OPERATIONS RESEARCH
Vol. 60, No. 3, May–June 2012, pp. 675–681
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.1120.1053

Technical Note—A Sampling-Based Approach to Appointment Scheduling

Mehmet A. Begen
Ivey School of Business, University of Western Ontario, London, Ontario N6A 3K7, Canada,
mbegen@ivey.uwo.ca

Retsef Levi
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139, retsef@mit.edu

Maurice Queyranne
Sauder School of Business, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada,
maurice.queyranne@sauder.ubc.ca

We consider the problem of appointment scheduling with discrete random durations but under the more realistic assumption that the duration probability distributions are not known and only a set of independent samples is available, e.g., historical data. For a given sequence of appointments (jobs, tasks), the goal is to determine the planned starting time of each appointment such that the expected total underage and overage costs due to the mismatch between allocated and realized durations is minimized. We use the convexity and subdifferential of the objective function of the appointment scheduling problem to determine bounds on the number of independent samples required to obtain a provably near-optimal solution with high probability.

Subject classifications: appointment scheduling; project management; surgery scheduling; discrete random durations; optimization; sample average approximation; nonparametric sampling approach.
Area of review: Optimization.
History: Received March 2010; revision received June 2011; accepted November 2011.

1. Introduction

We study the appointment (job, task) scheduling problem with discrete random durations as recently studied in Begen and Queyranne (2011). However, we assume that the probability distributions of job durations are not known and the only available information on the durations is a set of independent random samples. (Such samples may be obtained from historical data.) We determine bounds on the number of independent samples required to obtain a provably near-optimal solution with high probability by using the results developed in Begen (2010) on convexity and subdifferential of the objective function of the appointment scheduling problem.

1.1. Appointment Scheduling Problem (ASP)

We adapt the notation of Begen and Queyranne (2011). There are $n+1$ jobs numbered $1, 2, \ldots, n+1$ that need to be sequentially processed (in the order $1, 2, \ldots, n+1$) on a single processor. An appointment schedule (a vector of planned start times) needs to be determined before any processing can start. That is, each job i is assigned a planned start time A_i. In particular, job i will not be available before A_i. The process durations are a priori random and are realized only after all the appointment times are set. Therefore some jobs may finish earlier, whereas some others may finish later, than the appointment time of the next job. If job i ends earlier than the next job's appointment time then the system experiences *underage cost* at a rate of u_i due to underutilization of the processor. On the other hand, when job i finishes later than the next job's appointment time, the system is exposed to *overage cost* at a rate of o_i due to the wait of the next job and/or overtime for the processor. Therefore there is a trade-off between underutilization, waiting, and overtime, i.e., underage and overage. The goal is to find an appointment schedule $\mathbf{A} = (A_1, \ldots, A_{n+1})$ that minimizes the total expected (underage and overage) cost. (We write all vectors as row vectors.)

There are important real-world applications that fit this model (especially in healthcare), such as surgery scheduling, medical appointments, transportation, project management, and production. Specifically, in surgery scheduling, we can think of surgeries as the jobs/appointments, operating room/surgical team as the processor, and the hospital as the scheduler. As observed in practice, surgery durations show variability (Strum et al. 2000) and determining planned start times, i.e., setting appointment times of surgeries, is an important and challenging task (Erdogan and Denton 2011). A surgery appointment schedule has a direct impact on the amounts of overtime and idle time of operating room(s) (Peltokorpi et al. 2008). An operating

room's overtime can be costly because it involves staff overtime as well as additional overhead costs; on the other hand, idle time costs can also be high because of the opportunity cost of unused capacity (Erdogan and Denton 2011). A similar trade-off exists in scheduling container ships arrivals at a container terminal (Sabria and Daganzo 1989). Another example comes from a production system that has multiple stages and stochastic leadtimes and the objective is to determine planned leadtimes to minimize expected cost (Elhafsi 2002). Note that in most of the applications mentioned above (e.g., tasks in project management, planes at an airport gate/runway, container ships arriving to a terminal, certain surgeries/tests in a hospital) the sequence of jobs (tasks, appointments) is not controlled by the scheduler who makes the duration allocation decisions.

When there is only a single job the appointment scheduling reduces to the well-known newsvendor problem. This was first recognized by Weiss (1990). However, the problem departs from newsvendor characteristics and solution methods in the case of two or more jobs. In the multiperiod newsvendor problem, naturally, decisions are taken at each period sequentially. On the other hand, in appointment scheduling, one needs to have a schedule before any processing can start, i.e., one determines all the decision variables (i.e., appointment times) simultaneously at the beginning of the planning horizon (i.e., at time zero). In Begen (2010), a link between a class of inventory problems (advance multiperiod quantity commitments) and the ASP is established and it is shown that they can be solved as special cases of the ASP.

1.2. Sampling-Based Appointment Scheduling

Begen and Queyranne (2011) assume complete information on job duration distributions, i.e., there is an underlying discrete joint probability distribution for job durations, and this distribution is available and known fully. This may be the case for some applications. However, in other situtations, the true duration distribution is not known but its (past) realizations or some samples may be available. For example, hospitals and surgeons usually keep data on the length of previous surgeries, but no one can fully characterize the true distribution for a certain type of surgery.

In this paper, we assume that there is an underlying joint discrete distribution for the job durations but we only have access to a set of independent samples. For instance, this may correspond to historical data such as daily observations of surgery durations. Job durations need not be independent but samples are. In other words, each sample is a vector of durations where each component corresponds to a job duration, and these vectors are independent. Unlike the common assumption in the literature we do not require the job duration probability distributions to be independent of each other. Then the question becomes how to use these samples to find a "good solution." For a given set of samples we first form an empirical distribution. Then we solve the ASP, by using the algorithmic results developed in Begen and Queyranne (2011), with respect to the empirical distribution (i.e., as if this empirical distribution was the true distribution) and obtain a (sampling-based) solution. Next we establish a link between the number of samples and the quality of the sampling-based solution (with respect to the optimal solution relative to the true unknown distribution). In summary, we determine the number of independent samples required to obtain a provably near-optimal solution with high probability, i.e., the expected cost of the sampling-based solution is with probability at least $1 - \delta$ no more than $(1 + \epsilon)$ times the expected cost of the optimal schedule that is computed based on the true (unknown) distribution. We call this the sampling-based approach and refer to this problem, i.e., the problem of finding such a sampling-based solution, as the *sampling appointment scheduling problem* and will denote it as *sampling-ASP*. Our sample size bound is polynomial in the number of jobs, the accuracy level, the confidence level, and the cost coefficients. It does not depend on any parameter of the underlying job durations distribution.

The rest of this paper is organized as follows. In the remaining of this section, we present a brief literature review and highlight the contributions of this paper. In §2, we present notation and foundation needed for the sampling analysis. We present our sampling analysis in §3. Section 4 concludes the paper. We provide all the proofs in the (online) appendix. An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1053.

1.3. Brief Literature Review

Researchers have studied the ASP for the last 60 years, e.g., see Cayirli and Veral (2003). The existing literature only considers continuous processing durations with full probability characterization, i.e., the probability distributions of job processing durations are given as part of the input. Because of the continuous processing durations there are computational difficulties in the evaluation of the expected total cost. For a given sequence of jobs, only small instances can be solved to optimality; larger instances require heuristics. Recently Begen and Queyranne (2011) studied a discrete time version of the ASP, i.e., the processing durations are integer and given by a discrete probability distribution. This assumption fits many applications; for example, surgeries and physician appointments are at best scheduled on a minute-to-minute basis, and usually a block of certain minutes. (For instance, one 20 minute physician appointment could be two 10-minute blocks.) Begen and Queyranne (2011) establish the discrete convexity of the ASP objective function (under a mild condition) and show that an optimal schedule can be found in polynomial time.

There has been much interest for studying stochastic models with partial probabilistic characterization. For example, inventory models, especially the newsvendor problem and its multiperiod extension, have received a lot of attention. Depending on how much is known about the

true distribution(s) different approaches are possible, e.g., see Levi et al. (2007) and the references therein.

Levi et al. (2007) studied the classical newsvendor problem in the absence of a demand distribution. They use the *sampling average approximation* (SAA) method (e.g., see Shapiro 2007) and subgradient information from the (single period) newsvendor problem to determine the number of samples required for a provably near optimal solution. For the multiperiod case, they develop a dynamic programming framework using information obtained from the single period problem. Our objective, determining number of required samples, is similar to theirs. However we work with a nonseparable multivariable objective function. Furthermore, our computational approach is not dynamic programming.

Besides inventory models, researchers use sampling methods for stochastic programs, in particular the SAA method. SAA is one of the most popular approximation methods for stochastic programs. It works by replacing the true distribution with an empirical distribution obtained from random samples. Several papers, e.g., Shapiro (2007) and the references therein obtain results on convergence and the number of samples required for the SAA method to yield small relative errors with high probability.

1.4. Our Contributions

Our contributions in the present work are as follows:

1. To the best of our knowledge, the entire literature on appointment (job, task) scheduling focuses on cases with known duration distributions, almost in all cases independent variables with continuous distributions from a parameterized family with known or estimated parameters. Our analysis is nonparametric and does not assume independent durations. It is also the first work to address the ASP when the durations distributions are unknown, possibly correlated, and only sample information is available. The latter assumption makes the model significantly more realistic.

2. Our solution approach to the sampling-ASP is based on the well-known SAA method, which essentially replaces the original problem with a problem defined with respect to the empirical distribution induced by the samples. The resulting SAA of the ASP is a stochastic nonlinear and nonseparable integer program. However, our approach is not standard stochastic programming methodology. Based on a significant extension of the work of Begen and Queyranne (2011), we use the notion of discrete convexity to solve the SAA in polynomial time for (possibly) correlated job durations under very mild conditions.

3. In addition, we develop distribution-free bounds on the number of samples required to guarantee that the optimal SAA solution is arbitrarily close to the optimal solution that could be obtained if the duration distributions are known. Our analysis approach is based on a novel, complete characterization of the subdifferential of the ASP objective function. We also present a new version of a multidimensional bounding lemma in Levi et al. (2007) that is needed in our context. This is significantly different than the dynamic-programming based analysis of Levi et al. (2007) for the multiperiod stochastic inventory control problem. Unlike the ASP that is multivariable nonseperable stochastic optimization, the latter problem could be decomposed into a sequence of univariable problems.

2. Preliminaries

We present the results on the convexity and subdifferential of the ASP's objective function (developed in Begen 2010) used in the present sampling analysis. We start with additional notation needed (adapted from Begen and Queyranne 2011 and Begen 2010) on the ASP.

The random processing duration of job i is given by p_i and $\mathbf{p} = (p_1, p_2, \ldots, p_n, 0)$. The term $\bar{p}_i$ denotes the maximum possible value of processing duration p_i. The maximum of these $\bar{p}_i$'s is $\bar{p}_{\max} = \max(\bar{p}_1, \ldots, \bar{p}_n)$. The cost coefficients are together represented as $\mathbf{u} = (u_1, u_2, \ldots, u_n)$ and $\mathbf{o} = (o_1, o_2, \ldots, o_n)$.

As in Begen and Queyranne (2011), we assume that all cost coefficients and processing durations are nonnegative and processing durations are integer valued with respect to some given base duration unit. (For example, in the surgery scheduling we can think of a five-minute interval as the base unit.) We can restrict ourselves to integer appointment schedules, without loss of optimality by the appointment vector integrality theorem 5.1 of Begen and Queyranne (2011).

Job 1 starts on time, i.e., $A_1 = 0$, and there are n real jobs. The $(n+1)$th job is a dummy job with a processing duration of zero. We use the dummy job to compute the overage or underage cost of the nth job. The start time and completion time of job i are defined as $S_i = \max\{A_i, C_{i-1}\}$ and $C_i = S_i + p_i$ for $2 \leqslant i \leqslant n+1$, respectively. Because job 1 starts on time we have $S_1 = A_1 = 0$ and $C_1 = p_1$. Note that S_i and C_i are random variables that depend on the appointment vector $\mathbf{A}$, and the random duration vector $\mathbf{p}$. The total cost due to job i completing at C_i is $u_i(A_{i+1} - C_i)^+ + o_i(C_i - A_{i+1})^+$, where $(x)^+ = \max(0, x)$ is the positive part of real number x. The total cost of appointment vector $\mathbf{A}$ given processing duration vector $\mathbf{p}$ is defined as

$$F(\mathbf{A} \mid \mathbf{p}) = \sum_{i=1}^{n} (o_i(C_i - A_{i+1})^+ + u_i(A_{i+1} - C_i)^+).$$

The objective to be minimized is the expected total cost $F(\mathbf{A}) = \mathrm{E}_{\mathbf{p}}[F(\mathbf{A} \mid \mathbf{p})]$, where the expectation is taken with respect to random processing duration vector $\mathbf{p}$.

Lemma 3.3.1 of Begen (2010) shows that $F(\mathbf{A} \mid \mathbf{p})$, for any $\alpha_i \in \mathbb{R} (1 \leqslant i \leqslant n)$, can be rewritten as

$$\begin{aligned} F(\mathbf{A} \mid \mathbf{p}) &= \sum_{j=1}^{n} F_j(\mathbf{A} \mid \mathbf{p}) \\ &= \sum_{j=1}^{n} [\alpha_j L_j(\mathbf{A} \mid \mathbf{p}) + \beta_j T_j(\mathbf{A} \mid \mathbf{p}) + \gamma_j M_j(\mathbf{A} \mid \mathbf{p})], \end{aligned} \tag{1}$$

where $L_j(\mathbf{A} \mid \mathbf{p}) = (C_j - A_{j+1})$ (lateness of job j); $T_j(\mathbf{A} \mid \mathbf{p}) = (C_j - A_{j+1})^+$ (its tardiness); $M_j(\mathbf{A} \mid \mathbf{p}) = \max\{C_j, A_{j+1}\} - \sum_{k=1}^{j} p_k$ (total idle time of jobs $1, 2, \ldots, j$); $\beta_i = (o_i - \alpha_i)$; $\gamma_i = [(u_i + \alpha_i) - (u_{i+1} + \alpha_{i+1})]$ (with $\gamma_n = u_n + \alpha_n$); and the α_j are numbers that are assumed to satisfy Definition 1. Proposition 3.3.3 of Begen (2010) shows F is convex under a mild monotonicity condition (α-monotonicity) of cost coefficients. We recall the definition of α-monotonicity, Definition 3.3.2 in Begen (2010).

Definition 1. The cost coefficients $(\mathbf{u}, \mathbf{o})$ are *α-monotone* if there exist real numbers α_i $(1 \leqslant i \leqslant n)$ such that $0 \leqslant \alpha_i \leqslant o_i$ and $u_i + \alpha_i$ are nonincreasing in i.

The objective function is convex (under α-monotonicity) but nonsmooth. (It has kinks.) Because of the nondifferentiability of the objective function we work with its subdifferential ∂F, the set of all subgradients. Using Minkowski sums, subdifferential rules (Hiriart-Urruty and Lemarèchal 1993), and Equation (1), ∂F is characterized and expressed component by component in a closed-form formula in Begen (2010). Before we give the formula, we need to introduce more concepts and notation. The subdifferential of F is characterized by first obtaining subdifferentials $\partial L_j(\mathbf{A} \mid \mathbf{p})$, $\partial T_j(\mathbf{A} \mid \mathbf{p})$ and $\partial M_j(\mathbf{A} \mid \mathbf{p})$ and then finding the subdifferential of the corresponding expected values $\partial L_j(\mathbf{A})$, $\partial T_j(\mathbf{A})$, $\partial M_j(\mathbf{A})$. Finally, $\partial F(\mathbf{A})$ is obtained as the Minkowski sum of $\partial L_j(\mathbf{A})$, $\partial T_j(\mathbf{A})$ and $\partial M_j(\mathbf{A})$ over all jobs. To find $\partial L_j(\mathbf{A} \mid \mathbf{p})$ one needs to know which jobs k $(1 \leqslant k \leqslant j)$ maximize $\{A_k + P_{kj}\}$, where $P_{kj} = \sum_{t=k}^{j} p_t$. The set of such maximizers for job j is defined as $I_j = \arg\max_{k \leqslant j}\{A_k + P_{kj}\}$. Similarly, for $\partial T_j(\mathbf{A} \mid \mathbf{p})$ and $\partial M_j(\mathbf{A} \mid \mathbf{p})$ we define $I_j^{\varrho} = \{k \in I_j\colon A_k + P_{kj} \varrho A_{j+1}\}$, where the relation $\varrho \in \{>, =\}$.

Furthermore, the characterizations of $\partial L_j(\mathbf{A} \mid \mathbf{p})$, $\partial T_j(\mathbf{A} \mid \mathbf{p})$ and $\partial M_j(\mathbf{A} \mid \mathbf{p})$ include nonnegative variables representing convex combination weight terms. Let $[j] = \{1, 2, \ldots, j-1, j\}$ and $\mathcal{P}^*([j])$ denote all the nonempty subsets of $[j]$. Then for the job j the weight variables $\mathbf{X^j}$ and their feasible set Θ^j are defined as

$$\mathbf{X^j} = \big((X_{ij}^{v}(S)), (X_{kj}^{M^=}(S \cup \{j+1\}))\colon v \in \{L, T^>, T^=, M^>\}, \; 1 \leqslant i \leqslant j \leqslant n+1, 1 \leqslant k < j \leqslant n+1, S \in \mathcal{P}^*([j]), \; i \in S, k \in S\big);$$

$$\Theta^j = \Bigg\{\mathbf{X^j} \geqslant 0\colon \sum_{i \in S} X_{ij}^{v}(S) = 1, \sum_{i \in S} X_{ij}^{T^=}(S) \leqslant 1, \; \sum_{k \in S \cup \{j+1\}} X_{kj}^{M^=}(S \cup \{j+1\}) = 1, \forall v \in \{L, T^>, M^>\}, \; \forall S \in \mathcal{P}^*([j]), \forall i \in S, \forall k \in S\Bigg\}.$$

All $\mathbf{X^j}$ vectors can be collected into a single vector $\mathbf{X} = (\mathbf{X^j})_{j \in [n+1]}$ and we can then express the feasible set Θ of $\mathbf{X}$:

$$\Theta = \times_{j \in [n+1]} \Theta^j = \{\mathbf{X} = (\mathbf{X^j})_{j \in [n+1]}\colon \mathbf{X^j} \in \Theta^j \; \forall\, j \in [n+1]\}.$$

Now with the definitions of $\mathbf{X}$ and Θ, we can present $\partial F(\mathbf{A})$ component by component for a particular $\mathbf{X} \in \Theta$, i.e., each coordinate of the subgradient at point $\mathbf{A}$ defined for a particular $\mathbf{X} \in \Theta$. Let $g(\mathbf{X}, \mathbf{A})$ be the element of $\partial F(\mathbf{A})$ defined by the vector $\mathbf{X}$. Then $g(\mathbf{X}, \mathbf{A}) = (g_1(\mathbf{X}, \mathbf{A}), g_2(\mathbf{X}, \mathbf{A}), \ldots, g_{n+1}(\mathbf{X}, \mathbf{A}))$, where $g_k(\mathbf{X}, \mathbf{A})$ is the kth component of $g(\mathbf{X}, \mathbf{A})$. Corollary 3.4.9 of Begen (2010) gives an expression for the kth component of $g(\mathbf{X}, \mathbf{A})$ (i.e., $g_k(\mathbf{X}, \mathbf{A})$) as

$$\begin{aligned}
&\sum_{j=k}^{n} \alpha_j \sum_{S \in \mathcal{P}^*([j])} \text{Prob}\{I_j = S\} X_{kj}^{L}(S) - \alpha_{k-1} \\
&\quad \cdot \sum_{S \in \mathcal{P}^*([k-1])} \text{Prob}\{I_{k-1} = S\} \\
&\quad + \sum_{j=k}^{n} \beta_j \sum_{S \in \mathcal{P}^*([j])} \text{Prob}\{I_j^{>} = S\} X_{kj}^{T^>}(S) - \beta_{k-1} \\
&\quad \cdot \sum_{S \in \mathcal{P}^*([k-1])} \text{Prob}\{I_{k-1}^{>} = S\} \\
&\quad + \sum_{j=k}^{n} \beta_j \sum_{S \in \mathcal{P}^*([j])} \text{Prob}\{I_j^{=} = S\} X_{kj}^{T^=}(S) - \beta_{k-1} \\
&\quad \cdot \sum_{S \in \mathcal{P}^*([k-1])} \text{Prob}\{I_{k-1}^{=} = S\} \sum_{i \in S} X_{ik-1}^{T^=}(S) \\
&\quad + \sum_{j=k}^{n} \gamma_j \sum_{S \in \mathcal{P}^*([j])} \text{Prob}\{I_j^{>} = S\} X_{kj}^{M^>}(S) - \gamma_{k-1} \\
&\quad \cdot \sum_{S \in \mathcal{P}^*([k-1])} \text{Prob}\{I_{k-1}^{>} = S\} + \sum_{j=k}^{n} \gamma_j \sum_{S \in \mathcal{P}^*([j])} \text{Prob}\{I_j^{=} = S\} \\
&\quad \cdot X_{kj}^{M^=}(S \cup \{j+1\}) + \gamma_{k-1} \\
&\quad \cdot \Bigg(1 - \sum_{S \in \mathcal{P}^*([k-1])} \text{Prob}\{I_{k-1}^{=} = S\}\Bigg). \qquad (2)
\end{aligned}$$

Remark 2. Note that $\sum_{S \in \mathcal{P}^*([k-1])} \text{Prob}\{I_{k-1} = S\} = 1$. For our analysis in this paper, we do not require the values of $\text{Prob}\{I_j = S\}$, $\text{Prob}\{I_j^{>} = S\}$, and $\text{Prob}\{I_j^{=} = S\}$ (for $S \in \mathcal{P}^*([j])$ and $j \in [n+1]$). However these probabilities may be needed for other research, and indeed these probabilities are computed and used in Begen (2010) to compute subgradients of F.

Remark 3. There may be a given due date D, an integer satisfying $0 \leqslant D \leqslant \sum_{i=1}^{n} \bar{p}_i$ for the end of processing after which overtime is incurred, instead of letting the model choose a planned makespan A_{n+1}. Proposition 3.4.11 of Begen (2010) allows us to extend our results of the present paper to the case in which there is a given due date D.

3. Sampling Approach

In this section, we relax the perfect information assumption on the job durations distribution in Begen and Queyranne (2011). Recall our assumption that there exists an underlying (true) discrete joint distribution for the job durations but

the distribution is not known. Instead independent samples are available.

We first give a formal definition of the sampling-ASP. In words, it is the problem of finding an optimal appointment schedule with respect to the empirical distribution obtained from the available samples. Let N be the number of samples. Define $\mathbf{p}^k = (p_1^k, p_2^k, \ldots, p_n^k)$ as the kth observation in the N samples. We use "$\hat{}$" to denote quantities obtained from samples. Let $\hat{\mathbf{p}} = \hat{\mathbf{p}}(N)$ be the empirical joint probability distribution obtained from N independent observations of $\mathbf{p}$, i.e., $\text{Prob}\{\hat{\mathbf{p}} = \mathbf{p}^k\} = 1/N$ for $1 \leqslant k \leqslant N$. We denote a true optimal appointment vector with $\mathbf{A}^*$, i.e., $\mathbf{A}^*$ is a minimizer of $F_{\mathbf{p}}(\mathbf{A}) = \mathrm{E}_{\mathbf{p}}(F(\mathbf{A} \mid \mathbf{p}))$. (We use the subscript $\mathbf{p}$ to emphasize the fact that the quantities are obtained with respect to the true distribution $\mathbf{p}$.) Similarly, let $\hat{\mathbf{A}} = \hat{\mathbf{A}}(N)$ be a minimizer of $F_{\hat{\mathbf{p}}}(\mathbf{A}) = \mathrm{E}_{\hat{\mathbf{p}}}(F(\mathbf{A} \mid \hat{\mathbf{p}}))$. (Again we use the subscript $\hat{\mathbf{p}}$ to emphasize the fact that the quantities are obtained with respect to the sampling distribution $\hat{\mathbf{p}}$. We will omit the subscripts when the meaning is clear from the content.) Then the sampling-ASP is minimizing $F_{\hat{\mathbf{p}}}(\mathbf{A})$, i.e., finding a $\hat{\mathbf{A}}(N)$ for a given sample size N. We now present an overview of our sampling-based approach. Let ϵ be the accuracy level and $1-\delta$ the confidence level. We solve the sampling-ASP and determine the number $N = N(\epsilon, \delta, \mathbf{u}, \mathbf{o})$ of samples required such that for any $0 < \epsilon \leqslant 1$ and $0 < \delta < 1$ we have $F_{\mathbf{p}}(\mathbf{A}(\hat{\mathbf{N}})) \leqslant (1+\epsilon)F_{\mathbf{p}}(\mathbf{A}^*)$ with probability at least $1-\delta$. For a subgradient defined by vector $\mathbf{X} \in \Theta$ at point $\mathbf{A}$, we denote its kth component as $g_k(\mathbf{X}, \mathbf{A})_{\mathbf{p}}$ for $F_{\mathbf{p}}(\cdot)$ and $g_k(\mathbf{X}, \mathbf{A})_{\hat{\mathbf{p}}}$ for $F_{\hat{\mathbf{p}}}(\cdot)$. The formulas for $g_k(\mathbf{X}, .)_{\mathbf{p}}$ and $g_k(\mathbf{X}, .)_{\hat{\mathbf{p}}}$ are identical to (2) except that we now have $\text{Prob}_{\mathbf{p}}\{\cdot\}$ terms in $g_k(\mathbf{X}, .)_{\mathbf{p}}$ and $\text{Prob}_{\hat{\mathbf{p}}}\{\cdot\}$ terms in $g_k(\mathbf{X}, .)_{\hat{\mathbf{p}}}$.

We start our analysis by proving in Theorem 4 that we can minimize $F_{\hat{\mathbf{p}}}(\cdot)$ in polynomial time. Then, with an application of Hoeffding's inequality (Hoeffding 1963), we establish a connection between the probabilities of a given event with respect to $\mathbf{p}$ and $\hat{\mathbf{p}}$ as a function of sample size N for a given accuracy level ϵ' (absolute difference of the probabilities with respect to $\mathbf{p}$ and $\hat{\mathbf{p}}$) and a confidence level $1-\delta'$. After that, we provide a similar result for a family of events $\mathcal{F}$. Then we use the subdifferential characterization to show the existence of a subgradient $g \in \partial F_{\mathbf{p}}(\hat{\mathbf{A}})$ such that $|g_k| < \epsilon' K'$ with probability at least $1 - |\mathcal{F}|\delta'$, where $|\mathcal{F}| = |\mathcal{F}|(n)$ and $K' = K'(n, \mathbf{u}, \mathbf{o})$ are some constants. After that we prove that if there exists $g \in \partial F_{\mathbf{p}}(\hat{\mathbf{A}})$ such that $|g_k| < \epsilon\nu/3(n+1)n$, where

$$\nu = \min\{u_1, u_2, \ldots, u_n, o_1, o_2, \ldots, o_n\}$$

for all $1 \leqslant k \leqslant n+1$ then $F_{\mathbf{p}}(\hat{\mathbf{A}}) \leqslant (1+\epsilon)F_{\mathbf{p}}(\mathbf{A}^*)$. This is achieved with an application of Jensen's inequality and a new version of Lemma 5.1 of Levi et al. (2007) (Lemma 12). We conclude by stating our main result that determines the number of samples required to achieve a $(1+\epsilon)$ approximation with probability at least $1-\delta$.

THEOREM 4 (POLYNOMIAL TIME ALGORITHM). *If the cost vectors* $(\mathbf{u}, \mathbf{o})$ *are* α*-monotone and the processing durations are integer then* $F_{\hat{\mathbf{p}}}$ *can be minimized in* $O(n^8 N \log(\lceil \bar{p}_{\max}/2 \rceil))$ *time.*

Polynomial time algorithm Theorem 4 shows that for a given N-size sample of job durations we can solve the corresponding sampling-ASP efficiently. (One may also solve the sampling-ASP with nonsmooth convex optimization methods by using the subdifferential characterization as in Begen 2010.) The remaining task is, for a given accuracy level ϵ and confidence level $1-\delta$, to determine a sample size N such that the sampling-ASP optimal solution will have an expected cost no more than $(1+\epsilon)$ times the optimal expected cost with probability at least $1-\delta$.

Let O be any event depending on the processing times $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, that is, $O = O(p_1, p_2, \ldots, p_n) = O(\mathbf{p})$. Let $\text{Prob}_{\mathbf{p}}\{O(\mathbf{p})\}$ denote the true probability of O. Let $\text{Prob}_{\hat{\mathbf{p}}}\{O\}$ denote an estimate of $\text{Prob}_{\mathbf{p}}\{O(\mathbf{p})\}$ when true distribution of $\mathbf{p}$ is not known, and the empirical probability distribution $\hat{\mathbf{p}}$, based on N independent samples, is used in the estimation. We define an indicator function as

$$1\{O(\mathbf{p}^k)\} = \begin{cases} 1 & \text{if event } O \text{ occurs with realization } \mathbf{p}^k \\ 0 & \text{otherwise.} \end{cases}$$

Then $1\{O(\mathbf{p}^k)\}$ is Bernoulli distributed with parameter $\text{Prob}_{\mathbf{p}}\{O(\mathbf{p})\}$. We define our estimate $\text{Prob}_{\hat{\mathbf{p}}}\{O(\mathbf{p})\}$ as

$$\text{Prob}_{\hat{\mathbf{p}}}\{O(\mathbf{p})\} = \frac{1}{N}\sum_{k=1}^{N} 1\{O(\mathbf{p}^k)\}.$$

REMARK 5. Note that $N\,\text{Prob}_{\hat{\mathbf{p}}}\{O(\mathbf{p})\}$ is the sum of N independent Bernoulli random variables with parameter $\text{Prob}_{\mathbf{p}}\{O(\mathbf{p})\}$, therefore $N\,\text{Prob}_{\hat{\mathbf{p}}}\{O(\mathbf{p})\}$ is binomially distributed with parameters $\text{Prob}_{\mathbf{p}}\{O(\mathbf{p})\}$ and N.

We use Hoeffding's inequality to obtain the number of samples N required such that

$$\text{Prob}\{|\,\text{Prob}_{\mathbf{p}}\{O(\mathbf{p})\} - \text{Prob}_{\hat{\mathbf{p}}}\{O(\mathbf{p})\}| \leqslant \epsilon'\} > 1 - \delta'$$

for any given accuracy level $\epsilon' > 0$ and confidence level $0 < \delta' < 1$. A direct application of Hoeffding's inequality for Bernoulli random variables (Theorem 4.5 in Wasserman 2004) yields $N > (1/2)(1/(\epsilon')^2)\ln(2/\delta')$. Using union bounds we obtain a similar result for a family of events.

LEMMA 6. *Let* $\mathcal{F}$ *be a family of* (*possibly dependent*) *events,* $\mathcal{F} = \{O_1, O_2, \ldots, O_{|\mathcal{F}|-1}, O\}$, *where each* $O_k \in \mathcal{F}$ *depends on the processing times* $\mathbf{p} = (p_1, p_2, \ldots, p_n)$. *Let* $0 < \epsilon', \delta' < 1$. *If* $N > (1/2)(1/(\epsilon')^2)\ln(2/\delta')$ *then* $\text{Prob}\{|\,\text{Prob}_{\mathbf{p}}\{O_k(\mathbf{p})\} - \text{Prob}_{\hat{\mathbf{p}}}\{O_k(\mathbf{p})\}| \leqslant \epsilon' \ \forall k = 1, 2, \ldots, |\mathcal{F}|\} > 1 - |\mathcal{F}|\delta'$.

Recall that $\hat{A}$ is an optimal appointment vector for $F_{\hat{\mathbf{p}}}$. Therefore there exists $\hat{\mathbf{X}} \in \Theta$ such that $g_k(\hat{\mathbf{X}}, \hat{A})_{\hat{\mathbf{p}}} = 0$ for all $1 \leqslant k \leqslant n+1$. We show in Lemma 7 that if we

take enough samples then $|g_k(\mathbf{X}, \hat{\mathbf{A}})_{\mathbf{p}} - g_k(\mathbf{X}, \hat{\mathbf{A}})_{\hat{\mathbf{p}}}|$ will be small with high probability. This implies that there exists a small $g \in \partial F_{\mathbf{p}}(\hat{\mathbf{A}})$. Let $\varphi_{\max}$ be $\max\{\varphi_1, \ldots, \varphi_n\}$ for $\varphi \in \{o, u, \alpha, \beta, \gamma\}$.

Lemma 7. *If* $N > (1/2)(1/(\epsilon')^2)\ln(2/\delta')$ *then* $|g_k(\hat{\mathbf{X}}, \hat{A})_{\mathbf{p}}| < \epsilon' K'$ *for all* $k = 1, \ldots, n+1$ *with probability at least* $1 - |\mathcal{F}|\delta'$, *where* $\hat{\mathbf{X}} \in \Theta$, $g(\hat{\mathbf{X}}, \hat{A})_{\hat{\mathbf{p}}} = 0$ $|\mathcal{F}| = 5n^2 + 5$ *and* $K' = n(14o_{\max} + 6u_{\max})$.

Remark 8. If $u_i = u$ for all $i = 1, 2, ..., n$ then with $K' = n(6o_{\max} + 2u) + 4u$, we have

$$|g_k(\hat{\mathbf{X}}, \hat{A})_{\mathbf{p}} - g_k(\hat{\mathbf{X}}, \hat{A})_{\hat{\mathbf{p}}}| \leqslant \epsilon'(n(6o_{\max} + 2u) + 4u)(1 \leqslant k \leqslant n+1)$$

with probability at least $1 - |\mathcal{F}|\delta'$, where $|\mathcal{F}| = 5n^2 + 5$.

The remaining task is to establish $F_{\mathbf{p}}(\hat{\mathbf{A}}) \leqslant (1+\epsilon)F_{\mathbf{p}}(\mathbf{A}^*)$ when there exists a sufficiently small subgradient $g \in \partial F_{\mathbf{p}}(\hat{\mathbf{A}})$. We show this result in Lemma 13. But before doing so we need two more results and a definition.

Lemma 9. *Let* $\tilde{p}_i = \mathrm{E}[p_i]$, $\tilde{\mathbf{p}} = (\tilde{p}_1, \tilde{p}_2, \ldots, \tilde{p}_n)$, $\tilde{C}_1 = \tilde{p}_1$, $\tilde{C}_i = \max(\tilde{C}_{i-1}, A_i) + \tilde{p}_i$, *and* $\tilde{f}(\mathbf{A}) = \nu(\sum_{i=1}^n [(\tilde{C}_i - A_{i+1})^+ + (A_{i+1} - \tilde{C}_i)^+])$. *If cost coefficients* $(\mathbf{u}, \mathbf{o})$ *are* α*-monotone then* $\tilde{\mathbf{A}} = (0, \tilde{p}_1, \tilde{p}_1 + \tilde{p}_2, \ldots, \sum_{j=1}^n \tilde{p}_j) \in \arg\min_{\mathbf{A}} \tilde{f}(\mathbf{A})$ *and* $F_{\mathbf{p}}(\mathbf{A}) \geqslant \tilde{f}(\mathbf{A}) \geqslant (\nu/n)\|\mathbf{A} - \tilde{\mathbf{A}}\|_1$ *for all* $\mathbf{A}$.

Remark 10. The following example with $n = 2$ jobs shows that this lower bound is tight, that is we may have $F_{\mathbf{p}}(\mathbf{A}) = (\nu/n)\|\mathbf{A} - \tilde{\mathbf{A}}\|_1$. Let processing times $\mathbf{p} = (1, 4)$ be deterministic, $u_1 = u_2 = o_1 = o_2 = 1$ (therefore $\nu = 1$). Then $\tilde{\mathbf{A}} = (0, 1, 5)$. For $\mathbf{A} = (0, 4, 8)$, we have $F(\mathbf{A}) = 3 = (1/2)\sum_{i=1}^n |A_i - \tilde{A}_i|$.

Definition 11 (Definition 3.3 of Levi et al. 2007). Let $f\colon \mathbb{R}^m \mapsto \mathbb{R}$ be convex. A point y is an α-point if there exists a subgradient $g \in \partial f(y)$ such that $\|g\|_1 \leqslant \alpha$.

Lemma 12 (A New Version of Lemma 5.1 of Levi et al. 2007). *Let* $f\colon \mathbb{R}^m \mapsto \mathbb{R}$ *be convex, finite with a global minimizer* y^*. *Assume that there exists* $\bar{f}$ *such that* $f \geqslant \bar{f} = \lambda\|y - \tilde{y}\|_1$ *for some* $\lambda > 0$ *and* $\tilde{y} \in \mathbb{R}^m$. *If* $\hat{y}$ *is an* α*-point for* $\alpha = \lambda\epsilon/3$ *then* $f(\hat{y}) \leqslant (1+\epsilon)f(y^*)$, *where* $\epsilon \in [0, 1]$.

The last step we need before our main result is to prove that for a suitably chosen $\hat{\mathbf{A}} = \hat{\mathbf{A}}(N)$, where $N = N(\epsilon, \delta, \mathbf{u}, \mathbf{o})$, we have $F_{\mathbf{p}}(\hat{\mathbf{A}}) \leqslant (1+\epsilon)F_{\mathbf{p}}(\mathbf{A}^*)$ for any $0 < \epsilon \leqslant 1$. We derive the following result from Lemma 12.

Lemma 13. *Let* $0 < \epsilon \leqslant 1$. *If there exists* $g \in \partial F_{\mathbf{p}}(\hat{\mathbf{A}})$ *such that* $|g_k| < \epsilon\nu/(3(n+1)n)$ *for all* $1 \leqslant k \leqslant n+1$ *then* $F_{\mathbf{p}}(\hat{\mathbf{A}}) \leqslant (1+\epsilon)F_{\mathbf{p}}(\mathbf{A}^*)$.

Combining Lemmata 6, 7, and 13 yields our main result for the sampling-based approach.

Theorem 14. *Let* $0 < \epsilon \leqslant 1$ (*accuracy level*) *and* $0 < 1 - \delta < 1$ (*confidence level*) *be given. If* $N > (4.5(1/\epsilon)^2 (n^2(n+1)(14o_{\max} + 6u_{\max})/\nu)^2 \ln(2(5n^2+5)/\delta))$ *then* $F_{\mathbf{p}}(\hat{\mathbf{A}}) \leqslant (1+\epsilon)F_{\mathbf{p}}(\mathbf{A}^*)$ *with probability at least* $1 - \delta$.

Remark 15. In the case of uniform underage cost coefficients, i.e., $u_i = u$ for all i the bound in Theorem 14, $(4.5(1/\epsilon)^2(n^2(n+1)((6o_{\max} + 2u) + 4u)/\nu)^2 \ln(2(5n^2+5)/\delta)$, is similar but has a slightly higher polynomial with respect to the number of jobs n compared to the bound obtained for the multiperiod newsvendor problem in Levi et al. (2007), with respect to the number of periods T. This is expected because in the ASP one needs to make all the decisions, i.e., determine the planned start times of all jobs, at once (before any processing starts), whereas in the inventory problem one decides sequentially at each period.

4. Conclusion

We considered the ASP with discrete random durations, studied by Begen and Queyranne (2011), but without assuming any (prior) knowledge about the probability distribution of job durations. We assume that there is an underlying (true) joint discrete distribution for the job durations, and only independent samples are available, e.g., daily historical observations of surgery durations. Job durations need not be independent but samples are. We developed a sampling-based approach to solve the sampling problem and determine the number of independent samples required to obtain a provably near-optimal solution with high probability, i.e., the cost of the sampling-based optimal schedule is with high probability no more than $(1+\epsilon)$ times the cost of an optimal schedule if the true distribution were known. The bound on the samples is polynomial in the number n of jobs, accuracy level ϵ, confidence level $1 - \delta$, and (underage and overage) cost coefficients $\mathbf{u}$ and $\mathbf{o}$, and it does not depend on the underlying distribution.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/opre.1120.1053.

Acknowledgments

This research was supported by Discovery Grant from Natural Sciences and Engineering Research Council (NSERC) of Canada to the third author, and a PGSD3 NSERC scholarship to the first author. The research of the second author is partially supported by the National Science Foundation [Grants DMS-0732175 and CMMI-0846554] (CAREER Award), Air Force Office of Scientific Research [Award FA9550-08-1-0369], a Singapore-MIT Alliance grant, and the Buschbaum Research Fund of the Massachusetts Institute of Technology.

References

Begen MA (2010) Appointment scheduling with discrete random durations and applications. Ph.D. thesis, University of British Columbia, Vancouver, British Columbia, Canada.

Begen MA, Queyranne M (2011) Appointment scheduling with discrete random durations. *Math. Oper. Res.* 36(2):240–257.

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

Elhafsi M (2002) Optimal leadtime planning in serial production systems with earliness and tardiness costs. *IIE Trans.* 34(3):233–243.

Erdogan SA, Denton BT (2011) Surgery planning and scheduling. Cochran JJ ed. *Wiley Encyclopedia of Operations Research and Management Science* (John Wiley & Sons, Hoboken, NJ).

Hiriart-Urruty JB, Lemarèchal C (1993) *Convex Analysis and Minimization Algorithms I and II* (Springer-Verlag, Berlin).

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58(301):13–30.

Levi R, Roundy RO, Shmoys DB (2007) Provably near-optimal sampling-based policies for stochastic inventory control models. *Math. Oper. Res.* 32(4):821–838.

Peltokorpi A, Lehtonen J-M, Kujala J, Kouri J (2008) Operating room cost management in cardiac surgery: A simulation study. *Internat. J. Health Tech. Management* 9(1):60–73.

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

Shapiro A (2007) Stochastic programming approach to optimization under uncertainty. *Math. Programming* 112(1):183–220.

Strum DP, May JH, Vargas LG (2000) Modeling the uncertainty of surgical procedure times: Comparison of log-normal and normal models. *Anesthesiology* 92(4):1160–1167.

Wasserman L (2004) *All of Statistics: A Concise Course in Statistical Inference* (Springer, New York).

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

Mehmet A. Begen is an assistant professor of management science at the Ivey School of Business at the University of Western Ontario (UWO). He is also affiliated with the Epidemiology and Biostatistics Department at UWO. His current research interests are management science applications, healthcare operations management, and data-driven approaches. He is a recipient of the Canadian Operations Research Society Practice Prize.

Retsef Levi is the J. Spencer Standish (1945) Professor of Management, Associate Professor of Operations Management at the Sloan School of Management, Massachusetts Institute of Technology. He is a member of the Operations Management Group at Sloan and is affiliated with the Operations Research Center. His current research is focused on the design and the performance analysis of efficient algorithms for fundamental stochastic and deterministic optimization models arising in the context of supply chain and inventory management, revenue management, logistics, and healthcare management.

Maurice Queyranne is the Advisory Council Professor in Operations and Logistics at the Sauder School of Business at the University of British Columbia, Vancouver, British Columbia, Canada. His research interests include operations research methodology, in particular discrete optimization, and applications, in particular in the areas of supply chain management and healthcare management.