Optimal Capacity in a Coordinated Supply Chain

Xiuli Chao,[1] Sridhar Seshadri,[2] Michael Pinedo[2]

[1] *Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, Michigan 48109–2117*

[2] *Stern School of Business, New York University, New York, New York 10012*

Received 12 September 2004; revised 18 August 2007; accepted 2 November 2007
DOI 10.1002/nav.20271
Published online 9 January 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: We consider a supply chain in which a retailer faces a stochastic demand, incurs backorder and inventory holding costs and uses a periodic review system to place orders from a manufacturer. The manufacturer must fill the entire order. The manufacturer incurs costs of overtime and undertime if the order deviates from the planned production capacity. We determine the optimal capacity for the manufacturer in case there is no coordination with the retailer as well as in case there is full coordination with the retailer. When there is no coordination the optimal capacity for the manufacturer is found by solving a newsvendor problem. When there is coordination, we present a dynamic programming formulation and establish that the optimal ordering policy for the retailer is characterized by two parameters. The optimal coordinated capacity for the manufacturer can then be obtained by solving a nonlinear programming problem. We present an efficient exact algorithm and a heuristic algorithm for computing the manufacturer's capacity. We discuss the impact of coordination on the supply chain cost as well as on the manufacturer's capacity. We also identify the situations in which coordination is most beneficial. © 2008 Wiley Periodicals, Inc. Naval Research Logistics 55: 130–141, 2008

Keywords: optimal capacity; supply chain; coordination; dynamic programming; inventory systems

1. INTRODUCTION

In both the periodic and the continuous review stochastic inventory models that appear in the literature, it is standard to assume that the manufacturer either has an extremely large capacity or that they can increase or decrease their capacity at will at a negligible cost. In fact, this assumption is so strongly ingrained in the analysis that even if the manufacturer experiences difficulties in production (for example, due to the shortage of materials or due to a breakdown), the manufacturer is expected upon resumption of production not only to make up for the backlog immediately but also to restore the inventory to the prescribed levels at the very next shipment.

However, we know from the aggregate production planning literature that when overtime as well as undertime are allowed, there is an optimal rate at which production should be organized. In addition, there is anecdotal evidence that well-managed firms increase and decrease their capacity to match demand. Such fine tuning is achieved by resorting to overtime, short term subcontracting, capacity sharing with other manufacturers, etc. Presumably in these firms the optimal capacity is first chosen and adjustments through overtime and undertime are made from period to period in response to changes in the mix and the volume of demand. With increased visibility provided by collaborative planning technologies, as adopted by Cisco, Dell, and other companies, we argue that the manufacturer can plan and execute such short-term changes in capacity nowadays more easily than was possible in the past. However, as a consequence, suppliers are expected to fill orders completely even though their capacity is finite. This is in contrast to traditional production–inventory models in which the manufacturer is assumed to have the option to backlog an order, for example, see Holt et al. [12]. Therefore, in this article we focus on the case where the production facility has no option but to fill the entire order. We seek the optimal rate of production with the assumption that this rate can be deviated from, but increasing or decreasing this rate requires an effort as well as more expensive resources.

Several researchers have noted that joint capacity and inventory planning can reduce costs in a supply chain. The queueing theoretic work in this regard is summarized succinctly in Buzacott and Shanthikumar [4]. In contrast to our work, capacities in queueing network models for manufacturing systems are considered fixed at least for the short term. There are many articles that deal with capacity planning problems that consider stochastic demand and allow overtime; for example, see the survey in [24]. These studies do not infer or use the structure of the optimal coordinated

Correspondence to: Xiuli Chao (xchao@umich.edu)

retailer ordering policy to determine the optimal capacity. Production–inventory models with capacity constraints have attracted increased attention since the work of Federgruen and Zipkin [6, 7]. Models with stationary demand have been studied also by Tayur [23], Ciarallo et al. [5], and Glasserman and Tayur [8–10]. Kapuscinski and Tayur [14] study the periodic demand case. A survey of the work done in this area until 1999 can be found in Kapuscinski and Tayur [15]. In these articles, capacity is assumed to be given. Moreover, it has been shown that the base stock policy is optimal for the stationary case as well as for the periodic demand case.

The articles of Bradley and Arntzen [3] and Bradley and Glynn [2] are exceptions to this stream of work. They examine how the capacity and inventory decisions can be optimized jointly. The contributions in these two articles are to treat the capacity as well as the base stock level as decision variables. In the first article, overtime is allowed and capacity can thus be changed at a cost. In the second article it is presumed that the firm uses a base-stock policy to manage its inventory. The authors conclude that adding capacity and lowering inventory may be beneficial. In contrast to these articles, we assume that orders have to be filled completely, there are costs to changing the production rate and we determine the jointly optimal ordering policy and the optimal capacity. Parker and Kapuscinski [19] consider a two-stage serial system with finite capacity at both stages. Under some conditions on the capacity levels at the two stages, they show that the optimal inventory control strategy is modified echelon-base stock policy.

Another related article is that of Huggins and Olsen [13]. They consider a two-stage supply chain where ordering of stage one must be satisfied by stage two, using emergency shipping at a setup cost and a higher unit cost if needed, the optimal ordering policy at the first stage is determined by two numbers. There is infinite capacity at stage two and the leadtime is zero. They transform their problem using echelon inventory levels and obtain the optimal policy for each stage to minimize total discounted cost over an infinite horizon. The centralized optimization model and ordering policy at stage one in Huggins and Olsen are similar to ours after stage two fixes its production strategy to stationary base-stock level equal to the capacity. However, for the Huggins and Olsen result to hold, the demand distribution has to be assumed to be logconcave.

Two of the questions we intend to address in our model are: (i) What is the retailer's optimal ordering policy that minimizes the supply chain cost when there are overtime and undertime costs for the manufacturer? (ii) How to determine the optimal capacity under coordination and how does it compare to the decentralized case? The results in this article specify the structure of the optimal policy in answer to the first question. As for the second question, unlike the decentralized case for which there is a simple formula for the optimal capacity, the optimal centralized capacity requires the solution of a nonlinear program. We provide an efficient algorithm to solve the nonlinear program. Our algorithm uses a decomposition approach based on the structure of the control policy. It breaks up the search for the optimal capacity into a search for two numbers, namely an upper and a lower threshold limit; and a search for the capacity that depends on these limits. This solution procedure, even though it produces the optimal solution, does not yield additional insights into when coordination is beneficial. Therefore, we make a key simplification to obtain approximate formulae for the threshold limits. This procedure yields a heuristic method for computing the optimal coordinated capacity. The heuristic is efficient and yields results that are close to optimal. The arguments leading to the development of the heuristic provide combinations of problem parameters that might influence the benefit of coordination the most. These insights are used to design and carry out numerical experiments to study the extent of cost reduction from coordinating the ordering and capacity decisions. Finally, even though we are unable to formally establish a relationship between the centralized and decentralized optimal capacity, our extensive numerical studies suggest that the optimal capacity is lower under coordination.

The next section presents the model. Sections 3 and 4 study the cases without and with coordination, respectively. Section 5 discusses computational issues, and a heuristic computational method is proposed. Section 6 presents some numerical examples. The article concludes with a discussion in Section 7. Finally, some proofs are provided in the Appendix.

2. THE MODEL

Consider a two-stage supply chain that consists of a producer and a retailer. The retailer uses a periodic review system, with the length of review period being, without loss of generality, one. The customer demands during the review periods $D_1, D_2, \ldots$ are independent and identically distributed (i.i.d.) random variables. For simplicity we assume that the demands are continuous random variables with a common distribution function, and let D be a generic one period demand. Demand not filled at the retailer is backlogged. The cost of carrying inventory is h and the cost of backlogged inventory is b. These costs are assessed on the quantities (inventory and backlog) at the end of the review period. The retailer's problem is to find the optimal inventory control policy that minimizes the infinite horizon total discounted cost.

The retailer orders from the producer each period. The items are supplied by the producer who sets up the production upon receipt of the order and manufactures the quantity ordered. A key assumption in this article is that the producer must fill each order in its entirety in a finite time τ, either by using their own capacity or by outsourcing when necessary, see Manne [18] and Parker and Kapuscinski [19] for a similar assumption as well as Van Mieghem [24] for a discussion on modeling unsatisfied demand and capacity shortages. The leadtime τ includes the time to manufacture and ship the product. As discussed in the Introduction, the assumption of constant order fulfillment leadtime implies that capacity is infinite. The reality is that the producer incurs costs to provide such service. The combination of the assumptions that demand at the retailer may be backlogged whereas demand at the producer must be satisfied reflects situations in which the customer is willing to wait, there is cost incurred by the retailer due to backlogging, such as loss of goodwill and loss of future sales, and the manufacturer serves multiple retailers and offers a standard delivery contract including a fixed leadtime of τ and a fixed price.

We shall also assume that the manufacturer cannot build up inventory to satisfy future orders. For example, this could apply to products that have limited shelf-life or where freshness matters. The variable cost of production is c. In addition, the manufacturer incurs a capacity utilization related cost of production that is a convex function of the order quantity. This cost is based on the production capacity, a, of the supplier and the quantity ordered. The total cost of producing a quantity x, denoted as $C(x)$, is given by

$$C(x) = cx + c_{\mathrm{u}}(a-x)^+ + c_{\mathrm{o}}(x-a)^+, \tag{1}$$

where $x^+ = \max\{x, 0\}$. Thus, deviations of the production rate from the capacity, which is a function of the cost of resources used as well as the planned availability of resources, are costly. In this expression the coefficients, c_{u} and c_{o} stand for the cost of under- and over-utilization, respectively. When the unused capacity can be used for secondary work, c_{u} can be negative and the only assumption we need to make is $c_{\mathrm{u}} + c_{\mathrm{o}} > 0$. The cost due to under-utilization could arise even when there is an alternative spot market for the product or when it is possible to sell to other customers. In both cases, production and logistics related costs, namely, the cost of fulfilling the open market demand or the cost of switching production and fulfillment to another customer, might be higher due to the unplanned nature of work.

The manufacturer's problem is to choose the production capacity, a, to minimize her infinite horizon expected total discounted cost. We assume that the value of a is determined initially at time zero and remains unchanged from then on. We also assume that there is a cost of capacity C_a per unit of capacity per review period. In general, this comprises all costs including the fixed charges allocated to the production of the orders from the retailer, direct and indirect operating expenses, and maintenance expenses.

In the decentralized problem the retailer and the manufacturer determine their optimal strategy without knowing any cost information of the other party, which we call a system without coordination. In the centralized problem a central planner, with information from both the manufacturer and retailer, determines the optimal policy for both entities. In the following two sections we discuss the solution for the manufacturer and retailer without and with coordination, respectively.

3. THE CASE WITHOUT COORDINATION

Suppose the manufacturer charges the retailer a price c per unit, who then sells to the market at the price of p per unit. In the case without coordination the retailer manages a classical inventory problem and intends to minimize the total discounted holding and backorder costs. Let the demand over $\tau + 1$ periods be $D^{\tau+1}$ with cumulative distribution function $F(\cdot)$. The discounting factor is α. The optimal policy for the retailer is a base-stock policy with the base stock level S determined by (see for example, [1])

$$F(S) = \frac{b - (1-\alpha)c}{h+b}. \tag{2}$$

This result states that, at the beginning of each period always raise the inventory position, which is inventory on hand plus inventory on order minus backorders, to S. As a result, the demands for the retailer transfer from period to period to the manufacturer. Therefore, the manufacturer's optimization problem in this case can be recast nicely. The base stock policy essentially forces the manufacturer to manufacture the demand during the previous review period. The cost to the manufacturer is therefore given by

$$c_{\mathrm{u}}E[(a-D)^+] + c_{\mathrm{o}}E[(D-a)^+] + cE[D].$$

The manufacturer can optimally choose the value of a. The problem is to minimize

$$C_a a + c_{\mathrm{u}}E[(a-D)^+] + c_{\mathrm{o}}E[(D-a)^+] + cE[D].$$

The objective function is convex in a. Thus, the optimal solution is obtained by setting the derivative equal to zero, i.e.,

$$C_a + c_{\mathrm{u}}P\{D \le a\} - c_{\mathrm{o}}(1 - P\{D \le a\}) \equiv 0.$$

The optimal solution is given by solving for a in

$$P\{D \le a\} = \max\left\{\frac{c_{\mathrm{o}} - C_a}{c_{\mathrm{o}} + c_{\mathrm{u}}}, 0\right\}. \tag{3}$$

This expression suggests that for firms that find it very expensive to add capacity (high C_a) but who also wish to stay responsive to the retailer's needs, may find it worthwhile to simply subcontract the entire order quantity. For example, many firms moved to outsourcing their needs in the 1990's, see for example, [11]. On the other hand, firms that are adept in adding capacity for a family of products should be producing them.

What is nice about this solution is that the capacity decision is independent of the base stock level used by the retailer. Thus, we conclude that as long as the retailer continues to use a base stock policy the optimal capacity of the manufacturer remains the same! But, this is not the optimal capacity if the retailer is willing to relax the base stock policy. The question is whether this is an attractive proposition. What effect does the use of other order policy by the retailer have on the manufacturer's capacity?

4. THE CASE WITH COORDINATION

We now consider the case where a manager makes decisions for both the retailer and the manufacturer. The objective is to minimize the infinite horizon total discounted cost for the entire supply chain. Thus, the optimal policy might be to reduce the order quantity when the backlog is too high or to raise it when the backlog is low. The intriguing question is whether the optimal capacity will be smaller or will it be larger under joint optimization? Whom does the policy benefit? How should coordination mechanisms be designed to allocate the benefit?

To proceed, we start the analysis by considering the case with a finite horizon of N periods numbered from N to 1, and discounted cost. The decision maker wishes to minimize the N-period expected value of the discounted total cost. The total cost includes the cost of producing the item and the cost of backlog and carrying inventory. We first characterize the optimal coordinated ordering policy for the retailer. That is, for given a we want to find the optimal solution to the problem below. Assume that a central decision maker observes an inventory position of x units at the beginning of period n. Let $V_n(x)$ be the value function for the n-period problem starting at the beginning of period n. Again let $0 < \alpha < 1$ be the discount factor. Thus, $V_0(\cdot)$ is computed recognizing the fact that the retailer has no recourse for placing further orders. Then

$$V_n(x) = \min_{z \geq 0}\{C(z) + \alpha^\tau H(x+z) + \alpha E[V_{n-1}(x+z-D)]\},$$

where $C(\cdot)$ is given by (1), and $H(\cdot)$ is the one-period cost function, including holding and shortage cost for the retailer:

$$H(x) \equiv hE[(x - D^{\tau+1})^+] + bE[(D^{\tau+1} - x)^+].$$

Recall that $D^{\tau+1}$ stands for the demand over $\tau + 1$ periods. The cost $H(x + z)$ equals the cost $\tau + 1$ periods later: All orders that are currently outstanding will be received, including the present order of z, in τ periods. Therefore, the ending inventory in the $\tau + 1$-st period will be $x + z - D^{\tau+1}$. By induction, we can show that $V_n(x)$ is convex in x for all n (see, for example, Sobel [22]).

The following result may be of independent interest. Its proof, along with proofs for other results, is given in the Appendix. Let f and g be functions from R (the real line) to R.

LEMMA 1: Suppose $g(x)$ is a strictly increasing (decreasing) function. Let x_1^* and x_2^* be the finite minimizers of $f(x)$ and $g(x) + f(x)$, respectively. Then $x_1^* < x_2^*$ ($x_1^* > x_2^*$).

Let

$$G_n(y) \equiv cy + \alpha^L H(y) + \alpha E[V_{n-1}(y - D)],$$

and let U_n and L_n respectively minimize the two convex functions $-c_\mathrm{u}y + G_n(y)$ and $c_\mathrm{o}y + G_n(y)$. It follows from the Lemma 1 (by letting $f(y) = -c_\mathrm{u}y + G_n(y)$ and $g(y) = (c_\mathrm{o} + c_\mathrm{u})y)$ that $U_n > L_n$.

THEOREM 1: The optimal ordering strategy for the n-period problem is determined by two parameters L_n and $U_n (L_n < U_n)$ in such a way that

i. if $x \geq U_n$, then do nothing;
ii. if $U_n - a < x \leq U_n$, then order up to U_n (i.e., order $U_n - x$);
iii. if $L_n - a \leq x \leq U_n - a$, then order exactly a, and
iv. if $x \leq L_n - a$, then order up to L_n (i.e., order $L_n - x$).

Note that in (i) and (ii) undertime costs are incurred since the order quantity is less than a. In (iv) overtime costs are incurred since more than a is ordered. In (iii) neither undertime nor overtime costs are incurred.

Once we obtain the form of the optimal ordering strategy for the retailer, we are ready to analyze the optimal capacity for the manufacturer, i.e., a. Clearly, the optimal ordering strategy for the retailer, that is the two parameters, depends on the capacity level a. A moment of reflection shows that under the optimal ordering policy, the order process of the retailer, i.e., the demand process seen by the manufacturer, is a Markov process with transition rates that depend on the capacity level a. We can easily show that the cost function is convex in a. Thus finding the optimal a is straightforward. The remaining work will be to characterize the optimal capacity level a and compare it with the case when the retailer uses a base stock policy.

Since the value function V_n clearly depends on a, we shall also write it as $V_n(x, a)$.

THEOREM 2: The value function $V_n(x, a)$ is jointly convex in (x, a).

The result can then be extended to the infinite horizon case with discounted cost. To see why this is true, note that since the cost structure is nonnegative, we are dealing with a negative dynamic programming problem, and as a result, the method of successive approximation can be used to obtain the optimal value function for the infinite horizon problem Puterman [20]. This shows that $V_n(x, a)$ converges to $V(x, a)$, the discounted cost function of the infinite horizon problem. Since V_n is jointly convex in (x, a), $V(x, a)$ is also jointly convex in (x, a). Applying another result from negative dynamic programming, which states that if a stationary policy satisfies the optimality equation, it is the optimal policy for the infinite horizon problem, we conclude that the optimal policy for the infinite horizon problem is determined by two numbers, L, and U. We state this as a theorem.

THEOREM 3: The optimal inventory policy for the retailer for the infinite horizon problem with discounted cost is determined by two parameters L and U. The minimum value function $V(x, a)$ is jointly convex in (x, a).

Therefore, in the case of coordination the optimal inventory control policy for the retailer is no longer base stock, but is determined by two numbers L and U. A natural question is how to design a coordination mechanism under which the retailer will follow such an ordering policy.

To design such a mechanism, let γ_1 and γ_2 be the total discounted profits of the manufacturer and the retailer in the case with no coordination, γ be the total discounted profit of the supply chain with coordination. Note that the γ_1 and γ_2 are the total discounted revenues received by the manufacturer and retailer, by selling the item to the retailer and market respectively, subtract their minimum total discounted costs under the policies in Section 3; and γ is the total discounted revenue received from the market subtract the total supply chain discounted cost calculated in Section 4. Clearly, they satisfy

$$\gamma \geq \gamma_1 + \gamma_2.$$

Write

$$C(x) = cx + c_o(x-a)^+ + c_u(a-x)^+ = -A + Bx + s(a-x)^+,$$

where $A = ac_o$, $B = c + c_o$, and $s = c_o + c_u$. Since adding a constant κ to $C(x)$ will not change the optimal inventory control policy, this implies that the manufacturer could use a two-part tariff to coordinate the chain: For each unit of time the retailer is paid a fixed amount $A + \kappa$, each unit is charged a unit rate of B, and the retailer is penalized for ordering less than the capacity–for each unit of ordering below a, the retailer is charged an additional $s = c_o + c_u$, then the retailer would follow the optimal strategy as specified.

Since by choosing the value of κ the total discounted profit of the retailer can be any value, it is possible to reach any allocation of the total discounted profit γ between the manufacturer and the retailer by choosing the appropriate value of κ; in particular, for the retailer to receive a total discounted profit more than γ_2, and manufacturer receives a total discounted profit more than γ_1. Such a choice of κ will induce the retailer and the manufacturer to participate in the contract specified, and for the retailer to order following the optimal policy described in Theorem 3. The actual selection of κ, which determines the allocation of supply chain wide total discounted profit γ, will be determined by negotiation and relative market power of each partner.

5. COMPARISON, COMPUTATION, AND HEURISTIC

We now analyze the optimal capacity in the two models. For the purpose of comparisons, we consider average cost criterion for both models.

Recall that for the case without coordination, the demand for the manufacturer is the same as the customer demand, thus the optimal capacity is determined by

$$P\{D > a\} = \min\left\{\frac{C_a + c_u}{c_u + c_o}, 1\right\}.$$

As observed earlier, the optimal capacity is independent of the base stock level of the retailer. The capacity level for the manufacturer is always at a, the inventory position at the beginning of a period for the retailer is S, see (2) and (3). Thus the average cost per period is

$$c(S, a) = (C_a a + c_u E[(a-D)^+] + c_o E[(D-a)^+] + cE[D]) \\ + (bE[(D^{\tau+1} - S)^+] + hE[(S - D^{\tau+1})^+]),$$

where both a and S are determined, separately, by solving newsvendor problems.

For the case with coordination, let x be the inventory position at the beginning of a review period after the ordering decision is made. Let the demand during a review period be D. Given x, D, and the control parameters L and U, the order quantity z will be:

$$z = L - x + D \quad \text{when } x - D < L - a$$
$$z = U - x + D \quad \text{when } x - D > U - a$$
$$z = a \quad \text{when } L - a \leq x - D \leq U - a.$$

Note that in the first case the quantity z is greater than a and in the second case the quantity z is less than a. Thus, the probability that the retailer orders more than a is

$$P\{D > a + x - L\} \leq P\{D > a\}$$

because the inventory position at the beginning of the review period is between L and U. Similarly, the probability that the retailer orders less than a is

$$P\{D < a + x - U\} \le P\{D < a\}.$$

The two inequalities above immediately lead to the following result.

THEOREM 4: Orders are more likely to be equal to the capacity of the system in the coordinated supply chain.

We prove the following stronger result in the Appendix.

THEOREM 5: The ordering process $O_1, O_2, \ldots$ for the coordinated system and the ordering process $Y_1, Y_2, \ldots$ for the non-coordinated system satisfy

$$|O_n - a| \le |Y_n - a|, \quad n = 1, 2, \ldots. \tag{4}$$

REMARK: Notice that (assuming stationarity) $E[(O_n - a)^2] \le E[(Y_n - a)^2]$. The mean of both O_n and Y_n is $E[D]$. Therefore, the inequality can be re-written as $Var(O_n) \le Var(Y_n)$. Thus, the variance of the orders under the optimal policy is less than the variance of demand. This shows that Theorem 5 demonstrates the "anti-bullwhip effect" of the optimal policy. Typically, the inequality goes the other way and is termed the bullwhip effect—see for example Lee et al. [16,17].

Thus, on every sample path the coordinated order stream is closer to capacity than the order stream in the decentralized one. We now turn to determining the optimal parameters of the order policy. For computational purposes suppose that the demand distribution is discrete with probability mass function $d_i = P\{D = i\}$ (and if the demand is continuous, it can be discretized, as in most computational approaches, to discrete demand). Also for computational purposes, suppose the support for the demand is the finite set $[0, 1, \ldots, M]$. If this is not the case a common approach in computational Markov decision processes is to truncate the demand to a finite support for some large M, see for example Sennott [21]. Our calculations will be done for average cost.

Let $\bar{d}_i = P\{D \ge i\}$. The inventory position at the beginning of a period for the retailer is a function of $I_n - D_n$, where I_n is the stationary inventory position at the beginning of period n and D_n is the demand for period n. The process $\{I_n; n = 1, 2, \ldots\}$ is a Markov chain with state space $\{L, L+1, \ldots, U\}$. A moment of reflection shows that the stationary distribution of I_n, that is $P\{I_n = L+i\}$, is independent of L and depends only on $\Delta = U - L$. Let us consider the process $I_n - L$ with state space $\{0, 1, \ldots, \Delta\}$. The transition probabilities of $I_n - L$ are

$$p_{i,0} = P\{D \ge i + a\} = \bar{d}_{i+a}, \quad i = 0, \ldots, \Delta, \tag{5}$$

$$p_{i,j} = P\{i - D + a = j\} = d_{i-j+a}, \quad i = 0, 1, \ldots, \Delta, \quad j = 1, 2, \min\{\Delta, i + a\} - 1, \tag{6}$$

$$p_{i,\Delta} = P\{i - D \ge \Delta - a\} = \sum_{k=0}^{i+a-\Delta} d_k, \quad i = \Delta - a, \ldots, \Delta. \tag{7}$$

Recall that the policy prescribes that if the starting inventory position is less than $L - a$, it has to be replenished to L; if the starting inventory position is between $L - a$ and $U - a$, the amount a has to be ordered; and if the starting inventory position is higher than $U - a$, it has to be replenished to U. The stationary distribution, denoted by π_i, satisfies the balance equations

$$\pi_i = \sum_{j=0}^{\Delta} \pi_j p_{ji}, \quad i = 0, \ldots, \Delta, \tag{8}$$

$$\sum_{i=0}^{\Delta} \pi_i = 1. \tag{9}$$

Consider, for example, the case with $\Delta = 1$. Then

$$p_{0,0} = \bar{d}_a, \quad p_{0,1} = 1 - \bar{d}_a,$$
$$p_{1,0} = \bar{d}_{a+1}, \quad p_{1,1} = 1 - \bar{d}_{a+1}.$$

The stationary distribution is

$$\pi_0 = \frac{\bar{d}_{a+1}}{1 - d_a}, \quad \pi_1 = \frac{1 - \bar{d}_a}{1 - d_a}.$$

If the inventory position at the beginning of a period is i, then the optimal ordering quantity will exceed the capacity when demand is greater than $i + a$. Thus, the average overcapacity cost is $c_o \sum_{i=0}^{\Delta} \pi_i \sum_{j=i+a}^{M} (j - i) d_j$. Similarly, there will be a cost of underutilizing the capacity when $i - D \ge \Delta - a$. Therefore, the average under utilization cost is given by $c_u \sum_{k=0}^{i+a-\Delta} (\Delta + j - i) d_j$. The average cost for the system, for a given capacity level a, can be computed using the stationary distribution π_i by evaluating

$$\begin{aligned} f(L, U|a) = {} & c_o \sum_{i=0}^{\Delta} \pi_i \sum_{j=i+a}^{M} (j - i) d_j \\ & + c_u \sum_{j=0}^{i+a-\Delta} (\Delta + j - i) d_j \\ & + h \sum_{i=0}^{\Delta} \pi_i E_D[(i + L - D^{\tau+1})^+] \\ & + b \sum_{i=0}^{\Delta} \pi_i E_D[(D^{\tau+1} - i - L)^+], \end{aligned} \tag{10}$$

where E_D represents the expectation with respect to the demand $D^{\tau+1}$.

Hence, our optimization problem is a nonlinear program with objective function (10) and constraints (5), (6), (7), (8), and (9). The decision variables are L, U, and π_j, $j = 0, 1, \dots, \Delta$.

Note that for each given Δ, the p_{ij}'s are determined by (5), (6), and (7), and under mild conditions on the probability distribution of demand, the π_j are uniquely determined by the linear equations (8) and (9). That is, for a given Δ the optimization of L can be considered as an optimization problem with linear constraints. It is easily seen that for fixed Δ and a, the objective function (10) is convex in L, thus the optimization problem is a convex programming with linear constraints. The dependency on Δ, however, is in general not convex and exhaustive search methods will have to be used to find the optimal Δ.

Let $L(a)$ and $U(a)$ be the optimal policy for a given capacity level a. The cost function $C_a a + f(L(a), U(a)|a)$ is a convex function of a as seen from Theorem 3. Therefore, bisection search can be used to find the optimal capacity level a.

As noted, the main computational issue lies in the determination of L and U (and, in particular, U). In the Appendix we develop the following simple heuristic to compute L and U. Our numerical examples show that it performs very well.

We shall set L and U according to

$$P\{D^{\tau+1} \leq L\} = \max\left\{\frac{b - c_{\mathrm{o}}}{b + h}, 0\right\},$$
$$P\{D^{\tau+1} \geq U\} = \max\left\{\frac{h - c_{\mathrm{u}}}{h + b}, 0\right\}.$$

According to these formulas, if $b \leq c_{\mathrm{o}}$, then L should be made as small as possible, while if $h \leq c_{\mathrm{u}}$, then U should be made as large as possible. The interpretation is as follows. If $b \leq c_{\mathrm{o}}$, the shortage cost to the retailer is smaller than the cost of overtime to the manufacturer. In this case it is preferable not to exceed capacity. This in turn translates to making L as small as possible. On the other hand, if $h \leq c_{\mathrm{u}}$, then the holding cost of the retailer is smaller than the cost of undertime. In this case, we set U equal to a large value so that we rarely underload the manufacturer. Indeed, as seen from the numerical results in the next section, when these conditions are satisfied, even though the optimal L and U may not be close to the heuristic results, the optimal cost is very close to that achieved by the heuristic, see examples in the next section.

These expressions can be used to develop a heuristic algorithm to compute the optimal capacity level a. First, the values of L and U are computed using the two expressions above. Once L and U are obtained, the stationary distribution of the inventory (5), (6), (7) can be computed and the average cost, as a function of a, is obtained. Since the dependency of the cost on a is convex, bi-section search can be used to find the optimal a.

Note that the ratios, $(h + c_{\mathrm{o}})/(h + b)$ and $(h - c_{\mathrm{u}})/(h + b)$ play a role in determining L and U. This fact will be used in the design of the numerical experiments in the next section. The expressions for L and U can be plugged into (12) and (13) of the Appendix to obtain

$$-c_{\mathrm{o}} P\{I_a - U + a \leq D\} + c_{\mathrm{u}} P\{I_a - U + a \geq D\} + C_a \leq 0$$

and

$$-c_{\mathrm{o}} P\{I_a - L + a \leq D\} + c_{\mathrm{u}} P\{I_a - L + a \geq D\} + C_a \geq 0.$$

This suggests that the optimal capacity might deviate the most from the decentralized case when $U - L$ is large. Moreover, as seen above, the difference between U and L is related to the ratios $(h + c_{\mathrm{o}})/(h + b)$ and $(h - c_{\mathrm{u}})/(h + b)$.

6. NUMERICAL EXAMPLES

In this section, we evaluate the performance of the heuristic method for computing the policy parameters. We examine how the optimal capacity in the coordinated chain changes with the problem parameters. We also compare the optimal cost and capacity to those for the uncoordinated chain. On the basis of the analysis in the previous section we know that the ratios, $(h + c_{\mathrm{o}})/(h + b)$ and $(h - c_{\mathrm{u}})/(h + b)$, play an important role in determining the benefits from coordination. In our numerical analysis, the effect due to the former is more pronounced and we report only those. In addition to this, we know that inventory costs are affected by $h/(h + b)$ and the volatility of demand. The carrying cost of safety stock relative to the cost of capacity is affected by h/C_a. Therefore, we include these three parameters in our numerical study. The ranges of the four parameters in the experiments are as follows: the coefficient of variation (cv) demand is varied in the range $[0.25, 0.61]$, h/C_a in the range $[0.25, 2]$, $h/(h + b)$ in $[0.1, 0.5]$ and $(h + c_o)/(h + b)$ in $[0.25, 2]$. We modeled demand using both a negative binomial distribution and a truncated Normal distribution on $[1, \infty)$. By changing the parameters, we can obtain a range of values for cv. The insights obtained from the two distributions were very similar, so we report results only for the negative binomial distribution. The average demand is kept fixed throughout the experiments at 20.

The results are shown in Tables 1–4. We report the optimal capacity in the decentralized case, the optimal capacity in the coordinated case and the capacity given by the heuristic. We also depict the percentage cost increase with respect to the optimal cost for the centralized case and the L and $U - L$ values. The main inferences are given below.

Table 1. Change in coefficient of variation of demand; $n(1-p)/p = 20, h = 6, c_a = 4, c_u = 4, b = 30, c_o = 15$.

	Capacity			% increase in cost		Optimal		Heuristic	
cv of Demand	Optimal (decentr)	Optimal (centra)	Heuristic	Optimal (decentr)	Optimal heuristic	L	$U-L$	L	$U-L$
0.25	23	21	21	11.3	1.2	23	6	20	10
0.27	22	20	21	11.0	1.5	23	8	20	11
0.29	22	20	21	11.3	1.3	23	8	20	11
0.32	22	20	21	11.9	1.2	23	9	20	12
0.35	23	21	22	11.8	1.9	25	10	20	15
0.40	23	21	22	12.3	1.7	25	12	20	17
0.50	22	20	21	13.3	2.1	24	14	18	21
0.61	28	26	28	13.4	2.8	33	23	22	36

Table 2. Change in h/c_a; $n(1-p)/p = 20, c_a = 4, c_u = 4, b = 30, c_o = 15$.

	Capacity			% increase in cost		Optimal		Heuristic	
$\frac{h}{c\,a}$	Optimal (decentr)	Optimal (centra)	Heuristic	Optimal (decentr)	Optimal heuristic	L	$U-L$	L	$U-L$
0.25	23	20	20	24.8	8.3	27	21	21	60
0.50	23	20	19	19.0	10.9	26	13	21	60
0.75	23	20	19	15.9	9.1	25	10	21	60
1.00	23	20	19	13.7	8.7	24	9	21	60
1.25	23	20	20	12.1	1.3	24	8	21	11
1.50	23	21	21	11.3	1.2	23	6	20	10
1.75	23	20	21	10.1	0.9	23	7	20	9
2.00	23	21	21	9.8	0.7	22	6	20	8

Table 3. Change in $\frac{h}{h+b}$; $n(1-p)/p = 20, c_a = 4, c_u = 4, b = 54, c_o = 15$.

	Capacity			% increase in cost		Optimal		Heuristic	
$\frac{h}{h+b}$	Optimal (decentr)	Optimal (centra)	Heuristic	Optimal (decentr)	Optimal heuristic	L	$U-L$	L	$U-L$
0.10	23	21	20	9.8	0.9	25	6	23	9
0.18	23	21	20	6.4	0.2	24	4	23	5
0.25	23	21	20	5.2	0.3	23	3	22	4
0.31	23	21	20	4.4	0.8	22	3	22	3
0.36	23	21	20	3.6	0.5	21	3	21	3
0.40	23	21	20	3.9	0.7	21	2	21	2
0.44	23	21	20	2.9	0.8	21	2	20	3
0.47	23	21	20	3.3	0.6	20	2	20	2
0.50	23	21	20	3.9	1.4	20	2	20	2

Table 4. Change in $\frac{h+c_o}{h+b}$; $n(1-p)/p = 20, c_a = 4, c_u = 4, b = 54$.

			Capacity			% increase in cost		optimal		heuristic	
h	c_o	$\frac{h+c_o}{h+b}$	Decentr	Optimal	Heuristic	Decentr	Heuristic	L	$U-L$	L	$U-L$
6	9	0.25	20	19	19	6.8	0.1	26	6	25	7
12	21	0.5	24	22	21	7.6	0.3	23	5	22	6
18	54	1	27	24	24	10.3	2.7	19	7	1	25
24	132	2	30	25	24	14.1	1.5	11	13	1	24

6.1. Capacity

In all the examples, the capacity in the decentralized case is higher than the optimal capacity in the coordinated case. It is somewhat surprising to observe that the optimal capacity is close to the average demand in all but a few cases. The exceptions are when demand is very volatile and when the cost of overtime is very large.

We expect the decentralized optimal capacity to be higher based on our finding that the retailer's orders are less volatile in the coordinated system and therefore require a smaller capacity to provide the same service level. However, this line of reasoning is incomplete because changing the capacity changes the steady state distribution of the inventory which in turn affects the values of L and U. Thus, the effect of changing the capacity on total costs is ambiguous. Therefore, it is satisfying to find that the heuristic argument for requiring smaller capacity is borne out to be true in the experiments. An implication of this finding is that coordinated chains have a lower cost of switching to manufacture new products due to the lower investment in capacity and therefore face lower resistance to change.

The decentralized capacity deviates significantly from the coordinated optimal capacity when the value of $(h+c_o)/(h+b)$ is large—see the last two entries in Table 4. The deviation is large because when the cost of exceeding the ideal capacity relative to $h+b$ is high the "optimal" ordering policy cuts off large orders whereas the decentralized system passes on the large orders to the manufacturer. Thus, there is a need for a larger capacity in the decentralized system to cope with the higher variability in orders. The greatest difference between the optimal capacity and that given by the heuristic are seen when demand is extremely volatile (Table 1, cv equal to 0.61). In other cases the capacity suggested by the heuristic is in close agreement to the true optimal capacity.

6.2. Value of $U-L$

As argued in the previous section, the difference between L and U is expected to increase with a decrease in h/C_a (due to relatively more expensive capacity), as well as, with an increase in $(h+c_o)/(h+b)$ and a decrease in $(h-c_u)/(h+b)$ (relatively more expensive to keep plant idle and produce more than the ideal capacity). Subtracting the two quantities, we obtain a key insight that the magnitude of $(c_o+c_u)/(h+b)$ determines the gap between L and U. Note that in two of the examples in Table 3, we have $h < c_o$, and as a result the heuristic method automatically sets L to be as small as possible, which creates a large deviation from the optimal L. Despite this, as seen from these examples, the optimal cost is still extremely close to that of the heuristic policy.

What is somewhat surprising is that $U-L$ increases with increase in the volatility of demand and decrease in $h/(h+b)$ (see Tables 1 and 3). Both these parameters, one being large and the other small, signify that it is optimal for the retailer to provide a very good service. Therefore, the supply chain has to be more responsive. However, $U-L$ is very large in these cases suggesting that most demand variations are not passed on to the manufacturer! This apparent contradiction can be explained by noticing that the retailer carries more safety inventory and thus is able to buffer the manufacturer from the volatility of demand. Apparently, the need for the manufacturer to be responsive is mitigated by the isolation effect of the large safety inventory carried by the retailer.

The values of L and U produced by the heuristic deviate considerably from the corresponding optimal values when the ratio of h/C_a is small. The reason is quite obvious when we recall that the heuristic assumes that the steady state distribution of inventory does not change due to a small change in the capacity. This is no longer true when the values of L and U change very rapidly with a change in the value of h/C_a (alternatively when $U-L$ is large the assumption behind the heuristic does not hold)—see Table 2.

Finally, notice that even though we may have been expecting that the capacity a should be bracketed by L and U, it is not always the case.

6.3. Performance of the Heuristic

The heuristic performs extremely well with regard to cost increase over the optimal. In fact, if the holding cost and the cost of capacity are of the same order of magnitude, the increase in cost from using the heuristic is less than three percent in most cases. The greatest increase in cost (about 10%) occurs (in both demand scenarios) when the ratio of h/C_a is equal to 0.25, see Table 2. The reason once again is due to the rapid change in the steady state distribution of inventory with change in capacity when the value of $U-L$ is large. This suggests that careful calculation is required when the cost of capacity relative to the holding cost is high.

6.4. Summary of Decentralized Versus Coordinated Chains

In summary, it seems that the greatest benefit obtained in a coordinated chain is due to the fact that the retailer curtails the variations in demand. Because of this the capacity of the manufacturer can be kept at a lower level. The largest savings from coordinated ordering accrue when demand is highly volatile and the ratio h/C_a is low. We also note, when the slack in capacity is small relative to the demand, it will reinforce the need to coordinate between the supplier and the manufacturer. Thus, it is not just the absence of inventory as espoused by lean manufacturing proponents but also lower capacity relative to demand that leads to a greater need for coordination. This is to be expected, tighter capacity conditions imply greater benefits from coordination. It also points out to a possible conflict of interests, tighter capacity may motivate the manufacturer into a more opportunistic behavior at the expense of the retailer, thereby leading to distrust and lack of cooperation.

7. DISCUSSION

In this article we study the optimal capacity problem in a supply chain. In contrast to most models in the inventory literature, we consider a case where the manufacturer is required to satisfy the retailer's order, through overtime, outsourcing, etc., when demand exceeds capacity. The supplier does not use any unused capacity to build up inventory. Both the cases with and without coordination have been studied, and the optimal ordering and inventory policy for each case is characterized. For the case with coordination the computation of the optimal supplier capacity becomes a complicated nonlinear programming problem. We develop an efficient heuristic method to compute the optimal capacity for that case. We provide several insights into when and why coordinated supply chains have lower optimal capacity.

If inventory is allowed at the manufacturer, then as long as the installation holding cost at the manufacturer is the same as that at the retailer, the centralized optimal solution remains the same. This is because there exists no incentive for the system to keep inventory at the manufacturer, thus any production completed at the manufacturer is immediately shipped to the retailer, and it can be argued that, when the optimization is centralized, the retailer ordering quantity is precisely the manufacturer's production quantity. Another case where inventory is allowed but our result remains optimal is when the leadtimes are 0. In that case there is clearly no incentive in keeping inventory at the manufacturer either.

APPENDIX

PROOF OF LEMMA 1: We only prove the case that g is strictly increasing. We need to prove that $x_2^* < x_1^*$. First we prove $x_2^* \le x_1^*$. The following relationship is satisfied:

$$g(x_2^*) + f(x_2^*) \le g(x_1^*) + f(x_1^*) \le g(x_1^*) + f(x_2^*),$$

where the first inequality follows from the assumption that x_2^* is a minimizer of $g + f$, while the second inequality follows from x_1^* is a minimizer of f. Thus $g(x_2^*) \le g(x_1^*)$. By the assumption that g is strictly increasing we obtain $x_2^* \le x_1^*$.

Note that x_1^* and x_2^* satisfy $f'(x_1^*) = 0$ and $f'(x_2^*) + g'(x_2^*) = 0$. If $x_2^* = x_1^* = x^*$, then x^* satisfies $g'(x^*) + f'(x^*) = f'(x^*) = 0$, implying $g'(x^*) = 0$. This contradicts with the assumption that g is strictly increasing. Thus, $x_2^* < x_1^*$.

PROOF OF THEOREM 1: We need to compare between ordering less than or equal to capacity a with ordering more than capacity a, i.e., $z \le a$ and $z \ge a$. The cost functions for the two scenarios, defined on $z \le a$ and $z \ge a$ respectively, are

$$\begin{aligned} g_1(x,z) &= cz + c_{\mathrm{u}}(a-z) + \alpha^\tau H(x+z) + \alpha E[V_{n-1}(x+z-D)]\} \\ &= -c_{\mathrm{u}}(x+z) + G(x+z) + c_{\mathrm{u}}(a+x) - cx \end{aligned}$$

and

$$\begin{aligned} g_2(x,z) &= cz + c_{\mathrm{o}}(z-a) + \alpha^\tau H(x+z) + \alpha E[V_{n-1}(x+z-D)]\} \\ &= c_{\mathrm{o}}(x+z) + G(x+z) - c_{\mathrm{o}}(a+x) - cx. \end{aligned}$$

Note that g_1 and g_2 have a common term $-cx$ which is independent of the ordering decision z, thus this term will be ignored in the comparison of g_1 and g_2 below.

We consider several cases separately.

CASE 1: If $x \ge U_n$, then it follows from the convexity of $-c_{\mathrm{u}}y + G(y)$ and $c_{\mathrm{o}}(x+z) + G(x+z)$ that the minimum for g_1 is $z = 0$ and for g_2 is $z = a$. However, since $-c_{\mathrm{u}}y + G(y)$ is increasing on $y \ge U_n$ we have

$$\begin{aligned} [-c_{\mathrm{u}}x + G(x)] + c_{\mathrm{u}}(x+a) &\le [-c_{\mathrm{u}}(x+a) + G(x+a)] + c_{\mathrm{u}}(x+a) \\ &= g_2(x,a). \end{aligned}$$

Therefore, the optimal $y^* = x$ and $z^* = 0$.

CASE 2: If $U_n - a < x \le U_n$, then $x + a > U_n$. Hence g_2 is increasing on $z \ge a$ and its optimum is $z = a$ with cost function $G(x+a)$. The optimum of g_1 can be reached by setting $z = U_n - x$ with cost function $-c_{\mathrm{u}}U_n + G(U_n) + c_{\mathrm{u}}(a+x)$. Since

$$\begin{aligned} [-c_{\mathrm{u}}U_n + G(U_n)] + c_{\mathrm{u}}(x+a) &\le [-c_{\mathrm{u}}(x+a) + G(x+a)] + c_{\mathrm{u}}(x+a) \\ &= G(x+a), \end{aligned}$$

the optimal ordering quantity is $z^* = U_n - x < a$ and the replenishment level is U_n.

CASE 3: If $L_n - a < x \le U_n - a$, then $L_n < x + a \le U_n$. In this case the minimum of g_1 cannot be reached since $x + a \le U_n$, and the optimal solution for g_1 is to order a. Moreover, since $x + a \ge L_n$, the best for g_2 is to order a. This shows that the optimal $y^* = x + a$ and optimal ordering quantity is $z^* = a$.

CASE 4: If $x \le L_n - a$, then $x + a \le L_n \le U_n$. In this case within the range $z \le a$ the best is to order a since g_1 is decreasing on that range, and the cost function value is $G(x+a)$. The minimum of g_2 can be reached by ordering $L_n - x \ge a$. Since

$$\begin{aligned} [c_o L_n + G(L_n)] - c_o(x+a) &\le c_o(x+a) + G(x+a) - c_o(x+a) \\ &= G(x+a). \end{aligned}$$

Thus the optimal ordering quantity is $z = L_n - x > a$ and the optimal replenishment level is L_n.

This completes the proof of Theorem 1.

PROOF OF THEOREM 2: We prove it by induction. The result is clearly true for $n = 0$. Suppose $f_n(x,a)$ is jointly convex in (x,a). Since

$C(z,a)+\alpha^\tau H(x+z)+\alpha E[f_{n-1}(x+z-D,a)]$ is jointly convex in (x,z,a), where

$$C(z,a) = cz + c_{\mathrm{o}} \max\{z-a,0\} + c_{\mathrm{u}} \max\{a-z,0\},$$

we conclude that

$$f_n(x,a) = \min_{z\geq 0}\{C(z,a) + \alpha^\tau H(x+z) + \alpha E[f_{n-1}(x+z-D,a)]\}$$

is jointly convex in (x,a).

PROOF OF THEOREM 5: Let $\{X_n; n = 1,2,\}$ be the inventory position process for the coordinated system with state space $\{L, L+1,\ldots,U\}$. Clearly, $Y_n = D_n$, the demand process. Thus, we only need to prove $|O_n - a| \leq |D_n - a|$.

The optimal ordering quantity for the coordinated system is (i) $O_n = L - X_{n-1}+D_n$ if $X_{n-1}-D_n \leq L-a$, (ii) $O_n = a$ if $L-a < X_{n-1}-D_n \leq U-a$, and (iii) $O_n = U - X_{n-1} + D_n$ if $U-a < X_{n-1} - D_n \leq U$. We consider these three cases separately.

First, suppose $X_{n-1}-D_n \leq L-a$. Then $O_n-a = L-X_{n-1}+D_n-a \geq 0$. Furthermore, it follows from $X_{n-1} \geq L$ that

$$0 \leq O_n - a \leq D_n - a.$$

Hence (4) is satisfied.

Second, suppose $L-a < X_{n-1} - D_n \leq U-a$. Then $O_n = a$ and

$$0 = O_n - a \leq |D_n - a|,$$

thus (4) is also satisfied.

Finally, suppose $U-a < X_{n-1} - D_n \leq U$. Then

$$0 \geq O_n - a = (U - X_{n-1}) + (D_n - a) \geq D_n - a,$$

and it implies that

$$|O_n - a| \leq |D_n - a|.$$

This completes the proof of Theorem 5.

Development of heuristic policy in Section 5

For convenience we imagine that the inventory level is continuous and let $g(\cdot)$ stand for the *density* function of the stationary inventory distribution. To develop a heuristic algorithm, we parametrize the stationary density function of the inventory before an order (I'_a) is placed as g_a. The expected total cost per period is

$$\begin{aligned}&\int_{-\infty}^{L-a}[E_D[h(L-D^{\tau+1})^+ + b(D^{\tau+1}-L)^+] + c(L-x)\\&+ c_o(L-x-a)g_a(x)]dx + \int_{L-a}^{U-a}[E_D[h(a+x-D^{\tau+1})^+\\&+ b(D^{\tau+1}-a-x)^+] + ca\; g_a(x)]dx\\&+\int_{U-a}^{U}[E_D[h(U-D^{\tau+1})^+ + b(D^{\tau+1}-U)^+] + c(U-x)\\&+ c_u(a-U+x)g_a(x)]dx + C_a a.\end{aligned}$$

Assume that even though the distribution of inventory changes with a the effect of this change in distribution of inventory on the cost is negligible for small perturbations of a. Keep U and L fixed. Also assume that the expected value of production equals the average demand. In other words,

$$\begin{aligned}&\int_{-\infty}^{L-a} c(L-x)\; g_a(x)dx + \int_{L-a}^{U-a} ca\; g_a(x)dx\\&+ \int_{U-a}^{U} c(U-x)\; g_a(x)dx = cE[D].\end{aligned}$$

With this in mind, differentiate the expected one period cost with respect to a to get the first derivative of the total expected cost using Leibniz rule as

$$\begin{aligned}&- c_o P\{I'_a \leq L-a\} + c_u P\{I'_a \geq U-a\} + \int_{L-a}^{U-a}[hP\{D^{\tau+1} \leq a+x\}\\&- bP\{D^{\tau+1} \geq a+x\}\; g_a(x)]dx + C_a.\end{aligned}$$

Let I_a be the inventory position after placing the order. The above expression is equivalent to

$$\begin{aligned}&- c_o P\{I_a - D \leq L-a\} + c_u P\{I_a - D \geq U-a\}\\&+ \int_{L-a}^{U-a}(hP\{D^{\tau+1} \leq a+x\} - bP\{D^{\tau+1} \geq a+x\}\; g_a(x))dx + C_a.\end{aligned}$$

Setting the derivative equal to zero we get

$$\begin{aligned}&- c_o P\{I_a - L + a \leq D\} + c_u P\{I_a - U + a \geq D\}\\&+ \int_{L-a}^{U-a}(hP\{D^{\tau+1} \leq a+x\}\\&- bP\{D^{\tau+1} \geq a+x\}\; g_a(x))dx + C_a \equiv 0. \qquad (11)\end{aligned}$$

Also after some algebra we obtain the following bounds

$$\begin{aligned}&- c_o P\{I_a - L + a \leq D\} + c_u P\{I_a - U + a \geq D\}\\&+ P\{D^{\tau+1} \in (I_a - U + a, I_a - L + a)\}(hP\{D^{\tau+1} \leq L\}\\&- bP\{D \geq L\}) + C_a \leq 0, \qquad (12)\\&- c_o P\{I_a - L + a \leq D\} + c_u P\{I_a - U + a \geq D\}\\&+ P\{D^{\tau+1} \in (I_a - U + a, I_a - L + a)\}(hP\{D^{\tau+1} \leq U\}\\&- bP\{D^{\tau+1} \geq U\}) + C_a \geq 0. \qquad (13)\end{aligned}$$

These expressions can be used to interpret how the cost parameters of the retailer affect the capacity decision. Notice that, in all these expressions, the last term is of a "smaller" order of magnitude in comparison with the remaining terms because it equals the product of two probabilities. The term will be even smaller if h, b and $(U-L)$ are small. For example, if the values of h and b are relatively small compared to c_{o} and c_{u}, then the optimal capacity is less dependent on the retailer's cost parameters (not fully independent as it still depends on the values of L and U). On the other hand if $(U-L)$ is large then the last term is no longer negligible. Similarly, if the holding and backorder costs are much larger than the cost of deviating from the planned capacity, then the capacity decision is more dependent on the values of h and b. We return to this issue after solving for L and U.

Turning to the determination of L and U, if we differentiate the expression

$$\begin{aligned}&\int_{-\infty}^{L-a}[E_D[h(L-D^{\tau+1})^+ + b(D^{\tau+1}-L)^+]\\&+ c_o(L-x-a)\; g_a(x)]dx + \int_{L-a}^{U-a}[E_D[h(a+x-D^{\tau+1})^+\\&+ b(D^{\tau+1}-a-x)^+]\; g_a(x)]dx\\&+ \int_{U-a}^{U}[E_D[h(U-D^{\tau+1})^+ + b(D^{\tau+1}-U)^+]\\&+ c_u(a-U+x)\; g_a(x)]dx\end{aligned}$$

with respect to L and U to get respectively (using the same assumption that the distribution of inventory is unaffected due to changes in L or U) the first order conditions

$$hP\{D^{\tau+1} \le L\} - bP\{D^{\tau+1} \ge L\} + c_{\mathrm{o}} = 0,$$
$$hP\{D^{\tau+1} \le U\} - bP\{D^{\tau+1} \ge U\} - c_{\mathrm{u}} = 0.$$

Solving these equations yields

$$P\{D^{\tau+1} \le L\} = \frac{b - c_{\mathrm{o}}}{b + h}, \quad P\{D^{\tau+1} \ge U\} = \frac{h - c_{\mathrm{u}}}{h + b}. \tag{14}$$

This leads to the heuristic policy proposed in Section 5.

ACKNOWLEDGMENTS

The authors are grateful to the Associate Editor and the referees for detailed comments and suggestions. This research is partially supported by NSF under DMI-0200306 and DMI-0200406. The first author is also supported by a grant from National Natural Science Foundation of China under No. 70228001, and part of the work was done while he was with North Carolina State University. We thank Sean Zhou of Chinese University of Hong Kong for his computational assistance.

REFERENCES

[1] K. Arrow, S. Karlin, and H. Scarf, Studies in the mathematical theory of inventory and production, Stanford University Press, Stanford, CA, 1958.

[2] J. Bradley and P.W. Glynn, Managing capacity and inventory jointly in manufacturing systems, Working paper, S. C. Johnson Graduate School of Management, Cornell University, Ithaca, NY, 2000.

[3] J.R. Bradley and B.C. Arntzen, The simultaneous planning of production, capacity and inventory in seasonal demand environments, Working paper, S. C. Johnson Graduate School of Management, Cornell University, Ithaca, NY, 1999.

[4] J.A. Buzacott and J.G. Shanthikumar, Stochastic models of manufacturing systems, Prentice Hall, Englewood Cliff, New Jersey, 1993.

[5] F. Ciarallo, R. Akella, and T.E. Morton, A periodic-review, production planning model with uncertain capacity, Manag Sci 40 (1994), 320–332.

[6] A. Federgruen and P. Zipkin, An inventory model with limited production capacity and uncertain demands. I. The average-cost criterion, Math Oper Res 11 (1986), 193–207.

[7] A. Federgruen and P. Zipkin, An inventory model with limited production capacity and uncertain demands. II. The discounted-cost criterion, Math Oper Res 11 (1986), 208–215.

[8] P. Glasserman and S. Tayur, The stability of a capacitated, multi-echelon production-inventory system under base-stock level policy, Oper Res 42 (1994), 913–925.

[9] P. Glasserman and S. Tayur, Sensitivity analysis of base-stock levels in multi-echelon production-inventory systems, Manag Sci 42 (1995), 263–281.

[10] P. Glasserman and S. Tayur, A simple approximation for multi-stage capacitated production-inventory system, Nav Res Logist 43 (1996), 41–58.

[11] G.H. Hanson, R.J. Mataloni, and M. Slaughter, Vertical production networks in multinational firms, Working Paper 9723, National Bureau of Economic Research, 1050 Massachusetts Avenue, Cambridge, MA, 02138, 1993.

[12] C.C. Holt, F. Modigliani, and H.A. Simon, A linear decision rule for production and employment scheduling, Manag Sci 2 (1995), 1–30.

[13] E.L. Huggins and T.L. Olsen, Supply chain management with guaranteed delivery, Manag Sci 49 (2003), 1154–1167.

[14] R. Kapuscinski and S. Tayur, A capacitated production-inventory model with periodic demand, Oper Res 46 (1998), 899–911.

[15] R. Kapuscinski and S. Tayur, "Optimal policies and simulation-based optimization for capacitated production-inventory systems," in: S. Tayur, R. Ganeshan, and M. Magazine (Editors), Quantitative models for supply chain management, Kluwer, Norwell, MA, 2000, pp. 7–40.

[16] H. Lee, V. Padmanabhan, and S. Whang, The bullwhip effect in supply chains, Sloan Manag Rev 38 (1997), 93–102.

[17] H. Lee, V. Padmanabhan, and S. Whang, Information distortion in a supply chain: The bullwhip effect, Manag Sci 43 (1997), 546–558.

[18] A.S. Manne, Capacity expansion and probabilistic growth, Econometrica 29 (1961), 632–649.

[19] R.P. Parker and R. Kapuscinski, Policies for a capacitated two-echelon inventory system, Oper Res 52 (2004), 739–755.

[20] M.L. Puterman, Markov decision processes, Wiley, New York, 1994.

[21] I.L. Sennott, Stochastic dynamic programming and the control of queues (1998).

[22] M.J. Sobel, Production smoothing with stochastic demand. I. Finite horizon case, Manag Sci 16 (1969), 195–207.

[23] S. Tayur, Computing the optimal policy for capacitated inventory models, Stochastic Models 9 (1992), 585–598.

[24] J.A. Van Mieghem, Capacity management, investment and hedging: Review and recent developments, Manuf & Serv Oper Manag 5 (2003), 269–302.