Manufacturing & Service Operations Management

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

Sequencing Appointments for Service Systems Using Inventory Approximations

Ho-Yin Mak, Ying Rong, Jiawei Zhang

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 16, No. 2, Spring 2014, pp. 251–262
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs®
http://dx.doi.org/10.1287/msom.2013.0470

Sequencing Appointments for Service Systems Using Inventory Approximations

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

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

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

Managing appointments for service systems with random job durations is a challenging task. We consider a class of appointment planning problems that involve two sets of decisions: *job sequencing*, i.e., determining the order in which a list of jobs should be performed by the server, and *appointment scheduling*, i.e., planning the starting times for jobs. These decisions are interconnected because their joint goal is to minimize the expected server idle time and job late-start penalty costs incurred because of randomness in job durations. In this paper, we design new heuristics for sequencing appointments. The idea behind the development of these heuristics is the structural connection between such appointment scheduling problems and stochastic inventory control in serial supply chains. In particular, the decision of determining time allowances as *buffers* against random job durations is analogous to that of selecting inventory levels as *buffers* to accommodate random demand in a supply chain; having excess buffers in appointment scheduling and supply chain settings incurs idle time and excess inventory holding costs, respectively, and having inadequate buffers leads to delays of subsequent jobs and backorders, respectively. Recognizing this connection, we propose tractable approximations for the job sequencing problem, obtain several insights, and further develop a very simple sequencing rule of ordering jobs by duration variance to late-start penalty cost ratio. Computational results show that our proposed heuristics produce close-to-optimal job sequences with significantly reduced computation times compared with those produced using an exact mixed-integer stochastic programming formulation based on the sample-average approximation approach.

Keywords: appointment scheduling; service operations; stochastic inventory control; serial supply chains; stochastic programming
History: Received: September 26, 2011; accepted: September 29, 2013. Published online in *Articles in Advance* January 27, 2014.

1. Introduction

Managing appointments is an important problem for a large variety of service systems with limited capacity. In this paper, we study the following appointment planning problem for a single resource. Given a set of jobs with random durations, the planner determines the planned starting times. This is equivalent to making the decisions of *job sequencing*, i.e., determining the order in which the jobs will be performed, and *appointment scheduling*, i.e., determining the time allowances for jobs with a given sequence. The objective is to minimize the total expected costs of server idle time and late-start delays (i.e., waiting times) of jobs.

The problem studied in this paper is motivated by appointment planning systems that exhibit four distinctive features. First, there exists uncertainty in job durations, which makes it impossible to precisely predict the completion times of jobs. In such cases, scheduling appointments is a stochastic, rather than deterministic, problem. Second, when a job is completed ahead of schedule, it is not possible to start the next job early, possibly because the jobs (or customers) arrive to the system according to their scheduled appointment times and would not be ready even if the previous job is completed early. Therefore, the server will inevitably remain idle if a job is completed in less time than is allotted. Third, considering that customers are time sensitive, delays of job starting times can be costly, because customers have to wait until the server becomes available. Furthermore, a delay of one job can lead to a series of delays of subsequent jobs. Fourth, in the applications that we consider, the set of jobs to be performed is known to the planner when the schedule is determined. That is, the schedule is determined in a static, rather than dynamic, manner.

Appointment systems for some healthcare facilities exhibit the above features. The eye-care center studied by Kong et al. (2013) schedules arrival times of patients with uncertain testing and consultation times that inevitably trigger patient waiting times and doctor idle times. Appointments of elective surgeries in some countries, such as Australia, Belgium, Canada, and the United Kingdom, are based on similar systems. For example, the day-care center of the UZ Leuven Campus Gasthuisberg (Belgium) determines the sequence and schedule of surgeries to be performed one day in advance and informs patients accordingly (Cardoen and Demeulemeester 2011). Similarly, the BC Cancer Agency's Vancouver Centre generates schedules for cancer treatment several days in advance for waitlisted patients (Sauder School of Business 2011). In the public health systems of Australia, Canada, and the United Kingdom, patients are first entered (by surgeons) to waiting lists without being assigned specific dates for elective surgeries. Then, closer to the actual surgery dates, they are notified of the exact time slots (DeCoster et al. 1999). In each of these examples, the planner considers the pool of surgeries to be performed in a time window, and determines the schedule accordingly. We note that another popular practice (common in the United States) is that surgeons decide starting times for surgeries, in consultation with patients and operating room (OR) planners, one at a time. The model and results developed in this paper may not be directly applicable to such cases.

Besides, the aforementioned problem characteristics can be observed in a number of practical applications outside the healthcare domain. For example, cargo terminals need to schedule arrivals of vessels facing uncertain processing (i.e., unloading and loading) times of cargo on vessels (e.g., Sabria and Daganzo 1989). Similarly, for an assembly plant with just-in-time operations, replenishments of multiple parts are often scheduled to be delivered within a short period of time before they are used. The random unloading and inspection times make the schedule difficult to manage (e.g., Liao et al. 1993). Note that, unlike in the container terminal case in which the server is an expensive resource, idle time costs in this example arise from the missed opportunity to reduce holding costs by completing all deliveries earlier. Another example application in manufacturing is the scheduling of loading of parts onto the shop floor (e.g., Wang 1993).

In the appointment systems that we consider, the job sequencing and appointment scheduling decisions are closely connected. One may think of the combinatorial job sequencing problem as an outer problem, whose objective, as a function of sequencing decisions, is given by the *optimal* cost of the inner problem of appointment scheduling. Ideally, the combinatorial job sequencing problem and the appointment scheduling problem should be solved jointly. As will be discussed in the next section, this joint problem is known to be very difficult because of the uncertainty in job durations.

In this paper, we develop a conceptual model to capture the fundamental trade-offs involved in appointment planning using a two-step approximation approach. First, we approximate the expected optimal cost of the inner appointment scheduling problem with a tractable, deterministic function of the sequencing decisions. The outer sequencing problem is then solved with the approximate objective function. Then, once the sequence is determined using the approximate objective function, we solve the appointment scheduling problem with the exact formulation using known methods, such as sample-average approximation (SAA) for stochastic programming.

Our approximation of the inner appointment scheduling problem is based on the observation that its characteristics are similar to those exhibited in the well-studied stochastic inventory control problem for serial systems. In both settings, the planner determines buffer levels (time allowances for jobs or stocking levels) to accommodate stochastic factors (job durations or customer demand). Both problems exhibit overage (server idle time or inventory holding) and underage (late start or shortage) costs for misestimating the buffers needed to accommodate the random factors. Finally, the underage effects (late starts or shortages) can propagate along the system (series of jobs or multiechelon inventory system). Based on this connection, we apply existing results from inventory theory to approximate the objective function of the job sequencing problem. Our computational results show that this approximation yields high-quality solutions.

2. Literature Review

The design of appointment management systems has been studied extensively over the past 50 years. One focus of this stream of literature is on applications in healthcare, such as appointment sequencing and scheduling for ORs and clinics. For comprehensive literature reviews, see Cayirli and Veral (2003), Gupta (2007), Gupta and Denton (2008), Erdogan and Denton (2011), and Cardoen et al. (2010). In this section, we provide only a brief review of models that are most relevant to our research.

In particular, we focus mainly on heuristic algorithms and mathematical-programming-based solution approaches for appointment scheduling problems that minimize some weighted combination of expected late-start and idle time costs.

One line of research assumes that the job sequence is given and the focus is on finding the optimal time allowances for the jobs. Denton and Gupta (2003) model the problem as a two-stage stochastic linear program and develop a variant of the standard L-shaped algorithm (e.g., Chap. 5 of Birge and Louveaux 1997) to obtain optimal solutions. Robinson and Chen (2003) use computational results based on stochastic linear programming and SAA to characterize the form of the optimal policy over a test bed generated by varying the number of patients and the delay cost to idle time cost ratio, which further motivates a simple and easy-to-implement closed-form heuristic. These papers provide valuable insights regarding the optimal scheduling policies using computational examples.

Under the assumption that job durations follow exponential distributions, Kaandorp and Koole (2007) show that the objective function exhibits a certain discrete convexity property, called multimodularity. This result enables the authors to develop a local search algorithm that guarantees convergence to an optimal schedule. Begen and Queyranne (2011) prove the multimodularity result for the problem with general discrete duration distributions. They also show that, for the case with independent discrete probability distributions, the objective function can be computed in polynomial time for a given solution. The results of Begen and Queyranne (2011) are extended to the case with more general cost functions by Ge et al. (2013). Begen et al. (2012) study the SAA approach to the appointment scheduling problem when the duration distributions are unknown, and of which only independent samples are available. Their main result is a theoretical worst-case upper bound on the sample size needed for the solution to achieve desired accuracy and confidence levels.

The papers mentioned above assume either that the job duration distributions are known or that there is a black box that outputs independent samples from the distributions. In a recent paper, Kong et al. (2013) take a distributionally robust optimization approach by assuming that only the mean and covariance matrix of the joint distribution of durations are known. The objective is to minimize the worst-case expected cost, among all possible distributions with the given mean and covariance matrix. The authors show that the problem can be formulated as a copositive program, and propose a semidefinite programming relaxation to obtain near-optimal solutions.

The problem of jointly optimizing the job sequence and the time allowances is significantly more difficult. When there are only two jobs, Weiss (1990) proves that the optimal sequence is to order jobs by increasing duration variance. This result has been extended by Denton et al. (2007) to a model that includes overtime cost. When the unit late-start penalty costs are nonidentical, Gupta (2007) shows that, for two jobs, it is optimal to perform the job with less variable (measured by convex ordering) duration and higher late-start penalty cost first. We extend this insight to the case with a general number of jobs by proving an analogous partial ordering result for our approximate formulation.

To the best of our knowledge, the exact optimal sequencing policy is still unknown when there are three or more jobs. Mancilla and Storer (2012) formulate the joint sequencing and scheduling problem as a stochastic mixed-integer linear program and prove that it is NP-hard when the number of samples of job durations is finite. In the literature, various heuristics have been proposed and studied. Vanden Bosch and Dietz (2000, 2001) investigate a pairwise interchange local-search algorithm and demonstrate its effectiveness for instances with six jobs. Denton et al. (2007) study three simple heuristics of sequencing jobs in increasing order of expected value, variance, and coefficient of variation of duration, respectively. They show computationally that ordering by variance (OV hereafter) consistently outperforms the other two. Focusing on medical clinics, Chen and Robinson (2014) study the sequencing problem for hybrid appointments with routine patients and same-day patients. Assuming that patients of the same type have service times following identical distributions, their numerical results show that good sequences yield up to a 35% reduction in cost.

By utilizing the connection between a class of stochastic appointment scheduling problems and the classical serial supply chain inventory problem, we develop a mixed-integer second-order cone programming (MISOCP) approximation for the job sequencing problem. This conceptual model offers insights that enable us to further develop simple sequencing rules that generalize the OV heuristic under nonidentical late-start penalty costs. In particular, the easy-to-implement rule of sequencing jobs in increasing order of job duration variance to late-start penalty ratio provides close-to-optimal job sequences in minimal computation times.

3. Job Sequencing Problem

Consider the appointment sequencing and scheduling problem of deciding the sequence to perform N jobs and the time allowances for individual jobs. Let the jobs be indexed by $m=1,\ldots,N$. The positions of jobs in a sequence are indexed by $j=1,\ldots,N$, where $j=1$ indicates the first job to be performed and $j=N$ indicates the last. Therefore, determining the sequence is equivalent to assigning jobs $m=1,\ldots,N$ to positions $j=1,\ldots,N$. We first define the following notation:

Problem Parameters

- **d**: the vector $(d_1,\ldots,d_N)$, where $d_m(\geq 0)$ is the duration of job m, which is assumed to be independent, but not necessarily identically distributed for different m; we denote the mean and standard deviation of d_m by μ_m and σ_m, respectively;
- κ: the cost per unit of idle time due to finishing a job ahead of schedule;
- p_m: the late-start penalty cost for job m per unit time of delay.

Decision Variables

- **s**: the vector $(s_1,s_2,\ldots,s_{N-1})$, where s_j denotes the time allowance for the job in the jth position (note that the scheduled starting time for the job in the jth position is given by $\sum_{i=1}^{j-1} s_i$);
- **x**: the matrix $\{x_{jm}\}_{j=1,\ldots,N;\,m=1,\ldots,N}$, where x_{jm} is a binary indicator variable where $x_{jm}=1$ if job m is assigned the jth position.

Performance Indicators

For a given time allowance vector **s** and a sequence indicated by **x**, we define the following:

- $B_j(\mathbf{s},\mathbf{x})$: the delay of the completion of the job in the jth position;
- $I_j(\mathbf{s},\mathbf{x})$: the idle time due to the job in the jth position being completed ahead of schedule.

Additional Notation Defined for Brevity

$$\mathbf{X}=\left\{\mathbf{x}\,\middle|\,\sum_{j=1}^{N}x_{jm}=1,\,m=1,\ldots,N;\,\sum_{m=1}^{N}x_{jm}=1,\,j=1,\ldots,N;\right.$$
$$\left.x_{jm}\in\{0,1\},\,j=1,\ldots,N,\,m=1,\ldots,N\right\},$$

$$d_j(\mathbf{x})=\sum_{m=1}^{N}d_m x_{jm},\qquad p_j(\mathbf{x})=\sum_{m=1}^{N}p_m x_{jm}.$$

In other words, **X** represents the set of all feasible sequences of performing jobs. This can be obtained by requiring that all assignment variables x_{jm} sum up to 1 for each j (each position is occupied by some job) and for each m (each job is assigned to some position). The notations $d_j(\mathbf{x})$ and $p_j(\mathbf{x})$ denote the duration and the per unit late-start penalty of the job assigned to the jth position according to sequence **x**, respectively. Given any job sequence **x** and the realization of job durations **d**, the following transition equations hold:

$$I_j(\mathbf{s},\mathbf{x})=[s_j-d_j(\mathbf{x})-B_{j-1}(\mathbf{s},\mathbf{x})]^+,\quad 1\leq j\leq N-1,\quad (1)$$
$$B_j(\mathbf{s},\mathbf{x})=[s_j-d_j(\mathbf{x})-B_{j-1}(\mathbf{s},\mathbf{x})]^-,\quad 1\leq j\leq N-1,\quad (2)$$

where $B_0(\mathbf{s},\mathbf{x})=0$, $[\cdot]^+=\max\{\cdot,0\}$ and $[\cdot]^-=-\min\{\cdot,0\}$. The above holds because the job in the jth position is completed early if its time allowance s_j exceeds its duration $d_j(\mathbf{x})$ plus any delay of the previous job $B_{j-1}(\mathbf{s},\mathbf{x})$ (i.e., delay of the start of the jth job), and is completed late otherwise.

In some applications, such as the case of scheduling surgeries for ORs, reducing idle time is important when it results in a reduction in overtime, the scheduling of more jobs, or a reduction in the active number of ORs to be staffed on a given day. That is, small reductions of idle time may not directly improve costs or revenues. It may be possible that surgeons even prefer to have idleness such that a procedure is less likely to be delayed. Therefore, in developing a practical decision-support tool for OR scheduling, the exact form of idle time cost (e.g., strictly convex, linear, or piecewise linear) needs to be carefully selected by taking into account such considerations. Nevertheless, for our conceptual model, we opt for a linear form as an approximation and do not include such application-specific details in order to maintain tractability and draw qualitative insights. One may note that there exist convenient parameter estimation methods for the linear idle time cost (see Olivares et al. 2008 for the case of OR scheduling). Moreover, as is common in the literature (Robinson and Chen 2003, Kong et al. 2013), we assume that the idle time cost per unit time (κ) is uniform for all jobs and positions.

Similarly to Weiss (1990) and Robinson and Chen (2003), we do not consider overtime cost in the model. In Online Appendix F (online appendices available as supplemental material at http://dx.doi.org/10.1287/msom.2013.0470), we discuss how to extend our model to handle session length constraint and overtime costs. For any given $(\mathbf{s},\mathbf{x})$, the total expected idle and late-start cost is given by

$$\Pi(\mathbf{s},\mathbf{x})=\sum_{j=1}^{N-1}\kappa E[I_j(\mathbf{s},\mathbf{x})]+\sum_{j=1}^{N-1}p_{j+1}(\mathbf{x})E[B_j(\mathbf{s},\mathbf{x})].\quad (3)$$

The appointment scheduling problem is to minimize the above by varying the time allowances

$$C^*(\mathbf{x})=\min_{\mathbf{s}\geq\mathbf{0}}\Pi(\mathbf{s},\mathbf{x}).\quad (4)$$

As shown by Begen and Queyranne (2011, Lemma 4.5), the constraint $\mathbf{s}\geq\mathbf{0}$ can be removed in (4) without affecting the optimal solution. Thus, the job sequencing problem can be formulated as

$$\min_{\mathbf{x}\in\mathbf{X}}C^*(\mathbf{x})=\min_{\mathbf{x}\in\mathbf{X}}\min_{\mathbf{s}}\Pi(\mathbf{s},\mathbf{x}).\quad (5)$$

Problem (5) can be formulated as a stochastic mixed-integer linear program using the SAA approach; see

Online Appendix C for details. Mancilla and Storer (2012) prove that this approach gives rise to an NP-hard problem. Their computational tests show that it can possibly take days to solve instances with 10 jobs and 100 scenarios using the CPLEX commercial solver. Such computational difficulty inspires us to develop an alternative solution procedure, one based on inventory approximations. The key steps of our procedure are outlined below:

Step 1. Establish a connection between the appointment scheduling problem and a collection of serial supply chain inventory control problems (§3.1).

Step 2. Apply results from the inventory literature to obtain an MISOCP approximation for the job sequencing problem (§3.2). In particular, we apply a result due to Shang and Song (2003) to approximate the costs of serial supply chains by newsvendor cost functions. Then we utilize Scarf's (1958) result to approximate newsvendor cost functions using tractable expressions.

Step 3. Investigate the structural properties of the MISOCP formulation, which provide effective, easy-to-implement sequencing heuristics that generalize the OV rule (§3.3).

Step 4. Finally, with the job sequence determined either by solving the MISOCP formulation or the sequencing heuristics, we solve the appointment scheduling problem by the SAA approach, which involves solving a scenario-based stochastic linear program (a subproblem of the integer linear programming formulation in Online Appendix C).

We begin the discussion by presenting the connection with inventory problems (Step 1).

3.1. The Connection with Serial Supply Chain Inventory Control Problems

To begin, we note that, for any $\mathbf{h} = \{h_{ij}\}_{i=1,\ldots,N;\, j=1,\ldots,N}$ satisfying

$$\sum_{i=j}^{N-1} h_{ij} = \kappa, \quad j = 1, \ldots, N-1, \tag{6}$$

problem (4) can be reformulated as

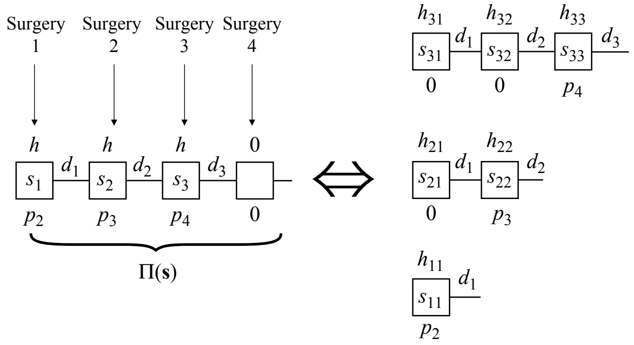
$$C^*(\mathbf{x}) = \min_{\mathbf{S}} \left\{ \sum_{i=1}^{N-1} \left(\sum_{j=1}^{i} h_{ij} E[I_{ij}(\mathbf{S}_i, \mathbf{x})] + p_{i+1}(\mathbf{x}) E[B_{ii}(\mathbf{S}_i, \mathbf{x})] \right) \right\} \tag{7}$$

$$\text{s.t.} \quad S_{ij} = S_{i+1,\,j}, \quad 1 \le j \le i \le N-2, \tag{8}$$

where $\mathbf{S}$ denotes the matrix $\{S_{ij}\}_{i=1,\ldots,N;\, j=1,\ldots,N}$ with $\mathbf{S}_i$ being its ith row. By (1) and (2), we have $B_{i0}(\mathbf{S}_i, \mathbf{x}) = 0$ for $i = 1, \ldots, N-1$ and

$$I_{ij}(\mathbf{S}_i, \mathbf{x}) = [S_{ij} - d_j(\mathbf{x}) - B_{i,\,j-1}(\mathbf{S}_i, \mathbf{x})]^+, \quad 1 \le j \le i \le N-1, \tag{9}$$

Figure 1 **Illustration of Job Block Decomposition with Variable Splitting Formulation**



$$B_{ij}(\mathbf{S}_i, \mathbf{x}) = [S_{ij} - d_j(\mathbf{x}) - B_{i,\,j-1}(\mathbf{S}_i, \mathbf{x})]^-, \quad 1 \le j \le i \le N-1. \tag{10}$$

The connection between formulations (4) and (7) is illustrated in Figure 1. With constraint (8), these two formulations are equivalent because the idle time and penalty costs associated with each job are the same in both formulations. Thus, the reformulation (7) may be potentially decomposed by i if constraint (8) is relaxed, and each of them can be interpreted as an inventory problem. The following proposition provides a lower bound on $C^*(\mathbf{x})$ by relaxing (8).

Proposition 1. *Define* $\underline{C}^*(\mathbf{x})$ *as follows*:

$$\underline{C}^*(\mathbf{x}) = \max_{\mathbf{h}\in\mathbf{H}} \min_{\mathbf{S}} \sum_{i=1}^{N-1} \tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x}) = \max_{\mathbf{h}\in\mathbf{H}} \sum_{i=1}^{N-1} \min_{\mathbf{S}_i} \tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x}), \tag{11}$$

where $\mathbf{H} = \{\mathbf{h} \ge 0 \mid \sum_{i=j}^{N-1} h_{ij} = \kappa,\ j = 1, \ldots, N-1;\ h_{i,\,j-1} \le h_{ij},\ 2 \le j \le i \le N-1\}$, *and*

$$\tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x}) = \sum_{j=1}^{i} h_{ij} E[I_{ij}(\mathbf{S}_i, \mathbf{x})] + p_{i+1}(\mathbf{x}) E[B_{ii}(\mathbf{S}_i, \mathbf{x})]. \tag{12}$$

Then, we have $C^*(\mathbf{x}) \ge \underline{C}^*(\mathbf{x})$.

The proofs of all analytical results are provided in Online Appendix A.

Remark 1. Formulation (11) is equivalent to the Lagrangian dual of (7) obtained by relaxing a constraint equivalent to (8). This alternative derivation is discussed in Online Appendix E.

Proposition 1 provides a formulation in which the jobs are decomposed into $N-1$ subproblems indexed by $i = 1, \ldots, N-1$, which we hereafter refer to as "blocks" of jobs. In particular, given $\mathbf{x}$ and $\mathbf{h}$, we may solve the decomposed subproblem for each i:

$$\min_{\mathbf{S}_i} \tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x}) = \min_{\mathbf{S}_i} \sum_{j=1}^{i} h_{ij} E[I_{ij}(\mathbf{S}_i, \mathbf{x})] + p_{i+1}(\mathbf{x}) E[B_{ii}(\mathbf{S}_i, \mathbf{x})]. \tag{13}$$

Formulation (13) is equivalent to the problem of optimizing local base-stock levels for a serial supply chain; see Zipkin (2000, equation 8.3.5). This problem has been extensively studied; for example, see Clark and Scarf (1960), Chen and Zheng (1994), Chen and Song (2001), Huh and Janakiraman (2008), Huh et al. (2010), Glasserman and Tayur (1994), and Gallego et al. (2005). More details regarding the connection between this decomposed job scheduling subproblem and the serial supply chain problem are provided in Online Appendix B. Based on this connection, we refer to subproblem (13) as both "job block i" and "supply chain i" interchangeably in the subsequent discussion. Maximization over $\mathbf{h}$ is equivalent to allocating the unit idle time cost κ among the i job blocks. Note that, in addition to (6), the set $\mathbf{H}$ imposes the property that holding costs in serial supply chains are nonnegative and increasing toward downstream stages. The monotonicity property naturally holds in practical supply chains, because the holding cost difference between two consecutive stages reflects the value added at the downstream stage, which is always nonnegative (see Zipkin 2000, p. 122, for a discussion). In the literature, this property is typically assumed in the development of solution procedures, and is required when we apply the results of Shang and Song (2003) to approximate problem (13) (in Step 2).

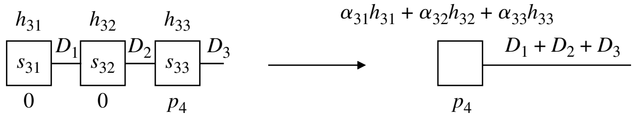
3.2. Inventory Approximations

In their seminal paper, Shang and Song (2003) propose a simple procedure (SS hereafter) to accurately approximate the optimal expected cost for the classical serial supply chain problem. The fundamental idea underlying their method is to aggregate the cost of a serial supply chain into a single newsvendor cost function whose overage cost equals a weighted sum of all local holding costs and whose demand equals the total lead time demand of all stages of the serial supply chain.

To aggregate the stages of supply chain i into a single newsvendor problem, let α_{ij} be the weight of the local holding cost h_{ij} for stage j. We require $0 \le \alpha_{ij} \le 1$ and $\sum_{j=1}^{i} \alpha_{ij} = 1$. A three-stage supply chain example is shown in Figure 2. As discussed by SS, selection of a good set of α_{ij} values can lead to a good approximation for $\tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x})$. We will discuss our choice of α_{ij} values in §4. Given α_{ij} values for $1 \le j \le i$, the SS approach suggests approximating $\tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x})$ by a newsvendor cost function $\tilde{c}_i^{SS}(\mathbf{S}_i, \mathbf{h}, \mathbf{x})$, defined as follows:

$$\begin{aligned}\tilde{c}_i^{SS}(\mathbf{S}_i,\mathbf{h},\mathbf{x}) = {} & \left(\sum_{j=1}^{i}\alpha_{ij}h_{ij}\right)E\left[\left(\sum_{j=1}^{i}(S_{ij}-d_j(\mathbf{x}))\right)^{+}\right] \\ & + p_{j+1}(\mathbf{x})E\left[\sum_{j=1}^{i}(d_j(\mathbf{x})-S_{ij}) \right. \\ & \left. + \left(\sum_{j=1}^{i}(S_{ij}-d_j(\mathbf{x}))\right)^{+}\right]. \qquad (14)\end{aligned}$$

Figure 2 Illustration of SS Heuristic



Shang and Song (2003) develop their approximation based on a serial supply chain with an *echelon* base-stock policy, whereas problem (13) is equivalent to one under a *local* base-stock policy. However, SS is also applicable to (13), because the echelon base-stock policy studied by Shang and Song can be transformed to an equivalent local base-stock policy (Zipkin 2000, p. 306). In particular, if the echelon base-stock levels are nonincreasing toward downstream, then the corresponding local base-stock level of a stage is simply equal to the difference between the echelon base-stock levels at the same stage and its immediate downstream stage. Then, the two systems yield the same costs. Because our purpose of applying the SS procedure is to approximate the cost function instead of the base-stock levels, we expect that the proven accuracy of the SS approximation in serial supply chains with echelon base-stock policies will carry over to our problem as well.

Note that, by setting $\alpha_{i1} = 1$ and $\alpha_{ij} = 0$ for $j = 2, \ldots, i$, Shang and Song (2003) prove that (14) provides a lower bound on $\tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x})$. This result holds because the resulting newsvendor problem is equivalent to a serial supply chain problem with local holding costs in all stages equal to h_{i1}, which is less than or equal to h_{ij} for all $j \ge 1$. When the local holding costs of all stages are equal in a serial supply chain, it is optimal to hold inventory at only the most downstream stage, so that the serial supply chain effectively becomes a single newsvendor. Similarly, when $\alpha_{ii} = 1$ and $\alpha_{ij} = 0$ for $j = 1, \ldots, i-1$, Shang and Song (2003) prove that (14) provides an upper bound on $\tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x})$. This result holds because the newsvendor function is equivalent to a serial supply chain with local holding costs in all stages equal to h_{ii}, which is greater than or equal to h_{ij} for all $j \le i$.

The difficulty of evaluating this newsvendor cost function depends on the demand (job duration) distributions. Because our goal is to propose practical heuristics that operate under a wide variety of settings, it is desirable to obtain a closed-form cost expression that does not require the exact distributional form. This approximation could be useful when planners do not have sufficient data to accurately estimate the exact distributional forms when faced with a large variety of job types (e.g., Macario 2010). Thus, we apply Scarf's (1958) results for approximating the expected cost of the newsvendor problem, using the information of only the mean and variance of demand. In particular, Scarf (1958) showed that, for a newsvendor problem with random demand D and stocking level q,

$$E[D-q]^+ \le \tfrac{1}{2}(\mu - q) + \tfrac{1}{2}\sqrt{(\mu-q)^2+\sigma^2}, \qquad (15)$$

where μ and σ are the mean and standard deviation of demand D, respectively. We note that there are several approaches to approximate the newsvendor function. Scarf's bound, which focuses on the worst-case expected cost, is merely one of them. We choose to use Scarf's bound because Scarf (1958) showed that the order quantity that optimizes the worst-case objective typically achieves close-to-optimal expected cost even when complete distributional information is given. Moreover, Perakis and Roels (2008) and Zhu et al. (2013) show that Scarf's solution is also promising under other objectives, such as absolute regret and relative regret. Furthermore, with structure that retains the basic newsvendor trade-off between underage and overage costs, Scarf's bound has been utilized in several works (e.g., Moon and Gallego 1994, See and Sim 2010) to obtain tractable approximations for inventory problems with limited distributional information.

Using Scarf's bound (15) to evaluate the newsvendor cost terms in (14), we may approximate $\tilde{c}_i^{SS}(\mathbf{S}_i, \mathbf{h}, \mathbf{x})$ by

$$\begin{aligned}\tilde{c}_i^{SB}(\mathbf{S}_i,\mathbf{h},\mathbf{x}) = \left(\sum_{j=1}^{i}\alpha_{ij}h_{ij}\right)&\bar{I}_i^{SB}(\mathbf{S_i},\mathbf{x})\\ &+p_{j+1}(\mathbf{x})\bar{B}_i^{SB}(\mathbf{S_i},\mathbf{x}), \qquad (16)\end{aligned}$$

where

$$\begin{aligned}&\bar{I}_i^{SB}(\mathbf{S_i},\mathbf{x})\\ &\quad=\tfrac{1}{2}\sum_{j=1}^{i}\left(S_{ij}-\sum_{m=1}^{N}\mu_m x_{jm}\right)\\ &\qquad+\tfrac{1}{2}\sqrt{\left(\sum_{j=1}^{i}S_{ij}-\sum_{j=1}^{i}\sum_{m=1}^{N}\mu_m x_{jm}\right)^2+\sum_{j=1}^{i}\sum_{m=1}^{N}\sigma_m^2 x_{jm}}, \qquad (17)\\ &\bar{B}_i^{SB}(\mathbf{S_i},\mathbf{x})\\ &\quad=\tfrac{1}{2}\sum_{j=1}^{i}\left(\sum_{m=1}^{N}\mu_m x_{jm}-S_{ij}\right)\\ &\qquad+\tfrac{1}{2}\sqrt{\left(\sum_{j=1}^{i}S_{ij}-\sum_{j=1}^{i}\sum_{m=1}^{N}\mu_m x_{jm}\right)^2+\sum_{j=1}^{i}\sum_{m=1}^{N}\sigma_m^2 x_{jm}}. \qquad (18)\end{aligned}$$

Then, we may replace $\tilde{c}_i(\mathbf{S}_i, \mathbf{h}, \mathbf{x})$ in formulation (11) with $\tilde{c}_i^{SB}(\mathbf{S}_i, \mathbf{h}, \mathbf{x})$, thus yielding the following approximate formulation for $\underline{C}_i^*(\mathbf{x})$, with a slight abuse of notation:

$$\max_{\mathbf{h}\in\mathbf{H}}\min_{\mathbf{S}}\sum_{i=1}^{N-1}\tilde{c}_i^{SB}(\mathbf{S}_i,\mathbf{h},\mathbf{x}). \qquad (19)$$

Then the optimal objective value of problem (19), as a function of $\mathbf{x}$, provides an approximate objective function for the job sequencing problem (5). In particular, we solve

$$\min_{\mathbf{x}\in\mathbf{X}}\max_{\mathbf{h}\in\mathbf{H}}\min_{\mathbf{S}}\sum_{i=1}^{N-1}\tilde{c}_i^{SB}(\mathbf{S}_i,\mathbf{h},\mathbf{x}). \qquad (20)$$

We next show that (20) can be formulated as an MISOCP.

Proposition 2. *Problem* (20) *is equivalent to the following MISOCP in the sense that the optimal objective values and the optimal values of* $\mathbf{x}$ *are the same in both problems*:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{a},\mathbf{c},\mathbf{r}}\quad&\left[\kappa\sum_{i=1}^{N-1}y_i+\sum_{i=2}^{N}\sum_{m=1}^{N}p_m r_{im}\right] && (21)\\ \text{s.t.}\quad& y_1\ge a_1,\\ & y_1+z_{i1}\ge\alpha_{i1}a_i,\quad i=2,\dots,N-1,\\ & y_i-z_{i,i-1}\ge\alpha_{ii}a_i,\quad i=2,\dots,N-1,\\ & y_j-z_{i,j-1}+z_{i,j}\ge\alpha_{ij}a_i,\\ &\qquad\qquad j=1,\dots,i-1,\ i=2,\dots,N-1,\\ & r_{im}\ge U_{im}x_{im}+a_{i-1}-c_{i-1}-U_{im},\\ &\qquad\qquad i=2,\dots,N,\ m=1,\dots,N, && (22)\\ & a_i\ge\tfrac{1}{2}c_i+\tfrac{1}{2}\sqrt{c_i^2+\sum_{j=1}^{i}\sum_{m=1}^{N}\sigma_m^2 x_{jm}^2},\\ &\qquad\qquad i=1,\dots,N-1,\\ & z_{ij}\ge 0,\quad j=1,\dots,i-1,\ i=1,\dots,N-1,\\ & r_{im}\ge 0,\quad i=2,\dots,N,\ m=1,\dots,N,\\ & \mathbf{x}\in\mathbf{X}.\end{aligned}$$

We provide interpretations of the new decision variables introduced in the above formulation. Variables a_i and c_i are the expected surplus inventory and the expected net inventory level of the ith newsvendor, which approximates the ith serial supply chain, respectively. Variable r_{im} is the expected shortage of the $(i-1)$th newsvendor if job m is assigned the ith position. The variables $\mathbf{y}$ and $\mathbf{z}$ arise from taking the dual of the maximization problem over holding cost allocations. Variable y_j is the shadow price of the constraint $\sum_i h_{ij} = \kappa$, i.e., the allocation of holding costs among supply chains. Variable z_{ij} is the dual multiplier associated with the constraints $h_{ij} \leq h_{i,\,j+1}$, i.e., a penalty for violating the monotonicity requirement in determining the holding cost allocation among serial supply chains. Finally, parameters U_{im} are linearization constants given by upper bounds on the expected backorders at the $(i-1)$st (most downstream) node of the $(i-1)$st serial supply chain, provided that job m is not assigned to position i, i.e., $x_{im} = 0$. Selecting tight values of U_{im} (e.g., using the results presented in Online Appendix D) can help speed up computation. It is also notable that the formulation does not include the time allowance variables $\mathbf{s}$. Recall that our solution approach involves solving for the job sequence by considering an approximate cost of the job scheduling problem. By applying the inventory results to approximate the cost function, we are able to obtain such a formulation without involving the $\mathbf{s}$ variables. After determining the sequence, the scheduling problem can be solved using SAA.

Remark 2. Our approach is based on approximating the objective function

$$\sum_{i}^{N-1}\left(\kappa E[I_i(\mathbf{s},\mathbf{x})] + p_{i+1}(\mathbf{x})E[B_i(\mathbf{s},\mathbf{x})]\right)$$

by $\sum_{i=1}^{N-1} \tilde{c}_i^{SB}(\mathbf{S_i}, \mathbf{h}, \mathbf{x})$. One may also consider a simpler and more straightforward approach. For example, one can show that $\sum_i^{N-1} E[I_i(\mathbf{s},\mathbf{x})]$, the total expected idle time, is equal to

$$E\left[\max\left(0, s_1 - d_1(\mathbf{x}), \ldots, \sum_{j=1}^{N-1}(s_j - d_j(\mathbf{x}))\right)\right];$$

and $E[B_i(\mathbf{s},\mathbf{x})]$, the expected waiting time of job i, is equal to $E[\max(0, d_1(\mathbf{x}) - s_1, \ldots, \sum_{j=1}^{i}(d_j(\mathbf{x}) - s_j))]$. Then, one can approximate the total idle time cost by $\kappa E[(\sum_{j=1}^{N-1}(s_j - d_j(\mathbf{x})))^+]$ and the total late-start penalty by $\sum_i^{N-1} p_{i+1} E[(\sum_{j=1}^{i}(s_j - d_j(\mathbf{x})))^-]$. By applying Scarf's approximation, one can obtain a different MISOCP formulation than (21). However, our formulation is a more refined approximation for the following reason. In each of the two steps of our approximation, i.e., approximating a serial supply chain with a single newsvendor cost function (using the SS result) and approximating the newsvendor cost function with Scarf 's model, both overage (idle time) and underage (late-start time) costs are approximated jointly, preserving the basic trade-off of the problem. However, in the alternative approximation above, the idle time and late-start time terms are disconnected. Using separate approximations of the two terms, the resulting formulation may introduce bias in the underlying trade-off of the problem.

3.3. Structural Properties and Sequencing Heuristics

As discussed in a number of studies in the literature (e.g., Denton et al. 2007, Gupta 2007, Mancilla and Storer 2012), obtaining the exact optimal job sequence is very difficult. When the per unit late-start penalty costs are identical for all jobs, Denton et al. (2007) suggest the use of the OV heuristic, which is proven to be optimal for two jobs (Weiss 1990). For the case in which there are two jobs with nonidentical per unit late-start penalty costs, Gupta (2007) shows that it is optimal to perform the job with less variable (measured by convex ordering) duration and higher late-start penalty cost first. In this section, we prove that the optimal sequence of jobs obtained from our approximate formulation exhibits an analogous partial ordering, for general number of jobs, and generalize the insight provided by Gupta (2007). Based on this result, we further propose effective heuristics that complement and generalize the widely used OV heuristic.

Proposition 3. *If $\sigma_{m_1} < \sigma_{m_2}$ and $p_{m_1} \geq p_{m_2}$ for jobs m_1 and m_2, then in the optimal solution to the MISOCP* (21), *job m_1 is sequenced earlier than job m_2.*

Proposition 3 proves that, in the optimal solution to the MISOCP formulation, a job with a smaller duration variance and higher penalty cost will be performed before one with a larger duration variance and lower penalty cost. Although this result only defines a partial ordering relationship, i.e., one in which some pairs of jobs may not be ranked, we may extend the qualitative insight and further develop complete ordering schemes. For example, we define the following measure:

$$w_m(\gamma) = \frac{\sigma_m}{p_m^{\gamma}}. \tag{23}$$

Thus, an intuitive heuristic is to sequence job m_1 ahead of job m_2 whenever $w_{m_1}(\gamma) \leq w_{m_2}(\gamma)$, i.e., perform jobs in increasing order of $w_m(\gamma)$. By varying the parameter γ, we may obtain different sequencing rules. We will focus on the two most intuitive choices: $\gamma = 1$, i.e., ordering by the standard deviation-to-penalty cost ratio (OSP hereafter), and $\gamma = 0.5$, i.e., ordering by the variance-to-penalty cost ratio (OVP hereafter). Heuristics of this class are easy to implement. To sequence N jobs, we only need to compute $w_m(\gamma)$ for $m = 1, \ldots, N$ and sort the list in increasing order. It is also notable that the class of new heuristics is a generalization of the common OV heuristic, which is the case in which $\gamma = 0$. When the per unit late-start penalty costs are identical, the OV heuristic has been shown by Denton et al. (2007) to be effective. Note that, by incorporating heterogeneity in late-start penalty costs, these heuristics are analogous to priority rules that help improve system performance in service queueing systems in which customers exhibit varying willingness to wait (e.g., Van Mieghem 2000). As we will demonstrate in our computational experiments, the two related heuristics of OSP and OVP are particularly applicable to instances with nonidentical per unit late-start penalty costs.

4. Computational Experiments

In this section, we perform computational experiments to study the performances of our heuristics. First, we study the impact of choosing different α_{ij} values in (14) on the performances of the resulting MISOCP formulations. Second, we evaluate the effectiveness of our approximate formulation (21) for the job sequencing problem (referred to as "MISOCP" in this section). To compare the different heuristics, we use two benchmark solution approaches: (i) solving the SAA formulation (provided in Online Appendix C) for the job sequencing problem with 2,000 samples ("SAA") and (ii) generating random sequences ("Rand") in which all jobs are equally likely to be assigned to any position. The purpose of considering the Rand benchmark is to study the consequences of not attempting to optimize the sequence. Finally, we also study the performances of the OVP, OSP, and OV heuristics discussed in §3.3. We attempt to demonstrate the importance of incorporating the information of late-start penalty costs when they are not identical. In all computational experiments, we first solve for sequences using the aforementioned heuristics and then compute the expected cost by solving the appointment scheduling problem using SAA with 2,000 samples.

We generate test instances with 8, 10, and 12 jobs, with job durations following uniform, normal, and lognormal distributions. For each combination of number of jobs and probability distribution, we test 10 instances. For uniform job durations, we draw the means from a uniform distribution between 0 and 2, denoted by $U[0,2]$, and we fix the lower support point to be 0. For normal job durations, we fix the means to be 1 and draw the standard deviations from $U[0,1/3]$. For lognormal job durations, we set the underlying normal distributions' means to 1 and draw their standard deviations from $U[0,1]$. For every instance, we generate the unit idle time cost (κ) from $U[0,5]$ and the per unit late-start penalty costs (p_i) for individual jobs from $U[0,10]$. These generic parameter settings are chosen for illustrating our key results and are not calibrated to data from specific scheduling applications. All linear programs, integer programs, and MISOCPs are solved using CPLEX 12.1 running on Windows 7 on a Dell Precision T7500 workstation with an Intel X5680 CPU (using six cores maximum) and 48 GB of memory. Finally, because larger instances of the job sequencing problem may potentially take days to solve to optimality (Mancilla and Storer 2012), we set the running time limit for CPLEX to be 18,000 seconds because of the need to solve large number of instances in our experiments.

4.1. Selection of Holding Cost Weights in the SS Procedure

As discussed in the previous section, varying the values of of α_{ij} in the SS procedure gives rise to different approximations, and thus different MISOCP formulations following Proposition 2. Ideally, one may attempt to optimize the α_{ij} values together with other decision variables of interest to obtain the best approximation. However, we note that such an approach gives rise to an intractable problem. Therefore, we focus on the simple choice of $\alpha_{ij} = 1/i$ for $j = 1, \ldots, i$, and perform the following experiment to test the performance of the resulting approximation.

In this experiment, we use 30 test instances (10 each for uniform, normal, and lognormal job duration distributions as discussed above) with eight jobs. In addition to our choice of setting $\alpha_{ij} = 1/i$, we randomly generate 10 other sets of different α_{ij} values for comparison. We first generate $\hat{\alpha}_{ij}$ values from a $U[0,1]$ distribution and then normalize them by setting $\alpha_{ij} = \hat{\alpha}_{ij} / \sum_{k=1}^{i} \hat{\alpha}_{ik}$. The reason to normalize the weights is that the SS heuristic requires $\sum_{j=1}^{i} \alpha_{ij} = 1$. Our results show that the differences in the performances (i.e., cost of the resulting sequence and schedule) of the resulting approximations arising from different α_{ij} values are less than 1% on average. Compared with each of the other 10 approximations over the 30 instances, the approximation based on setting $\alpha_{ij} = 1/i$ achieved better results in 268 out of 300 cases. In all instances tested, no other approximation could outperform this approximation by more than 4%. These results suggest that the choice of α_{ij} values has a relatively small impact on the quality of the resulting approximation of the job sequencing problem, and the set of values we choose typically outperform other chosen values. Based on our results, we set $\alpha_{ij} = 1/i$ in the remainder of the paper.

4.2. Evaluation of Performance of Heuristics

The next set of computation results, summarized in Table 1, demonstrates the performance of our MISOCP formulation by comparing with the SAA benchmark. For the SAA and MISOCP approaches, we report the average and maximum computation times among instances for each combination of number of jobs and duration distribution type. Because of the running time limit, the SAA formulation cannot

Table 1 Effectiveness of the MISOCP Heuristic

		SAA		MISOCP				Rand	
Distribution	No. of jobs	Avg. time	Max. time	Avg. time	Max. time	Avg. gap (%)	Max. gap (%)	Avg. gap (%)	Max. gap (%)
Uniform	8	16,905	18,000	0.25	0.50	1.91	5.01	32.43	32.43
	10	18,000	18,000	12.3	48.5	0.81	10.19	20.50	20.50
	12	18,000	18,000	3,492	18,000	−0.99	2.64	16.76	16.76
Normal	8	15,592	18,000	0.49	1.98	2.49	6.15	31.36	31.36
	10	18,000	18,000	11.7	33.9	0.60	4.68	24.65	24.65
	12	18,000	18,000	2,656	18,000	−2.65	2.40	21.40	21.40
Lognormal	8	12,415	18,000	0.48	1.72	1.87	4.70	65.33	65.33
	10	18,000	18,000	11.0	71.4	−0.48	5.26	62.03	62.03
	12	18,000	18,000	2,095	18,000	−5.36	2.15	36.82	36.82
Average		16,990	18,000	920	6,018	−0.20	4.80	34.58	34.58

be solved to optimality. Thus, it may not necessarily provide the best solution. In fact, there are 36 out of 90 instances in which SAA strictly performs better than all five other heuristics (i.e., MISOCP, Rand, OV, OSP, and OVP). Therefore, in the remaining cases, the gaps (measured by the percentage cost increase of MISOCP over SAA) can be negative because of the SAA method reaching the solution time limit before reaching the optimal solution. If one allows SAA to solve to optimality (which may take several days for each instance), the gap will eventually become nonnegative. In Table 1, we find that the MISOCP approach consistently generates job sequencing solutions that are comparable to those provided by SAA on average, with a maximum difference of approximately 10%. On the other hand, the running times for the MISOCP formulation, given the enhancements developed in Online Appendix D, are significantly shorter than those for SAA.

From this set of computational experiments, we find that job sequencing decisions carry significant implications regarding cost, even if the subsequent appointment scheduling problem can be solved to optimality. This significance can be illustrated by the large percentage gaps between the costs given by the randomized sequence (Rand) and the costs given by sequences generated by SAA and MISOCP approaches. Given sequences obtained by the corresponding approaches, these costs are evaluated by optimally solving the appointment scheduling problem (with the SAA approach). Therefore, these significant differences provide evidence that it is not possible to recover the suboptimality of an inferior sequence by optimally solving the appointment scheduling problem. In addition, if one attempts to recover the loss of optimality from selecting a bad sequence by varying the time allowances, we observe that the resulting total time allowances for all jobs can be much longer than those under the sequences generated by SAA and MISOCP. In our test bed, we found that the average (maximum, respectively) percentage increase in the total time allowances for a random sequence compared with the corresponding MISOCP sequence is 18.3% (63.48%, respectively). This result suggests that an inferior sequence often leads to the allocation of extra buffer times for jobs to hedge against the risk of delays propagating down the schedule, which is undesirable in practice even if there is no explicit overtime cost.

Finally, we discuss the effectiveness of the two simple heuristics of OVP and OSP introduced in §3.3, both of which are motivated by the partial ordering property of the MISOCP formulation. From the results shown in Table 2, we may draw several observations.

Table 2 Comparisons of the OV, OSP, and OVP Heuristics

		OV		OSP		OVP	
Distribution	No. of jobs	Avg. gap (%)	Max. gap (%)	Avg. gap (%)	Max. gap (%)	Avg. gap (%)	Max. gap (%)
Uniform	8	8.58	18.96	4.71	8.73	2.82	8.02
	10	4.39	14.85	2.46	11.34	1.38	9.97
	12	3.78	9.61	−0.08	4.35	−0.92	2.32
Normal	8	7.20	14.45	3.38	11.76	2.64	7.60
	10	5.61	18.14	2.78	9.84	0.85	4.05
	12	2.13	10.18	−1.56	2.65	−2.82	1.77
Lognormal	8	8.17	25.11	2.24	7.01	1.12	2.63
	10	6.13	21.32	0.97	3.96	−1.11	0.95
	12	−0.90	13.87	−4.98	1.94	−5.89	1.29
Average		5.01	16.28	1.10	6.84	−0.21	4.29

First, modifying the simple OV ranking by including the penalty cost information, the OVP and OSP solutions provide significant cost savings. The gaps are, on average, reduced by more than half. This result demonstrates that our inventory-approximation-based qualitative insight drawn from Proposition 3 helps identify high-quality solutions with minimal effort. Second, Table 2 shows that OVP performs better than OSP. Combining Tables 1 and 2, we observe that OVP demonstrates competitive performances when compared with the MISOCP approach. This result suggests the surprising effectiveness of such a simple heuristic. Finally, we observe that, in some cases, MISOCP produces solutions with higher costs than those produced by OVP, because MISOCP optimizes an approximate cost function instead of the exact one. We also note that OVP is motivated by the structural properties of the MISOCP formulation. Therefore, it is not surprising that the two heuristics exhibit comparable performances.

The effectiveness and efficiency of the simple heuristics of OVP and OSP carry practical significance. For example, in a practical OR planning setting, it may be difficult to estimate the precise distributions of job durations because of the lack of data. For the case of scheduling surgeries for ORs, it is noted that for approximately half of the surgeries scheduled in U.S. hospitals, only five or fewer data points of the same surgery type by the same surgeon are available from the preceding year (Macario 2010). Our heuristics can identify sequences with promising performances requiring only the variances of job durations. This property makes our heuristics easier to implement than the traditional SAA integer programming approach for the sequencing problem.

5. Conclusions and Future Research

Our research can be extended in several directions. First, we plan to explore alternative approximations of the serial supply chain cost. One important point to note is that the lower bound formulation (Proposition 1) is independent of the approximations we used for the serial supply chain problem. Therefore, by applying a better approximation (should one exist), it is possible to further improve the performance of our heuristics. We note that a possible source of error in the SS approximation is the substitution of the cost of a multistage supply chain by a single-stage newsvendor expression. As a result, the approximated cost depends only on the sum of local base-stock levels along chains and not the individual values.

Second, our paper mainly focuses on the aspect of job sequencing, i.e., determining the order in which to perform jobs, rather than appointment scheduling, i.e., determining the time allowances for jobs. It will be interesting to further explore the applicability of inventory-based approximations for the appointment scheduling problem.

Third, our MISOCP formulation can be a building block for addressing the important and difficult problem of assigning and sequencing jobs for multiple servers, which remains an open area of research (see Denton et al. 2007). To the best of our knowledge, multiple-server problems have only been addressed in two recent papers by Denton et al. (2010) and Zacharias and Pinedo (2013), both of which study settings that are different from ours. To tackle the multiple-server counterpart of our stochastic sequencing and scheduling problem, one potential approach is to develop an extension of our MISOCP approximation. One potential way to use our results in this paper is to derive valid inequalities based on imposing the partial ordering rule in Proposition 3, or the OVP heuristic, for jobs assigned to the same server to tighten the MISOCP formulation. However, several complicating factors require our special attention. This is particularly true in the case of OR scheduling. For instance, many surgeons perform multiple surgeries, which can be scheduled in different ORs, on the same day. Thus, the formulation needs to incorporate constraints to prevent the overlapping of jobs (surgeries) performed by the same surgeon. With such extra constraints and larger problem sizes, it may be necessary to apply recent advances in integer conic programming to ensure computational tractability.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2013.0470.

Acknowledgments

The authors gratefully acknowledge Stephen Graves, the associate editor, and four anonymous referees for their valuable comments that helped improve this paper considerably. The authors also thank Nan Liu (Columbia University) and Robert Storer (Lehigh University) for their helpful comments and discussions. The research of Ying Rong was partially sponsored by Shanghai Pujiang Program and the National Science Foundation of China [Grant 71202068]. The research of Jiawei Zhang was partially supported by the National Science Foundation of China [Grant 71331004].

References

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.

Birge JR, Louveaux F (1997) *Introduction to Stochastic Programming* (Springer, New York).

Cardoen B, Demeulemeester E (2011) A decision support system for surgery sequencing at UZ Leuven's day-care department. *Internat. J. Inform. Tech. Decision Making* 10(3):435–450.

Cardoen B, Demeulemeester E, Belien J (2010) Operating room planning and scheduling: A literature review. *Eur. J. Oper. Res.* 201(3):921–932.

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

Chen F, Song JS (2001) Optimal policies for multiechelon inventory problems with Markov-modulated demand. *Oper. Res.* 49(2):226–234.

Chen F, Zheng Y-S (1994) Lower bounds for multi-echelon stochastic inventory systems. *Management Sci.* 40(11):1426–1443.

Chen RR, Robinson LW (2014) Sequencing and scheduling appointments with potential call-in patients. *Production Oper. Management*. Forthcoming.

Clark AJ, Scarf H (1960) Optimal policies for a multi-echelon inventory problem. *Management Sci.* 6(4):475–490.

DeCoster C, Carriere KC, Peterson S, Walld R, MacWilliam L (1999) Waiting times for surgical procedures. *Medical Care* 37(6): JS187–JS205.

Denton BT, Gupta D (2003) A sequential bounding approach for optimal appointment scheduling. *IIE Trans.* 35(11):1003–1016.

Denton BT, Viapiano J, Vogl A (2007) Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Sci.* 10(1):13–24.

Denton BT, Miller AJ, Balasubramanian HJ, Huschka TR (2010) Optimal allocation of surgery blocks to operating rooms under uncertainty. *Oper. Res.* 58(4):802–816.

Erdogan SA, Denton BT (2011) Surgery planning and scheduling. Cochran JJ, Cox LA Jr, Keskinocak P, Kharoufeh JP, Smith JC, eds. *Wiley Encyclopedia of Operations Research and Management Science* (Wiley, Hoboken, NJ).

Gallego G, Viapiano J, Ozer O (2005) A new algorithm and a new heuristic for serial supply systems. *Oper. Res. Lett.* 33(4): 349–362.

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

Glasserman P, Tayur S (1994) The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Oper. Res.* 42(5):913–925.

Gupta D (2007) Surgical suites operations management. *Production Oper. Management* 16(6):689–700.

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

Huh WT, Janakiraman G (2008) A sample-path approach to the optimality of echelon order-up-to policies in serial inventory systems. *Oper. Res. Lett.* 36(5):547–550.

Huh WT, Janakiraman G, Nagarajan M (2010) Capacitated serial inventory systems: Sample path and stability properties under base-stock policies. *Oper. Res.* 58(4):1017–1022.

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

Kong Q, Lee C, Teo CP, Zheng Z (2013) Scheduling arrivals to a stochastic service delivery system using copositive cones. *Oper. Res.* 61(3):711–726.

Liao C-J, Pegden CD, Rosenshine M (1993) Planning timely arrivals to a stochastic production or service system. *IIE Trans.* 25(5): 63–73.

Macario A (2010) Is it possible to predict how long a surgery will last? *Medscape Anesthesiology* 108(3):681–685.

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

Moon I, Gallego G (1994) Distribution free procedures for some inventory models. *J. Oper. Res. Soc.* 45(6):651–658.

Olivares M, Terwiesch C, Cassorla L (2008) Structural estimation of the newsvendor model: An application to reserving operating room time. *Management Sci.* 54(1):41–55.

Perakis G, Roels G (2008) Regret in the newsvendor model with partial information. *Oper. Res.* 56(1):188–203.

Robinson LW, Chen RR (2003) Scheduling doctors's appointments: Optimal and empirically-based heuristic policies. *IIE Trans.* 35(3):295–307.

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

Sauder School of Business (2011) Innovative Sauder-BC Cancer Agency system reduces waitlisting for chemotherapy. (July 4), http://www.sauder.ubc.ca/News/2011/Sauder-BC_Cancer_Agency_system_reduces_waitlisting_for_chemotherapy.

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.

See CT, Sim M (2010) Robust approximation to multiperiod inventory management. *Oper. Res.* 58(3):583–594.

Shang KH, Song JS (2003) Newsvendor bounds and heuristic for optimal policies in serial supply chains. *Management Sci.* 49(5):618–638.

Van Mieghem JA (2000) Price and service discrimination in queuing systems: Incentive compatibility of $Gc\mu$ scheduling. *Management Sci.* 46(9):1249–1267.

Vanden Bosch PM, Dietz DC (2000) Minimizing expected waiting in a medical appointment system. *IIE Trans.* 32(9):841–848.

Vanden Bosch PM, Dietz DC (2001) Scheduling and sequencing arrivals to an appointment system. *J. Service Res.* 4(1):15–25.

Wang P (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.

Zacharias C, Pinedo M (2013) Appointment scheduling with no-shows and overbooking. *Production Oper. Management*, ePub ahead of print August 23, doi: 10.1111/poms.12065.

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

Zipkin P (2000) *Foundations of Inventory Management* (McGraw-Hill New York).