Manufacturing & Service Operations Management

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

Dynamic Capacity Expansion Problem with Deferred Expansion and Age-Dependent Shortage Cost

Vernon Ning Hsu,

Dynamic Capacity Expansion Problem with Deferred Expansion and Age-Dependent Shortage Cost

Vernon Ning Hsu

Department of Industrial Engineering and Engineering Management, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong and School of Management, George Mason University, Fairfax, Virginia 22030

vhsu@ust.hk

Deferring capacity expansion may be a cost effective decision when there is anticipation of cheaper capacity in the near future and/or the current demand is too low to justify an immediate expansion. This paper studies a finite-horizon capacity expansion problem (CEP) with deferred capacity expansion. The operating cost and the cost of holding unused capacity in each period depend on the time when the capacity is acquired, and the shortage cost depends on the time when the shortage occurred. Our model is a generalization of the Wagner-Whitin formulation of the CEP and an extension (with deferred expansion) of two other polynomially solvable CEPs in the literature. We explore structural properties of the problem and develop an efficient dynamic programming algorithm to solve the problem in polynomial time.

(*Capacity Planning; Capacity Expansion with Shortage; Dynamic Programming*)

1. Introduction

This paper studies a finite-horizon dynamic capacity expansion problem (CEP) where there is a known incremental demand for a single type of capacity/facility in each of the n periods. The incremental demand in each period can be met either by new capacity expansion in the same period or by unused capacity acquired in earlier periods. The incremental demand can also be left unsatisfied (or satisfied using temporary capacity such as leasing or outsourcing) until some future capacity expansion. We call this latter option *deferred capacity expansion.* The objective of the CEP is to determine the timing and the size of the capacity expansion in order to meet the incremental demands in a planning horizon at minimum total costs. Hereafter, whenever it is deemed not to cause confusion, we will use ''demand'' in place of ''incremental demand'' for easier exposition.

There are many practical situations where deferred capacity expansion may be a cost effective decision. A firm may delay the acquisition of a certain technology which is expected to be much cheaper and more advanced in the near future. For example, an organization that needs to purchase a number of personal computers (PCs) to expand its computing capacity may decide to wait for a few months to take advantages of the anticipated cheaper prices for certain PC models that are at the end of their lifespan in the market. Another situation in which a firm may consider deferring capacity expansion is when the current demand is too low to justify an immediate expansion. If the current demand is low and the needed capacity expansion involves large initial investment (large fixed cost), an immediate expansion may entail high cost of idle capacity. In this case, a firm may choose to delay the acquisition of certain capacity until a later time when the future demand growth justifies a new capacity expansion. Readers

Manufacturing & Service Operations Management © 2002 INFORMS
Vol. 4, No. 1, Winter 2002, pp. 44–54

1523-4614/02/0401/0044$05.00
1526-5498 electronic ISSN

are referred to Luss (1982) for additional discussion on motivations to defer capacity expansion by allowing temporary capacity shortage at certain shortage costs.

One of the basic versions of the CEP is described in Luss (1982) and is formulated as a variant of the well-known economic lot sizing (ELS) problem originally proposed by Wagner and Whitin (1958). Hereafter, we will call this basic CEP model the CWW model. It is well known that the CWW model is polynomially solvable (see for example Aggarwal and Park 1993). Though efficiently solvable, the CWW model makes the following three assumptions which may be unrealistic in some practical applications. First, the cost of holding unused capacity (called *holding* cost hereafter) in each period is *age-independent* (*AID*) in the sense that it depends only on the total volume of the capacity in the period and *not* on the time of the capacity's expansion/acquisition. Second, the CWW model also assumes AID operating cost; i.e., it depends only on the period when the capacity is first used and not the period when it is acquired. With this assumption, the operating cost is independent of the decision variables and is excluded from the CWW model (see Luss 1982). Finally, if capacity shortage is allowed, the CWW model assumes AID shortage cost in each period, i.e., the unit shortage cost in a period is the same for all shortages regardless of when they occur.

The AID holding and operating costs are also assumed in numerous past and recent literature on capacity expansion problems of multiple capacity/facility types—for example, Klincewicz et al. (1988), Lee and Luss (1987), Li and Tirupati (1994), Luss (1979, 1980, 1983), Rajagopalan et al. (1998).

There are many situations in which the AID holding and operating costs are not applicable. As pointed out by Luss (1982), ''in various applications the operating cost may depend on the technology available at the selected expansion time. Further, the operating cost may depend on the elapsed time since the facilities were installed because old facilities deteriorate over time and require more maintenance'' (p. 913). It is, however, further suggested by Luss that ''an explicit consideration of such variable operating costs is difficult.''

In recent years, there has been numerous research focusing on capacity expansion problems with *age-dependent* (*AD*) holding and operating costs, i.e., these costs depend on the period in which capacity is acquired. Some of the research also considers capacity disposal and deterioration (Rajagopalan and Soteriou 1994, Rajagopalan 1998, Chand et al. 2000). Most of these models are difficult to solve and rely on solution methods that are either heuristic or nonpolynomial procedures.

Rajagopalan (1992) considers a CEP with capacity deterioration and formulates the problem as an instance of the specially structured uncapacitated facility location problem (UFLP) proposed by Krarup and Bilde (1977), which is solvable in polynomial time. Rajagopalan's model includes fixed costs for capacity acquisition and variable costs associated with each unit of capacity acquired. The variable cost consists of unit capacity acquisition and holding cost, as well as the operating cost over the entire life of the unit capacity. Rajagopalan's model is later restated in Jones et al. (1995) where unit acquisition and operating costs are defined separately.

There have been very few capacity expansion models in the literature that consider capacity shortage. Luss (1980) and Lee and Luss (1987) formulate capacity shortage as negative inventory, an approach similar to that of the traditional CWW model. Thus, their shortage costs are AID. In a recent study on an ELS problem, Hsu and Lowe (2001) point out that when a demand is not satisfied, the penalty cost for backordering the demand could depend nonlinearly on the length of backorders. They propose AD backordering costs for backordered demands that occur in different periods. In the context of capacity shortage, for example, the cost of delayed expansion of a computing facility may accelerate due to deteriorating services or even project delay; the longer a project is delayed, the higher the unit time delay penalty cost may be. In this case, the cost of capacity shortage may be AD; i.e., it may depend on when the shortage occurred and for how long the shortage has occurred.

This paper presents a CEP with AD holding, operating, and shortage costs. The paper makes two main contributions. First, our model generalizes the

CWW model with more general AD holding, operating, and shortage costs. Our model also extends the models of Rajagopalan (1992) and Jones et al. (1995) by allowing deferred capacity expansion. Second, we develop a polynomial time dynamic programming (DP) algorithm to solve the proposed model. Furthermore, we show that the cost structure in our model satisfies a concavity property that allows us to reduce the computational complexity.

The remainder of this paper is organized as follows. Section 2 presents the CEP with deferred capacity expansion. Section 3 gives some structural properties of the proposed model and develops a polynomial time DP algorithm. Section 4 further explores the cost structures of the model and uses them to reduce the computational complexity of the proposed DP recursion in §3. Section 5 proposes another variant of the CEP with no speculative motive in the holding cost. Section 6 concludes the paper.

2. The Capacity Expansion Problem with Shortage

Suppose there are n periods with known incremental demand $d_t \geq 0$ of a certain capacity/facility in period t, $1 \leq t \leq n$. The demand in period t can be satisfied by capacity acquisition in the same period t or by unused capacity acquired in earlier periods that is carried to the end of period $t - 1$ (beginning of period t). Capacity expansion can also be deferred by allowing temporary capacity shortage.

Our model does not consider capacity disposal as a decision variable. Instead, we assume that a capacity, which is acquired and first used in a certain period, has a known useful life that is assumed to be no smaller than the planning horizon n. This useful life may depend on the durability of the capacity (e.g., a copy machine) and/or economical and technological usefulness of the capacity (e.g., a computer is too old to run new versions of software). At the end of its useful life, the disposed capacity creates a new incremental demand in a future period beyond period n. For $1 \leq t \leq n$, let

x_t = the amount of capacity acquired in period t;

f_t = the fixed cost in period t if $x_t > 0$; A period t is called an *expansion period* if $x_t > 0$;

c_t = the unit capacity acquisition cost in period t.

We assume that capacity acquisition occurs instantaneously at the beginning of each expansion period. We also assume that the demand in every period is satisfied at the beginning of the period.

The cost of holding unused capacity in each period depends on the period in which the capacity is acquired. The holding cost includes the costs of space, labor, and equipment used to maintain unused capacity. The cost of capacity shortage depends on the time when the shortage occurs. For $1 \leq i \leq t \leq n$, define

y_{it} = the amount of the unused capacity in period t that is acquired in period i;

z_{it} = the amount of period i demand that is unsatisfied in period t;

g_{it} = fixed cost in period t for carrying any unused capacity that is acquired in period i;

h_{it} = the cost of holding unit unused capacity in period t which is acquired in period i (note that the holding cost functions are fixed-plus-linear);

p_{it} = the cost of leaving unit unsatisfied period i demand in period t.

There are two approaches in dynamic capacity expansion literature to model the operating cost. One approach defines a per unit operating cost in each period (for example, Rajagopalan and Soteriou 1994, Rajagopalan 1998, Chand et al. 2000). The objective of the models using this approach is to minimize the total acquisition, holding, and operating costs *incurred within* the n-period planning horizon.

Another approach defines a per unit cost for operating a capacity in its entire useful life time (for example, Rajagopalan 1992 and Jones et al. 1995). With this second approach, the objective of the CEP is to minimize the total acquisition, holding, and operating costs *associated with* all capacity expansion incurred in the n-period planning horizon. This objective can be reasonable for a decision maker who needs to consider, at the time of capacity expansion, the impact of all the present and future costs associated with current acquisition. One example is the consideration of future costs of maintenance and repair (parts and labor) when purchasing an imported car. The disadvantage of this approach in modeling the operating cost, however, is that one has to estimate the operation cost beyond period n.

Since we model our capacity expansion problem as an extension of the problems studied in Rajagopalan (1992) and Jones et al. (1995), we choose to adopt the second approach by defining for $1 \leq i \leq t \leq n$,

o_{it} = the total cost of operating and maintaining unit capacity, which is acquired in period i and first used in period t, in its entire useful life.

All costs defined above are assumed to be discounted to the beginning of period 1. Thus, the holding cost excludes the cost of capital. We remark that it would not be difficult to incorporate the capacity deterioration explicitly in our model in a way similar to that of Rajagopalan (1992). We assume zero deterioration for the simplicity of the paper's exposition.

We assume that the cost of holding unused capacity is nondecreasing in the age of the capacity, i.e., we assume for $1 < i \leq t \leq n$,

$$g_{it} \leq g_{i-1,t} \quad \text{and} \quad h_{it} \leq h_{i-1,t}. \tag{1}$$

Similarly, we assume that the capacity shortage cost is nondecreasing in the age of the unsatisfied capacity, for $1 \leq j < t \leq n$,

$$p_{j+1,t} \leq p_{jt}. \tag{2}$$

For operating costs, we make the following two assumptions. First, we assume for $1 \leq i \leq t \leq n$,

$$o_{i,t+1} \geq o_{it}, \tag{3}$$

which says that the cost to operate unit capacity acquired in a certain period i is nondecreasing in its age. Note that $o_{i,t+1} - o_{it} \geq 0$, which can be interpreted as the marginal increase of operating cost with respect to the age of the capacity acquired in period i. We further assume for $1 \leq i < j \leq t < n$,

$$o_{j,t+1} - o_{jt} \leq o_{i,t+1} - o_{it}, \tag{4}$$

which implies that the marginal increase of operating cost for an older age capacity (acquired in period i) is no less than that for a younger age capacity (acquired in period j). This assumption of accelerating operating cost with respect to the age of capacity is reasonable in real world applications.

For $1 \leq i \leq n$ and $1 \leq j \leq n$, define

m_{ij} = the amount of capacity acquired in period i and first used to satisfy the demand in period j. Note that if $j < i$, m_{ij} is the amount of period j demand left unsatisfied until a deferred expansion in some future period i.

Assume without loss of generality that there is no unused capacity at the beginning of period 1 and no unused capacity is required at the end of period n. Denoting $\delta(x) = 1$, if $x > 0$; 0 otherwise, we are now ready to present our capacity expansion problem.

$$\text{minimize} \sum_{t=1}^{n} \Bigg(f_t\delta(x_t) + c_t x_t + \sum_{i=1}^{t} (g_{it}\delta(y_{it}) + h_{it}y_{it}) + \sum_{j=1}^{t} p_{jt}z_{jt} + \sum_{i=1}^{t} o_{it}m_{it} + \sum_{k=t+1}^{n} o_{kk}m_{kt} \Bigg) \tag{5}$$

subject to:

$$x_t - \sum_{i=1}^{t} m_{ti} = y_{tt}, \qquad 1 \leq t \leq n, \tag{6}$$

$$y_{i,t-1} - m_{it} = y_{it}, \qquad 1 \leq i < t \leq n, \tag{7}$$

$$z_{j,t}-1 - m_{tj} = z_{jt}, \qquad 1 \leq j < t \leq n, \tag{8}$$

$$\sum_{i=1}^{n} m_{it} = d_t, \qquad 1 \leq t \leq n, \tag{9}$$

$$x_t,\ y_{it},\ z_{jt},\ m_{it} \geq 0, \qquad 1 \leq i,\ j,\ t \leq n. \tag{10}$$

Constraint (6) says that after using $\Sigma_{i=1}^{t} m_{ti}$ out of a total of x_t units of newly acquired capacity to satisfy the demand increments in periods 1 through t, the remaining y_{tt} unused units are carried to the next period. Similarly, (7) is another supply-demand balance constraint: after taking m_{it} units out of $y_{i,t-1}$ at the beginning of period t (the end of period $t-1$) to satisfy the demand in period t, the remaining unused capacity is y_{it}. Constraint (8) shows that the period j capacity shortage in period t (z_{jt}), equals the difference of the shortage ($z_{j,t-1}$) carried from the previous period and the amount (m_{tj}) satisfied by acquisition in period t. Finally, the demand in a period t must be satisfied by capacities acquired in periods 1 through n (constraint (9)). (10) is the nonnegativity constraint. The objective of the problem is to minimize the total capacity acquisition, holding, operating, and shortage costs. Note that if $g_{it} = 0$, $h_{it} = h_t$, $p_{jt} = p_t$, and $o_{it} = 0$, the above CEP reduces to the classical CWW model.

3. Structural Properties and the DP Algorithm

Denote (CEP) as the capacity expansion problem defined by (5)–(10) which satisfies conditions (1)–(4). Denote Ω as a feasible solution to problem (CEP) and $V(\Omega)$ as the corresponding objective function value. We have the following results:

Theorem 1. *There is an optimal solution to (CEP) where the demand in a period t is met entirely by the capacity acquired in exactly one of the period i, $1 \leq i \leq n$.*

Proof. First note that problem (CEP) consists of a concave objective function and linear constraints. By Theorem 3.4.6 in Bazaraa and Shetty (1979), there exists an optimal solution to (CEP) which is an extreme solution (extreme points of the feasible solution region formed by (6)–(10)).

Also note that by moving the right-hand side terms in (6)–(8), each variable x_t, y_{it}, z_{it}, and m_{ij} in constraints (6)–(9) has exactly one $+1$ coefficient and other (possibly more than one) -1 coefficients. It follows from Veinott (1969) that if more than one variable appears in the same constraint, only one of these variables is positive. Applying this to constraint (9), we have the desired result. $\square$

Theorem 2. *There exists an optimal solution Ω^* to problem (CEP), where if $i < j$ are two expansion periods and $m^*_{jt} = d_t$ for some t, $1 \leq t \leq n$, then $m^*_{ik} = 0$ for all k, $1 \leq t \leq k \leq n$.*

Proof. The proof is in the Appendix. $\square$

It is easy to show that an optimal solution of (CEP) satisfying Theorem 2 has the following property:

Partition Property (PP). (i) *The capacity acquisition in each expansion period is used to satisfy all demands from a* **consecutive** *number of periods;* (ii) *Suppose $i < j$ are two expansion periods, and $k(i)$ and $k(j)$ are the smallest indexed periods whose demands are met by acquisition from periods i and j, respectively. We have $k(i) < k(j)$.*

Remark. The proof of Theorem 2 uses condition (4) on the operating cost. We give an example to demonstrate that PP may not hold for some instances of (CEP) if we only require a weaker condition (3) and not (4).

Example 1. Consider a 4-period problem where $d_t = 2$, $1 \leq t \leq 4$. $f_1 = f_2 = f_3 = f_4 = 0$, $c_1 = 1$, $c_2 = 4$, and $c_3 = c_4 = 20$. The holding costs is $h_{it}(y) = y$, $1 \leq i \leq t \leq 4$. The shortage costs are assumed to be very large. The operating costs are $o_{11} = 100$, $o_{12} = 105$, $o_{13} = 110$, $o_{14} = 120$, $o_{22} = 100$, $o_{23} = 109$, $o_{24} = 110$, $o_{33} = 100$, $o_{34} = 110$, and $o_{44} = 100$. Note that the operating costs satisfy condition (3) and not (4) because $o_{13} - o_{12} < o_{23} - o_{22}$. The unique optimal solution to the above problem is to acquire 4 units capacity in period 1 to satisfy demands in periods 1 and 3 and to acquire 4 units in period 2 to satisfy the demands in periods 2 and 4. It is clear that PP does not hold.

Without PP, (CEP) may not be efficiently solvable. In §5, we will relax condition (4) in another variant of the capacity expansion problem which satisfies a condition that demonstrates lack of speculative motive in holding cost.

Based on PP, we now present a DP algorithm to solve (CEP). For $1 \leq s \leq t < i \leq q \leq r \leq n$, suppose acquisition in period i is used to satisfy demands in periods q through r and it is also used to satisfy capacity shortages from periods s through t. Let $\mathrm{TH}_i(q, r)$ and $\mathrm{TO}_i(q, r)$ ($\mathrm{TP}_i(s, t)$) be the total holding and operating (shortage cost) to satisfy demands in periods q through r (s through t) by acquisition in period i. We have $\mathrm{TH}_i(q, r) = G_{ir} + \sum_{k=q}^{r} H_{ik} d_k$, where $G_{ir} = \sum_{l=i}^{r-1} g_{il}$ and $H_{ik} = \sum_{l=i}^{k-1} h_{il}$; $\mathrm{TO}_i(q, r) = \sum_{k=q}^{r} o_{ik} d_k$; and $\mathrm{TP}_i(s, t) = \sum_{k=s}^{t} BP_{ki}$, where $BP_{ki} = \sum_{l=k}^{i-1} p_{kl} d_k$. For $1 \leq i \leq n$ and $1 \leq p \leq q \leq n$, let $T_i(p, q)$ be the total cost to satisfy demands in periods p through q by acquisition in period i. Denoting $d_{ij} = \sum_{t=i}^{j} d_t$, $1 \leq i \leq j \leq n$, we have

$$T_i(p, q) = \begin{cases} f_i + c_i d_{p,q-1} + \mathrm{TH}_i(p,q) + \mathrm{TO}_i(p,q), & \text{if } i \le p; \\ f_i + c_i d_{pq} + \mathrm{TP}_i(p, i-1) + o_{ii} d_{p,i-1} \\ \quad + \mathrm{TH}_i(i,q) + \mathrm{TO}_i(i,q), & \text{if } p < i \le q; \\ f_i + c_i d_{pq} + \mathrm{TP}_i(p,q) + o_{ii} d_{pq}, & \text{if } q < i. \end{cases} \tag{11}$$

For $1 \le j \le n$, let $V(j)$ be the minimum cost of satisfying demands in periods 1 through j. The optimal objective value of (CEP) is $V(n)$. Defining $V(0) = 0$, we have the following DP recursion, for $1 \le j \le n$,

$$V(j) = \min_{1 \le i \le j, 1 \le k \le n} \{V(i-1) + T_k(i, j)\}. \tag{12}$$

To analyze the computational complexity of recursion (12), we note that for fixed i, $1 \le i \le n$, we can compute all $\{H_{ik} \mid i \le k \le n\}$ and $\{G_{ir} \mid i \le r \le n\}$ in $O(n)$ time via $H_{i,k+1} = H_{ik} + h_{ik}$ and $G_{i,r+1} = G_{ir} + g_{ir}$. Thus, the total time to precompute all $\{H_{ik}\}$ and $\{G_{ir}\}$ is $O(n^2)$. Also, note that $\mathrm{TH}_i(i, r) = \mathrm{TH}_i(i, r-1) + g_{ir} + H_{ir} d_r$ and $\mathrm{TO}_i(i, r) = \mathrm{TO}_i(i, r-1) + o_{ir} d_r$ for $r > i$. Thus, it takes additional $O(n^2)$ to obtain all $\{\mathrm{TH}_i(i, r)\}$ and $\{\mathrm{TO}_i(i, r)\}$. Once they are available, each $\mathrm{TH}_i(q, r)$ and $\mathrm{TO}_i(q, r)$ can be computed in constant time as needed via $\mathrm{TH}_i(q, r) = \mathrm{TH}_i(i, r) - \mathrm{TH}_i(i, q-1) + G_{i,q-1}$ and $\mathrm{TO}_i(q, r) = \mathrm{TO}_i(i, r) - \mathrm{TO}_i(i, q-1)$.

For fixed k and i, $1 \le k < i < n-1$, note that $BP_{k,i+1} = BP_{ki} + p_{k,i+1} d_k$. Thus, we can obtain all $\{BP_{ki} \mid 1 \le k < i \le n\}$ in $O(n^2)$ time. It then takes $O(n^2)$ additional effort to compute all $\{\mathrm{TP}_i(1, t) \mid 1 \le t < i \le n\}$ via $\mathrm{TP}_i(1, t+1) = \mathrm{TP}_i(1, t) + BP_{t+1,i}$. Once they are available, each $\mathrm{TP}_i(s, t)$, $1 < s \le t < i \le n$, can be evaluated in constant time since $\mathrm{TP}_i(s, t) = \mathrm{TP}_i(1, t) - \mathrm{TP}_i(1, s-1)$. Finally, it takes $O(n)$ time to get d_{1j}, for all j, $1 \le j \le n$. Each d_{ij}, $1 < i \le j \le n$, can then be computed in constant time as needed via $d_{ij} = d_{1j} - d_{1,i-1}$.

We now see that after an $O(n^2)$ time preprocessing step, each $T_i(p, q)$, where $1 \le i \le n$ and $1 \le p \le q \le n$, can be evaluated in constant time via (11). Once these values are available, all $\{V(j)\}$ can be obtained in $O(n^3)$ time. To summarize, a simple count of the computational effort for recursion (12) amounts to $o(n^3)$.

4. Reduced Computational Complexity

In this section, we will show that the DP recursion (12) in the last section can be implemented with a lower order complexity of $O(n^2)$. First, define for $1 \le j \le n$,

$$V^1(j) = \min_{1 \le i \le k \le j} \{V(i-1) + T_k(i, j)\}; \tag{13}$$

for $1 < j \le n$,

$$V^2(j) = \min_{1 \le k < i \le j} \{V(i-1) + D_k(i, j)\}, \tag{14}$$

where $D_k(i, j) = T_k(i, j)$ for $1 \le k < i \le j \le n$; and for $1 \le j < n$,

$$V^3(j) = \min_{1 \le i \le j < k \le n} \{V(i-1) + E_k(i, j)\}, \tag{15}$$

where $E_k(i, j) = T_k(i, j)$ for $1 \le i \le j < k \le n$. Clearly, $V(j) = \min\{V^1(j), V^2(j), V^3(j)\}$.

For each fixed k, $1 \le k \le n$, define a $k \times (n - k + 1)$ matrix $A_k = (a_{st})$, where $a_{st} = V(s-1) + T_k(s, t+k-1)$. For $k \le j \le n$, define $w_{kj} = \min_{1 \le i \le k} a_{i,j-k+1}$, which are the column minima of the matrix A_k. Note that $V^1(j) = \min_{1 \le k \le j}\{w_{kj}\}$.

Also, for fixed k and j, where $1 \le k < j \le n$, define

$$u_{kj} = \min_{k < i \le j} \{V(i-1) + D_k(i, j)\}. \tag{16}$$

For fixed k and j, $1 \le j < k \le n$, define

$$z_{kj} = \min_{1 \le i \le j} \{V(i-1) + E_k(i, j)\}. \tag{17}$$

We have $V^2(j) = \min_{1 \le k < j} \{u_{kj}\}$ and $V^3(j) = \min_{j < k \le n} \{z_{kj}\}$.

Our low-order solution procedure will take advantage of a certain concavity property in the above matrix A_k and the two functions $D_k(i, j)$ and $E_k(i, j)$. Similar approaches have also been used by Aggarwal and Park (1993) to solve the ELS problems and by Hsu et al. (1997) to solve a location problem. We first show that matrix A_k satisfies a so-called *Monge* condition (see Aggarwal et al. 1987 and Aggarwal and Park 1993) stated in the following lemma.

LEMMA 1. *If s, l, t, q satisfy $1 \le s < l \le k$ and $1 \le t < q \le n - k + 1$, then*

$$a_{sq} - a_{st} \geq a_{lq} - a_{lt}. \qquad (18)$$

PROOF. Since $s < l \leq k \leq t + k - 1 < q + k - 1$, by (11) we have

$$a_{sq} - a_{st} = c_k d_{t+k,q+k-1} + \mathrm{TH}_k(t + k, q + k - 1) - G_{k,t+k} + \mathrm{TO}_k(t + k, q + k - 1).$$

Note that the right-hand side of the expression above is independent of the index s. It follows that $a_{sq} - a_{st} = a_{lq} - a_{lt}$. □

Similarly, we now show that the functions $D_k(i, j)$ and $E_k(i, j)$ satisfy a *concavity* property stated in the following lemma.

LEMMA 2. *If s, l, t, q satisfy $s \leq l \leq t \leq q$, then*

$$D_k(s, q) - D_k(s, t) \geq D_k(l, q) - D_k(l, t) \qquad (19)$$

and

$$E_k(s, q) - E_k(s, t) \geq E_k(l, q) - E_k(l, t). \qquad (20)$$

PROOF. Note that for $t < q$,

$$D_k(s, q) - D_k(s, t) = c_k d_{t+1,q} + \mathrm{TH}_k(t + 1, q) - G_{k,t+1} + \mathrm{TO}_k(t + 1, q).$$

Again, the right-hand side of the expression above is independent of the index s. It follows that $D_k(s, q) - D_k(s, t) = D_k(l, q) - D_k(l, t)$. (20) can be shown similarly. □

For each fixed k, $1 \leq k \leq n$, by Lemma 1, we can use the algorithm in Aggarwal et al. (1987) to compute column minima of matrix A_k, i.e., all of the values $\{w_{kj} \mid k \leq j \leq n\}$, in $O(n - k)$ time, provided that any entry in matrix A_k can be looked up (or computed) in constant time.

By Lemma 2, we can apply algorithms given by Eppstein (1990) and Galil and Park (1990) to obtain all $\{u_{kj} \mid k < j \leq n\}$ ($\{z_{kj} \mid 1 \leq j < k\}$) in $O(n - k)$ ($O(k)$) time. The algorithm requires that at the time of computing value u_{kj} (z_{kj}), all $u_{k,k+1}, u_{k,k+2}, \ldots, u_{k,j-1}$ ($z_{k1}, z_{k2}, \ldots, z_{k,j-1}$) have been computed and all $V(1), \ldots, V(j - 1)$ are available.

Recall in our earlier discussion that, after an $O(n^2)$ time preprocessing, each value $T_k(i, j)$, $D_k(i, j)$ and $E_k(i, j)$ can be obtained as needed in constant time via (11). The overall DP algorithm consists of the following additional n iterations. In the jth iteration, we already have $V(1), \ldots, V(j - 1)$ available. The computation in each iteration has the following steps:

1. We first compute $w_{jj}, w_{j,j+1}, \ldots, w_{jn}$, the column minima of matrix A_j.
2. We then compute $V^1(j)$ via (13). Note that in the jth iteration, values $w_{1j}, w_{2j}, \ldots, w_{jj}$, which are needed in (13) to obtain $V^1(j)$, have already been computed in Step 1 of iterations $1, 2, \ldots, j$, respectively.
3. Since $V(1), \ldots, V(j - 1)$ are available in the jth iteration, we can compute values $u_{1j}, u_{2j}, \ldots, u_{j-1,j}$ via (16) and values $z_{j+1,j}, z_{j+2,j}, \ldots, z_{nj}$ via (17).
4. We then obtain $V^2(j)$ and $V^3(j)$ via (14) and (15), respectively. Finally, $V(j) = \{V^1(j), V^2(j), V^3(j)\}$ is obtained in constant time.

Since in every iteration, each of Steps 1, 2, and 4 can be executed in $O(n)$ time, the overall effort of these steps in the entire n-iteration DP algorithm is $O(n^2)$. To account for the total effort in Step 3 of the entire algorithm, note that in the jth iteration, we need to compute u_{kj} (z_{kj}) for every fixed k, $1 \leq k < j$ ($j < k \leq n$). Note that at this time (of jth iteration), we have already computed values $u_{k,k+1}, u_{k,k+2}, \ldots, u_{k,j-1}$ ($z_{k1}, z_{k2}, \ldots, z_{k,j-1}$) in Step 3 of iterations $k + 1, k + 2, \ldots, j - 1$ ($1, 2, \ldots, j - 1$), respectively. Thus by earlier discussion, for fixed k, $1 \leq k < j$ ($j < k \leq n$), the effort to compute all $\{u_{kj} \mid k < j \leq n\}$ ($\{z_{kj} \mid 1 \leq j < k\}$) is $O(n - k)$ ($O(k)$) time. It is now easy to see that the total effort of Step 3 throughout the entire DP algorithm is bounded by $O(n^2)$. We conclude that problem (CEP) can be solved in $O(n^2)$ time. We remark that it takes $O(n^2)$ time to process the input of problem (CEP). Thus, the $O(n^2)$ time we obtained for solving (CEP) is the lowest order of computational complexity.

5. The Model with No Speculative Motive in Holding Cost

In this section, we consider a version of the capacity expansion problem where the cost of holding unused capacity exhibits no speculative motive; i.e., the marginal cost of acquiring a unit of capacity in a certain period and holding this (unused) unit until a future period is no less than that of acquiring the unit in the future period. Formally, we assume for $1 \leq i < n$,

$$c_{i+1} \leq c_i + h_{ii}. \tag{21}$$

This condition of no speculative motive has been used by Wagner and Whitin (1958) and many others in their studies of ELS models. (See Aggarwal and Park 1993 for a thorough review of the ELS models.) In addition, we assume for $1 < i \leq t \leq n$,

$$o_{i-1,t} \geq o_{it}, \tag{22}$$

which, similar to condition (3) in §2, says that the operating cost is nondecreasing in the age of the capacity.

Denote (CEP1) as the CEP defined by (5)–(10) which satisfies conditions (1)–(3) and (21)–(22). Note that in problem (CEP1), we no longer require condition (4). It is now appropriate to discuss the relationships between our model (CEP1) and the models of Rajagopalan (1992) and Jones et al. (1995). Although the acquisition cost, holding cost, and operating cost are not defined separately in the above two studies (Jones et al. define acquisition cost separately from holding and operating cost), the assumptions on the variable costs in these two papers are similar to that of (CEP1). In addition, our model (CEP1) allows for deferred capacity expansion and includes a fix charge in each holding cost function, both of which are not modeled in Rajagopalan (1992) and Jones et al. (1995).

Remark. As mentioned in §1, Rajagopalan and Jones et al. formulate their CEPs as special instances of UFLPs that are polynomially solvable. We give an example in the Appendix to show that neither of their approaches can be used to solve our problem (CEP1).

Lemma 3. *For $i < j$, we have*

$$c_j \leq c_i + \sum_{l=i}^{j-1} h_{il}. \tag{23}$$

Proof. By (21), we have $c_i + \sum_{l=i}^{j-1} h_{ll} \geq c_j$. By condition (1), $h_{il} \geq h_{ll}$ for all l, $i \leq l \leq j-1$. Thus, $c_i + \sum_{l=i}^{j-1} h_{il} \geq c_j$. □

Theorem 3. *There exists an optimal solution Ω^* to problem (CEP1) where, if $i < j$ are two expansion periods in the solution, then,* (a) *$m^*_{ik} = 0$ for all k, $k \geq j$;* (b) *If $m^*_{jt} = d_t$ for some t, $1 \leq t < j$, then $m^*_{ik} = 0$ for all k, $t \leq k \leq j$.*

Proof. To show (a), note that by (22), $o_{it} \geq o_{jt}$ for all t, $t \geq j$. By (23) and (1), we have $c_i + \sum_{l=i}^{t-1} h_{il} \geq c_j + \sum_{l=j}^{t-1} h_{il} \geq c_j + \sum_{l=j}^{t-1} h_{jl}$. We see that the variable cost to satisfy unit demand in a period $t \geq j$ from acquisition in period i is no less than that from period j. (*a*) easily follows.

The proof of (b) is similar to that of Theorem 2. We note that in the proof of Theorem 2, Cases 1–3 are no longer relevant. Cases 4–6 use only conditions (1)–(3) and do not use condition (4). □

With property (a) of Theorem 3, there are optimal solutions to (CEP1) that satisfy the so-called *zero-inventory property* (*ZIP*), which was made famous by Wagner and Whitin (1958). In these solutions (satisfying ZIP), there is no unused capacity from earlier acquisition in any expansion period. Thus, these optimal solutions to (CEP1) satisfy the following property, which is a slight modification PP in §3:

PP1. (i) *The capacity acquisition in each expansion period is used to satisfy all demands from a consecutive number of periods;* (ii) *Suppose $i < j$ are two expansion periods, and $k(i)$ and $k(j)$ are the smallest indexed periods whose demands are met by acquisition from periods i and j, respectively. We have $k(i) < k(j)$, $i \geq k(i)$, and $j \geq k(j)$.*

Based on PP1, we use the following simpler DP recursion to solve (CEP1). Setting $V(0) = 0$, we have

$$V(j) = \min_{1 \leq i \leq k \leq n} \{V(i-1) + T_k(i, j)\}. \tag{24}$$

Similar to the complexity analysis for recursion (12), it is easy to see that the above DP recursion can be implemented in $O(n^2)$ time. Again, this complexity is the lowest possible for problem (CEP1).

6. Conclusion

This paper presents a dynamic capacity expansion problem with capacity shortage allowed. Our model generalizes the classical CWW model by including AD holding, operating and shortage costs. Our model also extends the two polynomially solvable instances (Rajagopalan 1992, Jones et al. 1995) of CEPs in the literature by allowing deferred capacity expansion. We explore special structures of the proposed model and solve it with an efficient DP algorithm which has the lowest possible computational complexity.

Possible extension of this paper would be to allow arbitrary demand and to consider the case where the useful life of a capacity could be a decision variable (e.g., capacity disposal/replacement). To include these issues in our model, however, would make the model difficult to solve. A more general model is currently under investigation.

Appendix

Proof of Theorem 2. Suppose $i < j$ are two acquisition periods in an optimal solution Ω^+ to problem (CEP). Suppose $m_{jt}^+ = d_t$ for some t, $1 \le t < n$, but $m_{ik}^+ = d_k$ for some k, $t < k \le n$. We construct a new feasible solution Ω^* where demand d_k is satisfied by acquisition in period j, instead of by acquisition in period i as in Ω^+. We now consider the following six cases:

Case 1. $i < j \le t < k$.

Assume without loss of generality that k is the only period with $k > t$ and $m_{ik}^+ = d_k$. Thus, $y_{il}^+ = d_k$ for all l, $t \le l < k$. We have

$$V(\Omega^*) - V(\Omega^+) \le c_j d_k + \sum_{l=t}^{k-1} g_{jl} + \sum_{l=j}^{k-1} h_{jl} d_k + o_{jk} d_k - c_i d_k - \sum_{l=t}^{k-1} g_{il} - \sum_{l=i}^{k-1} h_{il} d_k - o_{ik} d_k.$$

Since $m_{jt}^+ = d_t$ in optimal solution Ω^+, we must have

$$c_j + \sum_{l=j}^{t-1} h_{jl} + o_{jt} \le c_i + \sum_{l=i}^{t-1} h_{il} + o_{it}; \tag{A1}$$

otherwise, we can obtain a new easible solution to (CEP) where d_t is satisfied by acquisition in period i, instead of acquisition in period j. The new solution would have an objective function value *strictly* less than $V(\Omega^+)$, which would be a contradiction to the fact that Ω^+ is an optimal solution. Note that by (1), $\Sigma_{l=t}^{k-1} g_{jl} \le \Sigma_{l=t}^{k-1} g_{il}$ and $\Sigma_{l=t}^{k-1} h_{jl} \le \Sigma_{l=t}^{k-1} h_{il}$. By (A1) we have

$$V(\Omega^*) - V(\Omega^+) \le o_{jk} d_k - o_{ik} d_k + o_{it} d_k - o_{jt} d_k.$$

By (4), we have

$$o_{ik} - o_{it} = \sum_{l=t}^{k-1} (o_{i,l+1} - o_{il}) \ge \sum_{l=t}^{k-1} (o_{j,l+1} - o_{jl}) = o_{jk} - o_{jt},$$

which implies $V(\Omega^*) - V(\Omega^+) \le 0$.

Case 2. $i \le t < j < k$.

Again, we assume without loss of generality that k is the only period with $k > j$ and $m_{ik}^+ = d_k$. Thus, $y_{il}^+ = d_k$ for all l, $j \le l < k$. Similar to Case 1, we have

$$V(\Omega^*) - V(\Omega^+) \le c_j d_k + \sum_{l=j}^{k-1} g_{jl} + \sum_{l=j}^{k-1} h_{jl} d_k + o_{jk} d_k - c_i d_k - \sum_{l=j}^{k-1} g_{il} - \sum_{l=i}^{k-1} h_{il} d_k - o_{ik} d_k.$$

By (1), we have $\Sigma_{l=j}^{k-1} g_{jl} \le \Sigma_{l=j}^{k-1} g_{il}$ and $\Sigma_{l=j}^{k-1} h_{jl} \le \Sigma_{l=j}^{k-1} h_{il}$. Thus,

$$V(\Omega^*) - V(\Omega^+) \le c_j d_k + o_{jk} d_k - c_i d_k - \sum_{l=i}^{j-1} h_{il} d_k - o_{ik} d_k.$$

Similar to the argument for (A1), we have

$$c_j + \sum_{l=t}^{j-1} p_{tl} + o_{jj} \le c_i + \sum_{l=i}^{t-1} h_{il} + o_{it}.$$

Hence,

$$V(\Omega^*) - V(\Omega^+) \le \left(o_{jk} - o_{ik} + o_{it} - o_{jj} - \sum_{l=t}^{j-1} h_{il} d_k - \sum_{l=t}^{j-1} p_{tl}\right) d_k.$$

Since $i \le t < j$, by (3) $o_{it} \le o_{ij}$. Thus by (4),

$$V(\Omega^*) - V(\Omega^+) \le (o_{jk} - o_{ik} + o_{ij} - o_{jj}) d_k \le 0.$$

Case 3. $t < i < j < k$.

Similar to Case 2, we have

$$V(\Omega^*) - V(\Omega^+) \le \left(c_j + o_{jk} - c_i - \sum_{l=i}^{j-1} h_{il} - o_{ik}\right) d_k.$$

Again, similar to the argument for (A1), we have

$$c_j + \sum_{l=t}^{j-1} p_{tl} + o_{jj} \le c_i + \sum_{l=t}^{i-1} p_{tl} + o_{ii}.$$

Hence,

$$V(\Omega^*) - V(\Omega^+) \le (o_{jk} - o_{ik} + o_{ii} - o_{jj}) d_k.$$

By (3), $o_{ij} \ge o_{ii}$. Thus by (4),

$$V(\Omega^*) - V(\Omega^+) \le (o_{jk} - o_{ik} + o_{ij} - o_{jj}) d_k \le 0.$$

Case 4. $i \le t < k < j$.

In this case, we have

$$V(\Omega^*) - V(\Omega^+) \le \left(c_j + \sum_{l=k}^{j-1} p_{kl} + o_{jj} - c_i - \sum_{l=i}^{k-1} h_{il} - o_{ik}\right) d_k.$$

Similar to the argument for (A1), we have

$$c_j + \sum_{l=t}^{j-1} p_{tl} + o_{jj} \le c_i + \sum_{l=i}^{t-1} h_{il} + o_{it}.$$

Thus,

$$V(\Omega^*) - V(\Omega^+) \le \left(\sum_{l=k}^{j-1} p_{kl} + o_{it} - \sum_{l=t}^{j-1} p_{tl} - o_{ik}\right) d_k.$$

By (2), we have $p_{tl} \ge p_{kl}$ for $k \le l \le j - 1$. Also by (3), $o_{it} \le o_{ik}$. We clearly have $V(\Omega^*) - V(\Omega^+) \le 0$.

Case 5. $t < i \le k < j$.

Similar to Case 4, we have

$$V(\Omega^*) - V(\Omega^+) \le \left(c_j + \sum_{l=k}^{j-1} p_{kl} + o_{jj} - c_i - \sum_{l=i}^{k-1} h_{il} - o_{ik}\right) d_k.$$

Again, similar to the argument for (A1), we have

$$c_j + \sum_{l=t}^{j-1} p_{tl} + o_{jj} \le c_i + \sum_{l=t}^{i-1} p_{tl} + o_{ii}.$$

Thus,

$$V(\Omega^*) - V(\Omega^+) \le \left(\sum_{l=k}^{j-1} p_{kl} + o_{ii} - \sum_{l=i}^{k-1} p_{tl} - o_{ik}\right) d_k.$$

By (3), $o_{ii} \le o_{ik}$. Also, by (2), $p_{tl} \ge p_{kl}$ for $k \le l \le j - 1$. We have $V(\Omega^*) - V(\Omega^+) \le 0$.

Case 6. $t < k < i < j$.

In this case, we have

$$V(\Omega^*) - V(\Omega^+) \le \left(c_j + \sum_{l=i}^{j-1} p_{kl} + o_{jj} - c_i - o_{ii}\right) d_k.$$

Similar to the argument for (A1), we have

$$c_j + \sum_{l=t}^{j-1} p_{tl} + o_{jj} \le c_i + \sum_{l=t}^{i-1} p_{tl} + o_{ii}.$$

Thus, by (2) we have

$$V(\Omega^*) - V(\Omega^+) \le \left(\sum_{l=i}^{j-1} p_{kl} - \sum_{l=i}^{j-1} p_{tl}\right) d_k \le 0. \quad \square$$

An Example for the Remark in Section 5. To formulate problem (CEP1) as an instance of UFLP, we define n potential plant locations indexed by j and n customers indexed by i. The fixed cost for establishing plant at location j is set to be the fixed cost incurred in an expansion period j. The variable cost c_{ij} of satisfying demands from customer i by supply from the plant at location j is defined as follows: for $1 \le j \le n$,

$$c_{ij} = \begin{cases} c_j + o_{jj}, & \text{if } i = j; \\ c_j + o_{ji} + \sum_{t=j}^{i-1} h_{jt}, & \text{if } j < i; \\ c_j + o_{jj} + \sum_{t=i}^{j-1} p_{it}, & \text{if } i < j. \end{cases} \tag{A2}$$

The structured UFLPs used by Rajagopalan (1992) and Jones et al. (1995) to solve their CEPs both satisfy the following condition (see p. 530 in Rajagopalan 1992 and p. 662 in Jones et al. 1995): for any pair of potential plant locations $j_1 < j_2$,

$$\text{either } c_{ij_1} \le c_{ij_2}, \quad \text{for all } i, \quad 1 \le i \le n, \quad \text{or}$$
$$c_{ij_1} \ge c_{ij_2}, \quad \text{for all } i, \quad 1 \le i \le n. \tag{A3}$$

The following example shows that that there are instances of the problem (CEP1) that do **not** satisfy condition (A3) if they are formulated as UFLPs.

Example. Consider a 3-period problem where $d_t = 1$, $1 \le t \le 3$. $f_1 = f_2 = f_3 = 0$, $c_1 = 5$, $c_2 = 6$, and $c_3 = 1$. The holding costs are $g_{it} = 0$ and $h_{it}(y) = 2y$, $1 \le i \le t \le 3$. The shortage costs are $p_{11} = 4$, $p_{12} = 8$, and $p_{22} = 4$. The operating costs are $o_{11} = 100$, $o_{12} = 105$, $o_{13} = 110$, $o_{22} = 100$, $o_{23} = 105$, and $o_{33} = 100$. If we formulate the above problem as an UFLP using (A2), we have: $c_{11} = 105$, $c_{21} = 112$, $c_{31} = 119$; $c_{12} = 110$, $c_{22} = 106$, $c_{32} = 113$; and $c_{13} = 113$, $c_{23} = 105$, $c_{33} = 101$. Condition (A3) is clearly violated.

References

Aggarwal, A., J. Park. 1993. Improved algorithm for economic lot size problems. *Oper. Res.* **41** 549–571.

——, M. M. Klawe, S. Moran, P. W. Shor, R. Wilber. 1987. Geometric applications of a matrix-searching algorithm. *Algorithmica* **2** 195–208.

Bazaraa, M. S., C. M. Shetty. 1979. *Nonlinear Programming—Theory and Algorithms*. John Wiley & Sons, New York.

Chand, S., T. McClurg, J. Ward. 2000. A model for parallel machine replacement with capacity expansion. *Eur. J. Oper. Res.* **121** 519–531.

Eppstein, D. 1990. Sequence comparison with mixed convex and concave costs. *J. Algorithms* **11** 85–101.

Galil, Z., K. Park. 1990. A linear-time algorithm for concave one-dimensional dynamic programming. *Inform. Processing Lett.* **33** 309–311.

Hsu, V. N., T. J. Lowe. 2001. Dynamic economic lot size models with period-pair-dependent backorder and inventory costs. *Oper. Res.* **49** 316–321.

——, ——, A. Tamir. 1997. Structured p-facility location problems on the line solvable in polynomial time. *Oper. Res. Lett.* **21** 159–164.

Jones, P. C., T. J. Lowe, G. Muller, N. Xu, Y. Ye, J. L. Zydiak. 1995. Specially structured uncapacitated facility location problems. *Oper. Res.* **43** 661–669.

Klincewicz, J., H. Luss, C. S. Yu. 1988. A large-scale multilocation capacity planning model. *Eur. J. Oper. Res.* **34** 178–190.

Krarup J., O. Bilde. 1977. Plant location, set covering and economic lot size: An $O(mn)$-algorithm for structured problems. L. Collatz et al., eds. *Optimierung Bei Graphentheoretischen und Ganzzahligen Probleme.* ISNM 36, Birkhäuser Verlag, Basel, 155–179.

Lee, S. B., H. Luss. 1987. Multifacility-type capacity expansion planning: Algorithms and complexities. *Oper. Res.* **35** 249–253.

Li, S., D. Tirupati. 1994. Dynamic capacity expansion problem with multiple products: Technology selection and timing of capacity additions. *Oper. Res.* **42** 958–976.

Luss, H. 1979. A capacity expansion model for two facility types. *Naval Res. Logist. Quart.* **26** 291–303.

——. 1980. A network flow approach for capacity expansion problem with two facility types. *Naval Res. Logist. Quart.* **27** 597–608.

——. 1982. Operations research and capacity expansion problems: A survey. *Oper. Res.* **30** 907–947.

——. 1983. A multifacility capacity expansion model with joint expansion set-up costs. *Naval Res. Logist. Quart.* **30** 97–111.

Rajagopalan, S. 1992. Deterministic capacity expansion under deterioration. *Management Sci.* **38** 525–539.

——. 1998. Capacity expansion and equipment replacement: A unified approach. *Oper. Res.* **46** 846–857.

——, A. Soteriou. 1994. Capacity acquisition and disposal with discrete facility sizes. *Management Sci.* **40** 903–917.

——, M. R. Singh, T. E. Morton. 1998. Capacity expansion and replacement with uncertain technological breakthroughs. *Management Sci.* **44** 12–30.

Veinott, Jr., A. F. 1969. Minimum concave cost solution of Leontief substitution models of multifacility inventory systems. *Oper. Res.* **17** 262–291.

Wagner, R., T. M. Whitin. 1958. Dynamic version of the economic lot size model. *Management Sci.* **5** 89–96.

The consulting Senior Editor for this manuscript was Suresh Chand. This manuscript was received on June 11, 2001, and was with the author 95 days for 5 revisions. The average review cycle time was 29 days.