



Production, Manufacturing and Logistics

Integrated capacity and inventory management with capacity acquisition lead times

Gergely Mincsovcics^a, Tarkan Tan^{a,*}, Osman Alp^b^a Department of Technology Management, Technische Universiteit Eindhoven, P.O. Box 513, 5600MB Eindhoven, The Netherlands^b Industrial Engineering Department, Bilkent University, Bilkent, 06800 Ankara, Turkey

ARTICLE INFO

Article history:

Received 14 March 2007

Accepted 9 April 2008

Available online 2 June 2008

Keywords:

Inventory

Production

Capacity acquisition lead time

Capacity management

Flexible capacity

ABSTRACT

We model a make-to-stock production system that utilizes permanent and contingent capacity to meet non-stationary stochastic demand, where a constant lead time is associated with the acquisition of contingent capacity. We determine the structure of the optimal solution concerning both the operational decisions of integrated inventory and flexible capacity management, and the tactical decision of determining the optimal permanent capacity level. Furthermore, we show that the inventory (either before or after production), the pipeline contingent capacity, the contingent capacity to be ordered, and the permanent capacity are economic substitutes. We also show that the stochastic demand variable and the optimal contingent capacity acquisition decisions are economic complements. Finally, we perform numerical experiments to evaluate the value of utilizing contingent capacity and to study the effects of capacity acquisition lead time, providing useful managerial insights.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction and related literature

In a make-to-stock production system that faces volatile demand, system costs may be decreased by managing the capacity as well as the inventory in a joint fashion, in case there is some flexibility in the production capacity. In some production environments, it is possible to increase the production capacity temporarily while it may take some time to do so. We refer to this delay as capacity acquisition lead time. In this paper we consider such a make-to-stock production system subject to periodic review in a finite-horizon under non-stationary stochastic demand, where our focus is on the effects of capacity acquisition lead time.

The production capacity of a system can be permanent or contingent. We define permanent capacity as the maximum amount of production possible in regular work time by utilizing internal resources of the company such as existing workforce level on the steady payroll and the machinery owned or leased by the company. Total capacity can be increased temporarily by acquiring contingent resources, which can be internal or external. Some methods of acquiring contingent capacity are hiring temporary workers from external labor supply agencies, authorizing overtime production, renting work stations, and so on. We refer to the acquired additional capacity, which is only temporarily available, as contingent capacity. Contingent capacity can be acquired in any period provided that it is ordered in a timely manner, and corresponding costs are incurred only in the periods that contingent capacity is utilized.

Throughout this paper, we primarily consider the workforce capacity setting for ease of exposition. We use the temporary (contingent) labor jargon to refer to capacity flexibility. In that setting, we assume that the production quantity is mostly determined by the workforce size, permanent and contingent.

The availability of contingent capacity may be subject to a certain time lag associated with the operations of the capacity acquisition process, where we refer to this time as the capacity acquisition lead time. For example, an external labor supply agency (ELSA) may not be able to immediately send the temporary workers that a company asks for. The process of searching for and contacting the appropriate workers are the main drivers of this time delay, along with factors such as absenteeism, unavailability or skill requirements. The companies may make contracts with the ELSAs that guarantee the availability of temporary workers provided that they are requested some certain time in advance, which is the case according to our experience in the Netherlands. Naturally, the capacity acquisition lead time increases for jobs that require higher skills.

* Corresponding author. Tel.: +31 402473950; fax: +31 402465949.

E-mail addresses: G.Z.Mincsovcics@tue.nl (G. Mincsovcics), t.tan@tue.nl (T. Tan), osmanalp@bilkent.edu.tr (O. Alp).

Changing the level of permanent capacity as a means of coping with demand fluctuations, such as hiring and/or firing permanent workers frequently, is not only very costly in general, but it may also have many negative impacts on the company. Utilizing flexible capacity is a possible remedy to this problem and we consider it as one of the two main operational tools of coping with fluctuating demand, along with holding inventory. The ELSAs that employ their own workforce provide such flexibility to companies. Since the temporary workers are employed by the ELSAs, decreasing the temporary workforce levels is different than firing permanent workers for the companies and does not bring any additional costs. There exists a significant usage of flexible workforce in many countries. We refer the reader to [Tan and Alp \(in press\)](#) for statistical evidence.

Flexible capacity management refers to adjusting the total production capacity with the option of utilizing contingent resources in addition to the permanent ones. Since long-term changes in the state of the world can make permanent capacity changes unavoidable, we consider the determination of the permanent capacity level as a tactical decision that needs to be made only at the beginning of the planning horizon and not changed until the end of the horizon. The integrated inventory and flexible capacity management problem that we deal with in this paper refers to determining the contingent capacity to be ordered which will be available in a future period as well as determining the optimal production quantity in a certain period given the available capacity which has been determined in an earlier period. We note that this problem is essentially a stochastic version of the aggregate production planning problem.

The dynamic capacity investment/disinvestment problem has been investigated extensively in literature. This problem aims at optimizing the total production capacity of firms at a strategic level to meet long-term demand fluctuations. [Rocklin et al. \(1984\)](#) show that a target interval policy is optimal for this problem. This policy suggests investing in (expanding) the capacity if its current level is below a critical value, disinvesting in (contracting) the capacity if its current level is above another critical value, and doing nothing otherwise. [Eberly and van Mieghem \(1997\)](#) later extend this result to environments with multiple resources. Further multidimensional optimality results are shown by [Gans and Zhou \(2002\)](#) and [Ahn et al. \(2005\)](#). [Angelus and Porteus \(2002\)](#) show that target interval policy is still optimal for managing the capacity in the joint capacity and inventory management problem of a short-life-cycle product under certain assumptions. In general, the lead time for the realization of the capacity expansion and contraction decisions is neglected in this literature and [Angelus and Porteus \(2002\)](#) state that ‘this important generalization to the case of positive capacity lead time with inventory carry-over merits further research’. The lead time issue is considered in the capacity expansion literature to a certain extent. [Angelus and Porteus \(2003\)](#) show optimality of the echelon capacity target policy for multiple resources, which can have different investment lead times and for which investments can be deferred. [Ryan \(2003\)](#) presents a summary of the literature on dynamic capacity expansions with lead times. There are two main differences between the dynamic capacity investment/disinvestment problem and the integrated inventory and flexible capacity management problem that we consider: (i) investment results in possession of capital goods, which still has some value at the time of disinvestment, whereas flexible capacity is not possessed, but acquired only for a temporary duration, (ii) investment decisions are strategic, while integrated inventory and flexible capacity management is tactical and operational.

The problem that is addressed in this paper is closely related to the problems considered by [Tan and Alp \(in press\)](#), [Alp and Tan \(2008\)](#), and [Yang et al. \(2005\)](#). [Tan and Alp \(in press\)](#) deal with a similar problem environment where the lead time for capacity acquisitions is neglected and only the operational decisions are considered. [Alp and Tan \(2008\)](#) extend this analysis by including the tactical level decision of determination of the permanent capacity level. Both of these studies consider fixed costs that are associated with initiating production as well as acquiring contingent workers. We ignore such fixed costs in this study and focus on the effects of capacity acquisition lead time. When the fixed costs are ignored, these two studies reduce to a special case of our work for a capacity acquisition lead time of zero. We refer the reader to these two studies for analysis of this special case and also for a review of the literature on flexible capacity and inventory management for all aspects of the problem other than the capacity acquisition lead time. [Yang et al. \(2005\)](#) deal with a production/inventory system under uncertain permanent capacity levels and the existence of subcontracting opportunities. Subcontracting takes a positive lead time, which is assumed to be one period longer than or equal to the production lead time and a fixed cost is associated with subcontracting. The optimal policy on subcontracting is shown to be of capacity-dependent (s, S)-type. The authors also show that there is a complementarity condition between slack capacity and subcontracting: If subcontracting is more costly than production, no subcontracting will take place unless production capacity is fully utilized. There is a major operational difference between this form of subcontracting option and the use of contingent capacity as in our setting. Subcontracting affects the inventory level directly (any amount subcontracted increases the inventory position with full quantity), while contingent capacity gives extra flexibility as it allows under-utilization of capacity at the time of production.

The rest of the paper is organized as follows. We present our dynamic programming model in Section 2. The optimal policy and some of its properties are discussed in Section 3 and our computations that result in managerial insights are presented in Section 4. We summarize our conclusions and suggest some possible extensions in Section 5.

2. Model formulation

In this section, we present a finite-horizon dynamic programming model to formulate the problem under consideration. Unmet demand is assumed to be fully backlogged. The relevant costs in our environment are inventory holding and backorder costs, and the unit cost of permanent and contingent capacity, all of which are non-negative. There is an infinite supply of contingent capacity, and any number of contingent workers ordered become available with a given time lag. The notation is introduced as need arises, but we summarize our major notation in [Table 1](#) for ease of reference.

We consider a production cost component which is a linear function of permanent capacity in order to represent the costs that do not depend on the production quantity (even when there is no production), such as the salaries of permanent workers. That is, each unit of permanent capacity costs c_p per period, and the total cost of permanent capacity per period is Uc_p , for a permanent capacity of size U , independent of the production quantity. We do not consider material-related costs in our analysis, but it can easily be extended to accommodate this component. In order to synchronize the production quantity with the number of workers, we redefine the “unit production” as the number of actual units that an average permanent worker can produce; that is, the production capacity due to U permanent workers is U “unit”s per period. We also define unit production cost by contingent workers as c_c in the same unit basis. For ease of exposition we consider the productivity rates of contingent and permanent capacity to be the same, but our model can accommodate different productivity rates as explained in [Tan and Alp \(in press\)](#).

Table 1
Summary of Notation

T	Number of periods in the planning horizon
L	Lead time for contingent capacity acquisition
c_p	Unit cost of permanent capacity per period
c_c	Unit cost of contingent capacity per period
h	Inventory holding cost per unit per period
b	Penalty cost per unit of backorder per period
α	Discounting factor ($0 < \alpha \leq 1$)
W_t	Random variable denoting the demand in period t
$G_t(w)$	Distribution function of W_t
$g_t(w)$	Probability density function of W_t
U	Size of the permanent capacity
x_t	Inventory position at the beginning of period t before ordering
y_t	Inventory position in period t after ordering
θ_t	Contingent capacity available in period t (that is ordered in period $t - L$)
θ^t	$\begin{cases} (\theta_t, \theta_{t+1}, \dots, \theta_{t+L-2}, \theta_{t+L-1}) & \text{if } 0 < t \leq T - L \\ (\theta_t, \theta_{t+1}, \dots, \theta_{T-1}, \theta_T) & \text{if } T - L + 1 \leq t \leq T \\ 0 & \text{if } t = T + 1 \end{cases}$
$f_t(x_t, \theta^t, U)$	Minimum total expected cost of operating the system in periods $t, t + 1, \dots, T$, given the system state (x_t, θ^t, U)
$J_t(y_t, \theta^{t+1}, U)$	Cost-to-go function of period t excluding the period's capacity-related costs, given the system state (y_t, θ^{t+1}, U)
s_t	Slack capacity in period t , after production
$*$	Optimal solution
$\hat{\cdot}$	Unconstrained optimum
\bar{y}_t	Optimal inventory position after ordering in period t subject to $\theta_{t+L} = 0$
$\bar{\theta}_{t+L}^A$	Optimal contingent capacity ordered in period t subject to $y_t = x_t$
$\bar{\theta}_{t+L}^B$	Optimal contingent capacity ordered in period t subject to $y_t = x_t + \theta_t + U$

In every period, a decision is made to determine the number of contingent workers to be available in exactly L periods after the current period, as long as there are at least L periods before the end of planning horizon. If θ_t contingent workers are ordered in period $t - L$ then that many workers become available in period t at a total cost of $c_c \theta_t$ which is charged when they become available. There could be situations where modifications on pipeline of contingent workers are possible, however this was not the case in the applications we were involved with and we do not consider such flexibility in our model. In other words, we do not allow for cancellations of the previously ordered capacity or a carry-over of the currently available contingent capacity without prior notice. For example, when the capacity flexibility is by means of overtime, the workers need to be timely informed about overtime production in each and every occasion. Similarly, when the capacity flexibility is by means of temporary labor, the labor supply agencies need to plan where the workers will be sent, and in case the current employee of a given temporary worker wants to extend the employment of the worker for one more period without any prior notice, that may contradict the planned new employment of the temporary worker. In any period $t \leq T - L$, we keep a vector $\theta^t = (\theta_t, \theta_{t+1}, \dots, \theta_{t+L-1})$ which consists of the number of contingent workers that are ordered in periods $t - L, t - L + 1, \dots, t - 1$. In the next period, the vector θ^{t+1} consists of the information on the hired contingent workers available in periods $t + 1, t + 2, \dots, t + L - 1$, carried from the vector θ^t , as well as the decision made for period $t + L$, θ_{t+L} , in period t . Since no contingent workers are ordered after period $T - L$, $\theta^t = (\theta_t, \theta_{t+1}, \dots, \theta_{T-1}, \theta_T)$ for $T - L + 1 \leq t \leq T$ and $\theta^{T+1} := 0$. The size of the permanent workforce, U , is determined only at the beginning of the first period, and it is considered to be fixed during the whole planning horizon.

The order of events in a period is as follows. At the beginning of period t , the initial inventory level, x_t is observed, and the number of previously ordered contingent workers, θ_t , become available. The total amount of capacity in period t becomes $U + \theta_t$, which is the upper limit on the production quantity of this period. Then, the operational decisions, i.e. the production decision given the available capacity and the decision on the number of contingent workers to be available in period $t + L$, are made. According to the production decision, the inventory level is raised to $y_t \leq x_t + U + \theta_t$. We note that the optimal production quantity ($y_t - x_t$) may result in partial utilization of the available capacity, which is already paid for. At the end of period t , the realized demand w_t is met/backlogged, resulting in a starting inventory for period $t + 1$, $x_{t+1} = y_t - w_t$. The vector θ^{t+1} is constructed as explained above. We assume the demand to be independently but not necessarily identically distributed, and we denote the random variable corresponding to the demand in period t as W_t and its distribution function as G_t . Finally, denoting the minimum cost of operating the system from the beginning of period t until the end of the planning horizon as $f_t(x_t, \theta^t, U)$, we use the following dynamic programming formulation to solve the problem of integrated Capacity and Inventory Management with Capacity Acquisition Lead Times (CILT):

$$f_t(x_t, \theta^t, U) = Uc_p + \theta_t c_c + \begin{cases} \min_{y_t \in [x_t, x_t + \theta_t + U]} \{ \mathcal{L}_t(y_t) + \alpha E[f_{t+1}(y_t - W_t, \theta^{t+1}, U)] \} & \text{if } T - L + 1 \leq t \leq T, \\ \min_{\theta_{t+L} \geq 0, y_t \in [x_t, x_t + \theta_t + U]} \{ \mathcal{L}_t(y_t) + \alpha E[f_{t+1}(y_t - W_t, \theta^{t+1}, U)] \} & \text{if } 1 \leq t \leq T - L, \end{cases}$$

$$f_0(x_1) = \min_{U \geq 0, \theta^1 \geq 0} f_1(x_1, \theta^1, U)$$

where $f_{T+1}(\cdot) \equiv 0$, $0 \leq L \leq T$ and $\mathcal{L}_t(z) = h \int_0^z (z - \omega) dG_t(\omega) + b \int_z^\infty (\omega - z) dG_t(\omega)$.

We note that the number of contingent workers hired before the planning horizon begins, $\theta^1 = (\theta_1, \theta_2, \dots, \theta_L)$, is also optimized in the above formulation, assuming that those decisions are made in advance in an optimal manner. Nevertheless, all of our analytical results would hold for any given θ^1 as well.

When capacity acquisition lead time is zero ($L = 0$), the minimization operator, $\min_{\theta_{t+L} \geq 0, y_t \in [x_t, x_t + \theta_t + U]}$ is to be read as $\min_{\theta_t \geq 0} \min_{y_t \in [x_t, x_t + \theta_t + U]}$, the cost $\theta_t c_c$ gets inside the minimization, and θ^t disappears from the state space. Tan and Alp (in press) show that this two-dimensional minimization can be reduced to a single-dimensional one.

3. Analysis of the optimal policies

In this section, we first characterize the optimal solution to the problem that is modeled in Section 2. Then, we introduce some properties of the optimal solution, including those that regard the utilization of the available capacity.

Let J_t denote the cost-to-go function of period t excluding the period's capacity-related costs; $J_t(y_t, \theta^{t+1}, U) = \mathcal{L}_t(y_t) + \alpha E[f_{t+1}(y_t - W_t, \theta^{t+1}, U)]$. Accordingly, $f_t(x_t, \theta^t, U)$ can be rewritten as

$$f_t(x_t, \theta^t, U) = U c_p + \theta_t c_c + \begin{cases} \min_{y_t \in [x_t; x_t + \theta_t + U]} J_t(y_t, \theta^{t+1}, U) & \text{if } T - L + 1 \leq t \leq T, \\ \min_{\theta_{t+L} \geq 0, y_t \in [x_t; x_t + \theta_t + U]} J_t(y_t, \theta^{t+1}, U) & \text{if } 1 \leq t \leq T - L. \end{cases}$$

Let $(\hat{y}_t, \hat{\theta}_{t+L})$ be the unconstrained minimizer of the function $J_t(\cdot)$ for given state variables $\theta_{t+1}, \dots, \theta_{t+L-1}$, and U . We use the following definitions in our further discussion for $t \in \{1, \dots, T - L\}$:

$$\begin{aligned} \bar{y}_t &:= \arg \min_{y_t \in [x_t; x_t + \theta_t + U], \theta_{t+L} = 0} J_t(y_t, \theta^{t+1}, U), \\ \bar{\theta}_{t+L}^A &:= \arg \min_{\theta_{t+L} \geq 0} J_t(x_t, \theta^{t+1}, U), \text{ and} \\ \bar{\theta}_{t+L}^B &:= \arg \min_{\theta_{t+L} \geq 0} J_t(x_t + \theta_t + U, \theta^{t+1}, U). \end{aligned}$$

Let (y_t^*, θ_{t+L}^*) be the aggregate optimal production and contingent capacity hiring decision in period t given that the state variables are x_t, θ^t and U .

3.1. Optimal policy characterization

The optimal decisions at any period t (inventory level after production, y_t , and number of contingent workers hired, θ_{t+L}) are made by minimizing the function J_t over the feasible region. First, we characterize the solution of CILT in [Theorem 1](#).

Theorem 1. *The following hold for any capacity acquisition lead time $L = 0, 1, 2, \dots, T - 1$.*

- (a) *For any period t ($1 \leq t \leq T$), f_t and J_t are (jointly) convex functions.*
- (b) *For any period t such that $1 \leq t \leq T - L$, the optimal production and contingent capacity ordering policy is given by*

$$(y_t^*, \theta_{t+L}^*) = \begin{cases} (\hat{y}_t, \hat{\theta}_{t+L}) & \text{if } \hat{y}_t \in [x_t; x_t + \theta_t + U], \hat{\theta}_{t+L} \geq 0, \\ (x_t, \bar{\theta}_{t+L}^A) & \text{if } \hat{y}_t < x_t, \hat{\theta}_{t+L} \geq 0, \\ (x_t + \theta_t + U, \bar{\theta}_{t+L}^B) & \text{if } \hat{y}_t > x_t + \theta_t + U, \hat{\theta}_{t+L} \geq 0, \\ (\bar{y}_t, 0) & \text{if } \hat{y}_t \in [x_t; x_t + \theta_t + U], \hat{\theta}_{t+L} < 0, \\ (\bar{y}_t, \bar{\theta}_{t+L}^A) : (\bar{y}_t - x_t) \bar{\theta}_{t+L}^A = 0 & \text{if } \hat{y}_t < x_t, \hat{\theta}_{t+L} < 0, \\ (\bar{y}_t, \bar{\theta}_{t+L}^B) : (\bar{y}_t - x_t - \theta_t - U) \bar{\theta}_{t+L}^B = 0 & \text{if } \hat{y}_t > x_t + \theta_t + U, \hat{\theta}_{t+L} < 0. \end{cases}$$

Proof. See [Appendix](#). \square

The convexity of J_t as stated in part (a) implies that the production quantity should bring the inventory level to the base-stock level $\hat{y}_t(\theta^t, \theta_{t+L}, U)$ for a given θ_{t+L} – where $\hat{y}_t(\theta^t, \theta_{t+L}, U)$ is the minimizer of J_t for a given $(\theta^t, \theta_{t+L}, U)$ – as long as the base-stock level is in the interval $[x_t, x_t + \theta_t + U]$. Otherwise, $y_t^* = x_t$ if the base-stock level is less than x_t , meaning that no production should take place, and $y_t^* = x_t + \theta_t + U$ if the base-stock level is greater than $x_t + \theta_t + U$, meaning that all of the available capacity (permanent and contingent) should be utilized. With respect to the contingent capacity ordering decision, $\theta_{t+L}^* = \hat{\theta}_{t+L}(y_t, \theta^t, U)$ for any given y_t – where $\hat{\theta}_{t+L}(y_t, \theta^t, U)$ is the minimizer of J_t for a given (y_t, θ^t, U) – as long as $\hat{\theta}_{t+L}(y_t, \theta^t, U) \geq 0$. Otherwise, no contingent capacity should be ordered. We also note that for periods $T - L + 1$ to T , the optimal level of inventory after production is given by a state-dependent base-stock policy, due to convexity of J_t . Part (b) of [Theorem 1](#) characterizes the optimal integrated production and contingent capacity ordering decisions in terms of the unconstrained minimizer and $\bar{y}_t, \bar{\theta}_{t+L}^A$, and $\bar{\theta}_{t+L}^B$, which are the minimizers on the borders of the feasible domain ($\theta_{t+L} \geq 0, y_t \in [x_t; x_t + \theta_t + U]$). The first case corresponds to the situation where the unconstrained minimizer falls in the feasible region, and hence the unconstrained minimizer is the optimal solution. In the latter five cases the unconstrained minimizer is outside the feasible region, where the optimal solution is then on the boundary of the feasible region, due to convexity of J_t . The last two cases further characterize the optimal solution by imposing a condition (that we refer to as “complementary slackness property” in [Section 3.2](#)) when neither \hat{y}_t nor $\hat{\theta}_{t+L}$ is within its feasible interval. Finally, part (a) also states that the recursive minimum expected cost function of the dynamic programming formulation, $f_t(x_t, \theta^t, U)$ is convex. Therefore, finding the optimal permanent capacity level, U^* is a convex optimization problem.

Remark 1. If $c_c < c_p$, then $U^* = 0$.

[Remark 1](#) holds due to the fact that any solution with $U > 0$ would be dominated by the solution that has $U = 0$ and $\theta_t = U$ for all t .

In what follows we utilize the notion of supermodularity and submodularity to show properties on the pairwise relations of the variables and parameters in our model, which is a notion employed in economic theory often to explore economic complements and substitutes. A function which is supermodular (submodular) on two arguments implies that more of one of the arguments induces less (more) of the other ([Porteus, 2002](#)). In particular, [Theorem 2](#) identifies such relations in our problem environment between contingent capacity ordered, inventory position, permanent capacity, and demand.

Theorem 2. *For any period t ($1 \leq t \leq T$), and capacity acquisition lead time $L = 1, 2, \dots, T - 1$, the following hold:*

- (a) *$f_t(x_t, \theta^t, U)$ and $J_t(y_t, \theta^{t+1}, U)$ are supermodular functions.*

(b) J_t is submodular in $(W_t, (\theta^{t+1}, U))$ and in (W_t, y_t) , where $W_t \in D$, and D is the poset of discrete random variables with the first order stochastic dominance as partial order, $(\theta^{t+1}, U) \in \mathbb{R}^{L+1}$, on which the product order is the partial order.

Proof. See Appendix. \square

Supermodularity of $J_t(y_t, \theta^{t+1}, U)$ implies, for example, that y_t and θ^{t+1} are economic substitutes: in any element we increase in θ^{t+1} , the optimal y_t is non-increasing. Naturally, it implies as well that substitution holds between y_t and U , any element of θ^{t+1} and U , or any two elements of θ^{t+1} . Supermodularity of $f_t(x_t, \theta^t, U)$ allows similar interpretation as that of $J_t(y_t, \theta^{t+1}, U)$: the inventory, the pipeline contingent capacity and the permanent capacity are economic substitutes. For example, a higher starting inventory eliminates the necessity for a higher permanent capacity.

Submodularity of the function J_t in (W_t, θ_{t+L}) given in part (b) indicates that W_t and θ_{t+L} are economic complements. That is to say, stochastically larger demand distributions lead to hiring more contingent capacity. A similar relation also exists between W_t and y_t . Note that the sub- and supermodularity results in Theorem 2 do not apply only to optimal decisions. For example, supermodularity of J_t in y_t and θ_{t+L} implies that the marginal cost of increasing y_t increases in θ_{t+L} . The reader is referred to Porteus (2002) and Topkis (1998) for further details on sub- and supermodular functions, and Puterman (1994) for partial ordering of random variables.

The following corollary to the second part of Theorem 2(a) helps to reduce the search space by providing bounds on the decision variables y_t and θ_{t+L} , using the fact that they are economic substitutes.

Corollary 1. For any period t ($1 \leq t \leq T - L$), the (constrained) optimal solution of J_t is in the domain $\{(y_t, \theta_{t+L}) : y_t \in [x_t; \bar{y}_t], \theta_{t+L} \in [\bar{\theta}_{t+L}^B; \bar{\theta}_{t+L}^A]\}$.

3.2. Complementary slackness

In our model, we have two decision variables to be determined in every period: the inventory level after production and the contingent capacity ordered that will arrive L periods later. The former decision variable is bounded from above by the maximum amount of capacity available (the permanent capacity level plus the contingent capacity that was ordered L periods ago) whereas the latter decision variable is only constrained to be non-negative. Let s_t denote the slack capacity in period t after the production decision is implemented, $s_t = x_t + U + \theta_t - y_t$. We define the complementary slackness property as follows:

Definition 1. For any period t , there exists a Complementary Slackness Property (CSP) between slack capacity, s_t , and contingent capacity ordered, θ_{t+L} , only if $s_t \theta_{t+L} = 0$.

If a solution does not satisfy CSP, a positive contingent capacity is ordered for future use, while the current capacity which has already been paid for is not fully utilized. If such a solution is optimal then ordering contingent capacity to be available L periods later is preferred to utilizing currently available capacity fully which might lead to carrying inventory. In case the optimal solution is known to satisfy CSP, this helps not only to further characterize the optimal solution, but also to simplify the solution of CLT. In particular, whenever the optimal solution satisfies CSP, the problem reduces to one-dimensional optimization problems. In what follows, we present some special cases where the optimal solution satisfies CSP.

For the special case where the demand is deterministic, it is straightforward to show that the optimal solution satisfies CSP if $\sum_{i=0}^{L-1} \alpha^i h < \alpha^L c_c$. This condition simply implies that it is less costly to carry inventory than to order contingent capacity, which assures that contingent capacity is never ordered unless available capacity is fully utilized.

Theorem 3. When $L = 1$, the optimal solution satisfies CSP in the following cases:

- (a) In the infinite-horizon problem with stationary and positive demand ($T \rightarrow \infty, W_t \equiv W > 0$), when $h < \alpha c_c$.
- (b) In the two-period problem, when $h(1 + \alpha) < \alpha c_c$.

Proof. See Appendix. \square

Note that for the special case of $L = 1$, Theorem 3 is valid under reasonable cost parameter settings. Moreover, part (a) is valid for infinite-horizon and part (b) is valid for any demand distribution. Nevertheless, while an optimal solution which does not satisfy CSP might seem to be counter-intuitive, it turns out that in some cases this is true, as we illustrate in the following examples where CSP does not hold in the optimal solution.

Example 1. For $T = 15, L = 2, h = 1, b = 5, c_p = 2.4, c_c = 3.2, \alpha = 1, U = 10, x_1 = 0, \theta_1 \geq 0$, and $\theta_2 \geq 0$, consider the following demand stream: $P(W_1 = 0) = 1, P(W_2 = 30) = 0.4, P(W_2 = 0) = 0.6, P(W_3 = \dots = W_{11} = 0) = 1$ and $P(W_{12} = \dots = W_{15} = 10) = 1$. The optimal decision is $(y_1^*, \theta_2^*) = (0, 10)$. That is, 10 units of contingent capacity is ordered while the available capacity is not fully utilized which violates CSP. The intuition behind this solution is as follows: Because the uncertainty will be resolved in period 2, any production before that may result in holding inventory for a number of periods. On the other hand, if the production capacity in period 3 is not increased – which requires requesting 2 periods in advance – there may be high backordering costs in case positive demand in period 2 is materialized.

Example 2. For $T = 2, L = 1, h = 2.98, b = 5, c_p = 2.5, c_c = 3, \alpha = 0.99, U = 6, x_1 = 0$, and $\theta_1 = 0$, let the demand follow normal distribution with $E[W_1] = 3, \text{Var}[W_1] = 0.36$, and $E[W_2] = 21, \text{Var}[W_2] = 17.64$ (both yielding coefficient of variation = 0.2). In this case, $(y_1^*, \theta_2^*) = (4.8, 10.4)$, violating CSP.

4. Numerical results and discussion

The main goal of this section is to gain insights on how the value of flexible capacity and the optimal permanent capacity levels change as the following system parameters change: capacity acquisition lead time, unit cost of contingent capacity, backorder cost, and the

variability of the demand. For this purpose, we conduct some numerical experiments by solving CILT. We use the following set of input parameters, unless otherwise noted: $T = 12$, $b = 10$, $h = 1$, $c_c = 3$, $c_p = 2.5$, $\alpha = 0.99$, and $x_1 = 0$. We consider Normal demand with a coefficient of variation (CV) of 0.2 that follows a seasonal pattern with a cycle of 4 periods, where the expected demand is 10, 15, 10, and 5, respectively. Recall that the values of the pipeline of contingent capacity at the beginning of the first period are optimized in CILT, and accordingly the results containing different lead times are comparable.

Solution of CILT in a Pentium 4 with a 2.79 GHz CPU and 1 Gb RAM for the parameter set given above took less than 1 s for $L < 3$ and 14 s for $L = 3$. For longer lead times, the curse of dimensionality prevails and computational limitations become prohibitive.

In the results that we present, we use the term “increasing” (“decreasing”) in the weak sense to mean “non-decreasing” (“non-increasing”). We provide intuitive explanations to all of our results below and our findings are verified in several numerical studies. However, like all experimental results, one should be careful in generalizing them, especially for extreme values of problem parameters.

4.1. Value of flexible capacity

The option of utilizing contingent capacity provides additional flexibility to the system and leads to reduction of the total costs, even though there is a certain lead time associated with it. We measure the magnitude of cost reduction in order to gain insight on the value of flexible capacity. We compare a flexible capacity (FC) system with an inflexible one (IC), where the contingent capacity can be utilized in the former but not in the latter. We define the absolute value of flexible capacity, VFC , as the difference between the optimal expected total cost of operating the IC system, ETC_{IC} , and that of the FC system, ETC_{FC} . That is, $VFC = ETC_{IC} - ETC_{FC}$. We also define the (relative) value of flexible capacity as the relative potential cost savings due to utilizing the flexible capacity. That is, $\%VFC = 100 \cdot VFC / ETC_{IC}$. We note that both VFC and $\%VFC$ are always non-negative. We also note that the permanent capacity levels are optimized in both systems separately to ensure that the differences are not caused by the insufficiency of permanent capacity in the inflexible system.

We first test the value of flexibility with respect to the backorder and contingent capacity costs under different capacity acquisition lead times, by varying the value of one of the parameters while keeping the rest fixed. We present the results in Table 2, which verifies intuition in the sense that $\%VFC$ is higher when capacity acquisition lead time is shorter. These results also generalize the findings of Tan and Alp (in press) for $L = 0$ to the case of positive capacity acquisition lead times, such that $\%VFC$ is higher when contingent capacity cost is lower or backorders are more costly (equivalently, when a higher service level is targeted).

We note that, although $\%VFC$ decreases with an increasing lead time, the marginal decrease appears to be decreasing as L increases. Besides, we also observe that $\%VFC$ with higher lead times persists to be comparable with $\%VFC$ with lower lead times, meaning that flexibility is still valuable even when the capacity acquisition lead time is relatively long.

We also analyze the relation between the value of flexible capacity and the demand variability. The results presented by Alp and Tan (2008) indicate that the value of flexibility is not necessarily monotonic (i.e. it does not increase or decrease consistently) as the demand variability increases for the case where the lead time is zero. We find out that this continues to be true for the case where the lead time is strictly positive. The explanation is that the system has the ability to adapt itself to changes in coefficient of variation, CV , by optimizing the permanent capacity level accordingly. Nevertheless, for increasing values of the contingent capacity acquisition lead time we observe that the value of flexibility generally decreases when the demand variability increases as is the case in Fig. 1. A longer capacity acquisition lead time deteriorates the effectiveness of capacity flexibility. This effect is amplified in case of higher demand variability. In other words, since the capacity needs are more predictable for lower demand variability, use of contingent capacity – which has to be ordered one lead time ahead – becomes more effective as compared to the high variability case. This also explains why the decrease in the value of flexibility as lead time increases is steeper when the variability is higher.

4.2. Optimal level of permanent capacity

In this section, we investigate how the optimal level of permanent capacity changes as the problem parameters change. We present the data regarding some of our results in Table 3. We first note that the optimal permanent capacity decreases as the contingent capacity acquisition lead time decreases, in all of the cases that we consider. That is, since the decreased lead time makes the capacity flexibility a more powerful tool, it decreases the required level of permanent capacity. When c_c and L are small enough, the benefits of capacity flexibility becomes so prevalent that, even when $c_c > c_p$ the optimal permanent capacity level may turn out to be zero. We also note that the findings

Table 2
 $\%VFC$ as L , c_c , and b change

L	%VFC			
	0	1	2	3
c_c				
1.0	63.35	58.94	57.63	57.36
2.0	36.35	31.50	28.34	27.18
2.5	22.87	17.90	14.57	12.71
3.0	14.91	10.30	8.55	7.50
3.5	11.10	7.26	6.27	5.61
4.0	8.92	5.58	4.91	4.21
5.0	6.02	3.18	2.98	2.74
8.0	1.75	0.42	0.37	0.34
b				
5	11.79	7.91	6.49	5.54
10	14.91	10.30	8.55	7.50
20	17.50	12.22	10.22	9.07
50	20.51	14.63	12.31	11.09
250	24.82	18.06	15.49	14.22

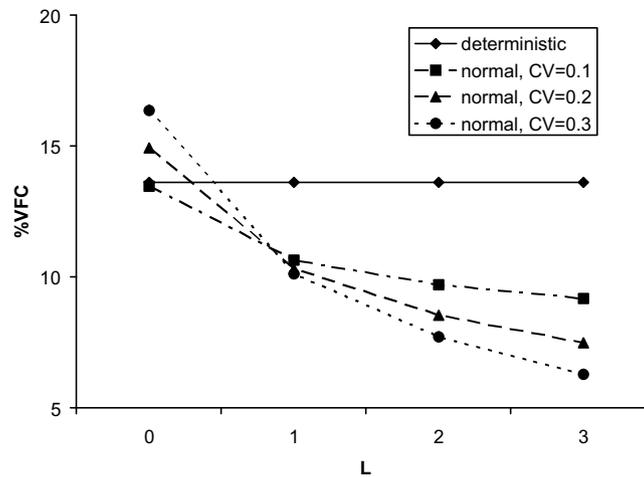


Fig. 1. %VFC as a function of L for different demand streams.

Table 3

U^* as a function of the lead time, L, for varied c_c , b and demand distribution streams

L	0	1	2	3
c_c	U^* for $b = 10$			
2.5	0	0	0	0
2.51	0	0	2	3
2.6	3	3	4	6
3.0	7	7	8	9
3.5	8	9	10	10
4.0	9	10	10	10
5.0	10	11	11	11
8.0	11	12	12	12
Demand	U^* for $b = 10$			
Deterministic	7	7	7	7
Normal, CV = 0.1	7	8	8	8
Normal, CV = 0.2	7	7	8	9
Normal, CV = 0.3	6	7	9	9
Demand	U^* for $b = 50$			
Deterministic	7	7	7	7
Normal, CV = 0.1	7	7	7	8
Normal, CV = 0.2	6	7	8	9
Normal, CV = 0.3	5	6	8	9

of Alp and Tan (2008) for the case of $L = 0$, which state that the optimal permanent capacity level decreases as contingent capacity cost decreases or backorder cost increases, also hold for positive capacity acquisition lead times.

Similar to the value of flexibility, we observe that the optimal permanent capacity level is not necessarily monotonic in demand variability. Nevertheless, for longer capacity acquisition lead times or higher costs of contingent capacity, optimal permanent capacity level in general increases as demand variability increases. On the contrary, for shorter capacity acquisition lead times and lower costs of contingent capacity, optimal permanent capacity level in general decreases as demand variability increases.

5. Conclusions and future research

In this paper, the integrated problem of inventory and flexible capacity management under non-stationary stochastic demand is considered, when lead time is present for flexible capacity acquisition. Permanent productive resources may be increased temporarily by hiring contingent capacity in every period, where this capacity acquisition decision becomes effective with a given time lag. Other than the operational level decisions (related to the production and capacity acquisition levels), we also keep the permanent capacity level as a tactical decision variable which is to be determined at the beginning of a finite planning horizon. We provide insights on the effects of capacity acquisition lead time.

We first prove that all of the decision making functions under consideration are convex. This result helps us to provide an optimal policy for the operational decisions and to find the optimal permanent capacity level. Moreover, we prove that the inventory (either before or after production), the pipeline contingent capacity, the contingent capacity to be ordered, and the permanent capacity are economic substitutes. We also show that the stochastic demand variable and the optimal contingent capacity acquisition decisions are economic complements; for stochastically larger demand streams, we observe higher contingent capacity levels in optimality. A similar interpretation is also true for stochastically larger demand streams and the optimal inventory levels obtained after production.

A policy that might seem to be optimal is never to order contingent capacity unless the permanent capacity is fully utilized, which we refer to as complimentary slackness property (CSP). We show through numerical examples that an optimal solution does not necessarily satisfy CSP. We also provide some cases where the optimal solution is assured to satisfy CSP.

By making use of our model, we develop some managerial insights. First of all, the value of flexibility naturally decreases with an increasing lead time. Consequently, there is a value in trying to decrease capacity acquisition lead time in the system through means such as negotiating with the external labor supply agency or forming a contingent labor pool perhaps within different organizations of the same company. This especially holds when the demand is highly variable. Nevertheless, the value of flexibility remains considerable even when the capacity acquisition lead time is relatively long. Therefore, the existence of a lead time in acquiring contingent capacity should not discourage the production company from making use of capacity flexibility, especially if the demand variability is not very high. Consequently, the managers should invest in higher levels of permanent capacity when capacity acquisition lead time and demand variability are high, and it is not wise to do so when the contingent capacity is a more “effective” tool in the sense that capacity acquisition lead time is short and the demand variability is high.

This research may be extended in several ways. Introducing an uncertainty on the permanent and contingent capacity levels would enrich the model (see Pac et al., in press, for the analysis of capacity uncertainty in a zero lead time environment). For example, the supply of contingent capacity may be certain for larger lead times whereas it may be subject to an uncertainty for shorter lead times. Some other extension possibilities include considering the fixed costs for production and/or acquisition of contingent capacity, including expansion and contraction decisions for the permanent capacity, considering alternative resources of capacity flexibility, considering the possibility to carry-over the contingent capacity and cancel previously ordered capacity, and developing efficient heuristic methods for the problem.

Appendix

Proof of Theorem 1. We prove part (a) by induction. Note that $f_{T+1}(\cdot) = 0$ and is convex. Assume that $f_{t+1}(\cdot)$ is also convex. The function $J_t(y_t, \theta^{t+1}, U) = L_t(y_t) + \alpha E[f_{t+1}(y_t - W_t, \theta^{t+1}, U)]$ is convex because (i) $L_t(y_t)$ is a convex function, (ii) $E[f_{t+1}(y_t - W_t, \theta^{t+1}, U)]$ is convex by the convexity preservation of the expected value operator (see Appendix A.5 in Bertsekas (1976)), and (iii) the convexity preservation of the linear combination with non-negative weights. Then note that the following minimization operators preserve the convexity of J .

$$g(x, \theta, U) = \min_{y \in [x, x+\theta+U]} J(y, U),$$

$$h(x, \theta, U) = \min_{\substack{y \in [x, x+\theta+U] \\ \delta \geq 0}} J(y, \delta, U).$$

From Proposition B-4 of Heyman and Sobel (2004), coupled with the convexity preservation of affine mappings (see Hiriart-Urruty and Lemaréchal, 1993) it follows that the resulting g and h functions are convex when J is convex. Finally,

$$f_t(x_t, \theta^t, U) = U c_p + \theta_t c_c + \begin{cases} g(x_t, \theta, U) & \text{for } T - L + 1 \leq t \leq T, \\ h(x_t, \theta, U) & \text{for } 1 \leq t \leq T - L \end{cases}$$

is convex, which completes the proof of part (a).

Part (a) implies directly part (b). □

Preliminaries to Proof of Theorem 2: We start with two lemmas that will help us with the proof of Theorem 2.

Lemma 1. For any cost parameters, the newsboy function (loss function)

$$L(W, y) = h \int_{-\infty}^y (y - w) dF_W(w) + b \int_y^{\infty} (w - y) dF_W(w)$$

is submodular with $y \in \mathbb{R}$ (real) and $W \in D$, where D is the poset of random variables with the first order stochastic dominance (\preceq) as partial order.

Proof. We need to show that $L(W, y)$ is submodular; that is $L(W^-, y^-) + L(W^+, y^+) \leq L(W^+, y^-) + L(W^-, y^+)$ for all $W^-, W^+ \in D$ and $y^-, y^+ \in \mathbb{R}$, for which $W^- \preceq W^+$ and $y^- \leq y^+$. We denote the cumulative distribution functions of W^- and W^+ by F^- and F^+ , respectively. Then by the definition of stochastic dominance, if $W^- \preceq W^+$ then we have $F^-(w) \geq F^+(w)$ for all $w \in \mathbb{R}$. In the first step, we split integration intervals in $L(W, y^+)$ and $L(W, y^-)$ by y^- and y^+ .

$$L(W, y^+) = h \int_{-\infty}^{y^-} (y^+ - w) dF_W(w) + h \int_{y^-}^{y^+} (y^+ - w) dF_W(w) + b \int_{y^+}^{\infty} (w - y^+) dF_W(w),$$

$$L(W, y^-) = h \int_{-\infty}^{y^-} (y^- - w) dF_W(w) + b \int_{y^-}^{y^+} (w - y^-) dF_W(w) + b \int_{y^+}^{\infty} (w - y^-) dF_W(w).$$

We denote the difference of the above standing two terms by $\Delta(W)$. One can show with the help of partial integration that $\Delta(W) := L(W, y^+) - L(W, y^-) = (h + b) \int_{y^-}^{y^+} F_W(w) dw$ holds.

In the final step, we subtract $\Delta(W^+)$ from $\Delta(W^-)$ and use the first order stochastic dominance of W^+ over W^- , meaning $F^-(w) \geq F^+(w)$ for all $w \in \mathbb{R}$.

$$\Delta(W^+) - \Delta(W^-) = [L(W^+, y^+) - L(W^+, y^-)] - [L(W^-, y^+) - L(W^-, y^-)] = (h + b) \int_{y^-}^{y^+} (F^+(w) - F^-(w)) dw \leq 0.$$

This completes the proof. □

Lemma 2. Convex minimization operators resulting in supermodular functions.

Assume that y, x, θ, c are real numbers, z is a real vector, and g is a real valued function.

- (a) If $g(y)$ is convex then $H(x, \theta) := \min_{y \in [x; x+\theta+c]} g(y)$ is supermodular ($\theta \geq 0, c \geq 0$).
- (b) If $g(y, z)$ is supermodular, then $H(x, z) = \min_{y \in [x; x+c]} g(y, z)$ is supermodular ($c \geq 0$).
- (c) If $g(y, z)$ is supermodular, then $H(\theta, z) = \min_{y \in [x; x+c+\theta]} g(y, z)$ is supermodular ($\theta \geq 0, c \geq 0$).

Proof. Proof of part (a): We define the global optimum point as $\hat{y} := \min_{y \in \mathbb{R}} g(y) \in \mathbb{R} \cup \{-\infty, +\infty\}$. For supermodularity, we aim to show for all $x^- \leq x^+, \theta^- \leq \theta^+$ that

$$H(x^+, \theta^-) + H(x^-, \theta^+) \leq H(x^-, \theta^-) + H(x^+, \theta^+).$$

The domain of $H(x, \theta)$ can be divided into three parts.

$$H(x, \theta) = \begin{cases} g(x) \text{ increasing in } x, & \text{if } \hat{y} \leq x \\ g(\hat{y}) \text{ constant,} & \text{if } \hat{y} - c \leq x + \theta \text{ and } x \leq \hat{y} \\ g(x + \theta + c) \text{ decreasing in } x + \theta, & \text{if } x + \theta \leq \hat{y} - c. \end{cases}$$

We can observe that $H(x^-, \theta^+) \leq H(x^-, \theta^-)$ holds for all x^-, θ^-, θ^+ , when $\theta^- \leq \theta^+$. We distinguish two cases by where x^+ is situated:

When $x^+ \geq \hat{y} - c$, then $H(x^+, \theta^-) = H(x^+, \theta^+)$ holds, which implies that supermodularity inequality holds.

When $x^+ \leq \hat{y} - c$, then we can define a function with a single variable $h(x + \theta) := H(x, \theta)$ which is convex (as discussed in Theorem 1(a)). Note that a function (H) is supermodular if it is defined as a single argument convex function (h) at its arguments' non-negative linear combination, due to Lemma 2.6.2.a in Topkis (1998). This completes the proof.

Proof of part (b): We introduce $y^A := \arg \min_{y \in [x^-, x^+ + c]} g(y, z^-)$ and $y^B := \arg \min_{y \in [x^+, x^+ + c]} g(y, z^+)$ with $x^- \leq x^+$ and $z^- \leq z^+$. Now we can express $H(x^-, z^-)$ and $H(x^+, z^+)$ as $g(y^A, z^-)$ and $g(y^B, z^+)$, respectively. Furthermore, $H(x^-, z^+) \leq g(y^A, z^+)$ and $H(x^+, z^-) \leq g(y^B, z^-)$ holds because $y^A \in [x^-; x^- + c]$ and $y^B \in [x^+; x^+ + c]$.

If $y^A \leq y^B$, then by supermodularity of g , we have

$$H(x^-, z^+) + H(x^+, z^-) \leq g(y^A, z^+) + g(y^B, z^-) \leq g(y^A, z^-) + g(y^B, z^+) = H(x^-, z^-) + H(x^+, z^+)$$

from which supermodularity of H follows.

If $y^B \leq y^A$, then both y^A and y^B are in the $[x^+, x^+ + c]$ interval, which imply $H(x^+, z^-) \leq g(y^A, z^-)$ and $H(x^-, z^+) \leq g(y^B, z^+)$. Therefore,

$$H(x^+, z^-) + H(x^-, z^+) \leq g(y^A, z^-) + g(y^B, z^+) = H(x^-, z^-) + H(x^+, z^+)$$

from which supermodularity of H follows.

Proof of part (c): We introduce $y^- := \arg \min_{y \in [x; x+c+\theta^-]} g(y, z)$ and $y^+ := \arg \min_{y \in [x; x+c+\theta^+]} g(y, z)$, for which $y^- \leq y^+$ obviously holds.

By supermodularity of $g(y, z)$, we have

$$H(\theta^-, z^-) + H(\theta^+, z^+) = g(y^-, z^-) + g(y^+, z^+) \geq g(y^+, z^-) + g(y^-, z^+) = H(\theta^+, z^-) + H(\theta^-, z^+) \text{ implying the supermodularity of } H(\theta, z). \quad \square$$

Proof of Theorem 2. Proof of part (a) is by induction. The base step consists of the following substeps:

$f_{T+1} \equiv 0$ and $J_T(y_T, U) = \mathcal{L}_T(y_T) + f_{T+1}$ are obviously supermodular. $\min_{y_T \in [x_T; x_T + \theta_T + U]} \mathcal{L}_T(y_T)$ is supermodular in (x_T, θ^T, U) by Lemma 2(a). Finally, $f_T(x_T, \theta^T, U) = U c_p + \theta_T c_c + \min_{y_T \in [x_T; x_T + \theta_T + U]} \mathcal{L}_T(y_T)$ and $J_{T-1}(y_{T-1}, \theta^T, U) = \mathcal{L}_{T-1}(y_{T-1}) + \alpha E[f_T(y_{T-1} - W_{T-1}, \theta^T, U)]$ are supermodular because of the supermodularity preservation of the non-negative linear combination and limit operators (see Lemma 2.6.1 and Corollary 2.6.2 in Topkis (1998)).

The general inductive step includes substeps as in the base step, and one additional substep. That is to prove

$$\begin{cases} \min_{y_t \in [x_t; x_t + \theta_t + U]} J_t(y_t, \theta^{t+1}, U) \text{ is supermodular in } (x_t, \theta^t, U) \text{ with } \theta^t = (\theta_t, \theta_{t+1}, \dots, \theta_{T-1}, \theta_T), & \text{if } T < t + L, \\ \min_{y_t \in [x_t; x_t + \theta_t + U], \theta_{t+L} \geq 0} J_t(y_t, \theta^{t+1}, U) \text{ is supermodular in } (x_t, \theta^t, U) \text{ with } \theta^t = (\theta_t, \theta_{t+1}, \dots, \theta_{t+L-1}, \theta_{t+L}), & \text{if } t + L \leq T \end{cases}$$

given that $J_t(y_t, \theta^{t+1}, U)$ is supermodular and convex in (y_t, θ^{t+1}, U) . The first branch follows directly from Lemma 2, the second branch follows from the supermodularity preservation of the projection operator (see Topkis, 1998), additionally.

Proof of part (b), first statement: From part (a), we have $f_{t+1}(x_{t+1}, \theta^{t+1}, U)$ being supermodular. By the definition of supermodularity, for $x^- \leq x^+, z^- \leq z^+$ we have

$$f_{t+1}(x^+, z^-) + f_{t+1}(x^-, z^+) \leq f_{t+1}(x^-, z^-) + f_{t+1}(x^+, z^+),$$

where $z^\pm = (\theta_t^\pm, \dots, \theta_{\min\{t+L, T\}}^\pm, U^\pm)$ are vectors such that $\theta_t^- \leq \theta_t^+, \dots, \theta_{\min\{t+L, T\}}^- \leq \theta_{\min\{t+L, T\}}^+$, and $U^- \leq U^+$.

We introduce new variables $w^- := y - x^+, w^+ := y - x^-$ with an arbitrary y . For $w^- \leq w^+, z^- \leq z^+$ we have $f_{t+1}(y - w^-, z^-) + f_{t+1}(y - w^+, z^+) \leq f_{t+1}(y - w^+, z^-) + f_{t+1}(y - w^-, z^+)$ for all y . This means that $H_t(w, z) := f_{t+1}(y - w, z)$ is submodular for all t . By submodularity preservation of the expected value and the non-negative linear combination operators (see Topkis, 1998), $J_t = \mathcal{L}_t(y_t) + \alpha E[H_t(W_t, (\theta^{t+1}, U))]$ is also submodular in $(W_t, (\theta^{t+1}, U))$.

Proof of part (b), second statement: We denote the first order stochastic dominance by \preceq . Since $f_{t+1}(x_{t+1}, \theta^{t+1})$ is convex for all t , we have $H_t(x) := f_{t+1}(x, \theta^{t+1})$ convex for all t . $H_t^-(w, y) := H_t(y - w)$ is also submodular in (w, y) for all t (due to Lemma 2.6.2.b in Topkis, 1998).

We introduce $Q(w) := H_t^-(w, y^-) - H_t^-(w, y^+)$ with some $y^- \leq y^+$. Because $H_t^-(w, y)$ is submodular, it has non-increasing differences (see Theorem 2.6.1 in Topkis (1998)), so $Q(w)$ is non-increasing. Therefore, for any $W^- \preceq W^+$, we have $Q(W^+) \preceq Q(W^-)$ implying $E[Q(W^+)] \leq E[Q(W^-)]$, as well (see Proposition 4.1.1 and Lemma 4.7.2 in Puterman, 1994). The latter expression means that $E[H_t^-(W_t, y_t)]$ has non-increasing differences, which is equivalent with its submodularity in (W_t, y_t) .

Finally, $J_t = \mathcal{L}(W_t, y_t) + \alpha E[H_t^-(W_t, y_t)]$ is submodular in (W_t, y_t) because of Lemma 1 and the submodularity preservation of the non-negative linear combination operator (see Corollary 2.7.2 in Topkis, 1998). \square

Proof of Theorem 3. We prove the first statement indirectly. The limiting (y_1^*, θ_2^*) for $T \rightarrow \infty$ exist because of the discountedness (see Puterman, 1994), and $y_1^* = \hat{y}_1$ holds. Let $W_{\min} > 0$ be the smallest possible realization of W . Assume, that (y_1^*, θ_2^*) does not satisfy the complementary slackness property. We can define another feasible strategy (y_1, θ_2) such that $y_1 := y_1^* + \varepsilon$ and $\theta_2 := \theta_2^* - \varepsilon$ with $\varepsilon := \min\{\frac{W_{\min}}{2}, x_1 + \theta_1 + U - y_1^*, \theta_2^*\} > 0$. We study the cost difference ΔJ_1 between the two strategies

$$\Delta J_1 := J_1(y_1^*, \theta_2^*, U) - J_1(y_1, \theta_2, U) = \Pr[\hat{y}_1 \in [x_2 + \varepsilon; \infty)](\mathcal{L}(y_1^*) - \mathcal{L}(y_1) + \alpha c_c \varepsilon) + \Pr[\hat{y}_1 \in [x_2; x_2 + \varepsilon)]C_1 + \Pr[\hat{y}_1 \in (-\infty; x_2)]C_2$$

with $x_2 = y_1^* - W$, and some C_1 and C_2 expected costs for the remaining periods. By the first term of the summation, the two strategies follow the same sample paths from the second period on, while none of the latter two terms are possible, as $0 < \varepsilon < W_{\min}$. Thus, we have $\Pr[\tilde{y}_1 \in [x_2; x_2 + \varepsilon]] = \Pr[\tilde{y}_1 \in (-\infty; x_2)] = 0$ and $\Pr[\tilde{y}_1 \in [x_2 + \varepsilon; \infty)] = 1$.

Therefore, $\Delta J_1 = \mathcal{L}(y_1^*) - \mathcal{L}(y_1^* + \varepsilon) + \alpha c_c \varepsilon$. Since \mathcal{L} is convex and $\mathcal{L}' < h$, we have $\mathcal{L}(y_1^* + \varepsilon) - \mathcal{L}(y_1^*) < h\varepsilon$. Using the required $h < \alpha c_c$ sufficiency condition, we find $\Delta J_1 > -h\varepsilon + \alpha c_c \varepsilon > 0$. However, the positive ΔJ_1 contradicts with (y_1^*, θ_2^*) being the optimum.

The proof of the second part is as follows. For the two-period problem, J_1 can be expressed explicitly. For a given U , the curve of the intersection of J_1 with the plane $y_1 + \theta_2 = 0$ defines a new function, \tilde{J}_1 , which we parameterize with variable y_1 .

$$\tilde{J}_1(y_1) = \mathcal{L}_1(y_1) + \alpha U c_p - \alpha y_1 c_c + \alpha \mathcal{L}_2(\hat{y}_2) [G_1(\omega)]_{y_1 - \hat{y}_2}^{U - \hat{y}_2} + \alpha \int_{-\infty}^{y_1 - \hat{y}_2} \mathcal{L}_2(y_1 - \omega) g_1(\omega) d\omega + \alpha \int_{U - \hat{y}_2}^{\infty} \mathcal{L}_2(U - \omega) g_1(\omega) d\omega.$$

We take the derivative of function $\tilde{J}_1(y_1)$ and look for negative values.

$$0 > \partial_{y_1} \tilde{J}_1(y_1) = +\mathcal{L}'_1(y_1) - \alpha c_c - \alpha \mathcal{L}_2(\hat{y}_2) g_1(y_1 - \hat{y}_2) + \alpha \partial_{y_1} \int_{-\infty}^{y_1 - \hat{y}_2} \mathcal{L}_2(y_1 - \omega) g_1(\omega) d\omega$$

As a result, we have the inequality,

$$\mathcal{L}'_1(y_1) + \alpha(h + b) \int_{-\infty}^{y_1 - \hat{y}_2} G_2(y_1 - \omega) g_1(\omega) d\omega < \alpha c_c + \alpha b G_1(y_1 - \hat{y}_2)$$

By increasing its LHS, we create a sufficient condition for this inequality to hold.

$$\mathcal{L}'_1(y_1) + \alpha(h + b) \int_{-\infty}^{y_1 - \hat{y}_2} G_2(y_1 - \omega) g_1(\omega) d\omega \leq h + \alpha(h + b) \int_{-\infty}^{y_1 - \hat{y}_2} \mathbf{1} g_1(\omega) d\omega = h + \alpha(h + b) G_1(y_1 - \hat{y}_2) \leq h(1 + \alpha) + \alpha b G_1(y_1 - \hat{y}_2).$$

When we check if the increased LHS is still below its RHS, we find

$$h(1 + \alpha) + \alpha b G_1(y_1 - \hat{y}_2) < \alpha c_c + \alpha b G_1(y_1 - \hat{y}_2)$$

which is equivalent to $h(1 + \alpha) < \alpha c_c$.

Consequently, for a given U , $\{y_1 + \theta_2 = 0, y_1 \rightarrow +\infty\}$ is an always decreasing ray for $J_1(y_1, \theta_2, U)$ when $h(1 + \alpha) < \alpha c_c$. Therefore, the constrained optimum of the first period satisfies the complementary slackness property.

References

- Ahn, H., Richter, R., Shanthikumar, J.G., 2005. Staffing decisions for heterogenous workers with turnover. *Mathematical Methods of Operations Research* 62, 499–514.
- Alp, O., Tan, T., 2008. Tactical capacity management under capacity flexibility in make-to-stock systems. *IIE Transactions* 40, 221–237.
- Angelus, A., Porteus, E.L., 2002. Simultaneous capacity and production management of short-life-cycle, produce-to-stock goods under stochastic demand. *Management Science* 48, 399–413.
- Angelus, A., Porteus E.L., 2003. On capacity expansions and deferrals, Technical Report, Graduate School of Business, Stanford University, CA.
- Bertsekas, D., 1976. *Dynamic Programming and Stochastic Control*. Academic Press, New York, NY.
- Eberly, J.C., van Mieghem, J.A., 1997. Multi-factor dynamic investment under uncertainty. *Journal of Economic Theory* 75, 345–387.
- Gans, N., Zhou, Y.P., 2002. Managing learning and turnover in employee staffing. *Operations Research* 50, 991–1006.
- Heyman, D.P., Sobel, M.J., 2004. *Stochastic Models in Operations Research*, vol. II. Dover Publications, Mineola, NY.
- Hiriart-Urruty, J.-B., Lemaréchal, C., 1993. *Convex Analysis and Minimization Algorithms*, vol. 1. Springer-Verlag, Berlin, Germany.
- Pac, M.F., Alp, O., Tan, T., in press. Integrated workforce capacity and inventory management under temporary labor supply uncertainty. *International Journal of Production Research*. doi:10.1080/00207540801930237.
- Porteus, E.L., 2002. *Foundations of Stochastic Inventory Theory*. Stanford University Press, Stanford, CA.
- Puterman, M.L., 1994. *Markov Decision Processes*. John Wiley and Sons Inc., Wiley, New York, NY.
- Rocklin, S.M., Kashper, A., Varvaloucas, G.C., 1984. Capacity expansion/contraction of a facility with demand augmentation dynamics. *Operations Research* 32, 133–147.
- Ryan, S.M., 2003. Capacity expansion with lead times and autocorrelated random demand. *Naval Research Logistics* 50, 167–183.
- Tan, T., Alp, O., in press. An integrated approach to inventory and flexible capacity management under non-stationary stochastic demand and setup costs. *OR Spectrum*. doi:10.1007/s00291-008-0122-y.
- Topkis, D.M., 1998. *Supermodularity and Complementarity*. Princeton University Press, Princeton, NJ.
- Yang, J., Qi, X., Xia, Y., 2005. A production-inventory system with markovian capacity and outsourcing option. *Operations Research* 53, 328–349.