first-stage problem, which is a technically challenging problem. Instead of using a specific service-time distribution to design the schedule, we employ a minimax approach so that the schedule is designed to minimize the maximum expected cost achieved by some distribution from a family of distributions. Next, we develop a conic optimization framework to transform the stochastic appointment scheduling problem into a single deterministic copositive programming problem (COP).[2]

Using the primal–dual optimality conditions, we prove several interesting structural properties of the optimal schedule. For instance, our analysis shows that when the appointment system is operating under the optimal schedule, other than the first slot and the last few (where the consultation intervals allocated are zero, i.e., patients are bunched together), the chances of waiting for service in the clinic is identical for patients assigned to all other slots. Furthermore, our model can also handle the correlations between patients' service durations, which has been largely overlooked in literature.

Computationally, we solve a tractable semidefinite approximation to the COP. Although the schedule obtained using our model is optimal for a set of canonical service-time distributions (called *worst-case distributions*), our numerical results show that this schedule also works reasonably well for several other service-duration distributions with the same moment conditions. We also find that the schedule obtained from solving the SDP approximation often satisfies the structural properties obtained from model analysis. Furthermore, with the help of existing semidefinite programming (SDP) packages, we can now obtain near optimal schedule for practical-size appointment-scheduling problems.

In a congested system with two types of patients, as in our eye clinic case, the optimal schedule often exhibits the pattern: "Bailey's Rule + Break"[3]—the optimal schedule allocates near-zero time slots to the first few patients, which resembles the well-known "Bailey's Rule," and a break is often inserted before switching from a class of patients with higher variability to another class of patients with lower variability. We use this observation and the solution from the SDP model to develop a simple and practical schedule for the eye clinic. Compared to the naive approach of allocating an equal interval to each patient with a break in-between (which is current practice in the clinic), our schedule can

reduce the total system waiting cost by around 36%. This approach thus has the potential of producing near-optimal appointment schedules that can be deployed in practice.

1.2. Structure of the Paper

In the next section, we briefly review the relevant literature for our problem. In §3, we describe the development of our conic optimization model in two steps, followed by several important extensions in §4 to address more practical issues. In §5, we analyze the structure and properties of the optimal scheduling policy, whereas in §6, numerical studies are presented to evaluate our approach under various circumstances as well as a case study of the eye clinic. We conclude in §7.

2. Literature Review

Since the pioneering work of Bailey (1952) and Welch and Bailey (1952), there have been extensive studies on the appointment design problem in the past six decades. In this section, we briefly review some key findings that are most relevant to our paper, but refer the readers to Cayirli and Veral (2003), Gupta (2007), Gupta and Denton (2008), and Erdogan and Denton (2010) for more thorough reviews.

There are several pertinent issues related to the design of a good appointment system. Some research considers the uncertainty of patient no-shows, i.e., a patient fails to show up for his or her appointment (Bailey 1952, Ho and Lau 1992, LaGanga and Lawrence 2007, and Chen and Robinson 2013). Some consider the possibility of (emergency) walk-in patients without appointments (Fetter and Thompson 1966, Vissers and Wijngaard 1979, Chen and Robinson 2013). Whereas the majority of research effort has been put into the analysis on the uncertainty of service durations (Bailey 1952, Wang 1993, Denton and Gupta 2003, Robinson and Chen 2003, Kaandorp and Koole 2007, Begen and Queyranne 2011), less is known about the optimal policy, even when the random service duration is the only source of uncertainty. Our work follows this line of research and tries to solve the problem from a different perspective. Using the primal–dual optimality conditions of our convex conic formulation, we prove several interesting structural properties of the optimal schedule.

Denton and Gupta (2003) formulated the appointment-scheduling problem as a two-stage stochastic linear program and used a sequential bounding approach to determine upper bounds of the problem. Kaandorp and Koole (2007) assumed that the service durations follow an exponential distribution and that the patient arrivals can only be scheduled at discrete intervals. They used results in queueing theory to calculate the objective function for a given schedule of starting times and used a local search algorithm to find the optimal solution. Begen and Queyranne (2011) went a step further and argued that under mild assumptions, the discrete-time version of the appointment-scheduling problem (i.e., the service-time distribution of each patient is given by a joint discrete probability distribution) could be solved in polynomial time, by showing that the objective function is an L-convex function. Unfortunately, the results are of theoretical nature, and no numerical solutions were presented. A recent paper by Begen et al. (2012) was based on the methodology developed in Begen and Queyranne (2011), but assumed no prior knowledge of probability distributions on job durations. They reconstructed an empirical distribution of the consultation durations from a set of historical data and then developed a sampling-based approach and established the cost (numbers of samples needed) to obtain a near-optimal solution with high probability. Thus far, simulation and stochastic programming remain the main solution methodologies for the appointment-scheduling problem.

For additional work closely related to this paper, see Boyd and Vandenberghe (2004), Burer (2009), Denton et al. (2010), Klassen and Rohleder (1996), Liu and Liu (1998), Robinson and Chen (2010), Rohleder and Klassen (1996), and Vanden and Dietz (2000).

In view of the analytical and computational difficulties of the problem, we address this issue from a different angle, utilizing the concept of robust optimization. Evolving from the *minimax theorem* established by John Von Neuman in 1928, the concept was first brought into the operations research area by Scarf (1958). Scarf solved an inventory problem with random demand by assuming only the mean and variance of the demand instead of a specific form of distribution. Noting that there could be multiple distributions that satisfy a given mean and variance, Scarf identified a worst-case distribution that would result in the highest expected total system cost, and found an inventory strategy to minimize this maximal cost. That is why another popular term describing the concept is *distributionally robust*. Recently, this concept has been extensively studied and developed. One stream of research is to exploit the connection between the theory of moments and semidefinite programming (SDP) (cf. Bertsimas et al. 2004, 2006, 2008; Vandenberghe et al. 2007). In a recent paper, Natarajan et al. (2010) showed that a robust mixed 0-1 linear program under objective uncertainty is equivalent to a convex conic program, which would be helpful in dealing with a second-stage recourse function in a two-stage stochastic programming framework.

In this paper, we exploit the ideas and tools of (distributionally) robust optimization to study the traditional appointment-scheduling problem. The concept of a distributionally robust model fits naturally in such a context because it is difficult to characterize the exact distribution of the service durations. Current research has assumed a wide range of distributions, like gamma distribution (Bailey 1952, Soriano 1966, Denton and Gupta 2003), uniform distribution (Ho and Lau 1992, Denton and Gupta 2003), exponential distribution (Ho and Lau 1992, Wang 1993, Kaandorp and Koole 2007), normal distribution (Denton and Gupta 2003), and log-normal distribution (Cayirli et al. 2008, Chen

and Robinson 2013), etc. On the other hand, estimating the moments of service durations is relatively easier, and the estimators are much more reliable. Therefore, finding a scheduling policy that could perform reasonably well against various distributions satisfying certain moment conditions appears to be a promising direction to solve the scheduling problem.

3. A Two-Stage Model with the Copositive Cone

3.1. Assumptions, Notation, and Problem Formulation

To isolate the impact of scheduling on the system performance, we rule out the presence of other disruptions in the system. The basic assumptions are listed as follows:

1. The sequence of patient arrivals is fixed. Service occurs in the same sequence.

2. Patients arrive punctually at the scheduled appointment times.[4]

3. There is a single doctor in the facility. The doctor arrives punctually and only serves the scheduled patients during the session. No break is taken during the time serving one patient.

4. Patients in the same class are homogenous in the distribution of consultation durations.

5. Walk-in and emergency patients are not considered.

Note that in a typical appointment-scheduling problem, it is common for the patients to choose the appointment slots in a dynamic fashion, and their characteristics, such as mean and standard deviation of service time, are known only at the time of booking. The problem described above matches more the surgery-scheduling environment. However, in certain appointment-scheduling environments, patients are classified into distinct classes, and each appointment slot in a single clinical session is preassigned to a dedicated class of patients. The slots are filled up when patients call in for appointments and their classifications are revealed. We assume that the clinic has enough volume to fill up the slots available in each day. In this way, the scheduling problem described here essentially addresses the design of the appointment system based on the patient classifications, not on the characteristics of individual patients.

Let $N = \{1, 2, \ldots, n\}$ be the index set for all patients, and the sequence of arrivals is $1, 2, \ldots, n$. Let $\tilde{u}_i$ be the random service time of patient i, $i = 1, 2, \ldots, n$. We define $\mathbf{s} = \{s_1, s_2, \ldots, s_n\}^T$, where s_i represents the length of the time slot scheduled for the ith patient in the sequence. Therefore, the appointment time of the patients in the sequence is given by $\{0, s_1, s_1 + s_2, \ldots, \sum_{i=1}^{n-1} s_i\}$.

We assume that $\tilde{u}_i$ follows a distribution with mean μ_i and standard deviation σ_i, and $\mathbf{P}(\tilde{u}_i \geqslant 0) = 1$, i.e., $\tilde{u}_i$ has nonnegative support. Let w_i denote the waiting time of the ith patient in the sequence. It is reasonable to assume that the first session starts at time zero, i.e., $w_1 = 0$. Define $\tilde{c}_i$ to be the difference between the actual consultation time and the allocated consultation interval of the ith patient in the sequence, i.e., $\tilde{c}_i = \tilde{u}_i - s_i$, $i = 1, \ldots, n$. Then the waiting time of subsequent patients are given by the following recursions:

$$w_i = \max\{0, w_{i-1} + \tilde{c}_{i-1}\}, \quad i = 2, 3, \ldots, n.$$

More precisely,

$$w_i = \max\left\{0, \tilde{c}_{i-1}, \tilde{c}_{i-1} + \tilde{c}_{i-2}, \ldots, \sum_{k=1}^{i-1} \tilde{c}_k\right\}, \quad i = 2, 3, \ldots, n. \tag{1}$$

If there were an additional "auxiliary" patient (i.e., the $(n+1)$st patient) arriving at the end of the consultation session, then the doctor's overtime would be exactly the waiting time of this patient, i.e., $w_{n+1} = \max\{0, w_n + \tilde{c}_n\}$. In this paper, we will use the total patients' waiting time and doctor's overtime (i.e., $\sum_{k=1}^{n} w_k$ and w_{n+1}) as the key performance indicators of the appointment system. The objective of the appointment-scheduling problem is to minimize the expectation of the weighted sum of the patients' waiting times and the doctor's overtime, i.e.,

$$\mathbf{E}\left[\sum_{i=1}^{n} \rho_i w_i + \rho_{n+1} w_{n+1}\right], \tag{2}$$

where ρ_i, $i = 1, 2, \ldots, n+1$ are the corresponding weights (or the unit waiting time/overtime cost). We first assume that $\rho_i = 1$ for all $i = 1, \ldots, n+1$, and then relax this assumption in §4.

Note that the doctor's total idle time during the session is also a crucial performance indicator of the appointment system. When the consultation interval (i.e., the session length, denoted as T) is predetermined, the total idle time is $T + w_{n+1} - \sum_{i=1}^{n} \tilde{u}_i$. Hence, we do not include the doctor's idle time in the objective because adding the expected total idle time can only cause the objective function to differ by a constant and the weight of w_{n+1} to increase by 1.

The technical difficulty associated with the scheduling problem is partially due to the computation of

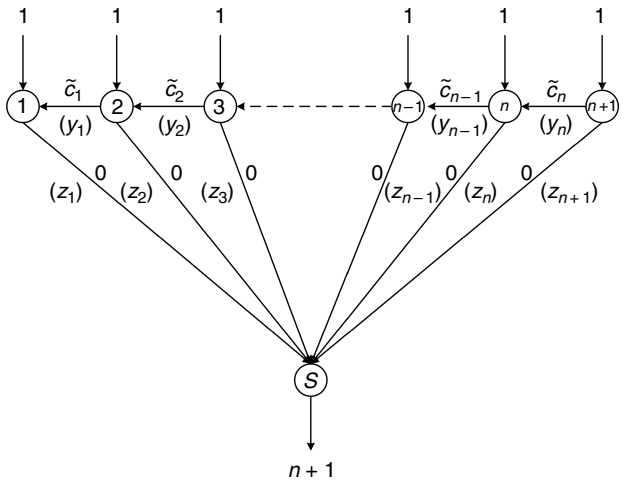
$$\mathbf{E}[w_i] = \mathbf{E}\left[\max\left\{0, \tilde{c}_{i-1}, \tilde{c}_{i-1} + \tilde{c}_{i-2}, \ldots, \sum_{k=1}^{i-1} \tilde{c}_k\right\}\right], \quad i = 2, 3, \ldots, n.$$

We introduce a two-stage stochastic optimization framework to tackle this problem. In the first stage, the appointment-scheduling decisions are made under the objective to minimize the expected total waiting time cost[5] defined in Equation (2). In the second stage, the patients' service durations are realized and the system performance is determined. Let us consider the second-stage problem first.

3.2. The Second-Stage Problem

Given the schedule of the patients (i.e., $\mathbf{s}$ is known), the total waiting time cost in Equation (2) can be computed

Figure 2. Network flow representation of the appointment-scheduling problem.



by solving a network flow problem on a directed acyclic graph shown in Figure 2, with $n+1$ supply nodes and a sink node s. The cost on arc (i, s) is 0, and the cost on arc $(i+1, i)$ is $\tilde{c}_i(\mathbf{s}) = \tilde{u}_i - s_i$, where the notation $\tilde{c}_i(\mathbf{s})$ is used here to emphasize the dependencies of $\tilde{c}_i$ on the given schedule $\mathbf{s}$ (not in the figure). The capacities for all the arcs are infinite. Let y_i, $i = 1, 2, \ldots, n$ be the flows on arc $(i+1, i)$, and z_i, $i = 1, 2, \ldots, n+1$ be the flows on arc (i, s).

Proposition 1. *Given the schedule* $\mathbf{s}$*, the optimal cost of the following maximum cost flow problem equals the total waiting time cost of the system under any realization of* $\tilde{\mathbf{u}}$:

$$
\begin{aligned}
f(\mathbf{s}, \tilde{\mathbf{u}}) := \max \quad & \sum_{i=1}^{n} \tilde{c}_i(\mathbf{s}) \cdot y_i = \sum_{i=1}^{n} (\tilde{u}_i - s_i) \cdot y_i \\
\text{s.t.} \quad & y_1 - z_1 = -1, \\
& y_i - y_{i-1} - z_i = -1, \quad \forall\, i = 2, 3, \ldots, n, \\
& -y_n - z_{n+1} = -1, \\
& y_i \geqslant 0, \quad \forall\, i = 1, 2, \ldots, n, \\
& z_i \geqslant 0, \quad \forall\, i = 1, 2, \ldots, n+1.
\end{aligned}
$$

Proof. The proposition can be easily verified through tracking the flow of each unit of supply at node $1, 2, \ldots, n+1$. A detailed argument can be found in EC.1 of the electronic companion to this paper (available as supplemental material at http://dx.doi.org/10.1287/opre.2013.1158). □

Remark 1. Note that Proposition 1 is developed in the deterministic situation. In the second stage, the patients' service durations are realized, i.e., they can be considered as deterministic. Then the network optimization problem in Proposition 1 is proposed to find out the total waiting time cost under this realization. When the patients' service durations ($\tilde{\mathbf{c}}(\mathbf{s})$) become stochastic, the optimal value of the network flow problem ($f(\mathbf{s}, \tilde{\mathbf{u}})$) also becomes stochastic and depends on $\tilde{\mathbf{c}}(\mathbf{s})$.

Removing one redundant network flow conservation constraint and using matrix notation, we rewrite $f(\mathbf{s}, \tilde{\mathbf{u}})$ as follows for the ease of exposition:

$$
\begin{aligned}
f(\mathbf{s}, \tilde{\mathbf{u}}) = \max \quad & \tilde{\mathbf{c}}^T(\mathbf{s})\mathbf{y} \\
\text{s.t.} \quad & \mathbf{a}(j)^T\mathbf{y} - \mathbf{e}(j)^T\mathbf{z} = -1, \quad \forall\, j = 1, 2, \ldots n \\
& \mathbf{y},\, \mathbf{z} \geqslant 0
\end{aligned}
$$

where $\tilde{\mathbf{c}}(\mathbf{s}) = (\tilde{c}_1(\mathbf{s}), \tilde{c}_2(\mathbf{s}), \ldots, \tilde{c}_n(\mathbf{s}))^T$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$, and $\mathbf{z} = (z_2, z_3, \ldots, z_{n+1})^T$; and $\mathbf{e}(j) \in \mathbb{R}^n$ is the unit vector with its jth entry being one; and $\mathbf{a}(j)_j = -1$ for $j = 1, \ldots n$, $\mathbf{a}(j)_{j+1} = 1$ for $j = 1, \ldots, n-1$, and $\mathbf{a}(j)_k = 0$ otherwise.

3.3. The First-Stage Problem

As mentioned before, we will deploy the minimax approach in our modeling framework, which we need to address before solving the scheduling problem. Under a fixed schedule $\mathbf{s}$, when the service durations become stochastic, but with given moment conditions, the maximal expected total waiting time cost can be written as:

$$
\text{(P)} \quad Z_P(\mathbf{s}) := \sup_{\tilde{\mathbf{u}} \sim (\boldsymbol{\mu}, \Sigma)^+} \left\{ \mathbf{E}_{\tilde{\mathbf{u}}}[f(\mathbf{s}, \tilde{\mathbf{u}})] \right\},
$$

where $\tilde{\mathbf{u}} \sim (\boldsymbol{\mu}, \Sigma)^+$ denotes that the distribution of $\tilde{\mathbf{u}}$ lies in the set of feasible multivariate distributions supported on $\mathbb{R}^n_+$ with finite first moment $\boldsymbol{\mu}$ and finite second moment Σ. We assume this set to be nonempty. The challenge to solve (P) reduces to the following: can one find a distribution for the random variable $\tilde{\mathbf{u}}$ in such a way that

$$
\mathbf{P}(\tilde{\mathbf{u}} \geqslant 0) = 1, \quad \mathbf{E}[\tilde{\mathbf{u}}] = \boldsymbol{\mu}, \quad \mathbf{E}[\tilde{\mathbf{u}}\tilde{\mathbf{u}}^T] = \Sigma,
$$

and a corresponding optimal solution $(\mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}), \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}))$ to $f(\mathbf{s}, \tilde{\mathbf{u}})$ in (P), so that $\mathbf{E}[\tilde{\mathbf{c}}(\mathbf{s})^T\mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}})]$ attains the maximum $Z_P(\mathbf{s})$? In general, if the maximum cannot be attained, can one find a sequence of random variables so that Z_P can be attained asymptotically?

It turns out that this problem can be reformulated into a conic programming problem through a moment decomposition approach. Before showing the main result, we introduce necessary notation and briefly review related subjects on the conic optimization problem.

3.3.1. Notation and a Brief Review of Conic Optimization. The trace of a matrix A, denoted by $\text{tr}(A)$, is the sum of the diagonal entries of A. The inner product between matrices A and B of the same dimensions is denoted as $A \cdot B = \text{tr}(A^T B)$. We use I_n to represent the identity matrix of dimension $n \times n$, whereas $\mathbf{0}_{m \times n}$ is used to denote the zero matrix of dimension $m \times n$. We may drop the subscript when it represents a zero vector of an appropriate dimension that is obvious.

For any cone $\mathcal{K}$, its dual cone is denoted as $\mathcal{K}^*$. Let $\mathcal{S}_n$ denote the cone of $n \times n$ *symmetric matrices*, and $\mathcal{S}_n^+$

denote the cone of $n \times n$ *positive semidefinite matrices*. $A \succeq 0$ indicates that the matrix A is positive semidefinite, and $B \succeq A$ is equivalent to $B - A \succeq 0$. Similarly, $A \geqslant 0$ indicates that the matrix A has nonnegative entries, and $B \geqslant A$ is equivalent to $B - A \geqslant 0$.

Two cones of special interest are the cone of *completely positive matrices* and the cone of *copositive matrices*. The cone of $n \times n$ completely positive matrices is defined as

$$\begin{aligned}\mathscr{CP}_n &:= \{A \in \mathscr{S}_n \colon \exists\, V \in \mathbb{R}_+^{n\times k}, \text{ such that } A = VV^T\} \\ &= \operatorname{conv}\{\mathbf{v}\mathbf{v}^T \colon \mathbf{v} \in \mathbb{R}_+^n\},\end{aligned}$$

where "conv" means the convex hull. The cone of $n \times n$ *copositive matrices* is defined as

$$\mathscr{CO}_n := \{A \in \mathscr{S}_n \colon \forall\, \mathbf{v} \in \mathbb{R}_+^n, \mathbf{v}^T A \mathbf{v} \geqslant 0\}.$$

$A \succeq_{\text{cp}} (\succeq_{\text{co}}) 0$ indicates that matrix A is completely positive (copositive). These two cones are both closed, convex, pointed, and have nonempty interior. Moreover, they are duals of each other (cf. Berman and Shaked-Monderer 2003). A linear program over the cone of copositive matrices is called a *copositive program* (*COP*), whose dual problem is a linear program over the cone of completely positive matrices known as a *completely positive program* (*CPP*).

Despite the nice properties of these two cones, it is widely believed that their membership status is $\mathscr{NP}$-hard to check. For instance, the problem of testing if a given matrix is copositive is known to be co-$\mathscr{NP}$-complete (cf. Murty and Kabadi 1987). In a recent paper, Dickinson and Gijben (2013) showed that the membership problems for both copositive and completely positive cones are $\mathscr{P}$-hard. Fortunately, there are well-known hierarchies of linear and semidefinite representable cones that approximate the copositive and completely positive cones (cf. Bomze et al. 2000, Klerk and Pasechnik 2002, Parrilo 2000). In this paper, we restrict our attention to the simplest relaxations of CPP and COP for the numerical experiments, i.e.,

$$\begin{cases} A \succeq_{\text{cp}} 0 \approx A \succeq 0, & \text{and } A \geqslant 0, \\ A \succeq_{\text{co}} 0 \approx \exists\, A_1 \succeq 0, & \text{and } A_2 \geqslant 0, \end{cases} \quad \text{such that } A = A_1 + A_2. \tag{3}$$

More information on CPP and COP can be found in Berman and Shaked-Monderer (2003).

3.3.2. Moment Decomposition and Conic Representation. For ease of exposition, we define

$$\mathbf{x} = \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix},$$

and rewrite the network flow constraints as $A\mathbf{x} = \mathbf{b}$, $\mathbf{x} \geqslant 0$, where

$$A = \begin{pmatrix} \mathbf{a}(1)^T & -\mathbf{e}(1)^T \\ \mathbf{a}(2)^T & -\mathbf{e}(2)^T \\ \vdots & \vdots \\ \mathbf{a}(n)^T & -\mathbf{e}(n)^T \end{pmatrix}, \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}.$$

Because A has full rank, the only feasible solution to $A\mathbf{x} = \mathbf{0}$ and $\mathbf{x} \geqslant 0$ is $\mathbf{x} = \mathbf{0}$.

Let

$$\mathscr{D} := \operatorname{conv}\left\{ \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix} \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix}^T : \pi \in \mathbb{R}_+, \ \mathbf{t} \in \mathbb{R}_+^n, \mathbf{v} \in \mathbb{R}_+^{2n}, A\mathbf{v} = \mathbf{b}\pi \right\}. \tag{4}$$

From the definition of $\mathscr{CP}_n$, we know that $\mathscr{D}$ is indeed the intersection of the completely positive cone, $\mathscr{CP}_{3n+1}$ with a hyperplane in $\mathbb{R}^{2n+1}$ projected onto $\mathbb{R}^{3n+1}$ (i.e., a polyhedral cone in $\mathbb{R}^{3n+1}$). Furthermore, if $\pi = 0$, then $A\mathbf{v} = \mathbf{0}$ and consequently $\mathbf{v} = \mathbf{0}$. Therefore, every $Z \in \mathscr{D}$ can be expressed as

$$Z = \sum_{k \in K_+} \pi(k)^2 \begin{pmatrix} 1 \\ \dfrac{\mathbf{t}(k)}{\pi(k)} \\ \dfrac{\mathbf{v}(k)}{\pi(k)} \end{pmatrix} \begin{pmatrix} 1 \\ \dfrac{\mathbf{t}(k)}{\pi(k)} \\ \dfrac{\mathbf{v}(k)}{\pi(k)} \end{pmatrix}^T + \sum_{k \in K_0} \begin{pmatrix} 0 \\ \mathbf{t}(k) \\ \mathbf{0}_{2n\times 1} \end{pmatrix} \begin{pmatrix} 0 \\ \mathbf{t}(k) \\ \mathbf{0}_{2n\times 1} \end{pmatrix}^T, \tag{5}$$

where K_+ and K_0 are the corresponding indicator sets, and they can be chosen to be finite[6] (c.f. Berman and Shaked-Monderer 2003).

If $Z_{1,1} = 1$, then $\pi(k)^2$ can be interpreted as the probability of the kth scenario with service duration $\tilde{\mathbf{u}} = \mathbf{t}(k)/\pi(k)$ and solution $\mathbf{x}(\mathbf{s}, \tilde{\mathbf{u}}) = (\mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}), \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}})) = \mathbf{v}(k)/\pi(k)$. The corresponding objective function in the kth scenario is given by $\sum_{i=1}^n (\tilde{u}_i - s_i) y(\mathbf{s}, \tilde{\mathbf{u}})_i$. Averaging over all the scenarios, each with probability $\pi(k)^2$, we get the objective function given by $Y(\mathbf{s}) \cdot Z$, where $Y(\mathbf{s})$ is a $(3n+1) \times (3n+1)$ symmetric matrix defined as

$$Y(\mathbf{s}) = \begin{pmatrix} 0 & \mathbf{0}_{1\times n} & -\dfrac{\mathbf{s}^T}{2} & \mathbf{0}_{1\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \dfrac{I_n}{2} & \mathbf{0}_{n\times n} \\ -\dfrac{\mathbf{s}}{2} & \dfrac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix}.$$

The second term in the expression for Z in (5) can be viewed as a characterization of the null set for the corresponding probability space. With such a moment decomposition interpretation, we get the following optimization problem by incorporating other moment conditions:

$$\begin{aligned} \text{(C)} \quad Z_C(\mathbf{s}) := \sup \ & Y(\mathbf{s}) \cdot Z, \\ \text{s.t.} \ & Z_{1,1} = 1, \quad Z_{1,i+1} = \mu_i, \\ & Z_{i+1,j+1} = \Sigma_{i,j}, \quad \forall\, i,\ j = 1, 2, \ldots, n, \\ & Z \in \mathscr{D}. \end{aligned}$$

Furthermore, we can prove that the above conic optimization problem is indeed equivalent to Problem (P).

PROPOSITION 2. *For any given schedule* $\mathbf{s}$, $Z_C(\mathbf{s}) = Z_P(\mathbf{s})$.

PROOF. There are two steps involved in the proof. Firstly, we show that Problem (C) provides an upper bound for (P), i.e., $Z_C(\mathbf{s}) \geqslant Z_P(\mathbf{s})$, $\forall \mathbf{s}$. Next, through a constructive approach, we find a sequence of random vectors, $\tilde{\mathbf{u}}^*_\epsilon$ that satisfies the moment conditions in the limiting sense and $\mathbf{E}[f(\mathbf{s}, \tilde{\mathbf{u}}^*_\epsilon)]$ converges to $Z_C(\mathbf{s})$ when ϵ converges to zero, i.e., the bound provided by (C) is tight. The technical details are omitted here but available in EC.2. □

REMARK 2. Note that our conic optimization model resembles the results of Natarajan et al. (2010), but is obtained from a different perspective. Instead of separating the moment requirement on $\tilde{\mathbf{u}}$ and the feasibility conditions on $\mathbf{x}$ and then enforcing their relationship through a lifting constraint, we directly characterize the cone $\mathcal{D}$ from the moment decomposition angle. There are several advantages to the new perspective—the dual program has a more intuitive interpretation with this approach, making the primal–dual relationship clear and transparent. Furthermore, strong conic duality follows directly via generalized Slater's constraint qualification (cf. Sturm 1999). This framework can also be easily extended to obtain conic relaxations for stochastic optimization problem with support conditions imposed on the random parameters.

Now we have a conic maximization problem that solves Problem (P) exactly. To incorporate the scheduling decision $\mathbf{s}$, we still need one more step, which is taking the dual of Problem (P).

3.3.3. Conic Duality and Copositive Program. Let $\mathcal{D}^*$ denote the dual cone of $\mathcal{D}$, i.e., $\mathcal{D}^* = \{W\colon Z \cdot W \geqslant 0, \forall Z \in \mathcal{D}\}$. Then the dual of Problem (C), denoted by $Z_D(\mathbf{s})$, can be written as follows:

$$
\begin{aligned}
Z_D(\mathbf{s}) := \inf\ & \Sigma \cdot \Gamma + \boldsymbol{\mu}^T\boldsymbol{\beta} + \alpha \\
\text{s.t. } & W = \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \mathbf{0}_{1\times 2n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & \mathbf{0}_{n\times 2n} \\ \mathbf{0}_{2n\times 1} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2n} \end{pmatrix} \\
& -Y(\mathbf{s}) = \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \frac{\mathbf{s}^T}{2} & \mathbf{0}_{1\times n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & -\frac{I_n}{2} & \mathbf{0}_{n\times n} \\ \frac{\mathbf{s}}{2} & -\frac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix}, \\
& W \in \mathcal{D}^*
\end{aligned}
$$

where $\alpha \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^n$ and $\Gamma \in \mathbb{R}^{n\times n}$ are the corresponding dual variables of the moment constraints.

By the definition of $\mathcal{D}^*$, for all $(1, \tilde{\mathbf{u}}, \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}), \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}))^T$ satisfying

$$
A\begin{pmatrix} \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}) \\ \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}) \end{pmatrix} = \mathbf{b}, \qquad \tilde{\mathbf{u}} \geqslant 0,\ \mathbf{y}(\mathbf{s},\ \tilde{\mathbf{u}}) \geqslant 0,\ \mathbf{z}(\mathbf{s},\ \tilde{\mathbf{u}}) \geqslant 0,
$$

we have

$$
\begin{aligned}
&\begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \frac{\mathbf{s}^T}{2} & \mathbf{0}_{1\times n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & -\frac{I_n}{2} & \mathbf{0}_{n\times n} \\ \frac{\mathbf{s}}{2} & -\frac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix} \cdot \left(\begin{pmatrix} 1 \\ \tilde{\mathbf{u}} \\ \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}) \\ \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}) \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\mathbf{u}} \\ \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}) \\ \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}) \end{pmatrix}^T \right) \\
&= \begin{pmatrix} 1 \\ \tilde{\mathbf{u}} \\ \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}) \\ \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}) \end{pmatrix}^T \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \frac{\mathbf{s}^T}{2} & \mathbf{0}_{1\times n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & -\frac{I_n}{2} & \mathbf{0}_{n\times n} \\ \frac{\mathbf{s}}{2} & -\frac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\mathbf{u}} \\ \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}) \\ \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}) \end{pmatrix} \geqslant 0,
\end{aligned}
$$

i.e.,

$$
\begin{pmatrix} 1 \\ \tilde{\mathbf{u}} \end{pmatrix}^T \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\mathbf{u}} \end{pmatrix} \geqslant (\tilde{\mathbf{u}} - \mathbf{s})^T \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}).
$$

Hence, for any distribution of the service durations, with probability 1,

$$
\begin{aligned}
&\begin{pmatrix} 1 \\ \tilde{\mathbf{u}} \end{pmatrix}^T \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\mathbf{u}} \end{pmatrix} \\
&\quad \geqslant \max \left\{ (\tilde{\mathbf{u}} - \mathbf{s})^T \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}})\colon A \begin{pmatrix} \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}) \\ \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}) \end{pmatrix} = \mathbf{b}, \tilde{\mathbf{u}} \geqslant 0, \right. \\
&\qquad\qquad \left. \mathbf{y}(\mathbf{s}, \tilde{\mathbf{u}}) \geqslant 0, \mathbf{z}(\mathbf{s}, \tilde{\mathbf{u}}) \geqslant 0 \right\}.
\end{aligned}
$$

Then the weak duality $Z_D(\mathbf{s}) \geqslant Z_C(\mathbf{s})$ follows immediately. Furthermore, because Problem (P) is obviously bounded, so is (C). Then, as long as there is a feasible solution to (C) that lies in the relative interior of $\mathcal{D}$, by the generalized Slater's constraint qualification, there is no duality gap between the primal $Z_C(\mathbf{s})$ and its dual $Z_D(\mathbf{s})$. Note that this condition is independent of the choice of $\mathbf{s}$. Furthermore, it is worthwhile to point out that $\mathcal{D}$ needs not be full dimensional for strong duality to hold. We use a simple example in EC.3 to illustrate this.

To convert $Z_D(\mathbf{s})$ into a copositive programming problem, we need to analyze the structure of the cone $\mathcal{D}$ and $\mathcal{D}^*$. Let $Z \in \mathcal{D}$, and

$$
\begin{aligned}
M_i &= \begin{pmatrix} b_i^2 & \mathbf{0}_{1\times n} & -b_i\mathbf{A}_i^T \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times 2n} \\ -b_i\mathbf{A}_i & \mathbf{0}_{2n\times n} & \mathbf{A}_i\mathbf{A}_i^T \end{pmatrix} \\
&= \begin{pmatrix} -b_i \\ \mathbf{0}_{n\times 1} \\ \mathbf{A}_i \end{pmatrix} \begin{pmatrix} -b_i \\ \mathbf{0}_{n\times 1} \\ \mathbf{A}_i \end{pmatrix}^T, \quad i = 1, 2, \ldots, n,
\end{aligned}
$$

where $\mathbf{A}_i^T$ is the ith row vector of A, i.e., $\mathbf{A}_i^T = (\mathbf{a}(i)^T - \mathbf{e}(i)^T)$. Note that

$$\begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix}^T M_i \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix} = (\mathbf{A}_i^T \mathbf{v} - b_i \pi)^2 \geq 0 \quad \text{if and only if } \mathbf{A}_i^T \mathbf{v} = b_i \pi.$$

Hence, we can rewrite $\mathcal{D}$ as

$$\mathcal{D} = \{Z\colon Z \cdot M_i = 0,\ \forall i = 1, 2, \ldots, n,\ Z \in_{3n+1}\}, \tag{6}$$

and it can be easily verified that

$$\mathcal{D}^* = \mathrm{cl}\left(\left\{W\colon W = V - \sum_{i=1}^{n} \gamma_i M_i,\ V \in \mathcal{CO}_{3n+1},\ \gamma_i \in \mathbb{R},\ i = 1, 2, \ldots, n\right\}\right), \tag{7}$$

where "cl" denotes "closure."

Under the assumption that there is a feasible solution to $Z_D(s)$ in the interior or $\mathcal{D}^*$, we obtain the following formulation for the appointment-scheduling problem:

$$\begin{aligned} Z_D(\mathbf{s}) := \inf\ & \Sigma \cdot \Gamma + \boldsymbol{\mu}^T \boldsymbol{\beta} + \alpha \\ \text{s.t.}\ & \begin{pmatrix} \alpha & \frac{\boldsymbol{\beta}^T}{2} & \frac{\mathbf{s}^T}{2} & \mathbf{0}_{1\times n} \\ \frac{\boldsymbol{\beta}}{2} & \Gamma & -\frac{I_n}{2} & \mathbf{0}_{n\times n} \\ \frac{\mathbf{s}}{2} & -\frac{I_n}{2} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \\ \mathbf{0}_{n\times 1} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n} \end{pmatrix} \\ & + \sum_{i=1}^{n} \gamma_i \begin{pmatrix} -b_i \\ \mathbf{0}_{n\times 1} \\ \mathbf{a}(i) \\ -\mathbf{e}(i) \end{pmatrix} \begin{pmatrix} -b_i \\ \mathbf{0}_{n\times 1} \\ \mathbf{a}(i) \\ -\mathbf{e}(i) \end{pmatrix}^T \succeq_{\mathrm{co}} 0, \end{aligned}$$

where the decision variables are $\alpha \in \mathbb{R}$, $\boldsymbol{\beta} \in \mathbb{R}^n$, $\Gamma \in \mathbb{R}^{n\times n}$, and $\boldsymbol{\gamma} \in \mathbb{R}^n$. We can now optimize the choice of $\mathbf{s} \in \mathbb{R}^n$, where $\mathbf{s}$ is constrained to be in a feasible set $\in \Omega_{\mathbf{s}}$. For example, $\mathbf{s} \in \Omega_{\mathbf{s}}$ in our case is

$$\sum_{i=1}^{n} s_i \leqslant T, \quad \text{and} \quad s_i \geqslant 0, \quad \forall i = 1, 2, \ldots, n, \tag{8}$$

which means the time slots must be nonnegative and the total scheduled time cannot exceed the session time T. We assume $T > 0$.

We have thus obtained the central result in this paper:

THEOREM 1. *Suppose there is a feasible solution to* (C) *that lies in the relative interior of $\mathcal{D}$, so that strong conic duality holds. Then*

$$\min_{\mathbf{s}\in\Omega_{\mathbf{s}}} \left\{ \sup_{\tilde{\mathbf{u}} \sim (\boldsymbol{\mu}, \Sigma)^+} \{\mathbf{E}[f(\mathbf{s}, \tilde{\mathbf{u}})]\} \right\} = \inf_{\mathbf{s}\in\Omega_{\mathbf{s}}} Z_D(\mathbf{s}). \tag{9}$$

Let (S) denote the convex conic programming problem on the right-hand side of (9). Dickinson (2010) showed that

$$\mathrm{int}(\mathcal{CP}_{3n+1}) = \left\{ \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T \colon \begin{array}{l} \mathbf{a}_i \in \mathbb{R}_+^{3n+1},\ \forall i = 1, 2, \ldots, m, \\ \mathrm{span}\{\mathbf{a}_1, \mathbf{a}_2 \ldots, \mathbf{a}_m\} = \mathbb{R}^{3n+1}, \\ \exists\, \mathbf{a} \in \{\mathbf{a}_1, \mathbf{a}_2 \ldots, \mathbf{a}_m\} \text{ such that } \mathbf{a} > 0 \end{array} \right\}.$$

Suppose $Z \in \mathrm{int}(\mathcal{CP}_{3n+1})$, with $Z = \sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T$ given by the above characterization, and $\mathbf{a}_i^T = (\pi_i, \mathbf{t}_i^T, \mathbf{v}_i^T)$. Because $\mathbf{a}_i \mathbf{a}_i^T \cdot M_k = (\mathbf{A}_k \mathbf{v}_i - b_k \pi_i)^2$, $Z \cdot M_k = 0$, $\forall k = 1, 2, \ldots, n$ if and only if $\mathbf{A}_k \mathbf{v}_i - b_k \pi_i = 0$, $\forall i,\ k = 1, 2, \ldots, n$. In this case, $\mathbf{v}_i \in \mathbb{R}^{2n}$ but $\dim(\mathrm{span}\{\mathbf{v}_i\colon i = 1, 2, \ldots, m\}) < 2n$. This is a contradiction, because it implies $\mathrm{span}\{\mathbf{a}_1, \mathbf{a}_2 \ldots, \mathbf{a}_m\} \neq \mathbb{R}^{3n+1}$. Thus the set $\mathcal{D}$ lies in the boundary of the completely positive cone.

To ensure that there is a feasible solution to (C) in the relative interior of $\mathcal{D}$, we assume that there is a set of scenarios $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_p\}$, with corresponding probabilities $\{\pi_1, \pi_2, \ldots, \pi_p\}$, so that

$$\sum_{i=1}^{p} \pi_i \begin{pmatrix} 1 \\ \mathbf{c}_i \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{c}_i \end{pmatrix}^T = \begin{pmatrix} 1 & \boldsymbol{\mu}^T \\ \boldsymbol{\mu} & \Sigma \end{pmatrix}.$$

Furthermore, we assume that the moments matrix lies in the interior of the completely positive cone, i.e., for any positively supported distribution $\tilde{\mathbf{s}}$ different from $\tilde{\mathbf{c}}$ and $\pi_{\tilde{\mathbf{s}}} \geqslant 0$, there exists a positively supported distribution $\tilde{\mathbf{t}}$, $\pi_{\tilde{\mathbf{t}}} \geqslant 0$ and $\lambda \in (0, 1)$ such that

$$\mathbf{E}\left[\begin{pmatrix} 1 \\ \tilde{\mathbf{c}} \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\mathbf{c}} \end{pmatrix}^T\right] = \lambda \mathbf{E}\left[\begin{pmatrix} \pi_{\tilde{\mathbf{s}}} \\ \tilde{\mathbf{s}} \end{pmatrix} \begin{pmatrix} \pi_{\tilde{\mathbf{s}}} \\ \tilde{\mathbf{s}} \end{pmatrix}^T\right] + (1 - \lambda)\mathbf{E}\left[\begin{pmatrix} \pi_{\tilde{\mathbf{t}}} \\ \tilde{\mathbf{t}} \end{pmatrix} \begin{pmatrix} \pi_{\tilde{\mathbf{t}}} \\ \tilde{\mathbf{t}} \end{pmatrix}^T\right].$$

In this case, we can assume without loss of generality that there is a point in the relative interior of $\mathcal{D}$, say $Z_{\mathbf{s}}$, because $\mathcal{D}$ is nonempty and convex. By the definition of $\mathcal{D}$, $Z_{\mathbf{s}}$ can be expressed as a sum of rank one matrices. We can apply the above assumption on $\tilde{\mathbf{c}}$ for every rank one matrix in $Z_{\mathbf{s}}$ to construct a point in $\mathcal{D}$ such that it satisfies all the moment conditions in Problem (C) and it can be written as a sum of $Z_{\mathbf{s}}$ and another point in $\mathcal{D}$.[7] Using the facts that $\mathcal{D}$ is a convex cone, and that adding a relative interior point of a convex cone with any point in the cone still results in a relative interior point of the cone, we can show that the point we found above lies in the relative interior of $\mathcal{D}$. Hence, the strong duality follows.

4. Extensions

In this section, we show that our model can be extended to capture more features of the practical appointment-scheduling problem, while still maintaining a formulation that is a compact convex conic optimization problem.

4.1. General Waiting Time Costs

In the earlier discussion, we have assumed that $\rho_i = 1$ for all patients. The network flow model used in the second-stage problem can be extended to cope with general waiting time costs ρ_i. This can be achieved by simply changing the inflow at each node i from 1 to ρ_i, and the outflow at node s from $n+1$ to $\sum_{i=1}^{n+1} \rho_i$. The reader can easily verify that the total waiting time cost is now mapped to the maximum cost flow problem in the network with the new supply and demand parameters.

4.2. Eye Test Before Consultation (Late Arrivals)

Suppose that the ith patient in the sequence has to undertake a test prior to the consultation. The test is often handled by a nurse and can be administered immediately upon arrival. The duration of the test is random and denoted by the random variable $\tilde{l}_i$. We define the waiting time of the patients to be *the waiting time needed to consult the doctor after the test is administered*. We also assume that the patients are seen by the doctor in the same sequence based on the appointment time, i.e., the sequence of the patients seen by the doctors is the same as the sequence of arrival. In this case, we can also use the network flow model to capture the impact of the test on the performance of the system. This is achieved by changing the cost on arcs (i, s), $i = 1, 2, \ldots, n$, from 0 to the random variables, $\tilde{l}_i$. Then the network flow solution in our model corresponds to the total waiting time cost in the system, offset by $\sum_{i=1}^{n} \tilde{l}_i$, i.e.,

$$
\begin{aligned}
f(\mathbf{s}, \tilde{\mathbf{u}}, \tilde{\mathbf{l}}) = \max \quad & \sum_{i=1}^{n} \tilde{c}_i(\mathbf{s}) \cdot y_i + \sum_{i=1}^{n} \tilde{l}_i z_i - \sum_{i=1}^{n} \tilde{l}_i \\
\text{s.t.} \quad & y_1 - z_1 = -1, \\
& y_i - y_{i-1} - z_i = -1, \quad \forall\, i = 2, 3, \ldots, n, \\
& -y_n - z_{n+1} = -1, \\
& y_i \geqslant 0, \quad \forall\, i = 1, 2, \ldots, n, \\
& z_i \geqslant 0, \quad \forall\, i = 1, 2, \ldots, n+1.
\end{aligned}
$$

To see this, note that when $z_i = 1$, the ith patient finishes the eye test and finds the doctor to be idling. This patient gets to consult the doctor at time $\tilde{l}_i$ after arrival. The waiting time is thus zero. This starts a new busy period, with the initial consultation duration given by $\tilde{l}_i + \tilde{c}_i(\mathbf{s})$, so we need to offset the objective by $\tilde{l}_i$. On the other hand, if after the test the patient finds the doctor to be busy, then $z_i = 0$ in the network flow solution, and hence the waiting time is simply the length of the longest path originating from node i deducted by $\tilde{l}_i$.

Then it is clear that we can extend the definition of the cone $\mathscr{D}^8$ to capture the impact of $\tilde{\mathbf{l}}$ just as $\tilde{\mathbf{u}}$, and finally we can still arrive at a convex conic optimization formulation for the appointment-scheduling problem with random prior tests. Note that the effect of such tests is exactly the same as late arrivals, i.e., patients arriving at a random time after the scheduled appointment. Thus, we can also address the issue of late arrivals with the same approach described above.

4.3. Relationship to Scenario Planning

In our model, we assume that only the moments and covariance parameters of the service durations are known. Then our model constructs a set of scenarios, the associated probability functions, and a solution that attains the (worst-case) performance objective under this set of scenarios. Our approach can be easily augmented to include specific scenarios when describing the uncertainty set for the service durations. More specifically, suppose that the system planner would like to construct the optimal schedule under the additional restrictions to include N scenarios $\mathbf{u}^L$ with probability p_L, such that $\sum_{L=1}^{N} p_L = p \leqslant 1$. Furthermore, the conditional first and second moments for the remaining scenarios are denoted by $(\boldsymbol{\mu}, \Sigma)^+$. Then our model reduces to

$$
Z_P(\mathbf{s}) = (1-p) \sup_{\tilde{\mathbf{u}} \sim (\boldsymbol{\mu}, \Sigma)^+} \{\mathbf{E}[f(\mathbf{s}, \tilde{\mathbf{u}})]\} + \sum_{L=1}^{N} p_L f_L(\mathbf{s}, \tilde{\mathbf{u}}),
$$

where $f(\mathbf{s}, \tilde{\mathbf{u}})$ is defined as before and

$$
\begin{aligned}
f_L(\mathbf{s}, \tilde{\mathbf{u}}) = \max \quad & \sum_{i=1}^{n} (u_i^L - s_i) \cdot y_i^L, \\
\text{s.t.} \quad & y_1^L - z_1^L = -1, \\
& y_i^L - y_{i-1}^L - z_i^L = -1, \quad \forall\, i = 2, 3, \ldots, n, \\
& -y_n^L - z_{n+1}^L = -1 \\
& y_i^L \geqslant 0, \quad \forall\, i = 1, 2, \ldots, n, \\
& z_i^L \geqslant 0, \quad \forall\, i = 1, 2, \ldots, n+1.
\end{aligned}
$$

In this way, we use a small set of scenarios to ensure that the optimal solution constructed will not perform too badly for these typical scenarios, and hence will not be overly conservative. Note that the dual to the above second-stage problem can be written using the approach described earlier, together with standard linear programming duality.

When $p = 1$, Z_P reduces to the conventional stochastic optimization problem solved via the sampling method. Hence, this framework can be viewed as a bridge between the traditional stochastic optimization and modern robust optimization.

4.4. Generalized Conic Relaxation Framework for More Support Information

For the random service time, except the moment conditions, we only require that they must be nonnegative. In general, there may be other conditions that the system planner would like to impose on the random service time, like a boundedness condition, etc. Our model provides a natural way to incorporate more support information through the construction of the cone $\mathscr{D}$. Recall that in Equation (6), we express $\mathscr{D}$ as

$$
\mathscr{D} = \left\{ Z \colon Z \cdot M_i = 0,\ \forall\, i = 1, 2, \ldots, n,\ Z \in \mathscr{C}_{3n+1} \right\}.
$$

We can view $\mathscr{D}$ as the intersection of the completely positive cone ${}_{3n+1}$ with

$$\mathscr{M}^i := \{Z\colon Z \cdot M_i = 0\}, \quad i = 1, 2, \ldots, n.$$

Whereas the network conservation constraints are embedded within $\mathscr{M}^i$, $\mathscr{CP}_{3n+1}$ captures both the nonnegativity constraints for the network flow variables and nonnegative support requirement of the random service time. Thus, it appears intuitive for us to augment $\mathscr{CP}_{3n+1}$ if we want to incorporate more support conditions. To develop a more general framework, we need the following lemma, which can be easily verified by the definition of a dual cone.

Lemma 1. *Suppose $\mathscr{K}^k \subseteq \mathbb{R}^{n\times n}$, $k = 1, 2, \ldots, m$, are closed convex cones. Let the dual cone of $\mathscr{K}^k$ be $\mathscr{K}^{k*}$. Then the dual cone of the following cone*

$$\mathscr{K}_n := \bigcap_{k=1}^{m} \mathscr{K}^k = \left\{A \in \mathbb{R}^{n\times n}\colon A \in \mathscr{K}^k,\ k = 1, 2, \ldots, m\right\}$$

is

$$\mathscr{K}_n^* := \mathrm{cl}\left(\sum_{k=1}^{m} \mathscr{K}^{k*}\right) = \mathrm{cl}\left(\left\{A \in \mathbb{R}^{n\times n}\colon \exists\, A_k \in \mathscr{K}^{k*},\ k = 1, 2, \ldots, m, \text{ such that } A = \sum_{k=1}^{m} A_k\right\}\right).$$

With Lemma 1, one can easily derive the expression of the dual cone of $\mathscr{D}$ as shown in Equation (7) by recognizing that the dual cone of $\mathscr{M}^i$ is $\mathscr{M}^{i*} := \{\gamma_i M_i\colon \gamma_i \in \mathbb{R}\}$. Thus, as long as the extra support conditions can be characterized with some conic constraints and their dual cones are compactly representable, we could still obtain a single conic optimization relaxation formulation for the appointment-scheduling problem.

For example, suppose the system planner would like to add some boundedness conditions for the random service time, which is characterized by the following ellipsoid constraint:

$$(\tilde{\mathbf{u}} - \bar{\mathbf{u}})^T \bar{Q} (\tilde{\mathbf{u}} - \bar{\mathbf{u}}) \leqslant r \text{ with probability 1,}$$

$$\text{for some } \bar{Q} \in S_n \subseteq \mathbb{R}^{n\times n},\ \bar{\mathbf{u}} \in \mathbb{R}^n \text{ and } r \in \mathbb{R}.$$

This constraint restricts the random service time to lie in an ellipsoid of size r centered at $\bar{\mathbf{u}}$. Using the probabilistic interpretation of $Z \in \mathscr{D}$, we can transform this condition into the following conic constraint on Z, together with the nonnegative and linear constraints,

$$Z \in \mathscr{E} := \mathrm{conv}\left\{ \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix} \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix}^T : \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix}^T \begin{pmatrix} r - \bar{\mathbf{u}}^T\bar{\mathbf{u}} & \bar{\mathbf{u}}^T\bar{Q} & \mathbf{0}_{1\times 2n} \\ \bar{Q}\bar{\mathbf{u}} & -\bar{Q} & \mathbf{0}_{n\times 2n} \\ \mathbf{0}_{2n\times 1} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2n} \end{pmatrix} \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix} \geqslant 0,\ \pi \in \mathbb{R}_+,\ \mathbf{t} \in \mathbb{R}_+^n,\ \mathbf{v} \in \mathbb{R}_+^{2n},\ A\mathbf{v} = b\pi. \right\}$$

Define

$$\Theta := \mathrm{conv}\left\{ \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix} \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix}^T : \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix}^T \begin{pmatrix} r - \bar{\mathbf{u}}^T\bar{\mathbf{u}} & \bar{\mathbf{u}}^T\bar{Q} & \mathbf{0}_{1\times 2n} \\ \bar{Q}\bar{\mathbf{u}} & -\bar{Q} & \mathbf{0}_{n\times 2n} \\ \mathbf{0}_{2n\times 1} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2n} \end{pmatrix} \begin{pmatrix} \pi \\ \mathbf{t} \\ \mathbf{v} \end{pmatrix} \geqslant 0,\ \begin{matrix} \pi \in \mathbb{R} \\ \mathbf{t} \in \mathbb{R}^n \\ \mathbf{v} \in \mathbb{R}^{2n} \end{matrix} \right\}.$$

Then it is obvious that $\mathscr{E} \subseteq \mathscr{D} \cap \Theta$, which provides a basis to get a relaxed conic optimization formulation for the problem. By Lemma 1, $(\mathscr{D} \cap \Theta)^* = \mathrm{cl}(\mathscr{D}^* + \Theta^*)$, where Θ^* is the dual cone of Θ. Furthermore, Θ^* can be easily obtained using S-Lemma (cf. Polik and Terlaky 2007), i.e.,

$$\Theta^* := \left\{ V \in \mathbb{R}^{(3n+1)\times(3n+1)}\colon \exists\, \tau \geqslant 0, \text{ such that } V - \tau \begin{pmatrix} r - \bar{\mathbf{u}}^T\bar{\mathbf{u}} & \bar{\mathbf{u}}^T\bar{Q} & \mathbf{0}_{1\times 2n} \\ \bar{Q}\bar{\mathbf{u}} & -\bar{Q} & \mathbf{0}_{n\times 2n} \\ \mathbf{0}_{2n\times 1} & \mathbf{0}_{2n\times n} & \mathbf{0}_{2n\times 2n} \end{pmatrix} \succeq 0 \right\},$$

which will translate into an extra semidefinite constraint in the final dual formulation of the problem. The resulted dual cone becomes

$$(\mathscr{D} \cap \Theta)^* = \mathrm{cl}\left(\left\{W\colon W = V_1 + V_2 + \sum_{i=1}^{n} \gamma_i M_i,\ V_1 \in \mathscr{CO}_{3n+1},\ V_2 \in \Theta^*,\ \gamma_i \in \mathbb{R},\ i = 1, 2, \ldots, n\right\}\right).$$

Therefore, we have come to a relaxed formulation for the appointment-scheduling problem with ellipsoidal support constraints.

5. Model Analysis

Our model provides a single deterministic convex formulation to solve a two-stage distributionally robust stochastic optimization problem. The formulation is exact under mild technical conditions so that strong conic duality holds. To the best of our knowledge, this model is the first of its kind. Furthermore, as shown in the development of the conic optimization model, the optimal solution to Problem (C) has a natural probabilistic interpretation under the worst-case distribution. Together with the network flow formulation of the waiting time experienced, they provide a new way to obtain some insights into the structure of the optimal appointment schedule. In the rest of this section, we show that the solution obtained from this deterministic model retains many of the intuitive properties of the optimal schedule under more realistic probabilistic consultation service distributions. To maintain the flow of this paper as well as to keep it succinct, we relegate all of the proofs in this section to EC.5. In terms of notation, we use the asterisk sign $(*)$ to indicate the respective optimal solution. For example, s_i^* denotes the optimal solution of s_i in Problem (S).

We show first that if there is a need to bunch the arrival of patients together, then it is optimal to bunch the arrivals at the end of the session. This is intuitive because whenever the consultation time is modeled by a nonnegative distribution, if bunching occurs for the $(i-1)$st and ith patient, but not the $(i+1)$st patient, then it is optimal to schedule the arrival of the ith patient slightly later and keep the schedule of the $(i+1)$st patient unchanged. The reason is obvious since the ith patient has to wait almost surely if she comes at the same time as the $(i-1)$st patient. The optimal schedule in our model retains this feature.

PROPOSITION 3. *Let the waiting time costs and overtime cost be strictly positive. In any optimal solution* $\mathbf{s}^*$ *to Problem* (S), *let* I *be the set of allocated service times, which are zero, i.e.,* $I := \{i\colon s_i^* = 0\}$. *Then* $I = \{n - |I| + 1, \ldots, n-1, n\}$, *i.e.,* I *is the last* $|I|$ *members of* $\{1, 2, \ldots, n\}$.

In practical settings, the nonnegativity constraints on the consultation slots (i.e., $s_i \geqslant 0, \forall\, i = 1, 2, \ldots, n$) enforce that all the appointment times are within the consultation session (T). Intuitively, if the system is heavily congested, it may be optimal to schedule some patients to arrive after time T, i.e., $\sum_{i=1}^{k} s_i > T$ for some $k < n$. To incorporate this into our model, we remove the nonnegativity constraints on the consultation slots. The next proposition shows that if these nonnegativity constraints are removed, only the last slot (s_n) can be negative in the optimal schedule as long as the costs of waiting time and overtime are strictly positive. Note that the scheduled arrival time of the nth patient is $\sum_{i=1}^{n-1} s_i$ and is therefore larger than T if $s_n < 0$, because $\sum_{i=1}^{n} s_i = T$. Furthermore, the constraint $\sum_{i=1}^{n} s_i = T$ ensures that the counting of the doctor's overtime starts from time T, and $s_n < 0$ in the network flow structure indicates that the doctor's overtime (i.e., the $(n+1)$th patient's waiting time) is at least $-s_n > 0$.

PROPOSITION 4. *Suppose the nonnegativity constraints on consultation slots* (*i.e., the second set of constraints in Equation* (8)) *are removed. When the waiting time costs and overtime cost are strictly positive, in the optimal solution to Problem* (S), *there is at most one negative slot. Furthermore, if this negative slot exists, it must be the last one, i.e.,* $s_i^* > 0, \forall\, i = 1, 2, \ldots, n-1$, *and* $s_n^* < 0$.

We investigate next the probability of a patient arriving at the scheduled time to find the system busy. From Figure 2, the flow y_i merges with ρ_i at node i. The probability that this combined flow goes through arc $(i, i-1)$ is exactly the probability that the ith patient has to wait. Otherwise, the flow on arc $(i, i-1)$ would be zero, which indicates that the waiting time cost is zero for the ith patient because arc (i, s) has zero flow cost. More precisely,

$$\begin{aligned}\mathbf{E}\big[y_{i-1}(\mathbf{s},\tilde{\mathbf{u}})\big] &= \mathbf{E}\big[\mathbf{E}[y_{i-1}(\mathbf{s},\tilde{\mathbf{u}}) \mid y_i(\mathbf{s},\tilde{\mathbf{u}})]\big]\\ &= \mathbf{E}\big[(y_i(\mathbf{s},\tilde{\mathbf{u}}) + \rho_i)\cdot \Pr\{i\text{th patient has to wait}\}\big]\\ &= (\mathbf{E}\big[y_i(\mathbf{s},\tilde{\mathbf{u}})\big] + \rho_i)\cdot \Pr\{i\text{th patient has to wait}\}\\ \Longrightarrow\quad \Pr\{i\text{th patient has to wait}\} &= \frac{\mathbf{E}[y_{i-1}(\mathbf{s},\tilde{\mathbf{u}})]}{\mathbf{E}[y_i(\mathbf{s},\tilde{\mathbf{u}})] + \rho_i}.\end{aligned}$$

Because the optimal $\mathbf{s}^*$ is selected to minimize

$$\begin{aligned}\mathbf{E}\Big[f(\mathbf{s},\tilde{\mathbf{u}})\Big] &= \mathbf{E}\Big[\sum_{i=1}^{n} \tilde{c}_i(\mathbf{s}) y_i(\mathbf{s},\tilde{\mathbf{u}})\Big]\\ &= \mathbf{E}\Big[\sum_{i=1}^{n} (\tilde{u}_i - s_i) y_i(\mathbf{s},\tilde{\mathbf{u}})\Big],\end{aligned}$$

From the first-order optimality conditions, we expect that at the optimal $\mathbf{s}^*$, if $s_{i-1}^* > 0$ and $s_i^* > 0$, then $\mathbf{E}[y_{i-1}(\mathbf{s}^*, \tilde{\mathbf{u}})] = \mathbf{E}[y_i(\mathbf{s}^*, \tilde{\mathbf{u}})]$. This indeed holds for the optimal schedule obtained using our model.

PROPOSITION 5. *If in the optimal solution to Problem* (S), *the allocated service-time slots are strictly positive,* (*i.e.,* $s_i^* > 0, \forall\, i \in I \subseteq \{1, 2, \ldots, n\}$), *then the network flow solution must satisfy* $\mathbf{E}[y_i(\mathbf{s}^*, \tilde{\mathbf{u}})] \equiv K, \forall\, i \in I$, *where* K *is some nonnegative constant.*

Combining the propositions established thus far, we can derive an important optimality condition for an appointment system:

THEOREM 2. *Suppose in the optimal solution to Problem* (S), *the allocated consultation slots are strictly positive for the first* k *patients* (*i.e.,* $s_i^* > 0, \forall\, i = 1, 2, \ldots, k$, *where* $0 < k \leqslant n$). *Furthermore, if* $\rho_i \equiv \rho$ *for some constant* $\rho > 0$, *for all* $i = 1, 2, \ldots, k$, *then the probabilities of waiting for the service are the same for all the patients from* $i = 2, \ldots, k$, *under the optimal worst-case distribution.*

REMARK 3. Note that the optimality condition stated in the above theorem is independent of the sequence of the patients. This property of the optimal schedule is particularly useful for the patients: there is little incentive to choose between the slots in the clinical session if the objective is to minimize the chances of waiting for the service.

6. Computational Results

All of the computational studies are carried out in MATLAB on a Dell desktop (Core 1.86 GHz and 3GB of RAM). We solve the simplest form of SDP relaxation of the COP and CPP as shown in Equation (3). In MATLAB, we use YALMIP as the programming interface with SDPT3 as the underlying SDP solver (cf. Löfberg 2004, Toh et al. 1999, Tutuncu et al. 2003).

Note that expressing a problem as a COP or CPP and relaxing it only partially resolve the difficulty of the problem, because even solving a large-scale SDP can be computationally prohibitive. Because our model lifts the original problem into a cone with higher dimensions, the current computational power limits the size of the problem instance we can solve to around 36 patients. While it is an interesting challenge to push the computational limit of this approach further, we leave this to future research. By "large-scale problem," we mean problems that involve hundreds or even thousands of variables. Fortunately, in practice we usually will not encounter such large-sized

Table 1. Near-optimal schedules from Denton and Gupta (2003) under different cost structures.

(ρ_1, ρ_{n+1})	(3, 14)	(5, 12)	(7, 10)	(3, 12)	(5, 10)	(7, 8)	(3, 10)	(5, 8)	(7, 6)
s_1	0.61	0.83	1.06	0.65	0.88	1.14	0.72	1.00	1.25
s_2	1.09	1.18	1.27	1.11	1.22	1.34	1.13	1.25	1.38
s_3	1.08	1.20	1.26	1.11	1.24	1.31	1.12	1.25	1.38
s_4	1.09	1.20	1.27	1.13	1.22	1.32	1.13	1.25	1.38
s_5	1.07	1.10	1.21	1.05	1.14	1.25	1.08	1.19	1.35
s_6	0.94	1.00	1.16	0.96	1.01	1.20	0.94	1.07	1.24
s_7	1.14	0.50	−0.23	1.01	0.31	−0.56	0.89	−0.01	−0.98

Table 2. Optimal schedules from our model under different cost structures.

(ρ_1, ρ_{n+1})	(3, 14)	(5, 12)	(7, 10)	(3, 12)	(5, 10)	(7, 8)	(3, 10)	(5, 8)	(7, 6)
s_1	0.35	0.87	0.94	0.52	0.89	0.99	0.76	0.92	1.05
s_2	1.32	1.09	1.16	1.22	1.10	1.20	1.08	1.13	1.26
s_3	1.05	1.17	1.25	1.08	1.19	1.30	1.11	1.22	1.38
s_4	1.12	1.29	1.38	1.16	1.31	1.44	1.21	1.35	1.53
s_5	1.20	1.31	1.36	1.23	1.31	1.42	1.26	1.33	1.50
s_6	1.17	1.27	1.20	1.20	1.20	1.25	1.24	1.18	1.33
s_7	0.79	0.00	−0.29	0.58	0.00	−0.61	0.33	−0.14	−1.04

problems. In the eye clinic case, we only need to schedule 36 patients for the whole morning session.

In what follows, we use extensive numerical experiments to provide a glimpse into the performance of the optimal scheduling solutions obtained using our model.

6.1. Comparison with Near-Optimal Solutions

In this section, we test the performance of our model against a set of near-optimal solutions given in Denton and Gupta (2003). Table 1 lists the near-optimal schedules given in that paper, for seven jobs with identically independent distributed service time ($Uniform(0, 2)$) under different cost structures and fixed session length $T = 7$. The waiting time costs are identical among all the patients. In their numerical results, the optimality gap is less than 1%. We solve our model to obtain the optimal schedule that minimizes the worst-case cost under all distributions with mean 1 and standard deviation $1/\sqrt{3}$. The results of our model are presented in Table 2. Note that in Denton and Gupta (2003), the objective function is the weighted sum of total waiting time, idle time and overtime of the doctor, whereas in our model the objective function does not include the cost of idle time. According to Proposition 1 in Denton and Gupta (2003) (similar to our argument in §3.1), we can transform the optimal scheduling problem in Denton and Gupta (2003) equivalently into our problem by combining the cost of idle time and overtime. Because Denton and Gupta (2003) allows negative schedules, we remove the nonnegativity constraints in Equation (8) when solving Problem (S) for a fair comparison.

Next, we compare the total waiting time costs under the schedules given in Tables 1 and 2 through Monte Carlo simulation. In evaluating our model, the service duration of each patient is generated under four common distributions used in practice: uniform, normal, two-point, and Gamma distribution, with mean 1 and standard deviation $1/\sqrt{3}$. All nine different cost structures are tested. Fifty thousand rounds of simulation are executed for each of the 36 scenarios (4 distributions × 9 costs structures).[9] The average total costs under different scenarios are then compared with the corresponding benchmark schedules given by Denton and Gupta (2003) under the uniform distribution. As shown in Table 3, the schedules obtained from our model work phenomenally well when evaluated against the benchmarks. The average total costs under our model is close to that of Denton and Gupta (2003) even under different distributions.

Table 3. Comparison of the average total costs between the schedules obtained by our model and Denton and Gupta (2003) under different distributions.

(ρ_1, ρ_{n+1})	(3, 14)	(5, 12)	(7, 10)	(3, 12)	(5, 10)	(7, 8)	(3, 10)	(5, 8)	(7, 6)
Benchmark	23.32	27.03	28.50	21.42	24.51	25.02	19.43	21.69	20.94
Uniform	23.55	27.62	28.89	21.55	24.79	25.48	19.60	21.94	21.48
Normal	23.57	27.77	28.98	21.63	24.92	25.55	19.72	22.03	21.53
Two-point	24.00	28.64	30.20	21.95	25.81	26.56	20.23	22.91	21.89
Gamma	22.73	27.53	28.87	20.93	25.08	25.84	19.48	22.10	22.21

The gaps are within 2%, and most of them are less than 1%. Moreover, it is worthwhile to point out that the average total costs of our schedules do not vary much under different distributions.

6.2. Eye Clinic

In this subsection, we present numerical results based on data collected from the eye clinic and discuss pertinent managerial insights from our model. We observed the consultation durations of 1,021 patients in the clinic for seven working days. The mean and standard deviation of the consultation time of the repeat patients are 6.24 minutes and 6.0 minutes, respectively, whereas the values for the new patients are noticeably higher, with a mean of 9.97 minutes and a standard deviation of 7.6 minutes.

In this experiment, we assume that one session lasts for 150 minutes. This mimics the current practice with a one-hour block, followed by a half-hour break, and then another one-hour block. During one session, 24 patients are scheduled to arrive in the clinic, with five new patients arriving before 19 repeat patients. The consultation durations follow the distributions with the mean and standard deviation as estimated by the empirical data. Note that the sum of mean service durations of all patients is 168.41 minutes, which is larger than the session length. This indicates that the system may be heavily congested.

The patient's waiting time cost (ρ_i) is assumed to be identical among all the patients and normalized to 1. We test various overtime costs, i.e., $\rho_{n+1} = 1$, 20, or 40. Figure 3 plots the optimal schedules obtained by our model under different ρ_{n+1}.

It is interesting to note that the optimal schedules exhibit the pattern of "Bailey's Rule + Break." First, the optimal schedule allocates a near-zero time slot to the first few patients. Although Proposition 3 indicates that all the zero time slots should be placed at the end of the session, the time slots for the first few patients are extremely small, although not zero—the smallest among these intervals are around 10^{-4}, well within the machine tolerance of being zero, i.e., 10^{-16} on a 32-bit machine.[10] This scheduling rule arises because the system is heavily congested and the overtime costs are large. The second outstanding feature is that after serving the group of new patients, a break is inserted before switching to the group of repeat patients with lower variability. To confirm this feature, we run another group of experiments with three classes of patients. Similar patterns are observed—breaks are inserted after serving the first and the second classes of patients.

One drawback of the optimal schedule is that it is generally not practical and is nonintuitive. To fix this problem, we try to use the above insights to develop a simple but effective appointment schedule. In the current practice, each patient is assigned with an equal interval of 5 minutes, and a 30-minute break is inserted after seeing 12 patients. We simply modify the "*Current Practice*" by reinserting the 30-minute break after the fifth patient, i.e., after serving all the new patients. We call this schedule "*Modified Practice*." Note that each patient is still assigned 5 minutes of consultation time.

In a more advanced system, we allow the allocated service intervals to vary accordingly. We denote this the "*Varying Interval*" schedule. To resemble the optimal schedule (under $\rho_{n+1} = 1$), we assign zero time slots to the first patient and the last six patients (this is a characteristic of the optimal schedule obtained from the conic program). Other patients are assigned with time slots by rounding up

Figure 3. Optimal schedule when ρ_{n+1} is equal to 1, 20, and 40, given $\rho_1 = 1$.

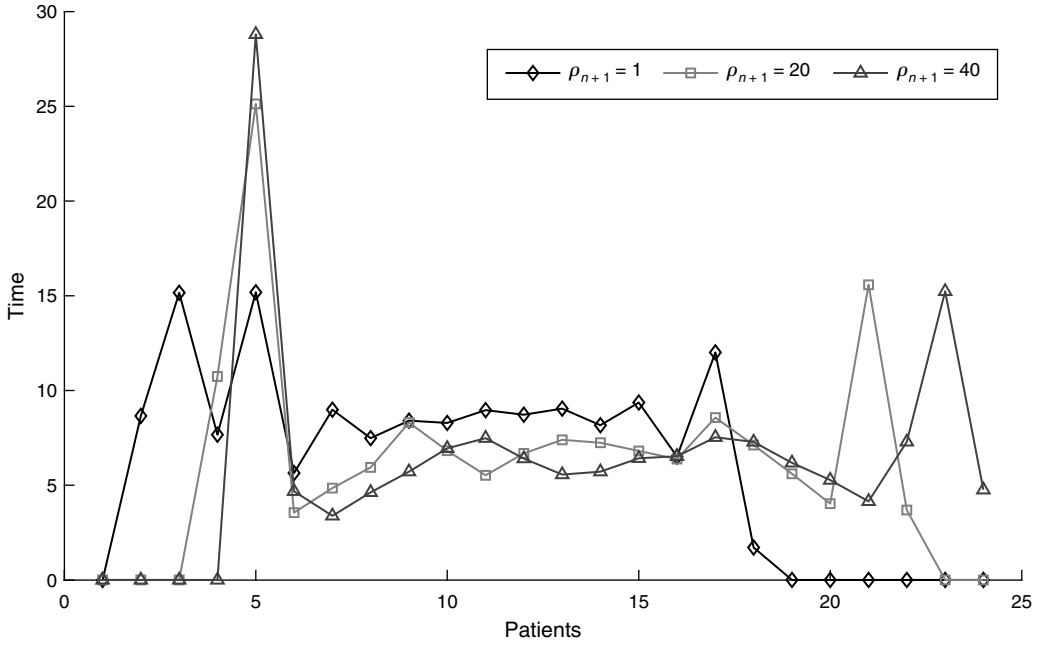


Table 4. Average total waiting time cost under different scheduling policies when $\rho_1 = 1$ and $\rho_{n+1} = 1$.

	Uniform	Normal	Two-point	Gamma
Optimal schedule	352.58	349.80	355.37	352.78
Current practice	564.13	560.18	570.31	535.37
Modified practice	485.36	479.95	491.95	462.44
Varying interval	358.24	353.61	363.60	354.83

their mean service durations, i.e., 10 minutes for a new patient and 7 minutes for a repeat patient. The remaining time is combined and inserted after the fifth patient as a break.

The simulated performance of various policies under different service-time distributions are shown in Table 4. Implementing a schedule resembling the optimal solution dramatically decreases the total waiting time cost by about 35% as compared to the current practice. Interestingly, it seems that one can significantly improve the performance of the system by simply inserting a break after serving one class of patients in the optimal scheduling. The easily implemented "Varying Interval" strategy makes it quite attractive for practical considerations.

Note that the above simulation results are obtained under $\rho_{n+1} = 1$. In most environments, the overtime cost ρ_{n+1} is likely to be large, and should be proportional to the number of patients seen in the clinic. The choice of $\rho_{n+1} = 1$ is thus a conservative estimate and assumes that the doctor places a small penalty on overtime work. In what follows, we summarize the features of optimal schedules when ρ_{n+1} increases. The pattern of "Bailey's Rule + Break" seems to be quite robust no matter how the overtime cost changes. Besides this, Figure 3 also illustrates several interesting features: As ρ_{n+1} increases,

- more patients are assigned with near-zero consultation time slots at the beginning of the session;
- fewer patients are assigned with zero time slots at the end of the session;
- a longer time slot is assigned to the last patient.

Intuitively, all these features benefit a clinic that prefers a shorter overtime. Consequently, patients may suffer from longer waiting times as a result.

As we can see from Figure 3, the optimal properties persist as the overtime cost ρ_{n+1} increases. One question is whether we can still design efficient appointment systems with the help of the optimal properties under different overtime costs. To answer this question, we first solve our model with $\rho_{n+1} = 2$, 5, 10, 20, 50, and 100, and then create the "Varying Interval" schedules using the following heuristics: allocating zero time slots to those clustering patients (with zero or close to zero time slots) at the beginning and the end of the session, assigning the rest of the patients their mean consultation durations, and inserting the remaining time as a break after the fifth patient. We simulate the expected total costs of "Varying Interval" schedules and compare with current practice. Table 5 records the efficiency gains under different overtime costs ρ_{n+1}. The percentage savings decrease as ρ_{n+1} increases. The efficiency gain drops to around 10% when $\rho_{n+1} = 100$. Because higher overtime cost indicates larger total cost, a 10% efficiency gain when $\rho_{n+1} = 100$ can save around 360 minutes in total waiting time. Hence, although efficiency gain drops as ρ_{n+1} increases, employing the "Varying Interval" schedule can still ensure significant efficiency improvement in the clinic.

Table 5. Efficiency gains under different overtime costs.

	Percentage increase (%)			
Overtime	Uniform	Normal	Two-point	Gamma
1	36.5	37.1	36.3	33.7
2	31.7	32.0	31.4	28.5
5	24.0	24.3	23.9	21.0
10	13.3	13.4	13.3	10.1
20	7.1	7.3	7.3	4.2
50	7.3	6.8	7.9	5.5
100	11.4	11.5	12.2	9.7

7. Conclusion

We propose a novel approach to deal with the difficult appointment-scheduling problem. Instead of planning against a fixed service distribution, we plan against a canonical set of service distributions with the same mean and covariance parameters (may include more support constraints). The canonical distribution is "constructed" via a general conic programming framework. In this way, we reduce a difficult two-stage stochastic programming problem into a single-stage convex programming problem. Through extensive simulations we show that the optimal schedules obtained under the "worst case" give near-optimal solutions when the objective is to minimize expected total waiting time cost. This approach allows us to shed some light on the structure of the optimal schedules, which can be readily modified to obtain more practical and efficient scheduling policies.

Our model is able to handle the correlations between the service durations of different patients. It has been a standard assumption that patients' service durations are independent of each other. However, in reality, this assumption may not hold due to the common resource—the doctor, who serves all the patients in a clinical session. The doctor could impact the service durations of all patients in the same session uniformly. Nevertheless, we leave the study of the impact of correlations for future research, while focusing on developing the methodology to solve the appointment scheduling problem.

The approach can be generalized to deal with the situation when the patients need to undergo a test (with random duration) prior to the consultation, which is a pertinent feature in many eye clinics. The network flow approach can also be conceivably extended to deal with other practical considerations in a clinical environment. There is,

however, some limitation with this approach: the computational difficulty associated with solving large-scale SDP limits our ability to solve large-scale appointment scheduling problems.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.2013.1158.

Endnotes

1. See http://www.iimahd.ernet.in/publications/data/2005-08-02nravi.pdf for a thorough discussion.
2. A copositive programming problem is a linear programming problem over the convex cone of the copositive matrices. Details of this optimization problem are discussed later in this paper.
3. "Bailey's Rule" refers to the scheduling strategy proposed in the seminal paper by Bailey (1952). It states that in a highly congested system, "*an optimum system seems to be as follows*: *the patients are given appointments at regular intervals equal to the average consultation time, and the consultant arrives at the same time as the second patient*" Bailey (1952, p. 198). That means the first two patients are scheduled to come at the beginning of the consultation session at the same time.
4. This assumption can be relaxed. In §4.2, we demonstrate how to extend our model to incorporate late arrivals.
5. In the rest of this paper, we use the phrase "the total waiting time (cost)" to include both the waiting time (costs) of all the patients and the overtime (cost) of the doctor.
6. Indeed, not only could they be finite, but also bounded. This is related to the concept of *cp-rank*, details of which can be found in Berman and Shaked-Monderer (2003).
7. Details of the decomposition and construction are available in EC.4.
8. More precisely, the new dimension of $\mathscr{D}$ is $(4n+1)\times(4n+1)$.
9. We obtain similar results through the test under a larger set of distributions as well, but only the four most commonly used distributions are reported in this paper.
10. We would like to thank one referee for pointing this out.

Acknowledgments

The authors thank Alexander Shapiro, the associate editor, and other anonymous reviewers for their valuable comments and suggestions on improving this manuscript. They are also grateful to John Buzacott, Gideon Weiss, and Diwakar Gupta for their constructive comments on earlier versions of the paper. The authors are particularly indebted to Jiajie Liang for providing the clinical data (Liang 2006), to Toh Kim Chuan for his help and guidance in tackling the computational difficulties of the problem, and to Peter J. C. Dickinson for identifying a glitch in the conic duality of the model.

References

Bailey NTJ (1952) A study of queues and appointment systems in hospital outpatient departments with special reference to waiting times. *J. Royal Statist. Soc., B* 14:185–199.

Begen MA, Queyranne M (2011) Appointment scheduling with discrete random durations. *Math. Oper. Res.* 36:240–257.

Begen MA, Levi R, Queyranne M (2012) A sampling-based approach to appointment scheduling. *Oper. Res.* 60:675–681

Berman A, Shaked-Monderer N (2003) *Completely Positive Matrices* (World Scientific, Singapore, Republic of Singapore).

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

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

Bertsimas D, Doan XV, Natarajan K, Teo CP (2008) Models for minimax stochastic linear optimization problems with risk aversion. *Math. Oper. Res.* 35:580–602.

Bomze IM, Dür M, Klerk ED, Roos C, Quist AJ, Terlaky T (2000) On copositive programming and standard quadratic optimization problems. *J. Global Optim.* 18:301–320.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press, Cambridge, UK).

Burer S (2009) On the copositive representation of binary and continuous nonconvex quadratic programs. *Math. Programming* 120:479–495.

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

Cayirli T, Veral E, Rosen H (2008) Assessment of patient classification in appointment system design. *Production Oper. Management* 17:47–58.

Chen RR, Robinson LW (2013) Sequencing and scheduling appointments with potential call-in patients. Submitted.

de Klerk E, Pasechnik DV (2002) Approximation of the stability number of a graph via copositive programming. *SIAM J. Optim.* 12:875–892.

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

Denton B, Miller A, Balasubramanian H, Huschka T (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. *Oper. Res.* 58:802–816.

Dickinson PJC (2010) An improved characterisation of the interior of the completely positive cone. *Electronic J. Linear Algebra* 20:723–729.

Dickinson PJC, Gijben L (2013) On the computational complexity of membership problems for the completely positive cone and its dual. Submitted.

Erdogan S, Denton B (2010) Surgery planning and scheduling: A literature review. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Hoboken, NJ).

Fetter R, Thompson J (1966) Patients' waiting time and doctors' idle time in the oupatient setting. *Health Services Res.* 1:66–90.

Gupta D (2007) Surgical suites' operations research. *Production Oper. Management* 16:689–700.

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

Ho CJ, Lau HS (1992) Minimizing total cost in scheduling outpatient appointments. *Management Sci.* 38:1750–1764.

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

Klassen KJ, Rohleder TR (1996) Scheduling outpatient appointments in a dynamic environment. *J. Oper. Management* 14:83–101.

LaGanga LR, Lawrence SR (2007) Clinical overbooking to improve patient access and increase provider productivity. *Decision Sci.* 38: 541–545.

Liang JJ (2006) Intelligent appointment scheduling to reduce turnaround time. Master's thesis, National University of Singapore, Singapore.

Liu L, Liu X (1998) Dynamic and static job allocation for multi-server systems. *IIE Trans.* 30:845–854.

Löfberg J (2004) YALMIP: A toolbox for modeling and optimization in MATLAB. *Proc. CACSD Conf., Taipei, Taiwan*. http://control.ee.ethz.ch/~joloef/yalmip.php.

Murty KG, Kabadi SN (1987) Some $\mathscr{NP}$-complete problems in quadratic and nonlinear programming. *Math. Programming* 39:117–129.

Natarajan K, Teo CP, Zheng Z (2010) Mixed zero-one linear programs under objective uncertainty: A completely positive representation. *Oper. Res.* 59:713–728.

Parrilo PA (2000) Structured semidefinite programs and semi-algebraic geometry methods in robustness and optimization. Ph.D. thesis, California Institute of Technology, Pasadena, CA. Accessed August 1, 2009, http://www.cds.caltech.edu/~pablo/.

Polik I, Terlaky T (2007) A survey on the *S*-lemma. *SIAM Rev.* 49: 371–418.

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

Robinson LW, Chen RR (2010) A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing Service Oper. Management* 12:330–346.

Rohleder TR, Klassen KJ (1996) Using client-variance informaion to improve dynamic appointment scheduling performance. *Omega* 28: 293–302.

Scarf H (1958) A min–max solution of an inventory problem. Arrow K, Karlin S, Scarf H, eds. *Studies in the Mathematical Theory of Inventory and Production* (Stanford University Press, Stanford, CA), 201–209.

Soriano A (1966) Comparison of two scheduling systems. *Oper. Res.* 14:388–397.

Sturm JF (1999) Theory and algorithms of semidefinite programming. Frenk H, Roos K, Terlaky T, Zhang S, eds. *High Performance Optimization, Part I, Applied Optimization*, Vol. 33 (Kluwer Academic Publishers, Boston), 21–60.

Toh KC, Todd MJ, Tutuncu RH (1999) SDPT3—A Matlab software package for semidefinite programming. *Optim. Methods Software* 11: 545–581.

Tutuncu RH, Toh KC, Todd MJ (2003) Solving semidefinite-quadratic-linear programs using SDPT3. *Math. Programming* 95:189–217.

Vanden B, Dietz CD (2000) Minimizing expected waiting in a medical appointment system. *IIE Trans.* 32:841–848.

Vandenberghe L, Boyd S, Comanor K (2007) Generalized Chebyshev bounds via semidefinite programming. *SIAM Rev.* 49:52–64.

Vissers J, Wijngaard J (1979) The outpatient appointment system: Design of a simulation study. *Eur. J. Oper. Res.* 3:459–463.

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

Wang PP (1999) Sequencing and scheduling *N* customers for a stochastic server. *Eur. J. Oper. Res.* 119:729–738.

Welch JD, Bailey N (1952) Appointment systems in hospital outpatient departments. *The Lancet* 259:1105–1108.

Qingxia Kong is an assistant professor at the UAI Business School, Universidad Adolfo Ibáñez, Chile. She works on medical decision making, behavioral decision making, and sport analytics.

Chung-Yee Lee is Cheong Ying Chan Chair Professor of Engineering at Hong Kong University of Science and Technology. He is the founding director and current director of the Logistics and Supply Chain Management Institute. His research areas are in logistics and supply chain management, scheduling, and inventory management.

Chung-Piaw Teo is professor and head of the Department of Decision Sciences at the NUS Business School, National University of Singapore. His research areas are in operations and supply chains, and optimization under uncertainty.

Zhichao Zheng is a Ph.D. candidate in the Department of Decision Sciences at the National University of Singapore. His current research interests lie in the area of prediction and planning under uncertainty using conic programming techniques. He applies his research in various industrial domains, including healthcare operations management, aircraft spares logistics service planning and contracting, and liner shipping scheduling.