*IIE Transactions* (2003) **35**, 1003–1016
Copyright © "IIE"
ISSN: 0740-817X print
DOI: 10.1080/07408170390230169

# A sequential bounding approach for optimal appointment scheduling

BRIAN DENTON[1,*] and DIWAKAR GUPTA[2]

[1]*IBM, 1000 River Road, Essex Junction, VT 05452, USA*
*E-mail: bdenton@us.ibm.com*
[2]*Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN 55455, USA*

Received June 2001 and accepted January 2003

This study is concerned with the determination of optimal appointment times for a sequence of jobs with uncertain durations. Such appointment systems are used in many customer service applications to increase the utilization of resources, match workload to available capacity, and smooth the flow of customers. We show that the problem can be expressed as a two-stage stochastic linear program that includes the expected cost of customer waiting, server idling, and a cost of tardiness with respect to a chosen session length. We exploit the problem structure to derive upper bounds that are independent of job duration distribution type. These upper bounds are used in a variation of the standard *L-shaped* algorithm to obtain optimal solutions via successively finer partitions of the support of job durations. We present new analytical insights into the problem as well as a series of numerical experiments that illustrate properties of the optimal solution with respect to distribution type, cost structure, and number of jobs.

## 1. Introduction

Appointment systems are used in many customer service industries to increase the utilization of resources, match workload to available capacity, and smooth the flow of customers. A common problem faced by decision-makers is how to determine the scheduled start times of each service when their durations are uncertain. This problem is materially different from say a machine scheduling problem (for example, see Forst (1993)) in the sense that once appointments are set, customers are not available prior to the scheduled start time, even if the server becomes free at an earlier time. Thus, choosing an early start time will lead to better server utilization at the cost of additional waiting by customers, whereas a late start time will reduce customer waiting at the cost of additional server idling. What we propose in this article is a model that can be used to find optimal start times under different cost structures associated with server idling, customer waiting, and tardiness.

The appointment scheduling problem arises in many contexts in which appointment decisions are economically significant. For example, Sabria and Daganzo (1989) consider it from the viewpoint of scheduling the arrival of cargo ships at a seaport. In their treatment of the problem the costs of underutilization of a seaport are traded off against the cost of cargo ship waiting. Wang (1993) discusses the problem in a manufacturing setting where the objective is to schedule the arrival of parts on the shop floor such that work-in-process inventory and server idling are minimized. Also, there have been numerous other studies presented in operations research, statistics, and health care journals over the past three decades on the problem of assigning appointments for arrivals at outpatient clinics (for example Bailey (1952), Welch (1964), Jansson (1966), Soriano (1966), Mercer (1973), Charnetski (1984), Ho and Lau (1992), Dexter (1999) and references therein). We begin by motivating the problem in another light by providing a specific example in the context of allocating resources for elective surgeries at hospitals.

A typical urban hospital in North America has operating expenditures measured in hundreds of millions of dollars. Operating rooms (ORs) are estimated to account for between one-third to one-half of the total costs incurred by a hospital (Redelmeier and Fuchs, 1993; Macario *et al.*, 1995). As a result they represent the area with the highest potential for cost savings. Even small relative improvements in efficiency translate into significant dollar savings and benefits to society. Major components of OR costs are fixed costs. These consist of salaries of staff, i.e., surgeons, anesthesiologists and nurses, and fixed cost of facilities and equipment. Thus, effective delivery of surgical services requires an OR manager, or similar governing body, to schedule surgeries efficiently so as to trade-off high utilization of the OR staff and other resources with low OR idling and overtime costs. At many large hospitals in North America, a

*Corresponding author

*block-booking* strategy is used to allocate time for a sequence of surgeries within a department (usually performed by the same surgeon) at a particular OR. Based on the allocation of block times a particular time of day is specified for the arrival of the staff and material resources to the OR. Since surgery durations are not known with certainty it is a common practice to schedule block sizes based on estimates of the sum of mean durations of surgeries in the block. In order to avoid large overtime costs, an empty block is scheduled at the end of the day to absorb fluctuations in finish times of all blocks scheduled that day.

Using a numerical example, we show in this article that an optimal block schedule can effectively increase the capacity of an OR and at the same time lower the sum of expected waiting, overtime and idling costs. Thus, the optimization model discussed in this article can be the source of significant cost savings. Other issues surrounding the scheduling of elective surgeries at hospitals are discussed in detail in papers by Goldman *et al.* (1970), Pierskalla and Brailer (1994), Dexter (1999), and Strum *et al.* (1999).

We shall use the term *customers* to refer to resources (e.g., surgical teams in the block-booking example above) which become available only at the assigned start time and *facility* or *server* to refer to fixed resources, such as an OR. The evaluation of a given schedule of appointment times requires the calculation of expected customer waiting times and facility idle times. Exact calculation of these quantities is problematic when there are many jobs because it requires the evaluation of multi-dimensional integrals. Accordingly, a number of previous studies have used simulation to study the performance of heuristic rules for setting appointments. According to one heuristic regime (Bailey, 1952, 1954; Welch and Bailey, 1952; Welch, 1964), if there are $n$ customers to be scheduled, $m$ of these are scheduled to arrive at the beginning of the session and the remaining $n-m$ appointment times are spaced by their mean job durations (denoted by $\mu_i$'s). Thus, if $a_i$ represents the appointment time for job $i$ then

$$a_i = 0, \quad i = 1, \ldots, m, \tag{1}$$

$$a_i = a_{i-1} + \mu_i, \quad i = m+1, \ldots, n. \tag{2}$$

Alternatively, in the block appointments regime (White and Pike, 1964; Soriano, 1966), a session of length $d$ is broken up into $k$ blocks and $n_j = n/k$ customers are scheduled to arrive at the start of each block. Thus, if we let index $i$ denote the customer, and $j(i)$ the block to which customer $i$ belongs, then

$$a_{ij} = j(i)d/k, \quad i = 1, \ldots, n, \text{ and } j = 1, \ldots, k. \tag{3}$$

Heuristics for assigning individual appointment times to customers have also been explored. For example, Charnetski (1984) considered a heuristic that assigns a job allowance of $\mu_i + h\sigma_i$ to job $i$, regardless of its place in the sequence where $\mu_i$ and $\sigma_i$ denote the mean and standard deviation of the $i$th job duration, respectively. He experimented with different values of $h$ using a simulation model while assuming that job durations are normally distributed. Ho and Lau (1992) also used simulation to compare the performances of a number of heuristics. More sophisticated heuristics are proposed by Robinson and Chen (2000), who provide an excellent review of heuristics.

Articles by Mercer (1960, 1973), Jansson (1966), Sabria and Daganzo (1989), and Brahimi and Worthington (1991) use queuing analysis to study the same problem. This literature generally assumes that job durations are independent and identically distributed (i.i.d.) and appointment times are equally spaced. With respect to the latter assumption, it has been shown that the optimal spacing of appointments (job allowances) when service times are i.i.d. is not in general uniform (Wang, 1993). Also, a majority of queuing theoretic models obtain expected customer waiting times and expected facility idle times under the assumption of a steady state. Often, however, appointments need to be set for finite session length $d$ during which the facility is up and running, and a steady state is never reached in such cases.

Another line of research is the study of optimization models for appointment systems. Weiss (1990), and Robinson *et al.* (1996), solve two and three customer problems, respectively, which can be solved relatively easily owing to the low dimensionality. Robinson *et al.* (1996) and Robinson and Chen (2000) report a method based on Monte Carlo simulation for computing appointment times. Wang (1993) considered the case in which job durations are exponentially distributed and showed that for this special case the probability density function (p.d.f.) for customer waiting times is phase-type. He then exploited the computational advantages associated with phase-type distributions to find the optimal appointment times. Vanden Bosch and Dietz (2000) present an algorithm for a similar problem for the case of phase-type distributions in which appointment slots are integer multiples of a discrete slot parameter.

In this article we formulate the Appointment Scheduling Problem (ASP) as a two-stage Stochastic Linear Program (2-SLP). We show how the problem structure can be exploited using a decomposition-based approach for solving the large-scale deterministic equivalent problem. Next, we develop general upper bounds that are independent of distribution type and cost parameters. This leads to a version of the L-shaped algorithm that exploits these bounds and obtains epsilon-optimal solutions. Since variability in job durations is a major source of inefficiency in the use of OR time, we also present analytic results for the effect of changing variance of job-durations, while keeping their means fixed, on the total expected cost of running an appointments-based service system.

The article is organized as follows. In the next section, we discuss different performance criterion for an appointment scheduling system, and introduce the SLP model for determining individual appointment times. In Section 3, we present the algorithm and bounds on its performance. Analytical insights and numerical examples which illustrate the practical importance of the model are provided in Section 4. Finally, in Section 5, we summarize implications for policy makers and discuss future research directions.

## 2. Formulation and preliminary analysis

We consider a single server system at which customers arrive punctually at scheduled appointment times, and are served in the order of their arrival. Job sequence is thus assumed fixed. We use the following notation throughout the paper:

$n$ = number of jobs to be scheduled;
$\mathbf{x}$ = vector of job allowances for the first $n-1$ jobs;
$\mathbf{a}$ = vector of scheduled start times for $n$ jobs;
$\mathbf{Z}$ = vector of random job durations;
$\boldsymbol{\mu}$ = vector of mean durations for each job;
$\mathbf{W}$ = vector of customer waiting times for given $(\mathbf{x}, \mathbf{Z})$;
$\mathbf{S}$ = vector of facility/server idle times between consecutive jobs for given $(\mathbf{x}, \mathbf{Z})$, e.g., $S_2$ is the idle time between jobs 1 and 2;
$d$ = time allotted for a given sequence of jobs (session length);
$L$ = tardiness for a given sequence of jobs with respect to $d$ for a given $(\mathbf{x}, \mathbf{Z})$;
$G$ = earliness for a given sequence of jobs with respect to $d$ for a given $(\mathbf{x}, \mathbf{Z})$;
$\mathbf{c}^w$ = vector of cost coefficients for customer waiting;
$\mathbf{c}^s$ = vector of cost coefficients for facility idle time;
$c_\ell$ = cost coefficient for tardiness with respect to $d$.

Bold face and upper case notation indicates vectors and random variables respectively (to avoid confusion with lower case $L$ we use script $\ell$). The vector of job allowances $\mathbf{x} \in \Re^{n-1}$ (we need to specify only the job allowances for the first $n-1$ jobs), the vectors $\mathbf{a}$, $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{S}$, $\mathbf{c}^w$, $\mathbf{c}^s \in \Re^n$, and $d$, $L$, $G$, and $c_\ell$ are scalar quantities. The vector of random job durations, $\mathbf{Z}$, has support $\Xi \subseteq \Re^n$ and probability distribution $P$ on $\Re^n$ and it is assumed that $\mathbf{Z}$ has finite first moments. The scheduled start time for a given job is equal to the sum of the job allowances of its predecessors. We assume that the first job commences at time zero, i.e., $a_1 = 0$ and $a_i = \sum_{j=1}^{i-1} x_j$ for $i = 2, ..., n$. The vectors of cost coefficients, $\mathbf{c}^w$ and $\mathbf{c}^s$, can be different for each job. For example, if there are different customer classes then this can be modeled by having customer-dependent waiting time costs. Similarly, if the value of idle time of the server is dependent on the customer (e.g., if different customers require different resources) then this can be modeled by having different idle time costs.

Three commonly used metrics for the performance of an appointment system are customer waiting time, server idle time, and tardiness of a collection of jobs with respect to the alloted time for the session. Whereas, early arrival increases customer waiting, late arrival results in increased idle time of the facility and greater overtime costs. A manager of an appointments-based service system needs to balance efficient server utilization against the cost of customer waiting and overtime. The relative weights of the different metrics may vary from one system to another. For a given realization of job durations, $\mathbf{Z}$, and job allowances, $\mathbf{x}$, these metrics can be written as the following recursions:

$$W_i = (W_{i-1} + Z_{i-1} - x_{i-1})^+, \quad i = 2, \ldots, n, \tag{4}$$

$$S_i = (-W_{i-1} - Z_{i-1} + x_{i-1})^+, \quad i = 2, \ldots, n, \tag{5}$$

$$L = \left( W_n + Z_n + \sum_{i=1}^{n-1} x_i - d \right)^+, \tag{6}$$

$$G = \left( -W_n - Z_n - \sum_{i=1}^{n-1} x_i + d \right)^+. \tag{7}$$

Note that we use $(\cdot)^+$ to indicate $\max(\cdot, 0)$ and that waiting and idling, and tardiness and earliness, satisfy a parity relationship, i.e., $W_i \times S_i = 0$, $i = 2, \ldots, n$, and $L \times G = 0$.

Assuming linear costs for waiting, idling and tardiness, the ASP is to find a schedule of times for customer arrivals that minimize the following function:

$$\min_{\mathbf{x}} \left\{ \sum_{i=1}^{n} c_i^w E[W_i] + \sum_{i=1}^{n} c_i^s E[S_i] + c_\ell E[L] \right\}, \tag{8}$$

where the expectations are over $\mathbf{Z}$. Note that the objective function in Equation (8) is independent of the earliness, $G$. This is assumed because in typical applications there is no additional value or cost to having the server become free earlier than the allotted time, $d$. However, adding a linear reward or cost for earliness does not pose any difficulty for the methodology described in the remainder of the paper.

Using conventional non-linear optimization techniques for solving Equation (8) is problematic when there are many jobs because evaluation of the objective function, and its gradient, necessitates the computation of multidimensional integrals which typically have no known closed-form expressions. Our approach is to use a stochastic linear programming formulation to overcome this difficulty.

An assumption of the above formulation is that the session length, $d$, and the number of jobs $n$, are not decision variables. They are treated as being exogenously determined. However, determining the number of jobs to schedule during a given session length may be an important aspect of some ASPs. This parameter determines the frequency with which overtime will be required, or conversely facility idle time will be experienced, to complete all $n$ jobs. In fact, our model can be used to understand how total cost (given optimally scheduled appointments) depends on the number of jobs, $n$ for each fixed session length, $d$. In Section 4 we provide some numerical examples and calculate expected session lengths for certain problems. We also demonstrate how the total cost and optimal appointment times depend on $n$.

### 2.1. *Stochastic linear program formulation of the ASP*

In this section, we formulate the ASP as a stochastic linear program. First, Equation (8) is written as the following deterministic equivalent of a 2-SLP. (See Dempster (1986), Ermoliev and Wets (1988), Kall and Wallace (1994), and Birge and Louveaux (1997) and references therein for more details.)

$$\min \quad E\left\{ \sum_{i=2}^{n} c_i^w w_i + \sum_{i=2}^{n} c_i^s s_i + c_\ell \ell \right\}, \tag{9}$$

subject to

$$\begin{array}{llll} +w_2 & & -s_2 & = Z_1 - x_1, \\ -w_2 + w_3 & & -s_3 & = Z_2 - x_2, \\ \ddots \quad \ddots & & \ddots & \vdots \\ & -w_{n-1} + w_n \quad -s_n & & = Z_{n-1} - x_{n-1}, \\ & -w_n \quad +\ell & & -g = Z_n - d + \sum_{j=1}^{n-1} x_j, \end{array} \tag{10}$$

$$\mathbf{x} \geq 0, \; w_i \geq 0, \; s_i \geq 0 \, \forall i = 1, \ldots, n, \text{ and } \ell, g \geq 0.$$

The summation in Equation (9) begins with an index of two because $W_1$ and $S_1$ are set to zero, leaving $n-1$ decision variables, $(x_1, x_2, \ldots, x_{n-1})$, for an $n$-job problem. The first stage decision variables, $\mathbf{x}$, and second stage decision variables $(\mathbf{w}, \mathbf{s}, \ell, g)$ are written in lower case and the dependence of second stage decisions on the random variables, $\mathbf{Z}$, is implied. The constraints in Equation (10) enforce the piecewise linearity of waiting, idling and tardiness functions. The first $n-1$ constraints correspond to waiting (idling) time of customers (jobs) two through $n$, and the $n$th constraint corresponds to tardiness (earliness). (Note that the structure of the second stage problem is equivalent to that of a simple network flow problem.) We can rewrite the 2-SLP above more compactly as shown below:

$$\min_{\mathbf{x}} \{\mathcal{Q}(\mathbf{x})\}, \tag{11}$$

where $\mathcal{Q}(\mathbf{x}) = E[Q(\mathbf{x}, \mathbf{Z})]$ and

$$Q(\mathbf{x}, \mathbf{Z}) = \min_{\mathbf{y}} \{\mathbf{cy} \mid \mathcal{T}\mathbf{x} + \mathcal{W}\mathbf{y} = \mathbf{h}, \mathbf{y} \geq 0\},$$

$$\mathbf{c} = \begin{bmatrix} \mathbf{c}^w \\ c_\ell \\ \mathbf{c}^s \\ 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{w} \\ \ell \\ \mathbf{s} \\ g \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_n - d \end{bmatrix}. \tag{12}$$

In the stochastic programming literature $\mathcal{Q}(\mathbf{x})$ is called the *recourse function*, whereas $\mathcal{T}$ and $\mathcal{W}$ ($n \times n-1$ and $n \times 2n$ matrices, respectively) are called the *technology* and *recourse* matrices. We can write the recourse matrix for this problem as $\mathcal{W} = [\mathcal{W}' \mid -I]$ where $I$ is the identity matrix. The matrix $\mathcal{T}$ and the submatrix $\mathcal{W}'$ have the following form (empty spaces indicate zeroes):

$$\mathcal{T} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & 1 \\ -1 & \cdots & -1 \end{bmatrix}, \quad \mathcal{W}' = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}. \tag{13}$$

When viewed as a 2-SLP the two-job problem is a *simple recourse problem* which follows from the fact that the recourse matrix has the form $\mathcal{W} = [1 \mid -1]$. In general, problems where the recourse matrix is of the form $[\boldsymbol{I} \mid -\boldsymbol{I}]$ can be solved very efficiently due to the separability of the second stage constraints. In fact, for $n = 2$, the problem was recognized by Weiss (1990) to be equivalent to a newsvendor problem. The multi-job problem, Equation (10), differs from this because of a sub-diagonal of $-1$'s in $\mathcal{W}'$ resulting in a nonseparable second stage. However, the second stage is feasible for any $\mathbf{x} \in \Re^{n-1}$, i.e., $\text{pos}(\mathbf{W}) = \Re^{n-1}$, and the 2-SLP is said to have *complete recourse*.

The dual of the ASP can be written as

$$\max\{E[\boldsymbol{\pi}\mathbf{h}] \mid E[\boldsymbol{\pi}]\mathcal{T} \leq 0, \boldsymbol{\pi}\mathcal{W} \leq \mathbf{c}\}. \tag{14}$$

where, $\boldsymbol{\pi}$, the dual decision variables, are unrestricted in sign. Decomposition methods (e.g., Bender's decomposition, Dantzig-Wolfe decomposition), whether applied to the primal, Equation (11), or dual, Equation (14), rely on efficient solution of the *subproblems*, $Q(\mathbf{x}, \mathbf{Z})$. We focus on solution methods that take advantage of the structure of the primal problem. In our case the solution of subproblems, defined by Equation (12), is trivial because it requires only the evaluation of piecewise linear functions (waiting, idling and tardiness). Similarly, the dual of Equation (12),

$$\max\{(\mathbf{h} - \mathcal{T}\mathbf{x})\pi \mid \mathcal{W}^T \pi \leq \mathbf{c}\}, \tag{15}$$

can also be solved without actually solving a linear program. It has random cost coefficients but the constraints are fixed. Thus, the feasible region in Equation (15) is the same for all $\mathbf{x}$ and $\mathbf{Z}$ and it is easily shown to be compact. The optimal solution to Equation (15) for a given choice of $\mathbf{x}$ and realization of $\mathbf{Z}$ follows from the recursive structure of waiting, idling and tardiness functions. If there is nonzero waiting for jobs $i$ and $(i+1)$ then any increase in job $i$'s waiting time (i.e., an increase in the right-hand side (RHS) of constraint $i$ in Equation (12)) results in a subsequent increase in job $(i+1)$'s, and so on. On the other hand if a job has a nonzero idle time then it is decoupled from future ones because the next job starts at the scheduled time.

As a result we can write the solution to Equation (15) as the following backwards recursion

$$\pi_i^*(\mathbf{x},\mathbf{Z}) = \begin{cases} -c_{i+1}^s & w_i = 0, \\ c_{i+1}^w + \pi_{i+1}^*(\mathbf{x},\mathbf{Z}) & w_i > 0, \end{cases} \tag{16}$$

where

$$\pi_n^*(\mathbf{x},\mathbf{Z}) = \begin{cases} 0 & \ell = 0, \\ c_\ell & \ell > 0. \end{cases} \tag{17}$$

The dual solution, $\boldsymbol{\pi}$, is the subgradient of $Q(\mathbf{x},\mathbf{Z})$ with respect to the RHS of the equality constraints in Equation (12). If $\Xi$ is continuous then $\boldsymbol{\pi}$ is continuous everywhere except on a set of points of measure zero. For example, $w_i = 0$ implies that $w_{i-1} + Z_{i-1} - x_{i-1} \leq 0$. When strict equality is satisfied, i.e., $w_i = s_i = 0$, there is degeneracy in the second stage problem and $\pi_i^*(\mathbf{x},\mathbf{Z}) \in [-c_{i+1}^s, c_{i+1}^w + \pi_{i+1}^*(\mathbf{x},\mathbf{Z})]$.

Before closing this section, we present a property of the ASP which allows us to establish an equivalence between two common cost structures.

**Proposition 1.** *Expected idle time is equal to the difference between two sums: the sum of expected tardiness and the session length, and the sum of average job durations and expected earliness, i.e.,*

$$\sum_{i=1}^n E[S_i] = \left[E[L] + d] - [E[G] + \sum_{i=1}^n \mu_i\right]. \tag{18}$$

**Proof.** This is easily seen by adding the equality constraints in Equation (10) and taking the expectation of both sides. If the session length $d$ is zero then expected tardiness is equal to expected makespan and expected earliness is zero. ■

Having established Proposition 1, it is easy to see that the solution to the ASP is the same under the following two cases:

- idle time costs are identical, for example if $c_i^s = \alpha$, $\forall i$, and tardiness cost is zero, i.e, $c_\ell = 0$;
- idle time costs are zero for all jobs, i.e., $c_i^s = 0$ and $c_\ell = \alpha$.

In the two situations described above, the objective function differs only by a constant, i.e., the sum of the first moments of the job durations. Robinson *et al.* (1996) do not consider tardiness cost in their formulation, and use a similar argument to net out the sum of the first moments of job durations from the objective function.

## 3. Solution method and aggregation bounds

If $\Xi$ is finite, then Equation (10) is a linear program with the *block-diagonal* structure. Each block corresponds to the evaluation of (second stage) waiting, idling and tardiness costs that result from a given (first stage) set of job allowances, $\mathbf{x}$, and realization of surgery durations, $\mathbf{z}$. We refer to a set of realizations of job durations, $\{z_1, z_2, \ldots, z_n\}$, as a scenario, $\omega_k$, where $k = 1, \ldots, K$ indexes the $K$ scenarios. Because of the simple form of the second stage problem in the ASP, decomposition algorithms are very efficient at solving large problem instances. However, if independent finite service time distributions are specified, the number of scenarios grows exponentially with respect to $n$, or alternatively, if service time distributions are continuous then the number of scenarios is infinite. Approximate methods for such cases are typically based on partitioning $\Xi$ and solving the resulting large scale linear program. For example, statistical sampling is used to obtain a discrete set of scenarios that define an approximate problem, which is then solved using the L-shaped method (Dantzig and Glynn, 1990; Infanger, 1992). Alternatively, quasi-gradient methods (Ermoliev, 1988) use general purpose non-linear optimization techniques where the objective and its gradient are obtained from sampled data. These methods rely on statistical estimates of the objective and its gradient for solution. In this section we provide an efficient algorithm for computing approximate (near-optimal) job allowances and deterministic bounds on the accuracy loss due to the approximation.

### 3.1. *Sequential bounding algorithm*

This algorithm is an adaptation of the standard L-shaped algorithm (Van Slyke and Wets, 1969). It is based on successively partitioning the space of the random job durations in an effort to improve the solution obtained at each iteration, as measured by the gap between the lower and upper bounds on the optimal solution.

The basic idea of the L-shaped algorithm with Sequential Bounding (LSB) is to use constraints, based on *lower bounding functionals*, and upper bounds for $Q(\mathbf{x})$ to approximate the optimal solution, $\mathbf{x}^*$. Classic bounds such as the Jensen lower bound and the Edmundson-Madansky upper bound can be generalized over a partition of the support (Huang *et al.*, 1977). For example, the Jensen bound can be written as

$$E[f(\mathbf{x},\mathbf{Z})] \geq \sum_{k=1}^{\nu} p^k f(\mathbf{x},\mathbf{z}^k), \tag{19}$$

where $f(\cdot)$ must be a convex function of the components of random vector $\mathbf{Z}$, $k$ indexes cells of a partition of $\Xi$, $p^k$ is the probability on the cell, and $\mathbf{z}^k$ is the vector of expectations of the job durations on the cell. We denote the partition of $\Xi$ by $\mathcal{S}^{(\nu)}$ where $\mathcal{S}^{(\nu)} = \{S^k,\ k = 1, \ldots, \nu\}$ and for simplicity we assume a rectangular partition, i.e.,

$$S^k = \left[a_1^{(k)}, b_1^{(k)}\right] \times \left[a_2^{(k)}, b_2^{(k)}\right] \times \cdots \times \left[a_n^{(k)}, b_n^{(k)}\right], \tag{20}$$

where, recall that, $n$ is the number of jobs. If job durations are independent then the integrals for $p^k$ and $\mathbf{z}^k$ are separable. When the partition is refined in such a way that the approximate distribution $\{p^k,\ k = 1, \ldots, \nu\}$ converges in expectation to the true distribution, $P$, then the bound in Equation (19) converges to the expectation of the function as $\nu \to \infty$. Readers should note that when partitions are successively refined, parameters $S^k$, $a_i^{(k)}$, and $b_i^{(k)}$ are also functions of $\nu$.

We first outline the LSB and postpone the discussion of how to obtain upper bounds used in the algorithm to the next section. In our description the lower bounding functionals used to outer linearize the recourse function are hyperplanes that are obtained using Jensen bounds. As such the discrete approximation is

$$\min \quad \sum_{k=1}^{\nu} p^k \mathbf{c}\mathbf{y}^k \tag{21}$$

subject to

$$\begin{aligned} &\mathcal{T}\mathbf{x} + \mathcal{W}\mathbf{y}^k = \mathbf{h}^k, \quad k = 1, \dots, \nu,\\ &\mathbf{x} \geq 0,\ \mathbf{y}^k \geq 0, \quad k = 1, \dots, \nu. \end{aligned}$$

We refer to the optimal solution of Equation (21) as $(\mathbf{x}^{(\nu)}, (\mathbf{y}^{k*}, k = 1, \dots \nu))$ and its objective function at the optimum as $Q^{(\nu)}$, where $Q^{(\nu)} = \sum_{k=1}^{\nu} p^k Q(\mathbf{x}^{(\nu)}, \mathbf{z}^k)(Q^{(\nu)} = \theta^{(\nu)}$ in the description below). Similarly, $Q^{\mathrm{UB}}(\mathbf{x}^{(\nu)})$ is an upper bound on the current solution as well as the optimal recourse function available at iteration $\nu$. At each iteration the objective, $Q^{(\nu)}$, is a lower bound on the optimal solution. The basic form of the algorithm is

**L-Shaped Algorithm with Sequential Bounding:**

*Step 1.* Let $\nu$ index the iteration. Set $\nu = 0$.

*Step 2.* Set $\nu = \nu + 1$. Solve the discrete problem, (21) defined by partition $\mathcal{S}^\nu$ using the standard L-shaped method and let $(\mathbf{x}^{(\nu)}, \theta^{(\nu)})$ be the optimal solution.

*Step 3.* Evaluate $Q^{\mathrm{UB}}(\mathbf{x}^{(\nu)})$. If $Q^{\mathrm{UB}}(\mathbf{x}^{(\nu)}) - \theta^{(\nu)} \leq \textit{tolerance}$ then stop. Otherwise go to Step 4.

*Step 4.* Refine the current partition $\mathcal{S}^{(\nu)} \to \mathcal{S}^{(\nu+1)}$ and return to Step 2.

The stopping criterion in Step 3 is based on the absolute difference between the upper and lower bounds which bounds the accuracy loss due to solving the discrete approximation.

The above algorithm generates bounds on the gap between the solution obtained from solving the discrete version of the ASP at each step and the optimal solution. The gap depends on how the partition is refined in Step 4 at each iteration. If the partition is refined in such a way that the discrete distribution converges to the true distribution as $\nu \to \infty$ then $\mathbf{x}^{(\nu)} \to \mathbf{x}^*$ (Birge and Wets, 1986). We use a refinement method which is based on the work of Frauendorfer and Kall (1988). Details of the approach can be found in Appendix A.

### 3.2. *Aggregation bounds*

Standard methods for obtaining upper bounds of convex expectational functionals rely on determining an approximate discrete distribution composed of extreme points of $\Xi$. For example, the Edmundson-Madansky (E-M) upper bound is a weighted average of the function at the extreme points. It requires evaluation of $Q(\mathbf{x}, \mathbf{z})$ at each vertex of each cell in the partition. Since the number of vertices increases exponentially with the number of dimensions the computation time quickly becomes prohibitive as the number of random variables increases. Also, when $\Xi$ is not bounded the bound is not defined. A method based on solving a generalized moment problem, which is applicable to the case in which $\Xi$ is not compact, was developed by Birge and Wets (1995) and was subsequently applied to the problem of computing upper bounds on tardiness in a project network (Birge and Maddox, 1995). However, it is not well suited for use in an optimization algorithm.

We now show how efficient upper bounds can be obtained which are independent of service duration distributions by using a dual representation of the ASP. Using Equation (14) we can write the recourse function in the following form

$$Q(\mathbf{x}) = \int_{\Xi} \boldsymbol{\pi}(\mathbf{x}, \mathbf{z})(\mathbf{h} - \mathcal{T}\mathbf{x})P(\mathrm{d}\mathbf{z}). \tag{22}$$

Of course it is generally not possible to compute the integral in Equation (22) exactly. However, we can obtain an upper bound on $Q(\mathbf{x})$ by using an approach similar to the aggregation bounds described by Zipkin (1979) for deterministic linear programs, and subsequently extended by Birge (1985) to the case of multi-stage stochastic linear programs. As before we let $\mathcal{S}^\nu = \{S^k, k = 1, \dots, \nu\}$ denote a rectangular partition of the support, $\Xi$, where the $\mathbf{h}^k$ are expectations, and $p^k$ are probabilities on cells. We start with an application of aggregation bounds to the ASP.

**Proposition 2.**

$$Q^{(\nu)} \leq Q(\mathbf{x}^*) \leq Q^{(\nu)} + \epsilon_1(\nu), \tag{23}$$

*where* $Q^{(\nu)} = \sum_{k=1}^{\nu} p^k \mathbf{c}\mathbf{y}^{k*}$ *and*

$$\begin{aligned} \epsilon_1(\nu) = \sum_{k=1}^{\nu} \sum_{i=2}^{n} \int_{S^k} \Bigg( \Bigg( \sum_{j=i}^{n} c_j^w + c_\ell \Bigg) \big(h_i - h_i^k\big)^+ \\ + c_i^s \big(h_i^k - h_i\big)^+ \Bigg) P(\mathrm{d}\mathbf{z}). \end{aligned} \tag{24}$$

**Proof:** See Appendix B. ■

The bounds in Proposition 2 are on the optimal solution to the ASP and do not necessarily provide any information about the recourse function at $\mathbf{x}^{(\nu)}$, i.e., on the value of $Q(\mathbf{x}^{(\nu)})$. To bound the accuracy of the LSB algorithm we need bounds on the recourse function at the current iterate $\mathbf{x}^{(\nu)}$. We now derive a similar bound on $Q(\mathbf{x}^{(\nu)})$ which can be improved by applying it on a partition of $\Xi$.

**Proposition 3.**

$$Q^{(\nu)} \leq Q\big(\mathbf{x}^{(\nu)}\big) \leq Q^{(\nu)} + \epsilon_2(\nu), \tag{25}$$

*where*

$$\epsilon_2(\nu) = \sum_{k=1}^{\nu}\sum_{i=2}^{n}\int_{S^k}\left(\pi_i^{k,\mathrm{UB}}\left(h_i - h_i^k\right)^+ - \pi_i^{k,\mathrm{LB}}\left(h_i^k - h_i\right)^+\right)P(d\mathbf{z}) \tag{26}$$

*and* $\pi_i^{k,\mathrm{UB}}(\mathbf{x}^{(\nu)})$, $\pi_i^{k,\mathrm{LB}}(\mathbf{x}^{(\nu)})$ *are upper and lower bounds given* $\mathbf{x}^{(\nu)}$ *and* $\mathbf{Z} \in S^k$.

**Proof.** See Appendix B ■

The lower bounds in Propositions 2 and 3 are identical. That the upper bound in Proposition 3 is tighter than the bound in Proposition 2 follows from the fact that

$$\pi_i^{k,\mathrm{UB}} \leq \pi_i^{\mathrm{UB}} \quad \text{and}\, \pi_i^{k,\mathrm{LB}} \geq \pi_i^{\mathrm{LB}}, \quad k = 1, \ldots, \nu. \tag{27}$$

The upper bound in Proposition 3 is an upper bound on $Q(\mathbf{x}^{\nu})$ and thus also on $Q(\mathbf{x}^*)$. Thus, from Propositions 2 and 3 the solution to Equation (21) provides upper and lower bounds on the accuracy loss due to the approximation, i.e., $Q(\mathbf{x}^{(\nu)}) - Q(\mathbf{x}^*) \leq \epsilon_2(\nu)$. It remains to be shown how upper and lower bounds for the dual solution on a partition of $\Xi$ can be obtained. The following simple dynamic programming procedure can be used.

### 3.2.1. *Bounding the dual multipliers on a cell*

For each job we must determine whether nonzero waiting or idling is possible and whether nonzero tardiness or earliness is possible for $\mathbf{x}^{(\nu)}$ and $\mathbf{Z} \in S^k$. From Equations (4) and (6) waiting times and tardiness are nondecreasing in $\mathbf{Z}$ and hence we can bound them by evaluating them at the extreme points of the cell as follows:

$$W_i^{k,\mathrm{UB}} = \left(W_{i-1}^{k,\mathrm{UB}} + b_{i-1}^k - x_{i-1}\right)^+, \quad i = 2, \ldots, n, \tag{28}$$

$$W_i^{k,\mathrm{LB}} = \left(W_{i-1}^{k,\mathrm{LB}} + a_{i-1}^k - x_{i-1}\right)^+, \quad i = 2, \ldots, n, \tag{29}$$

and

$$L^{k,\mathrm{UB}} = \left(W_n^{k,\mathrm{UB}} + b_n^k + \sum_{i=1}^{n} x_i - d\right)^+, \quad i = 2, \ldots, n, \tag{30}$$

$$L^{k,\mathrm{LB}} = \left(W_n^{k,\mathrm{LB}} + a_n^k + \sum_{i=1}^{n} x_i - d\right)^+, \quad i = 2, \ldots, n. \tag{31}$$

Note that when $\Xi$ is not compact the upper bounds in Equations (28) and (30) may be infinite, however, we are concerned only with knowing whether they are greater than zero. A nonzero upper bound on waiting time indicates that zero is the tightest lower bound on the corresponding idle time and *vice versa* (due to the parity relationship). The same is true for tardiness and earliness. The upper and lower bounds on the dual solution can therefore be obtained using the following backward recursions

$$\pi_i^{k,\mathrm{UB}}(\mathbf{x}) = \begin{cases} -c_{i+1}^s & \text{if } W_{i+1}^{k,\mathrm{UB}} = 0, \\ \max\left\{-c_{i+1}^s, c_{i+1}^w + \pi_{i+1}^{k,\mathrm{UB}}(\mathbf{x})\right\} & \text{if } W_{i+1}^{k,\mathrm{LB}} = 0, \\ & \quad \text{and } W_{i+1}^{k,\mathrm{UB}} > 0, \\ c_{i+1}^w + \pi_{i+1}^{k,\mathrm{UB}}(\mathbf{x}) & \text{if } W_{i+1}^{k,\mathrm{LB}} > 0, \end{cases} \tag{32}$$

and

$$\pi_i^{k,\mathrm{LB}}(\mathbf{x}) = \begin{cases} -c_{i+1}^s & \text{if } W_{i+1}^{k,\mathrm{UB}} = 0, \\ \min\left\{-c_{i+1}^s, c_{i+1}^w + \pi_{i+1}^{k,\mathrm{LB}}(\mathbf{x})\right\} & \text{if } W_{i+1}^{k,\mathrm{LB}} = 0, \\ & \quad \text{and } W_{i}^{k,\mathrm{UB}} > 0, \\ c_{i+1}^w + \pi_{i+1}^{k,\mathrm{LB}}(\mathbf{x}) & \text{if } W_{i+1}^{k,\mathrm{LB}} > 0, \end{cases} \tag{33}$$

for $i = 1, \ldots, n-1$, and the upper and lower bounds for $\pi_n$ are

$$\pi_n^{k,\mathrm{UB}}(\mathbf{x}) = \begin{cases} 0 & \text{if } L^{k,\mathrm{UB}} = 0, \\ c_\ell & \text{if } L^{k,\mathrm{UB}} > 0, \end{cases} \tag{34}$$

and

$$\pi_n^{k,\mathrm{LB}}(\mathbf{x}) = \begin{cases} 0 & \text{if } L^{k,\mathrm{LB}} = 0, \\ c_\ell & \text{if } L^{k,\mathrm{LB}} > 0. \end{cases} \tag{35}$$

The bounds in Proposition 3 can be viewed as having a penalty term on the discrepancy between the approximate discrete problem and the continuous problem. This penalty for a given cell is expressed as $E_{S^k}[\pi_i^{k,\mathrm{UB}}(h_i - h_i^k)^+ - \pi_i^{k,\mathrm{LB}}(h_i^k - h_i)^+]$ above. When it is used in LSB, refinement of the partition results in a reduction of this measure of discrepancy on the chosen cell at $\mathbf{x}^{(\nu)}$.

## 4. Insights and examples

It can be shown for the case of $c_\ell = 0$ that some simple transformations of the random job durations in the ASP result in linear transformation of $\mathbf{x}^*$ and $Q(\mathbf{x}^*)$. For instance, consider changing each job duration by a constant factor $\mathbf{Z} \mapsto \mathbf{Z} + \mathbf{b}$ where $\mathbf{b} \in \Re^n$. It is easy to show that the effect on the optimal solution of the ASP in this case is $\mathbf{x}^* \mapsto \mathbf{x}^* + \mathbf{b}$ and $Q(\mathbf{x}^*)$ is unchanged. When $c_\ell = 0$ the following can also be shown.

**Proposition 4.** *The effect of the transformation* $\mathbf{Z} \mapsto a\mathbf{Z} + \mathbf{b}$, *where* $a \in \Re$ *and* $\mathbf{b} \in \Re^n$, *on the optimal solution, is* $Q(\mathbf{x}^*) \mapsto aQ(\mathbf{x}^*)$ and $\mathbf{x}^* \mapsto a\mathbf{x}^* + \mathbf{b}$.

**Proof.** See Appendix B. ■

**Table 1.** Results for seven jobs with $U(0, 2)$ job durations after 50 iterations

| $(c^s, c^w)$ | $(9, 1)$ | $(8, 2)$ | $(7, 3)$ | $(6, 4)$ | $(5, 5)$ | $(4, 6)$ | $(3, 7)$ | $(2, 8)$ | $(1, 9)$ |
|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 0.360 | 0.624 | 0.838 | 1.035 | 1.165 | 1.313 | 1.461 | 1.631 | 1.807 |
| $x_2$ | 0.876 | 1.093 | 1.162 | 1.259 | 1.349 | 1.437 | 1.549 | 1.669 | 1.817 |
| $x_3$ | 0.969 | 1.070 | 1.201 | 1.255 | 1.361 | 1.446 | 1.552 | 1.670 | 1.819 |
| $x_4$ | 0.952 | 1.065 | 1.174 | 1.255 | 1.351 | 1.443 | 1.543 | 1.670 | 1.819 |
| $x_5$ | 0.911 | 1.060 | 1.125 | 1.228 | 1.300 | 1.426 | 1.528 | 1.665 | 1.821 |
| $x_6$ | 0.784 | 0.871 | 0.970 | 1.087 | 1.203 | 1.344 | 1.479 | 1.639 | 1.813 |
| $Q(\mathbf{x})^{\mathrm{LB}}$ | 8.963 | 13.423 | 15.726 | 16.656 | 16.400 | 15.139 | 12.906 | 9.674 | 5.394 |
| $Q(\mathbf{x})^{\mathrm{UB}}$ | 9.985 | 14.619 | 17.326 | 17.789 | 17.259 | 15.812 | 13.345 | 9.896 | 5.501 |
| $\bar{Q}(\mathbf{x})$ | 9.077 | 13.498 | 15.855 | 16.858 | 16.551 | 15.278 | 12.983 | 9.768 | 5.412 |
| $E[SL]$ | 7.310 | 7.714 | 8.111 | 8.548 | 9.032 | 9.586 | 10.210 | 10.988 | 11.903 |

From Proposition 4 it follows that under the mean preserving transformation, $\mathbf{Z} \mapsto a\mathbf{Z} - (a-1)\boldsymbol{\mu}$, $Q(\mathbf{x}^*)$ is increasing linearly with respect to the standard deviation of job durations. Also, if the solution for a particular $a$ and $\mathbf{b}$ is known then the solution for any other $a$ and $\mathbf{b}$ can be obtained by a simple transformation. Note also that several distributions have the property that they can be completely described through the above linear transformations (e.g., normal, uniform, exponential).

### 4.1. *Experimental design*

The numerical results presented below fall into two categories: (i) numerical experiments that give general insights and illustrate the quality of the aggregation bounds; (ii) examples that reinforce the fact that OR scheduling problem is economically important. The partitioning method described in Section 3 was used to compute the solutions and deterministic bounds. In all cases 50 iterations were carried out and 500 additional cells were added at each iteration. Solution times for the largest examples are typically less than 10 minutes on a modest workstation (Sun Ultra 10 with 128 MB Ram). The master problem was solved using CPLEX 5.0 (Anon, 1998) at each iteration and the majority of computation time is spent in updating the partition. To simplify comparison of results the job distributions are assumed i.i.d. in each case. For each example we also compute a statistical estimate of the recourse function at the approximate solution, $\bar{Q}(\mathbf{x}^{(\nu)})$. (Since the ASP is a convex minimization problem these are statistical upper bounds on $Q(\mathbf{x}^*)$.) The statistical estimates were obtained using a sample size of $10^4$, which is consistent with results of the simulation study by Ho and Lau (1992) that indicate an accuracy of ±1% at the 95% confidence level for $n \leq 30$.

### 4.2. *Accuracy of LSB and parametric variations*

Tables 1 and 2 contain results for problems with a variety of different cost structures. In each example we assume waiting and idling cost coefficients are the same for each job ($c_i^w = c_j^w$ and $c_i^s = c_j^s$, $\forall i, j$). In Table 1 there are no overtime costs, i.e., $c_\ell = 0$, and the relative costs of waiting and idling are varied. In Table 2 results for nonzero overtime costs are reported when the session length is assumed to be equal to the sum of the mean job durations ($d = 7.0$). The uniform distribution was chosen as a test case for these experiments. For distributions with bounded support it represents an extreme condition with respect to the application of a partitioning method, in the sense that no particular region of the support $\Xi$ has a higher probability mass. Later in this section we provide results from numerical experiments with distributions with unbounded support (e.g., normal and gamma). Note that Proposition 4 shows how the results in Table 1 can be transformed to correspond to any mean and variance of the job durations.

**Table 2.** Results for seven jobs with $U(0, 2)$ job durations after 50 iterations

| $(c_\ell, c^s, c^w)$ | $(7, 7, 3)$ | $(7, 5, 5)$ | $(7, 3, 7)$ | $(5, 7, 3)$ | $(5, 5, 5)$ | $(5, 3, 7)$ | $(3, 7, 3)$ | $(3, 5, 5)$ | $(3, 3, 7)$ |
|---|---|---|---|---|---|---|---|---|---|
| $x_1$ | 0.606 | 0.831 | 1.063 | 0.645 | 0.875 | 1.136 | 0.719 | 0.997 | 1.250 |
| $x_2$ | 1.085 | 1.175 | 1.267 | 1.113 | 1.217 | 1.337 | 1.125 | 1.250 | 1.375 |
| $x_3$ | 1.080 | 1.197 | 1.264 | 1.106 | 1.236 | 1.308 | 1.120 | 1.250 | 1.375 |
| $x_4$ | 1.091 | 1.196 | 1.266 | 1.125 | 1.216 | 1.321 | 1.131 | 1.251 | 1.383 |
| $x_5$ | 1.067 | 1.104 | 1.208 | 1.049 | 1.137 | 1.252 | 1.077 | 1.194 | 1.351 |
| $x_6$ | 0.936 | 0.997 | 1.164 | 0.956 | 1.009 | 1.203 | 0.935 | 1.069 | 1.242 |
| $Q(\mathbf{x})^{\mathrm{LB}}$ | 22.167 | 26.546 | 28.236 | 20.486 | 24.225 | 24.829 | 18.716 | 21.507 | 20.812 |
| $Q(\mathbf{x})^{\mathrm{UB}}$ | 25.045 | 29.114 | 30.699 | 22.754 | 26.457 | 26.438 | 20.547 | 23.282 | 21.891 |
| $\bar{Q}(\mathbf{x})$ | 22.743 | 27.047 | 28.888 | 20.801 | 24.644 | 25.258 | 18.928 | 21.853 | 20.921 |
| $E[SL]$ | 7.732 | 8.124 | 8.616 | 7.805 | 8.266 | 8.890 | 7.881 | 8.491 | 9.230 |

**Table 3.** Comparison of job allowances for different problem sizes for three, five, and seven jobs with $U(0, 2)$

| | *n* = *3* | | | *n* = *5* | | | *n* = *7* | | |
|---|---|---|---|---|---|---|---|---|---|
| $(c^s, c^w)$ | *(9, 1)* | *(5, 5)* | *(1, 9)* | *(9, 1)* | *(5, 5)* | *(1, 9)* | *(9, 1)* | *(5, 5)* | *(1, 9)* |
| $x_1$ | 0.262 | 1.094 | 1.809 | 0.311 | 1.160 | 1.808 | 0.335 | 1.168 | 1.808 |
| $x_2$ | 0.674 | 1.203 | 1.811 | 0.849 | 1.336 | 1.818 | 0.882 | 1.349 | 1.818 |
| $x_3$ | | | | 0.881 | 1.313 | 1.818 | 0.985 | 1.358 | 1.818 |
| $x_4$ | | | | 0.769 | 1.210 | 1.809 | 0.955 | 1.345 | 1.823 |
| $x_5$ | | | | | | | 0.914 | 1.310 | 1.818 |
| $x_6$ | | | | | | | 0.784 | 1.219 | 1.812 |
| $\triangle x/\mu$ | 0.412 | 0.109 | 0.002 | 0.570 | 0.176 | 0.01 | 0.650 | 0.190 | 0.015 |

The results in Tables 1 and 2 illustrate a common behavior of solutions to the ASP for i.i.d. job durations. The solution exhibits a dome shape, i.e., job allowances are initially increasing and then decreasing. Numerical experiments for other types of distributions confirm this is a typical property of solutions with i.i.d. distributions and equal waiting and idling costs for all jobs. The dome shape is most pronounced when the ratio of idling to waiting cost coefficients is high. Job allowances tend to be more uniform when the opposite is true. If, however, the waiting and idling cost coefficients are not equal for all jobs, and/or job duration distributions are not i.i.d., then the solution to the ASP does not share the dome-shape property.

The model assumes a fixed session length $d$ and number of surgeries $n$. For the examples in Tables 1 and 2 the session length equals the sum of the mean job durations ($d = 7$). This is a commonly used method for determining how many surgeries to schedule during any session. The examples show that the optimal solution can result in a expected session length, $E[SL]$, which exceeds $d$. Increasing the cost of tardiness, $c_\ell$, in Table 2 significantly reduces the expected session length, however it is still nonzero in all cases. In fact, for all examples in which the session length is equal to the sum of the mean job durations, the expected tardiness is zero only if $a_i = 0, i = 1, \dots, n$. These examples help to explain why in applications such as OR scheduling, when mean durations are used to set session lengths it is necessary to allow for some buffer time at the end of the session to avoid tardiness.

For the sake of brevity, results for all the numerical experiments performed are not presented. In all cases, it was found that job allowances increase for all jobs as $n$ increases when $c_\ell = 0$. However, when nonzero overtime costs are included, changes in job allowances with respect to problem size are not necessarily monotonic. Table 3 illustrates the effect for the case of zero overtime costs for different problem sizes. In the table $\triangle x/\mu$ is the ratio of the difference between the maximum and minimum job allowance to the mean job duration. It is useful as a representative measure of the non-uniformity of the job allowances. The results show significant increases in job allowances as $n$ increases when waiting cost coefficients are low compared to idling cost coefficients, but, relatively small changes when the opposite is true. As idling cost coefficients decrease $\triangle x/\mu$ decreases. Also, the results indicate that the change in job allowance for a particular job as $n$ increases is increasing at a decreasing rate.

Figure 1 shows the dependence of the aggregation bounds, and the statistical upper bound, on problem size. The results indicate that the actual performance, based on the statistical estimate, is typically much better than the worst-case bound as the problem size grows. Numerical experiments indicate that this linear dependence is not sensitive to relative changes in the waiting and idling cost coefficients. Also, from Proposition 4, the slope is proportional to the standard deviation of the job durations for the case of i.i.d. job durations.

The results in Table 4 illustrate the importance of solving the ASP for different values of the cost coefficients. For each case $\bar{Q}(\mathbf{x})$ and $\bar{Q}(\boldsymbol{\mu})$ are estimated by sampling. Together these are used to approximate the *Value of the Stochastic*

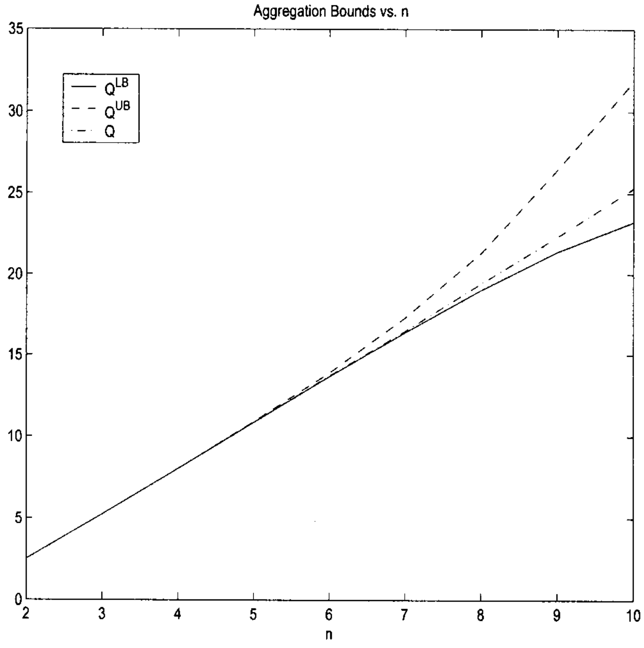

**Fig. 1.** Dependence of the upper and lower bounds on the problem size for $c_i^w = 5$, $c_i^s = 5$, $\forall i$, $c_\ell = 0$, and $U(0, 2)$ job durations.

**Table 4.** Results for seven jobs with $N(5.0, 1.0)$ job durations

| $(c^w, c^s)$ | $(9, 1)$ | $(8, 2)$ | $(7, 3)$ | $(6, 4)$ | $(5, 5)$ | $(4, 6)$ | $(3, 7)$ | $(2, 8)$ | $(1, 9)$ |
|---|---|---|---|---|---|---|---|---|---|
| $Q(\mathbf{x})^{\mathrm{LB}}$ | 10.602 | 17.251 | 22.097 | 25.369 | 27.129 | 27.496 | 25.970 | 22.215 | 15.030 |
| $Q(\mathbf{x})^{\mathrm{UB}}$ | 11.540 | 19.324 | 24.804 | 28.453 | 30.661 | 30.892 | 29.340 | 25.553 | 18.365 |
| $\bar{Q}(\mathbf{x})$ | 10.736 | 17.516 | 22.484 | 25.816 | 27.842 | 27.993 | 26.577 | 22.800 | 15.499 |
| $\bar{Q}(\boldsymbol{\mu})$ | 54.509 | 49.321 | 45.041 | 40.888 | 36.959 | 32.031 | 27.727 | 23.363 | 18.953 |
| Relative VSS% | 407.72 | 181.58 | 100.32 | 58.38 | 37.74 | 14.42 | 4.32 | 2.46 | 22.28 |

*Solution* (VSS), the difference between the optimal solution and the mean value solution, which can be interpreted as a measure of the importance of solving the ASP. In the table the *relative* VSS (difference between the LSB algorithm solution and the mean value solution, shown as a percentage of the LSB solution) is reported. From the table it is clear that relative VSS is high when waiting cost coefficients are high or similar to the idling cost coefficients. As waiting time cost coefficients decrease with respect to idle time cost coefficients the VSS is initially decreasing and then increasing again.

Many heuristic rules suggested in the literature utilize only the mean and variance of the job durations. Table 5 contrasts different distribution types and shows that optimal schedules may depend on higher moments. The results are for i.i.d. gamma, uniform, and normally distributed job durations with a mean of 3.0 and a variance of 0.5 for each distribution. Both the uniform and normal distribution are symmetric, whereas the Gamma distribution is skewed (skewness of $\Gamma(6, 18)$ is 0.55). Comparison of the solutions for the different distribution types indicates that dependence of $Q(\mathbf{x})$ on the distribution type is most pronounced when waiting cost coefficients are high relative to idle cost coefficients. For instance, the relative difference between $\Gamma(6, 18)$ and $U(1.775, 4.225)$ for $(c^w, c^s) = (9, 1)$ is approximately 25%. For the other cost structures in Table 3, $(c^w, c^s) = (5, 5)$ and $(1, 9)$, the solutions vary somewhat with respect to changes in the distribution type but the relative changes in $Q(\mathbf{x})$ are typically less than 5%. Another observation that can be made from Table 5 is that the relative gap between the upper and lower bound is lower for the uniform distribution, 0.31%, than for the normal and gamma distributions, 2.05 and 1.21% respectively. This is evidence of an increase in the relative gap for distributions with long tails.

### 4.3. *Allocating block time for deferrable surgeries*

In this example we illustrate the benefits of solving the ASP in the context of optimizing the allocation of block times for sequences of deferrable surgeries. We assume that there are five blocks scheduled at a given OR in an 8 hour day and that the session lengths are i.i.d. with distribution $\Gamma(1.0, 1.5)$ (gamma distributed with $\lambda = 1.0$, and $a = 1.5$). A common heuristic used by OR managers is to allocate block times such that the total time allocated is equal to the mean of the sum of the surgery durations in the session and to reserve time at the end of the day to avoid overtime costs. For example, setting block sizes equal to the mean in this example results in 1.5 hours for each block and a total of 7.5 of the 8 hours available. Thus the surgical team, patients and other resources for the first scheduled block would be coordinated to arrive at the beginning of the day. The start time for the second block would be scheduled 1.5 hours later, and subsequent blocks are scheduled in a similar fashion. Typically there are no direct costs for OR idling, rather, the goal is to trade-off the number of surgeries scheduled, costs of idling surgical teams and material resources, and overtime costs. In this example we assume that an equal weight is assigned, i.e., $c_i^w = 1\ \forall i$ and $c_\ell = 1$. The stopping criteria for the LSB algorithm was set at a tolerance of 0.01. Under this cost structure the optimal block sizes are

$$x_1^* = 1.563,\ x_2^* = 2.065,\ x_3^* = 2.022,\ x_4^* = 1.703, \quad (36)$$

**Table 5.** Comparison of distributions with $\mu = 3.0$ and $\sigma^2 = 0.5$

| | $\Gamma(6, 18)$ | | | $U(1.775, 4.225)$ | | | $N(3, 0.5)$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $(c^w, c^s)$ | $(9, 1)$ | $(5, 5)$ | $(1, 9)$ | $(9, 1)$ | $(5, 5)$ | $(1, 9)$ | $(9, 1)$ | $(5, 5)$ | $(1, 9)$ |
| $x_1$ | 3.990 | 3.120 | 2.285 | 3.992 | 3.196 | 2.155 | 3.952 | 3.167 | 2.268 |
| $x_2$ | 4.056 | 3.352 | 2.784 | 4.002 | 3.412 | 2.818 | 3.995 | 3.369 | 2.818 |
| $x_3$ | 4.055 | 3.340 | 2.830 | 4.002 | 3.382 | 2.853 | 3.989 | 3.348 | 2.844 |
| $x_4$ | 3.994 | 3.199 | 2.726 | 3.991 | 3.259 | 2.718 | 3.954 | 3.227 | 2.752 |
| $Q(\mathbf{x})^{\mathrm{LB}}$ | 5.577 | 13.059 | 6.341 | 4.421 | 13.286 | 6.736 | 5.056 | 12.734 | 6.657 |
| $Q(\mathbf{x})^{\mathrm{UB}}$ | 5.779 | 13.225 | 6.422 | 4.422 | 13.323 | 6.779 | 5.073 | 12.830 | 6.736 |

and $Q(\mathbf{x}^*) = 4.14$. Conversely the mean value approach ($\bar{x}_i = 1.5\ \forall i$) yields $\bar{Q}(\boldsymbol{\mu}) = 4.84$. Thus, there is approximately a 14.5% reduction in total overtime and surgical team idling associated with using an optimal schedule. The improvement can alternatively be viewed as allowing an increase in the effective capacity of the OR. For example, assume that each session has three scheduled surgeries distributed as $\Gamma(1.0, 0.5)$ (thus the sum of durations is distributed as $\Gamma(1.0, 1.5)$). Increasing the number of surgeries in the last session by one-third corresponds to a session duration distributed as $\Gamma(1.0, 2.0)$. In this case solving the ASP yields the following block sizes

$$x_1^* = 1.550,\ x_2^* = 2.051,\ x_3^* = 2.013,\ x_4^* = 1.701, \quad (37)$$

and $Q(\mathbf{x}^*) = 4.65$. The result is an increase of about 6.7% in the number of surgeries scheduled while still maintaining a reduction in cost compared to the heuristic approach. Due to the high costs of delivering surgical care at hospitals discussed in Section 1 such improvements can represent significant savings.

## 5. Summary and conclusions

This article provides a new formulation of an important and common problem. The stochastic linear programming model allows considerable flexibility in modeling different types of cost considerations. For example, the formulation and algorithm can easily be generalized to accommodate piecewise linear cost structures, not only for overtime costs, but also for waiting and idling costs. It can also be extended to include application specific aspects of systems such as customer "no shows" through adjustments to the modeling of job duration distributions. Furthermore, it is a natural starting point to more complex project-scheduling-based applications in which the second stage problem is a general project network.

The LSB algorithm bounds the optimal solution with high accuracy in a small number (50 iterations with 500 new cells added in each iteration) of iterations for problem sizes of interest in OR block-booking applications. Although the generation of simultaneously tight upper and lower bounds may be computationally intensive, numerical results indicate that solutions are much closer to the optimum than the worst-case upper bound would indicate. Thus, the algorithm is also a good approach for quickly approximating solutions to large-scale instances of the problem. Also, the algorithm is generalizable to any two-stage stochastic linear program for which upper bounds on dual multipliers can be computed on a partition of the space of random variables.

We conclude by summarizing some new and important insights for policy makers:

- Numerical experiments indicate that VSS is in general high, particularly when waiting costs are high relative to idling/overtime costs, indicating the need for solving the ASP. However, for certain ranges of cost parameters, choosing job allowances equal to mean job durations is near optimal (when unit cost of waiting are about 10 to 50% of the unit cost of idling).
- The cost of operating an appointments-based service system are: (i) linearly increasing in the standard deviation of job durations; and (ii) appear to be approximately linearly increasing in the number of jobs scheduled.
- For the case of i.i.d. location scale distributions, all instances of the problem can be solved effectively by solving any one instance of the problem due to the transformation (Proposition 4).
- Optimal scheduling of jobs can both increase the effective utilization of expensive resources as well as lower overall costs associated with the solution.
- Optimal job allowances for a particular position in the job sequence are increasing at a decreasing rate with respect to the number of jobs scheduled.
- Optimal job allowances exhibit a pronounced dome shaped structure when idling costs are high relative to waiting costs, and are roughly uniform when idling costs are low relative to waiting costs.
- Initial experiments indicate that the first two moments are sufficient for computing appointment times when idling costs are high relative to waiting costs, however, higher moments may be necessary, when idling costs are low relative to waiting costs.

## Acknowledgements

The authors are grateful to two anonymous referees for their helpful suggestions to improve the manuscript. This research was supported in part by the Social Sciences and Humanities Research Council of Canada through a research grant to DG. The work was done, in part, when both authors were at McMaster University, Canada.

## References

Anon (1998) *CPLEX Installation and Use-Notes*, ILOG Inc., CPLEX Division, Incline Village, NV.

Bailey, N. (1952) A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society*, **A14**, 185–189.

Bailey, N. (1954) Queuing for medical care. *Applied Statistics*, **3**, 137–145.

Birge, J.R. (1985) Aggregation bounds in stochastic linear programming. *Mathematical Programming*, **31**, 25–41.

Birge, J.R. and Louveaux, F. (1997) *Introduction to Stochastic Programming*, Springer, New York, NY.

Birge, J.R. and Maddox, M.J. (1995) Bounds on expected project tardiness. *Operations Research*, **5**, 838–850.

Birge, J.R. and Wets, R.J.-B. (1986) Designing approximation schemes for stochastic optimization problems. *Mathematical Programming Study*, **27**, 54–102.

Brahimi, M. and Worthington, D.J. (1991) Queuing models for outpatient appointment systems – a case study. *Journal of the Operational Research Society*, **42**, 733–746.

Charnetski, J. (1984) Scheduling operating room surgical procedure with early and late completion penalty costs. *Journal of Operations Management*, **5**, 91–102.

Dantzig, G.B. and Glynn, P. (1990) Parallel processors for planning under uncertainty. *Annals of Operations Research*, **22**, 1–21.

Dempster, M.A.H. (1986) *Stochastic Programming*, Academic Press, New York, NY.

Dexter, F. (1999) Design of appointment systems to minimize patient waiting times: a review of computer simulation and patient survey studies. *Anesthesia and Analgesia*, **89**, 925–931.

Ermoliev, Y. (1988) Stochastic quasigradient methods, in *Numerical Techniques for Stochastic Optimization*, Ermoliev, Y. and Wets, R.J.-B. (eds.), Springer Verlag, Berlin, 141–185.

Ermoliev, Y. and Wets, R.J.-B. (1988) Stochastic programming, an introduction, in *Numerical Techniques for Stochastic Optimization*, Ermoliev, Y. and Wets, R.J.-B. (eds.) Springer Verlag, Berlin, 1–32.

Forst, F.G. (1993) Stochastic scheduling on one machine with earliness and tardiness penalties. *Probability in Engineering and Informational Sciences*, **7**, 291–300.

Frauendorfer, K. and Kall, P. (1988) A solution method for SLP recourse problems with arbitrary multivariate distributions – the independent case. *Problems in Control and Information Theory*, **17**, 177–205.

Goldman, J., Knappenberger, H.A. and Shearson, W.J. (1970) A study of the variability of surgical estimates. *Hospital Management*, **110**, 46–46D.

Ho, C.-J. and Lau, H.-S. (1992) Minimizing total cost in scheduling outpatient appointments. *Management Science*, **38**, 750–764.

Huang, C.C., Ziemba, W.T. and Ben-Tal, A. (1977). Bounds on the expectation of a convex function of a random variable: with applications to stochastic programming. *Operations Research*, **25**, 315–325.

Infanger, G. (1992) Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. *Annals of Operations Research*, **39**, 69–95.

Jansson, B. (1966) Choosing a good appointment system—a study of queues of the type (D,M,1). *Operations Research*, **14**, 292–312.

Kall, P. and Wallace, S.W. (1994) *Stochastic Programming*, Wiley, New York, NY.

Macario, A., Terry, V.S., Dunn, B. and McDonald, T. (1995) Where are the costs in perioperative care? *Anesthesiology*, **83**, 1138–1144.

Mercer, A. (1960) A queueing problem in which the arrival times of the customers are scheduled. *Journal of the Royal Statistical Society, Series B*, **22**, 108–113.

Mercer, A. (1973) Queues with scheduled arrivals: a correction simplification and extension. *Journal of the Royal Statistical Society, Series B*, **35**, 104–116.

Pierskala, W.P. and Brailer, D.J. (1994) Applications of operations research in health care delivery. in *Operations Research and the Public Sector*, Pollock, S.M. Rothkopf, M.H. and Barnett, A. (eds.), North-Holland, Amsterdam, 469–505.

Redelmeier, D.A. and Fuchs, V.R. (1993) Hospital expenditures in the United States and Canada. *New England Journal of Medicine*, **11**, 772–778.

Robinson, L. W. and Chen, R. (2000) Scheduling doctors'; appointments: optimal and empirically-based heuristic policies, Working Paper, Cornell University, Ithaca, NY 14850, USA.

Robinson, L.W., Gerchak, Y. and Gupta, D. (1996) Appointment times which minimize waiting and facility idleness. Working Paper. McMaster University, Hamilton, Ontario, Canada.

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

Soriano, A. (1966) Comparison of two scheduling systems. *Operations Research*, **14**, 388–397.

Strum, D.P., Vargas, L.G. and Jerrold, M.H. (1999) Surgical subspecialty block utilization and capacity planning. *Anesthesiology*, **4**, 1176–1185.

Vanden Bosch, P.M. and Dietz, D.C. (2000) Minimizing expected waiting time in a medical appointment system. *IIE Transactions*, **32**, 841–848.

Van Slyke, R.M. and Wets, R.J.-B. (1969) L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal of Applied Mathematics*, **17**, 638–663.

Wang, P.P. (1993) Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics*, **40**, 345–360.

Weiss, E.N. (1990) Models for determining estimated start times and case orderings in hospital operating rooms. *IIE Transactions*, **22**, 143–150.

Welch, J. (1964) Appointment systems in hospital outpatient departments. *Operations Research Quarterly*, **15**, 224–237.

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

White, M. and Pike, M. (1964) Appointment systems in outpatient clinics and the effect of patients' unpunctuality. *Medical Care*, **2**, 133–145.

Zipkin, P.H. (1979) Bounds for row-aggregation in linear programming. *Operations Research*, **28**, 903–916.

## Appendices

### Appendix A: The partition refinement technique

Refinement of a partition is typically described as involving three decisions. The first is the choice of a cell to split, $k^*$, the second is the direction along which to make the split, $i^*$, and the third is the point at which to make the split, $c_{i^*}^{k^*}$. After the split is made the old and new cells, $S^{k^*}$ and $S^{\nu+1}$ respectively, are

$$\begin{aligned} S^{k^*} &= \left[a_1^{k^*}, b_1^{k^*}\right] \times \left[a_2^{k^*}, b_2^{k^*}\right] \times \cdots \times \left[a_{i^*}^{k^*}, c_{i^*}^{k^*}\right] \\ &\quad \times \cdots \times \left[a_n^{k^*}, b_n^{k^*}\right] \qquad \text{(A1)} \\ S^{\nu+1} &= \left[a_1^{k^*}, b_1^{k^*}\right] \times \left[a_2^{k^*}, b_2^{k^*}\right] \times \cdots \times \left[c_{i^*}^{k^*}, b_{i^*}^{k^*}\right] \\ &\quad \times \cdots \times \left[a_n^{k^*}, b_n^{k^*}\right]. \qquad \text{(A2)} \end{aligned}$$

The aim is to obtain solutions to the approximate problem which converge to the optimum quickly. We describe a simplified version of the method proposed by Frauendorfer and Kall (1988) (which is suitable to guarantee convergence of the probability distribution in the limit). Choose the cell, $k^*$, which has the largest difference between the upper and lower bound, i.e.,

$$k^* = \underset{k=1,\ldots,\nu}{\operatorname{argmax}} \left\{ Q^{k,\mathrm{UB}}\left(\mathbf{x}^{(\nu)}\right) - Q^{k,\mathrm{LB}}\left(\mathbf{x}^{(\nu)}\right)\right\}. \qquad \text{(A3)}$$

Note that the upper and lower bounds in (A3) are for a particular cell. For example, the conditional Jensen bound on a given cell, $k$, is $Q^{k,\mathrm{LB}}(\mathbf{x}^{(\nu)}) = p^k Q(\mathbf{x}^{(\nu)}, \mathbf{z}^k)$. The rationale behind choosing the cell with the largest difference is that it has the highest potential for improvement. In choosing the direction, $i^*$, it is desirable to choose one along which there is a high degree of nonlinearity of the recourse function.

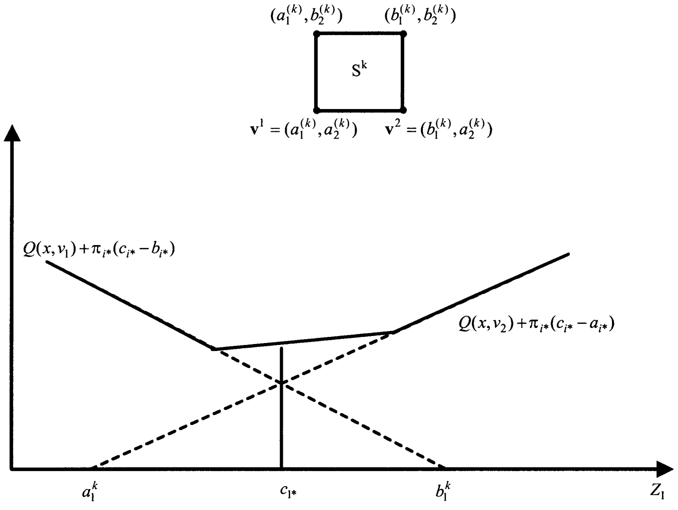
$(a_1^{(k)}, b_2^{(k)})$
$(b_1^{(k)}, b_2^{(k)})$
$S^k$
$\mathbf{v}^1 = (a_1^{(k)}, a_2^{(k)})$
$\mathbf{v}^2 = (b_1^{(k)}, a_2^{(k)})$
$Q(x, v_1) + \pi_{i*}(c_{i*} - b_{i*})$
$Q(x, v_2) + \pi_{i*}(c_{i*} - a_{i*})$
$a_1^k$
$c_{1*}$
$b_1^k$
$Z_1$

**Fig. A1.** Illustration of the method for choosing the split point for cell $k^*$ and direction $i^*$.

Along a given direction $i$ evaluate

$$\epsilon_1^i = Q\big(\mathbf{x}^{(\nu)}, \mathbf{v}^2\big) - Q\big(\mathbf{x}^{(\nu)}, \mathbf{v}^1\big) - \pi(\mathbf{z}^k, \mathbf{v}^1)(\mathbf{v}^2 - \mathbf{v}^1), \quad \text{(A4)}$$
$$\epsilon_2^i = Q\big(\mathbf{x}^{(\nu)}, \mathbf{v}^1\big) - Q\big(\mathbf{x}^{(\nu)}, \mathbf{v}^2\big) - \pi(\mathbf{z}^k, \mathbf{v}^2)(\mathbf{v}^1 - \mathbf{v}^2), \quad \text{(A5)}$$

where $(\mathbf{v}^1, \mathbf{v}^2)$ is a pair of adjacent vertices of $S^k$, $\mathbf{v}^1 = (a_1^k, \ldots, a_i^k, \ldots, a_n^k)$, $\mathbf{v}^2 = (a_1^k, \ldots, b_i^k, \ldots, a_n^k)$. From the subgradient inequality $\epsilon_1$ and $\epsilon_2$ are nonnegative since $Q(\mathbf{x}^{(\nu)}, \mathbf{z})$ is convex. The direction is chosen such that $i^* = \text{argmax}\{\min\{\epsilon_1^i, \epsilon_2^i\}\}$. The point at which to split, $c_{i^*}$, is then chosen so that

$$\begin{aligned} &Q\big(\mathbf{x}^{(\nu)}, \mathbf{v}^2\big) + \pi_{i^*}(\mathbf{z}^k, \mathbf{v}^2)(c_{i^*} - b_{i^*}) \\ &\quad = Q\big(\mathbf{x}^{(\nu)}, \mathbf{v}^1\big) + \pi_{i^*}(\mathbf{z}^k, \mathbf{v}^1)(c_{i^*} - a_{i^*}). \qquad \text{(A6)} \end{aligned}$$

The choice of split point for a two dimensional example is illustrated in Fig. A1.

## Appendix B: Proofs

**Proof of Proposition 2:** The lower bound follows from the fact that the objective in Equation (21) is a Jensen bound for any $\mathbf{x}$ and hence the optimal solution to Equation (21) is a lower bound on the optimal solution to the ASP. The second inequality can be proved in the following way

$$\begin{aligned} Q(\mathbf{x}^*) &= \int_\Xi \boldsymbol{\pi}(\mathbf{z}, \mathbf{x}^*)\mathbf{h}P(\mathrm{d}\mathbf{z}) && \text{(B7)} \\ &\le \int_\Xi \boldsymbol{\pi}(\mathbf{z}, \mathbf{x}^*)\mathbf{h}P(\mathrm{d}\mathbf{z}) + \sum_{k=1}^{\nu} \int_{S^k} \big((\mathbf{c} - \boldsymbol{\pi}(\mathbf{z}, \mathbf{x}^*)\mathcal{W})\mathbf{y}^{k*} \\ &\quad - \boldsymbol{\pi}(\mathbf{z}, \mathbf{x}^*)\mathcal{T}\mathbf{x}^{(\nu)}\big)P(\mathrm{d}\mathbf{z}), && \text{(B8)} \end{aligned}$$

where the inequality in (B8) is due to non-negativity of the second term which results from the dual constraints in Equation (14). Reorganizing the terms we can write the RHS as follows:

$$\begin{aligned} &= Q^{(\nu)} + \sum_{k=1}^{\nu} \int_{S^k} \boldsymbol{\pi}(\mathbf{z}, \mathbf{x}^*)\big(\mathbf{h} - \mathcal{W}\mathbf{y}^{k*} - \mathcal{T}\mathbf{x}^{(\nu)}\big)P(\mathrm{d}\mathbf{z}) && \text{(B9)} \\ &\le Q^{(\nu)} + \sum_{k=1}^{\nu}\sum_{i=2}^{n} \int_{S^k} \big(\pi_i^{\mathrm{UB}}\big(h_i - \mathcal{W}_{(i\cdot)}\mathbf{y}^{k*} - \mathcal{T}\mathbf{x}^{(\nu)}\big)^+ \\ &\quad - \pi_i^{\mathrm{LB}}\big(\mathcal{W}_{(i\cdot)}\mathbf{y}^{k*} + \mathcal{T}\mathbf{x}^{(\nu)} - h_i\big)^+\big)P(\mathrm{d}\mathbf{z}), && \text{(B10)} \\ &= Q^{(\nu)} + \sum_{k=1}^{\nu}\sum_{i=2}^{n} \int_{S^k} \big(\pi_i^{\mathrm{UB}}\big(h_i - h_i^k\big)^+ \\ &\quad - \pi_i^{\mathrm{LB}}\big(h_i^k - h_i\big)^+\big)P(\mathrm{d}\mathbf{z}), && \text{(B11)} \end{aligned}$$

where $\pi_i^{\mathrm{UB}}$ and $\pi_i^{\mathrm{LB}}$ are upper and lower bounds on the dual solution over all $\mathbf{x} \in \Re^{n-1}$ and $\mathbf{Z} \in \Xi$, and the positive and negative parts of the integrand in (A10) have been separated to give an overall upper bound on $Q(\mathbf{x}^*)$. The feasible region of the dual is compact and the bounds can be obtained from Equations (16) and (17) as follows

$$\begin{aligned} \pi_i^{\mathrm{UB}} &= \max\{\pi_i(\mathbf{x}, \mathbf{Z}) \mid \mathbf{x} \in \Re^{n-1}, \mathbf{Z} \in \Xi\}, && \text{(B12)} \\ &= c_{i+1}^w + \max\{\pi_{i+1}(\mathbf{x}, \mathbf{Z}) \mid \mathbf{x} \in \Re^{n-1}, \mathbf{Z} \in \Xi\}, && \text{(B13)} \\ &= \sum_{j=i+1}^{n} c_j^w + c_\ell, && \text{(B14)} \end{aligned}$$

and

$$\pi_i^{\mathrm{LB}} = \min\{\pi_i(\mathbf{x}, \mathbf{Z}) | \mathbf{x} \in \Re^{n-1}, \mathbf{Z} \in \Xi\} = -c_{i+1}^s. \qquad \text{(B15)}$$

■

**Proof of Proposition 3.** The lower bound in Equation (25) follows from the fact that $Q^{(\nu)}$ is a Jensen lower bound on the optimum. The upper bound is proved as follows:

$$\begin{aligned} Q\big(\mathbf{x}^{(\nu)}\big) &= \int_\Xi \boldsymbol{\pi}\big(\mathbf{x}, \mathbf{z}^{(\nu)}\big)\big(\mathbf{h} - \mathcal{T}\mathbf{x}^{(\nu)}\big)P(\mathrm{d}\mathbf{z}) && \text{(B16)} \\ &\le \int_\Xi \boldsymbol{\pi}\big(\mathbf{x}, \mathbf{z}^{(\nu)}\big)\big(\mathbf{h} - \mathcal{T}\mathbf{x}^{(\nu)}\big)P(\mathrm{d}\mathbf{z}) \\ &\quad + \sum_{k=1}^{\nu} \int_{S^k} \big(\mathbf{c} - \boldsymbol{\pi}\big(\mathbf{x}, \mathbf{z}^{(\nu)}\big)\mathcal{W}\big)\mathbf{y}^{k*}P(\mathrm{d}\mathbf{z}), && \text{(B17)} \end{aligned}$$

where the inequality follows from the dual constraints on the discrete approximation, Equation (21). Rearranging the terms as in Proposition 2

$$\begin{aligned} Q\big(\mathbf{x}^{(\nu)}\big) \le Q^{(\nu)} &+ \sum_{k=1}^{\nu}\sum_{i=2}^{n} \int_{S^k} \big(\pi_i^{k,\mathrm{UB}}\big(h_i - h_i^k\big)^+ \\ &- \pi_i^{k,\mathrm{LB}}\big(h_i^k - h_i\big)^+\big)P(\mathrm{d}\mathbf{z}), \qquad \text{(B18)} \end{aligned}$$

where $\pi_i^{k,\mathrm{UB}} = \max\{\pi_i(\mathbf{x}^{(\nu)}, \mathbf{Z}) \mid \mathbf{Z} \in S^k\}$ and $\pi_i^{k,\mathrm{LB}} = \min\{\pi_i(\mathbf{x}^{(\nu)}, \mathbf{Z}) \mid \mathbf{Z} \in S^k\}$. ■

**Proof of Proposition 4:** The proof follows from the fact that

$$\min_{\mathbf{y}} \{\mathbf{cy} \mid \mathcal{T}\mathbf{x} + \mathcal{W}\mathbf{y} = a\mathbf{h} + \mathbf{b},\ \mathbf{y} \geq 0\}, \qquad \text{(B19)}$$

is equivalent to

$$\min_{\mathbf{y}} \{a\mathbf{cy}' | \mathcal{T}\mathbf{x}' + \mathcal{W}\mathbf{y}' = \mathbf{h},\ \mathbf{y}' \geq 0\}, \qquad \text{(B20)}$$

where $\mathbf{y}' = a^{-1}\mathbf{y}$ and $\mathbf{x}' = a^{-1}\mathbf{x} - a^{-1}\mathbf{b}$. ■

## Biographies

Brian Denton is an engineer in IBM's Microelectronics Division. He is involved in development and implementation of heuristics and linear programming models and methods for solving large-scale supply chain planning problems for semiconductor manufacturing and data storage device manufacturing. He completed his doctorate in Management Science in 2001 at McMaster University. Prior to joining IBM he worked on applications of stochastic programming and combinatorial optimization to appointment scheduling systems in the health care industry, and inventory design and deployment problems in the steel industry.

Diwakar Gupta teaches in the Graduate Program in Industrial Engineering at the University of Minnesota, where he holds the rank of Mayhugh Associate Professor of Mechanical Engineering. His research interests are in the area of stochastic modeling with applications to manufacturing, supply chain management and health care delivery systems. Before joining the University of Minnesota in 1999, he was an Associate Professor of Production and Management Science at McMaster University, Canada.

*Contributed by the Supply Chains/Production-Inventory Systems Departments*