

# Inventory Control with Generalized Expediting

Eric Logan Huggins

huggins\_e@fortlewis.edu

School of Business Administration

Fort Lewis College

Durango, CO 81301-3999

Tava Lennon Olsen

olsen@wustl.edu

Olin Business School

Washington University in St. Louis

St. Louis, MO 63130-4899

Last Revised: November 2007

## Abstract

We consider a single-item, periodic review inventory control problem where discrete stochastic demand must be satisfied. When shortages occur, the unmet demand must be filled by some form of expediting. We allow a very general form for the cost structure of expediting, which might include costs associated with in-house rush production or outsourcing. We explicitly consider the case where expedited production is allowed to produce up to a positive inventory level. We also considered the case where expedited production beyond the shortfall is not permitted; an alternate application for this model is an inventory system with general lost sales costs. For the infinite horizon discounted problem, we characterize the structure of the optimal expediting policy and show that an  $(s, S)$  policy is optimal for regular production. In certain cases, we demonstrate that it may indeed be optimal to use expedited production to build up inventory. For the special cases where the expediting cost function is concave or consists of a fixed and linear per-unit cost, we show that the optimal expediting policy is generalized  $(s, S)$  or order-up-to, respectively. An explicit heuristic for policy calculation is given; a numerical study tests the heuristic and allows us to gain insight into when expediting above zero is cost-effective. We find that, while excess expediting above zero is frequently optimal (particularly when the per-unit costs are close to those of regular production), the actual cost savings from the additional expediting are minimal.

# 1 Introduction

Shutting down an assembly line costs money in the form of both opportunity costs from reduced supply and actual costs of unused labor, facilities, etc. When the shut down is due to missing parts, the supplier is often considered responsible and may be asked to pay for the losses, either explicitly (e.g., penalty payments per unit time) or implicitly (e.g., in loss of future business). Anecdotally, suppliers have historically protected against causing parts shortages by keeping large inventories. However, the shift towards lean inventory has caused many managers to reduce inventories, which in turn may increase the likelihood of stockouts. We consider a problem where the cost of holding inventory may be relatively high and backorders are strictly forbidden. Of course, shortages will occur with stochastic demand; in our problem, these shortages must be filled by some form of expediting.

Our problem is concerned with meeting shortages in a timely fashion. The literature on this subject appears under various names, such as expediting, emergency orders, and dual supply modes. Papers on expediting most related to our work include Daniel (1963), Fukuda (1964), Whittmore and Saunders (1977), Moinzadeh and Schmidt (1991), Chiang and Gutierrez (1998), Johansen and Thorstenson (1998), Tagaras and Vlachos (2001), Groenevelt and Rudi (2002), Sethi et al. (2003), Veeraraghavan and Scheller-Wolf (2006), and Song and Zipkin (2007). (See references therein for further literature.)

One motivating environment for our work was the inventory control problem of a large automobile parts supplier in Michigan, who faced a scenario similar to that of the introductory paragraph. They avoided backorders at the automobile assembly line by either running overtime at the end of a shift (and then shipping overnight by using their regular truck route) or by producing at the beginning of a new shift and then shipping parts by air so that they arrived in a few hours or less. Particularly in the first form of “expediting” above, there is the possibility of production above and beyond the backlog. We refer to this type of expediting as “rush production.” In the second case it seems likely that excess material beyond the backlog cannot be pushed onto the assembler. Further, there will be other scenarios, by managerial decree or because the shortage is actually “met” by lost sales, where no expedited production in excess of the backlog is allowed. We refer to this second form of “expediting” as “penalty production.” The bulk of this paper deals with rush

production; however, penalty production is also studied, as is the combination of the two forms of expediting.

In the above example, rush production corresponds to in-house production. However, another possible application includes outsourced or subcontracted production. Bradley (1997), (2004), and (2005) has looked at (and provided motivation for) subcontracting in a number of forms. There has also been some work that looks at the interaction between subcontracting and capacity investment (e.g., Van Mieghem, 1999) as well as some continuous time models of outsourcing (e.g., Arslan et al., 2001, and Zheng, 1994). Overtime, or a “vendoring option,” in the context of inventory systems with production quotas was considered in Hopp et al. (1993), Duenyas et al. (1993), and Duenyas et al. (1997). Each paper provides structural results for a number of different models. Of these models, the most related to ours is Model 2 in Duenyas et al. (1997). In that model, an order-up-to policy is assumed for regular production and the paper concentrates on evaluating the amount of overtime to use in a system where backlogs are allowed. To the best of our knowledge, our work is the first to determine the structure of the optimal stationary inventory policy for a repeated two-stage decision (i.e., how much to produce and how much to expedite), where at each stage the firm can produce beyond demand backlog.

Although rush production beyond zero is allowed, it may not be optimal. We study its frequency using a numerical study (Section 4.1). We also provide an equation for estimating whether it is likely to be optimal (Section 4.3). Further, we numerically study the cost savings associated with expediting beyond zero inventory. We find that, while excess expediting above zero is frequently optimal (particularly when the per-unit costs are close to those of regular production), the actual cost savings from the expediting are minimal.

We provide explicit heuristic approximations for the parameters in the model with fixed and linear per-unit expediting costs for rush production. These approximations are heavily based on traditional continuous-time lost sales models with an adjustment for fixed and linear costs and a further adjustment for expediting to a positive inventory level. We have focused on explicit approximations that can be easily implemented in a spreadsheet because exact numerical methods (e.g., solving the dynamic programming problem) are already available. See Lee and Nahmias (1993) for a broad discussion on related heuristics.

Our second option for expediting, penalty production, does not allow for production beyond

filling the backlog. As such, it may also be viewed as a shortage or lost sales cost. Relevant papers on this topic include Smith (1977), Çetinkaya and Parlar (1998), Lovejoy and Sethuraman (2000), and Mohebbi and Posner (1999). Closest to our work, Aneja and Noori (1987) consider a problem where unmet demand is met by “some external arrangement” with both per-unit and fixed costs. They assume that if a shortage occurs, the inventory level will be brought up to zero (i.e., penalty production only) and they show that  $(s, S)$  policies are optimal over the finite horizon when the demand density is non-increasing. Ishigaki and Sawaki (1991) extend this work to give a condition based on the problem parameters for  $(s, S)$  policies to be optimal for a finite horizon model with both fixed and per-unit holding and lost sales costs. We believe that our work is the first to extend lost sales models with fixed and per-unit costs to the infinite horizon and the first to consider general non-decreasing lost sales costs.

We study a single-item, periodic review inventory control problem where discrete stochastic demand cannot be backlogged. The traditional problem studied in Scarf (1960) and Veinott (1966) allows for backorders when shortages occur; under this condition,  $(s, S)$  policies are optimal. Similarly, Zheng (1991) allows for backorders and shows that  $(s, S)$  policies are optimal for the infinite horizon case. As in these papers, we do not consider capacity constraints. We use Zheng’s results extensively for our main theorem with a modification to exclude backorders (i.e., we require that  $s \geq -1$ ). However, due to the dual nature of the decision when considering rush production (i.e., we must decide both how much to expedite and how much to produce during regular production), we must embed Zheng’s results within a proof that relies on two viewpoints for the sequence of events. This, in turn, requires us to assume that the demand distribution is logconcave (see Rosling, 2002, for an overview of logconcavity and related results).

As described above, the contribution of our paper is fivefold. First, we provide a model of two different modes of expediting, namely rush production and penalty production. Second, in the first mode of expediting (rush production), we prove structural results for a repeated two-stage decision, where at each stage one can produce beyond demand backlog. We also study the policy values numerically. Third, with the second mode of expediting (penalty production), we generalize previous work on lost sales models to the infinite horizon with general cost functions. Fourth, we provide a novel methodology (using two viewpoints for the sequence of events) for dealing with the two-stage nature of the decisions. Finally, we provide heuristics for the explicit calculation of

problem parameters.

This paper is organized as follows. In Section 2, we present our basic model involving rush production and explain the two time sequence viewpoints that will be used throughout the paper. Section 3 characterizes the structure of optimal expedited production policies and shows that the optimal regular production policy is an  $(s, S)$  policy. It also extends the model to the two modes of expediting (i.e., rush production and penalty production). In Section 4, we develop explicit heuristic approximations and numerically analyze both the frequency and costs savings associated with expediting beyond demand backlog and the performance of the heuristic. Finally, Section 5 concludes the paper.

## 2 Model, Notation, and Assumptions

We consider a model where a manager must make two inventory decisions each day (or period)  $t$ , over an infinite horizon,  $t = 0, 1, 2, \dots$ . At the beginning of day  $t$ , the current inventory level,  $x_t$ , is known and the manager must then decide what inventory level,  $y_t$ , to produce up to with regular production. After regular production is determined, discrete stochastic demand,  $D_t$ , is realized and inventory updated accordingly to  $\tilde{x}_t = y_t - D_t$ . The manager must now decide how many expedited orders to produce given that backlogged demand is not allowed for the following day. Whatever inventory level,  $\tilde{y}_t \geq 0$ , is chosen becomes the starting inventory level for the next day (i.e.,  $x_{t+1} = \tilde{y}_t$ ) and the cycle continues. This sequence of events is depicted in Figure 1, where the time frame outlined is shown on the top half of the graphic. We refer to this sequence of events as the original timing, or the original viewpoint.

We will also use an alternative perspective on the sequence of events because the separation between the time periods is actually artificial. In the second perspective, we consider each cycle of events to begin after demand has been realized and write the model in terms of the overlapping periods shown in the bottom half of Figure 1. We will use the original viewpoint when proving the structure of the optimal stationary policy for regular production and the overlapping period viewpoint when proving the structure of the optimal stationary policy for expedited production.

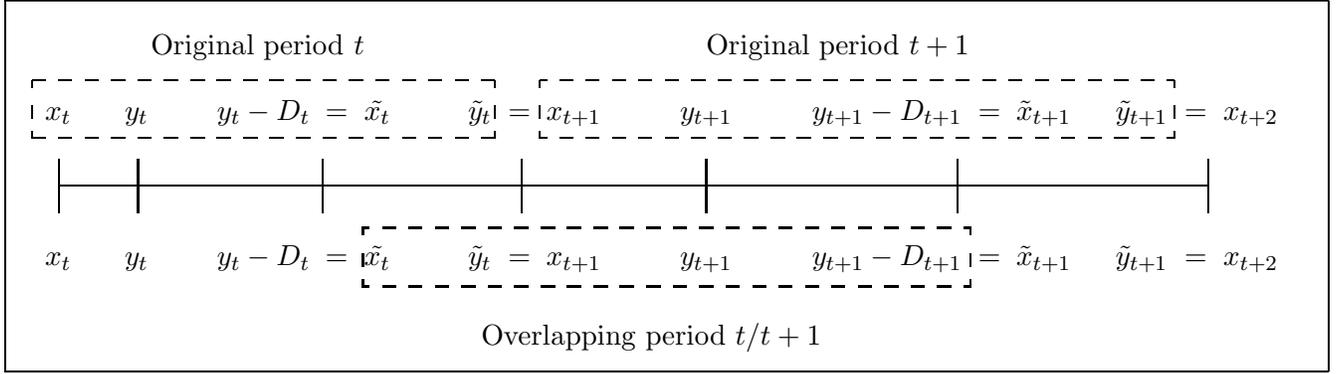


Figure 1: Original and Overlapping Periods

To ensure quasi-convexity of the optimal discounted cost function, we require the demand distribution to be logconcave. More specifically, we assume that the natural log of the probability mass function (pmf)  $\phi(\cdot)$  is a concave function (we will actually be able to relax this assumption to one on the cumulative distribution function (cdf)  $F(\cdot)$  in some of our results). Logconcavity is equivalent to assuming that a function is a Pólya frequency function of order 2 (PF2) (e.g., Theorem 4.1 of Barlow and Proschan, 1965). Note that this assumption is not terribly restrictive because most commonly assumed distributions are, in fact, logconcave. In particular, the normal, uniform, and exponential are all continuous logconcave distributions and the binomial, Poisson, and discrete uniform are all discrete logconcave distributions. The gamma distribution has a logconcave probability density function for appropriate shape parameter choices and a logconcave cdf for all parameter values (e.g., Rosling 2002).

We assume that all costs are stationary and are discounted per period by a factor  $\alpha$ . Regular production incurs a fixed cost of  $K_r \geq 0$  and a per-unit cost of  $c_r \geq 0$ . Expedited production incurs a cost  $e(z)$  for  $z$  units. A general holding cost function  $h(x)$  is assessed to all positive inventory,  $x$ , *after* expediting; thus, holding costs are assessed on the values  $\tilde{y}_t$ ,  $t = 0, 1, 2, \dots$ . We define the following normalized cost functions:

$$h_a(x) \equiv h(x) + (1 - \alpha)c_r x$$

and

$$e_a(z) \equiv e(z) - c_r z.$$

The normalized holding cost incorporates the monetary cost of having produced the unit in

this period rather than the following period and the normalized expediting cost represents the additional cost beyond the regular production cost for the items incurred by expediting.

Define  $\delta(x) = 1$  if  $x > 0$  and  $\delta(x) = 0$  otherwise. We make the following assumptions throughout the paper. Unless otherwise noted, Assumption (A2a) is assumed to hold. However, the (weaker) Assumption (A2b) will sometimes be sufficient (see Lemma 3).

(A1) Demand across periods is independent and identically distributed (i.i.d.). The per period demand is non-negative, and discrete.

(A2a) The per period demand has a logconcave pmf,  $\phi(\cdot)$ .

(A2b) The per period demand has a logconcave cdf,  $F(\cdot)$ .

(A3) Let  $D$  be a random variable with pmf  $\phi(\cdot)$ . Assume  $0 < E[D] < \infty$ ,  $Var[D] < \infty$ , and  $E[e_a(D)] < \infty$ .

(A4)  $0 < \alpha < 1$ .

(A5)  $e_a(z) \geq \alpha K_r \delta(z)$  for  $z \geq 0$ .

(A6)  $e_a(z)$  is non-negative, non-decreasing in  $z$  with  $e_a(0) = 0$  and  $\lim_{z \rightarrow \infty} e_a(z) = \infty$ .

(A7)  $h_a(x)$  is non-negative, non-decreasing in  $x$  with  $h_a(0) = 0$  and  $\lim_{x \rightarrow \infty} h_a(x) = \infty$ .

Note that (A5) implies that the fixed costs associated with expediting are at least as great as those associated with performing regular production in the following period. It also precludes a convex  $e_a(\cdot)$  function. This assumption will only be necessary if expediting above zero is allowed. For the problem with lost sales or rush production only, expediting need not have a fixed cost component. Assumption (A6) implies that the marginal cost of expediting is at least as great as the marginal cost of regular production. Clearly, a fixed and per-unit cost associated with expediting of  $K_e$  and  $c_e$ , respectively, fits this model if  $K_e \geq \alpha K_r$  and  $c_e > c_r$ . However, other more general cost functions, such as ones incorporating quantity discounts, also fit the assumptions. Fixed and linear per-unit costs or concave cost functions will automatically satisfy  $E[e_a(D)] < \infty$  if  $E[D] < \infty$ .

Define  $x^+ = x$  if  $x \geq 0$  and  $x^+ = 0$ , otherwise;  $x^- = -x$  if  $x \leq 0$  and  $x^- = 0$ , otherwise. During period  $t$ , the costs incurred are  $K_r \delta(y_t - x_t) + c_r(y_t - x_t) + e(\tilde{y}_t - \tilde{x}_t) + h(\tilde{y}_t)$ , where  $y_t \geq x_t^+$  and

$\tilde{y}_t \geq \tilde{x}_t^+$ . By substituting  $x_t = \tilde{y}_{t-1}$  and  $y_t = \tilde{x}_t + D_t$ , and observing that  $E[c_r D_t]$  is a finite constant (by Assumption A3) that will not affect the eventual optimization, the costs incurred during period  $t$  (from the perspective of the original sequence of events) can be redefined as (cf., Veinott, 1966):

$$g(x_t, y_t, \tilde{x}_t, \tilde{y}_t) \equiv K_r \delta(y_t - x_t) + e_a(\tilde{y}_t - \tilde{x}_t) + h_a(\tilde{y}_t). \quad (1)$$

Observe that  $g$  is non-negative (by Assumptions A6 and A7).

Let  $\pi$  be an admissible policy if  $y_t \geq x_t^+$  and  $\tilde{y}_t \geq \tilde{x}_t^+$  for all  $t$  and  $y_t$  and  $\tilde{y}_t$  are chosen in a non-anticipatory fashion. In other words,  $y_t$  may only depend on  $x_t$  and  $(x_i, y_i, D_i, \tilde{x}_i, \tilde{y}_i)$  where  $i < t$ ;  $\tilde{y}_t$  may only depend on  $\tilde{x}_t$  and  $(\tilde{x}_{i-1}, \tilde{y}_{i-1}, x_i, y_i, D_i)$  where  $i \leq t$ . Let  $\Pi$  be the set of all such policies. We define the optimal discounted expected cost function as:

$$f^*(x_0) \equiv \min_{\pi \in \Pi} E_{\mathcal{D}_0} \left[ \sum_{t=0}^{\infty} \alpha^t g(x_t, y_t, \tilde{x}_t, \tilde{y}_t) \right], \quad (2)$$

for  $x_0 \in \mathcal{Z}$  where  $\mathcal{D}_0 = \{D_0, D_1, D_2, \dots\}$ . (The right-hand-side of equation 2 will be shown to be well-defined in Lemma 1.) Note that, while  $x_t \geq 0$  for  $t = 1, 2, \dots$ , it will be expositionally convenient to allow  $x_0 < 0$ . As alluded to above, this cost function does not include  $-c_r x_0$  or  $c_r E[D]/(1 - \alpha)$ , but because these are fixed finite costs they do not affect the minimization.

As discussed above, we will use  $f^*(\cdot)$  to prove structural results on the optimal stationary regular production policy. However, to prove structural results on the optimal expediting policy it will be helpful to view the system from the overlapping period perspective depicted in the lower half of Figure 1. From this perspective, the costs per overlapping period may be defined as:

$$\tilde{g}(\tilde{x}_t, \tilde{y}_t, y_{t+1}) \equiv e_a(\tilde{y}_t - \tilde{x}_t) + h_a(\tilde{y}_t) + \alpha K_r \delta(y_{t+1} - \tilde{y}_t). \quad (3)$$

Then, from the overlapping period perspective, the optimal expected discounted cost may be written as:

$$\tilde{f}^*(\tilde{x}) = \min_{\pi \in \Pi} E_{\mathcal{D}_1} \left[ \sum_{t=0}^{\infty} \alpha^t \tilde{g}(\tilde{x}_t, \tilde{y}_t, y_{t+1}) \right], \quad (4)$$

where  $\mathcal{D}_1 = \{D_1, D_2, D_3, \dots\}$  and  $\tilde{x} \in \mathcal{Z}$ . Since  $\tilde{g}(\tilde{x}, \tilde{y}, y) \geq 0$ , Proposition 1.1 of Bertsekas (1995, p. 137) holds and the optimal discounted expected cost function  $\tilde{f}^*$  satisfies

$$\tilde{f}^*(\tilde{x}) = \min_{\tilde{y} \geq \tilde{x}^+, y \geq \tilde{y}^+} E_D \left[ \tilde{g}(\tilde{x}, \tilde{y}, y) + \alpha \tilde{f}^*(y - D) \right]. \quad (5)$$

While equation (5) is sufficient to numerically compute the optimal policy, in order to prove structural results we need to make use of *both* the original time sequence optimization (represented by  $f^*(\cdot)$ ) *and* the overlapping time period optimization (represented by  $\tilde{f}^*(\cdot)$ ).

As would be expected, there is a strong relationship between  $f^*(\cdot)$  and  $\tilde{f}^*(\cdot)$ ; this relationship is given in Lemma 1, which also provides other properties on the cost functions that will be used in the following section. Its proof is provided in the appendix. Note that Lemma 1 proves that  $f^*(x)$  is finite valued for any  $x \in \mathcal{Z}$ , which will allow it to be used as a well-defined and finite-valued function in definition (12).

**Lemma 1**

$$f^*(x) = \min_{y \geq x^+} \left\{ K_r \delta(y - x) + E_D[\tilde{f}^*(y - D)] \right\} < \infty \quad (6)$$

and

$$\tilde{f}^*(\tilde{x}) = \min_{\tilde{y} \geq \tilde{x}^+} \left\{ e_a(\tilde{y} - \tilde{x}) + h_a(\tilde{y}) + \alpha f^*(\tilde{y}) \right\} < \infty, \quad (7)$$

where the corresponding policy that solves equation (5) or jointly solves (6) and (7) is the optimal stationary policy for the system. Further, define

$$f_-^* = \min_{y \geq 0} \left\{ K_r + E_D[\tilde{f}^*(y - D)] \right\},$$

so that, by (6),  $f^*(x) = f_-^*$  for  $x < 0$ . For any  $\tilde{x} < 0$ ,

$$\tilde{f}^*(\tilde{x}) \geq \alpha f_-^*. \quad (8)$$

Finally,

$$K_r + f^*(0) \geq f_-^*. \quad (9)$$

As noted in the lemma,  $f^*(x) = f_-^*$  for  $x < 0$ . Thus,  $f_-^*$  reflects the optimal discounted expected cost were we allowed to start a regular production period in a deficit (excluding the per-unit production costs for this deficit). Equivalently,  $f_-^*$  reflects the optimal discounted expected cost when starting with zero inventory and assuming the fixed cost of regular production must be incurred (in rare cases it may be optimal to produce using expediting alone and hence the fixed cost of regular production is not incurred even if inventory is zero). If we were to assume that regular production always occurs when starting inventory is zero then  $f_-^* = f^*(0)$ . However, it appears that proving that regular production is indeed optimal when inventory is depleted to zero (hence not all production is done using expediting) requires some extra (non-obvious) conditions on the cost functions. In what follows, reading  $f_-^*$  as  $f^*(0)$  may yield the reader added intuition.

### 3 Optimal Production with Expediting

Because our problem has two decision variables (one for expedited production and the other for regular production), our proof is divided into two main steps. In the first step (in Section 3.1), we characterize the structure of the optimal expediting production policy. In the second step (in Section 3.2), we show that an  $(s, S)$  policy is optimal for regular production. To prove our results we rely on the relationship given in Lemma 1 between the two optimal cost functions  $f^*(\cdot)$  and  $\tilde{f}^*(\cdot)$ . Section 3.3 considers an extension to the model where two modes of expediting (rush production and penalty production) are possible.

#### 3.1 Optimal Expediting Policy

This subsection characterizes the optimal expedited production policy and shows how that characterization may be used to derive results for the optimal regular production policy. The following theorem shows that if inventory is non-negative when the expediting decision is made, then expedited production will not be used; this is a direct consequence of Assumptions (A5) and (A6) and, almost certainly, could also be proven directly from first principles (e.g., using a sample path argument). Theorem 1 also provides further structure for the optimal expediting policy when the normalized expediting cost function is either concave (i.e., there are economies of scale) or consists of a fixed and per-unit cost.

**Theorem 1** *Let  $\tilde{y}^*(\tilde{x})$  be the smallest minimizer of (7).*

1. *If  $\tilde{x} \geq 0$ , then  $\tilde{y}^*(\tilde{x}) = \tilde{x}$ .*
2. *If  $e_a(\cdot)$  is concave and  $\tilde{x} < 0$ , then  $\tilde{y}^*(\tilde{x})$  is non-increasing in  $\tilde{x}$ .*
3. *If  $e(z) = K_e\delta(z) + c_e z$  for some fixed  $K_e$  and  $c_e$ , then  $\tilde{y}^*(\tilde{x})$  is constant across  $\tilde{x} < 0$ .*

**Proof:**

*Proof of 1:* From (5), for  $\tilde{x} \geq 0$ ,

$$\begin{aligned} \tilde{f}^*(\tilde{x}) &= \min_{\tilde{y} \geq \tilde{x}, y \geq \tilde{y}^+} \{e_a(\tilde{y} - \tilde{x}) + h_a(\tilde{y}) + \alpha K_r \delta(y - \tilde{y}) + \alpha E[\tilde{f}^*(y - D)]\} \\ &\geq \min_{y \geq \tilde{x}} \left\{ \min_{\tilde{y} \geq \tilde{x}} \{e_a(\tilde{y} - \tilde{x}) + h_a(\tilde{y}) + \alpha K_r \delta(y - \tilde{y})\} + \alpha E[\tilde{f}^*(y - D)] \right\}, \end{aligned} \quad (10)$$

by definition of the minimum. For  $\tilde{y} \geq \tilde{x}$ ,  $e_a(\tilde{y} - \tilde{x}) + h_a(\tilde{y}) + \alpha K_r \delta(y - \tilde{y}) \geq h_a(\tilde{x}) + \alpha K_r \delta(y - \tilde{x})$  by Assumptions (A5) - (A7). Thus,

$$\tilde{f}^*(\tilde{x}) \geq \min_{y \geq \tilde{x}} \{h_a(\tilde{x}) + \alpha K_r \delta(y - \tilde{x}) + \alpha E[\tilde{f}^*(y - D)]\}. \quad (11)$$

Therefore, because (11) is (10) evaluated at  $\tilde{y} = \tilde{x}$ , the optimal expediting policy for  $\tilde{x} \geq 0$  is not to produce and therefore  $\tilde{y}^*(\tilde{x}) = \tilde{x}$ .

*Proof of 2:* Assume  $e_a(\cdot)$  is concave and suppose that  $\tilde{x}_1 < \tilde{x}_2 < 0$  and  $\tilde{y}^*(\tilde{x}_1) < \tilde{y}^*(\tilde{x}_2)$ ; we wish to find a contradiction. By definition of  $\tilde{y}^*(\tilde{x}_2)$  (since  $\tilde{y}^*(\tilde{x}_1) < \tilde{y}^*(\tilde{x}_2)$ ),

$$e_a(\tilde{y}^*(\tilde{x}_1) - \tilde{x}_2) + h_a(\tilde{y}^*(\tilde{x}_1)) + \alpha f^*(\tilde{y}^*(\tilde{x}_1)) > e_a(\tilde{y}^*(\tilde{x}_2) - \tilde{x}_2) + h_a(\tilde{y}^*(\tilde{x}_2)) + \alpha f^*(\tilde{y}^*(\tilde{x}_2)).$$

But, by the concave non-decreasing nature of  $e_a(\cdot)$ ,

$$e_a(\tilde{y}^*(\tilde{x}_1) - \tilde{x}_1) - e_a(\tilde{y}^*(\tilde{x}_1) - \tilde{x}_2) - e_a(\tilde{y}^*(\tilde{x}_2) - \tilde{x}_1) + e_a(\tilde{y}^*(\tilde{x}_2) - \tilde{x}_2) \geq 0$$

so

$$e_a(\tilde{y}^*(\tilde{x}_1) - \tilde{x}_1) + h_a(\tilde{y}^*(\tilde{x}_1)) + \alpha f^*(\tilde{y}^*(\tilde{x}_1)) > e_a(\tilde{y}^*(\tilde{x}_2) - \tilde{x}_1) + h_a(\tilde{y}^*(\tilde{x}_2)) + \alpha f^*(\tilde{y}^*(\tilde{x}_2)),$$

which contradicts the optimality of  $e_a(\tilde{y}^*(\tilde{x}_1))$ .

*Proof of 3:* Suppose  $e(z) = K_e \delta(z) + c_e z$ . For  $\tilde{x} < 0$  expedited production is required and, therefore,

$$e_a(\tilde{y} - \tilde{x}) + h_a(\tilde{y}) + \alpha f^*(\tilde{y}) = K_e + (c_e - c_r)(\tilde{y} - \tilde{x}) + h_a(\tilde{y}) + \alpha f^*(\tilde{y}).$$

Thus,

$$\tilde{f}^*(\tilde{x}) = K_e - (c_e - c_r)\tilde{x} + \min_{\tilde{y} \geq 0} \{(c_e - c_r)\tilde{y} + h_a(\tilde{y}) + \alpha f^*(\tilde{y})\},$$

where the term inside the minimization is independent of (and hence constant in)  $\tilde{x}$ .  $\square$

Theorem 1 shows that if inventory is negative and the expediting cost function is concave, then the expedite-up-to amount is non-decreasing in shortage. In other words, a generalized  $(s, S)$  policy (as defined in Porteus, 1971) is optimal for expedited production when there are concave expediting costs. According to Porteus's definition, the ordering policy  $\tilde{y}$  is "generalized  $(s, S)$ " if there exists  $(\tilde{s}, \tilde{S})$  such that  $\tilde{y}(\tilde{x}) = \tilde{x}$  for  $\tilde{x} \geq \tilde{s}$  and  $\tilde{y}(\tilde{z}) \geq \tilde{y}(\tilde{x}) \geq \tilde{S} \geq \tilde{s}$  for  $\tilde{z} < \tilde{x} < \tilde{s}$ . In our case,  $\tilde{s} = 0$  and  $\tilde{S} = \tilde{y}(-1) \geq 0$  (recall that inventory is assumed to be discrete). For the special case of fixed plus linear costs  $\tilde{S}$  is the fixed base-stock level that is optimal when expediting (i.e., when  $\tilde{x} < \tilde{s} = 0$ ).

In Assumption (A5) we assumed that  $e_a(z) \geq \alpha K_r \delta(z)$  for  $z \geq 0$ . This was used in Theorem 1 to show that expediting is not used if inventory is non-negative entering the expedited production period. Without this restriction it may well be optimal for expedited production to be used even if inventory is positive. In this case, we would no longer expect an  $(s, S)$  policy to be optimal for regular production (as will be shown to be the case in the following subsection).

The following lemma sets the stage (mathematically) for incorporating the optimal expediting policy into the optimal regular production policy. The given function  $p(z)$  represents the penalty associated with a backlog of size  $z$  following regular production and the realization of demand. Note that a function  $g(x)$  is quasi-convex if and only if  $-g(x)$  is unimodal.

**Lemma 2** *Define, for  $z \geq 0$ ,*

$$p(z) \equiv \min_{y \geq 0} \{e_a(z+y) + h_a(y) + \alpha f^*(y)\} - \alpha(f^*(0) - \delta(z)(f^*(0) - f_-^*)), \quad (12)$$

*which is well-defined by Lemma 1, then  $p(z)$  is non-decreasing in  $z$  with  $p(0) = 0$  and*

$$\tilde{f}^*(\tilde{x}) = h_a(\tilde{x}^+) + p(\tilde{x}^-) + \alpha f^*(\tilde{x}), \quad (13)$$

*where  $h_a(\tilde{x}^+) + p(\tilde{x}^-)$  is quasi-convex in  $\tilde{x}$  with  $\lim_{|\tilde{x}| \rightarrow \infty} (h_a(\tilde{x}^+) + p(\tilde{x}^-)) = \infty$ .*

**Proof:** *Proof that  $p(0) = 0$ :* By definition  $\tilde{y}^*(-z)$  is the minimizer of the right hand-side of (12) for  $z \geq 0$ . from Theorem 1,  $\tilde{y}^*(0) = 0$ , so that:

$$p(0) = e_a(0) + h_a(0) + \alpha f^*(0) - \alpha f^*(0) = 0,$$

since, by Assumptions (A6) and (A7),  $e_a(0) + h_a(0) = 0$ .

*Proof that  $p(z)$  is non-decreasing in  $z$ :* For  $z_1 > 0$ , if  $\tilde{y}^*(-z_1) > 0$ ,

$$\begin{aligned} p(z_1) &= \tilde{f}(\tilde{y}^*(-z_1)) - \alpha f_-^* \\ &\geq 0 = p(0) \end{aligned}$$

from Lemma 1. If, for  $z_1 > 0$ ,  $\tilde{y}^*(-z_1) = 0$ , then,

$$\begin{aligned} p(z_1) &= e_a(z_1) + \alpha(f^*(0) - f_-^*) \\ &\geq 0 = p(0), \end{aligned}$$

from Lemma 1, since  $e_a(z_1) \geq \alpha K_r$  by Assumption (A5). Thus, in both cases,  $p(z_1) \geq p(0)$  for  $z_1 > 0$ . Now for  $0 < z_1 \leq z_2$ ,

$$\begin{aligned} p(z_1) &= e_a(z_1 + \tilde{y}^*(-z_1)) + h_a(\tilde{y}^*(-z_1)) + \alpha(f^*(\tilde{y}^*(-z_1)) - f_-^*) \} \\ &\leq e_a(z_1 + \tilde{y}^*(-z_2)) + h_a(\tilde{y}^*(-z_2)) + \alpha(f^*(\tilde{y}^*(-z_2)) - f_-^*) \} \\ &\leq e_a(z_2 + \tilde{y}^*(-z_2)) + h_a(\tilde{y}^*(-z_2)) + \alpha(f^*(\tilde{y}^*(-z_2)) - f_-^*) \} = p(z_2), \end{aligned}$$

where the first inequality follows from the definition of minimum and the second from the fact that  $e_a(x)$  is non-decreasing in  $x$ . Thus  $p(z)$  is non-decreasing in  $z$ .

*Proof of (13):* From equation (7),

$$\tilde{f}^*(\tilde{x}) = e_a(\tilde{y}^*(\tilde{x}) - \tilde{x}) + h_a(\tilde{y}^*(\tilde{x})) + \alpha f^*(\tilde{y}^*(\tilde{x})).$$

If  $\tilde{x} \geq 0$ , then, from Theorem 1, this implies

$$\tilde{f}^*(\tilde{x}) = e_a(0) + h_a(\tilde{x}) + \alpha f^*(\tilde{x}) = h_a(\tilde{x}^+) + p(\tilde{x}^-) + \alpha f^*(\tilde{x}),$$

where the final equality follows since  $p(0) = 0$ . If  $\tilde{x} < 0$ , then

$$h_a(\tilde{x}^+) + p(\tilde{x}^-) + \alpha f^*(\tilde{x}) = p(\tilde{x}^-) + \alpha f_-^* = \tilde{f}^*(\tilde{x}),$$

where the final equality follows from (7) and the definition of  $p(\cdot)$ .

*Proof that  $h_a(\tilde{x}^+) + p(\tilde{x}^-)$  is quasi-convex in  $\tilde{x}$  with  $\lim_{|\tilde{x}| \rightarrow \infty} (h_a(\tilde{x}^+) + p(\tilde{x}^-)) = \infty$ :* Since  $p(z)$  is non-decreasing in  $z$ ,  $p(x^-)$  is non-increasing in  $x$ . Further, since  $p(0) = h_a(0) = 0$  and, for  $x \geq 0$ ,  $h_a(x)$  is non-decreasing in  $x$  we have that  $h_a(\tilde{x}^+) + p(\tilde{x}^-)$  is quasi-convex. That  $\lim_{\tilde{x} \rightarrow \infty} (h_a(\tilde{x}^+) + p(\tilde{x}^-)) = \infty$  follows immediately from Assumption (A7). Further, by an argument similar to (9),  $f^*(y) - f_-^* \geq -2K_r$ , for any  $y \geq 0$ . Therefore, by Assumption (A6),  $\lim_{\tilde{x} \rightarrow -\infty} (h_a(\tilde{x}^+) + p(\tilde{x}^-)) = \infty$ .

This completes the proof.  $\square$

In Lemma 2, if expedited production above zero is not allowed (i.e., penalty production only), then  $p(z)$  simply equals  $e_a(z) - \alpha \delta(z)(f^*(0) - f_-^*)$ . Further, if regular production occurs when initial inventory is zero (as will usually be the case), then  $p(z) = e_a(z)$  and the intuition given prior to the lemma, that  $p(z)$  represents the penalty associated with a backlog of size  $z$ , becomes transparent.

### 3.2 Optimal Regular Time Production Policy

Combining (13) and (6) we have that

$$f^*(x) = \min_{y \geq x^+} \{K_r \delta(y - x) + G(y) + \alpha E_D[f^*(y - D)]\}, \quad (14)$$

where we define

$$G(y) \equiv E_D[h_a((y - D)^+) + p((y - D)^-)].$$

**Lemma 3** *If either (i) assumption (A2a) holds, or (ii)  $e(z) = K_e \delta(z) + c_e z$ ,  $h_a(x) = h_a x$  for some  $h_a > 0$ , and assumption (A2b) holds, then  $G(y)$  is quasi-convex in  $y$ .*

**Proof:** For (i), using the fact that the pmf of demand is logconcave and  $h_a(\tilde{x}^+) + p(\tilde{x}^-)$  is quasi-convex in  $\tilde{x}$ , we have that  $G(y)$  is quasi-convex in  $y$  (Ibragimov, 1956).

For (ii), using the same reasoning as in the proof of Theorem 2 part 3,

$$\begin{aligned} p(z) &= \delta(z)(K_e + \alpha(f^*(0) - f_-^*)) + \min_{\tilde{y} \geq 0} \{(c_e - c_r)\tilde{y} + h_a(\tilde{y}) + \alpha f^*(\tilde{y})\} - \alpha f^*(0) \\ &= \delta(z)\tilde{K}_e + C^*, \end{aligned}$$

where we define  $\tilde{K}_e \equiv K_e + \alpha(f^*(0) - f_-^*)$  and  $C^* \equiv \min_{\tilde{y} \geq 0} \{(c_e - c_r)\tilde{y} + h_a(\tilde{y}) + \alpha f^*(\tilde{y})\} - \alpha f^*(0)$ .

Hence

$$G(y) = h_a(y - E_D[D]) + h_a E_D[(D - y)^+] + \tilde{K}_e P(D > y) + C^*.$$

Dropping the  $C^*$ , which is constant, this is equivalent to equation (11) in Rosling (2002) by setting  $L = 1$ ,  $h = h_a$ ,  $\pi = p = 0$ , and  $b = \tilde{K}_e$ . Therefore, by Proposition 2.1 in Rosling,  $G(y)$  is quasi-convex in  $y$ .  $\square$

We can therefore characterize the optimal regular production policy as follows.

**Theorem 2** *Under the conditions of Lemma 3, the optimal regular production policy is an  $(s, S)$  policy with  $-1 \leq s < S$ .*

**Proof:** Equation (14) is the optimization equation from Zheng (1991) except for the  $y \geq x^+$  rather than  $y \geq x$ . From Lemma 3,  $G(\cdot)$  is quasi-convex. Further,  $\lim_{|\tilde{x}| \rightarrow \infty} G(y) = \infty$  since  $h(\cdot)$  and  $p(\cdot)$  are monotone and therefore the interchange of limit and expectation follows from the monotone convergence theorem. Thus,  $G(\cdot)$  satisfies the required assumptions in Zheng. For  $x < 0$ , redefine  $G(x) = \infty$  and allow backordering. Note that  $G(\cdot)$  remains quasi-convex with  $\lim_{|x| \rightarrow \infty} G(x) = \infty$ .

This is an equivalent optimization and the optimality of an  $(s, S)$  policy follows directly from Zheng (1991). This infinite cost immediately implies it is better to order than not order when  $x < 0$ , and hence  $s \geq -1$ .  $\square$

Note that Theorem 2 applies both to models with the possibility of expediting above zero and to models with penalty production or lost sales. In the latter case the expediting or lost sales cost function need only satisfy Assumption (A6), which guarantees quasi-convexity of  $G(\cdot)$ ; there are no structural results to prove for the expediting or lost sales policy and hence, as mentioned earlier, Assumption (A5) is unnecessary.

Corollary 1 provides further structure on the optimal policy. In particular, it shows that if expedited production is used to build up inventory in some period, then regular-time production will not be used in the following period. It gives a (relatively restrictive) condition on when expedited production above zero is not used. Finally, it shows that the up-to amount when expediting will never exceed the optimal regular-time produce-up-to amount if the latter amount is positive.

**Corollary 1** *Let  $(s^*, S^*)$  be the optimal regular production controls and let  $\tilde{y}^*(\tilde{x})$  be the optimal produce-up-to amount when expediting (i.e., when  $\tilde{x} < 0$ ).*

1. *If, for some  $\tilde{x} < 0$ ,  $\tilde{y}^*(\tilde{x}) > 0$ , then  $s^* < \tilde{y}^*(\tilde{x})$ .*
2. *If, for some  $\tilde{x} < 0$ ,  $e_a(y - \tilde{x}) - e_a(-\tilde{x}) + h_a(y) \geq \alpha K_r$  for all  $y > 0$ , then  $\tilde{y}^*(\tilde{x}) = 0$ .*
3. *If  $S^* > 0$ , then  $\tilde{y}^*(\tilde{x}) \leq S^*$  for any  $\tilde{x} \leq S^*$ .*

**Proof:** See Appendix.  $\square$

Note that Condition 2 of Corollary 1 implies that, if  $K_r = 0$  (i.e., there are no fixed costs associated with regular time production), then expedited production beyond zero will never be used. Both this and Condition 2 could likely also be proven directly from first principles because the intuition behind the condition (and associated proof) is simply that the per period fixed cost of regular production in the next period is always less than the extra expediting cost of producing beyond zero plus the holding cost of the extra units of inventory.

Our proofs yield results for the infinite horizon case, bypassing any finite horizon results. The dual decision nature of the problem, combined with end of horizon effects, make finite horizon results difficult to prove. We have found examples via numerical analysis where, near the end of

the horizon, the parameters vary non-monotonically, preventing us from using monotonicity results for the finite horizon. Further, the infinite horizon model coupled with a quasi-convex cost function fits naturally into the framework of Zheng (1991).

### 3.3 Multiple Modes of Expediting

In many cases one may have two (or more choices) for the expediting options, each having differing cost structures. In this section we assume that there are two modes of expedited production: rush production, which may occur up to any amount, and penalty production, which may only be used to produce up to zero inventory. The motivation for the former is in-house production, while the motivation for the latter is either an outside source (which may be problematic in non-commodity environments) or simply lost sales. In the original environment which motivated this work, rush production was overtime and penalty production was expedited shipping, where the managerial policy was that expedited shipping was not to be used beyond the shortage. While general concave cost functions are easily modeled, here we restrict our attention to fixed and linear costs (for both rush and penalty production) for ease of exposition.

We assume that rush production of  $z$  units costs  $K_o\delta(z) + c_o z$  for some  $K_o \geq \alpha K_r$  and  $c_o > c_r$  and penalty production of  $z$  units costs  $K_p\delta(z) + c_p z$  for some  $K_p \geq \alpha K_r$  and  $c_p > c_r$ . For  $w \geq z \geq 0$ , it can readily be shown that,

$$K_o + c_o z + K_p + c_p(w - z) \geq K_o + K_p + \min(c_o, c_p)w.$$

The left side of the equation is the cost of rush production of  $z$  units and penalty production of  $w - z$  units. Since  $\min(K_o + c_o w, K_p + c_p w) \leq K_o + K_p + \min(c_o, c_p)w$ , it is never optimal for both modes of expediting to occur simultaneously. A formal proof is trivial and would mirror Theorem 1 (or possibly even a more direct argument).

Define

$$C^* \equiv \min_{y \geq 0} \{(c_o - c_r)y + h_a(y) + \alpha f^*(y)\} - \alpha f^*(0)$$

and

$$\tilde{S} \equiv \arg \min_{y \geq 0} \{(c_o - c_r)y + h_a(y) + \alpha f^*(y)\}.$$

Then, when  $\tilde{S} > 0$ ,  $C^*$  represents the added cost (or, more precisely,  $-C^* \geq 0$  represents the added benefit) in using expediting up-to  $\tilde{S}$  instead of up-to 0. Using this,  $p(z)$  in Lemma 2, should now

be redefined as, for  $z \geq 0$ ,

$$p(z) = \min\{K_o + (c_o - c_r)z + C^*, K_p + (c_p - c_r)z\} + \delta(z)(f^*(0) - f_-^*),$$

where, again,  $p(z)$  represents the penalty associated with a backlog of size  $z$  following regular production and demand (but now when there are two expediting modes as described above). Hence, the values of  $K_o$ ,  $c_o$ ,  $K_p$ ,  $c_p$ , and  $C^*$  determine which mode of expedited production is best, and when. In particular, if  $c_o \neq c_p$ , define

$$z^* \equiv \left\lfloor \frac{C^* + K_o - K_p}{c_o - c_p} \right\rfloor,$$

where, for any  $x$ ,  $\lfloor x \rfloor$  is the largest integer less than or equal to  $x$ . Note that  $z^*$  remains undefined if  $c_o = c_p$  because it is unnecessary. The optimal expediting policy is represented in Table 1.

	$K_o + C^* > K_p$	$K_o + C^* = K_p$	$K_o + C^* < K_p$
$c_o < c_p$	PP for $z^* < \tilde{x} < 0$ RP for $\tilde{x} \leq z^*$	RP for all $\tilde{x} < 0$	RP for all $\tilde{x} < 0$
$c_o = c_p$	PP for all $\tilde{x} < 0$	Any	RP for all $\tilde{x} < 0$
$c_o > c_p$	PP for all $\tilde{x} < 0$	PP for all $\tilde{x} < 0$	RP for $z^* < \tilde{x} < 0$ PP for $\tilde{x} \leq z^*$

Table 1: Regions in Which to Use Rush Production (RP) Versus Penalty Production (PP)

**Theorem 3** *An optimal stationary expediting production policy exists and has structure as in Table 1 where, under rush production, the optimal produce-up-to amount equals  $\tilde{S}$ , and, under penalty production, the up-to amount is zero by definition. Regular production follows an  $(s, S)$  policy.*

**Proof:** Follows directly from the above, Theorem 2, and simple algebra.  $\square$

Note that when  $C^* = 0$  (for example, when  $K_r = 0$ ), the optimal expediting policy depends only on the relative expediting costs and is independent of the demand distribution.

## 4 Numerical and Heuristic Analysis

In this section, we consider a single method of expediting with fixed and linear per-unit costs of  $K_e$  and  $c_e$ , respectively; that is,  $e_a(z) = K_e\delta(z) + (c_e - c_r)z$  for  $z \geq 0$ . We also assume that holding

costs are linear with  $h_a(x) = (h + (1 - \alpha)c_r)x \equiv h_a x$ , where (in a slight abuse of notation)  $h_a$  is defined as  $h + (1 - \alpha)c_r$ .

In this case, Theorems 1 and 2 imply that the optimal regular production policy is  $(s, S)$  with optimal policy parameters  $s^*$  and  $S^*$ , say, and the optimal expediting policy is to expedite up to some fixed level  $\tilde{S}^* \geq 0$ . In Section 4.1, we use numerical analysis to show that excess expedited production is frequently optimal (i.e.,  $\tilde{S}^* > 0$ ) when the expediting per-unit cost is close to the regular per-unit cost; however, we also determine that the cost savings of excess expediting are limited when compared to the case where expediting only fills the shortage. In Section 4.2, we develop explicit heuristics to estimate the optimal policy parameters  $s^*$ ,  $S^*$ , and  $\tilde{S}^*$ . Finally, in Section 4.3, we numerically analyze the heuristic policy and show that it performs reasonably well; we also discuss a simple method to determine whether excess expediting is worth considering.

#### 4.1 Excess Expediting

For our numerical analyses, we set the regular per-unit cost  $c_r$  to 10 and vary all other cost parameters. For each combination, we determine the total discounted cost and optimal policy parameters for regular and expedited production using value iteration programmed with C++ code. We also solve for these values assuming that excess expedited production is not allowed (i.e., fixing  $\tilde{S} = 0$ ).

The regular setup cost,  $K_r$ , was tested at the values of 50, 100 and 200. We ignored the case where  $K_r = 0$  because excess expedited production is never optimal in that case; larger regular setup costs seemed unnecessary. We let the (unadjusted) holding cost  $h = 0.0025, 0.005$ , and  $0.01$ . For expedited production, we let the per-unit cost  $c_e = 10.1, 10.25, 10.5, 11$ , and  $15$ , where the smaller costs correspond to the expediting being performed by a second shift or some other form of inexpensive expediting, rather than the largest cost which represents traditional “time-and-a-half.” The expedited setup cost,  $K_e$ , was tested at 75, 150 and 300.

Finally, we let  $\alpha = 0.99, 0.999$ , and  $0.9999$ , roughly corresponding to quarterly, weekly, and daily discounting. We assume that our holding cost is a cost per unit *per unit time*, so to accurately compare scenarios we alter  $h$  when we change  $\alpha$ . We anchor our holding costs to the case with  $\alpha = 0.9999$ . We adjust our holding costs for  $\alpha = 0.99$  by multiplying  $h$  by  $\ln(0.99)/\ln(0.9999)$ ; for  $\alpha = 0.999$ , we multiply  $h$  by  $\ln(0.999)/\ln(0.9999)$ . In doing so, we test annual per-unit inventory

costs roughly equivalent to 10%, 20% and 40% of the regular production cost per unit.

The above variations lead to  $3^5 4 = 405$  combinations, of which 270 (two-thirds) are feasible because some of the combinations violate Assumption (A5) with  $K_e < \alpha K_r$ . Requiring discrete, logconcave probability distributions, we ran the experiment for the discretized normal distribution (with  $\mu = 25$  and  $\sigma = 5$ ) truncated on  $[0,50]$  and for the discrete uniform distribution on  $[0,50]$ . The results for both distributions were similar; we focus our discussion below on the experimental outcomes of the normal distribution, with a brief discussion of the uniform results at the end of the section.

For the normal distribution, excess expedited production is frequently optimal, 36% of the time. However, the benefit of excess expedited production is minimal, saving on average only 0.10% on inventory/expediting costs (with median benefit of 0.00% and a maximum benefit of 1.18%) when compared to the model where expedited production just fills the shortage. Next, we examined how the different parameters affected the frequency of excess expediting (FEE) and the average savings (AS) on inventory/expediting costs. The FEE and AS decrease as the per-unit cost of expediting  $c_e$  and the discount factor  $\alpha$  increase. Conversely, the FEE and AS increase as the regular setup cost  $K_r$  increases. The holding cost  $h$  (after adjusting for  $\alpha$ ) shows little or no effect as does expediting setup cost  $K_e$  since it is basically a sunk cost when the expediting decision is made. Tables 2 and 3 list the FEE and AS for small, medium and large values of each correlated parameter (for  $c_e$ , small implies  $c_e \leq 10.5$ , medium is  $c_e = 11$  and large is  $c_e = 15$ ).

	Small	Medium	Large
$c_e$	59%	6%	0%
$\alpha$	48%	39%	19%
$K_r$	25%	42%	57%

Table 2: Frequency of Excess Expediting by Parameter

Excess expediting occurs less often as the per-unit cost of expediting increases and the cost savings become negligible. [In fact, excess expediting never occurs when the linear cost of expediting represents traditional “time-and-a-half” with  $c_e = 15$ .] This calls into question whether excess expedited production is practical or not. Inventory managers would have to answer this question

	Small	Medium	Large
$c_e$	0.16%	0.01%	0.00%
$\alpha$	0.21%	0.07%	0.01%
$K_r$	0.04%	0.12%	0.22%

Table 3: Average Inventory/Expediting Cost Savings by Parameter

themselves, but in some cases (where perhaps a “free” production day could be used to produce a different product, for plant maintenance, etc.), it may make sense. On the other hand, since the savings are generally low, managers may choose to forego the option of excess expedited production. In Section 4.3 we present a simple calculation that helps determine whether excess production may have value.

For the uniform distribution, excess expedited production is optimal 32% of the time and the average benefit is 0.15% (with a median benefit of 0.00% and a maximum benefit of 2.96%). The parameters had similar effects on the frequency and cost savings of excess expediting. Lastly, if the linear cost of expediting is very close to the regular linear cost ( $c_e = 10.1$ ), excess expediting saves on average about 0.5% under the assumption of uniform demand.

## 4.2 Heuristic Development

This section provides explicit heuristics that produce estimates  $s^H$ ,  $S^H$ , and  $\tilde{S}^H$  for the optimal policy parameters  $s^*$ ,  $S^*$ , and  $\tilde{S}^*$ , respectively. We focus on explicit heuristics that are readily implemented (on a spreadsheet, for example) rather than more complex (although possibly more accurate) iterative procedures.

Throughout our heuristic development, we will approximate our discounted model by an average cost model. This has been shown in simpler settings to provide a reasonable approximation (e.g., p. 80, Hadley and Whitin, 1963, Porteus, 1985b). We will use equation (1) for the per period cost function and will therefore continue to use the adjusted cost functions  $h_a(\cdot)$  and  $e_a(\cdot)$ , despite not having any other model of discounting. We believe this is more likely to yield accurate results than ignoring discounting altogether because by using these functions the time value of production costs is being accounted for.

The basic steps to derive our heuristics are as follows. First, we begin by approximating the system using a continuous time  $(Q, r)$  model of production where  $r = s$  and  $Q \approx S - s$ . Under such an approximation, the leadtime for replenishment is taken as one period (see, e.g., Porteus, 1995a). Using this model, and ignoring the possibility of shortages, we find the heuristic order quantity  $Q_r^H$ . Second, we ignore the possibility of excess ordering to derive the value of  $s^H$ . These two values are then used to compute  $S^H$ . Finally, the values of  $s^H$  and  $S^H$  are used to compute  $\tilde{S}^H$ .

Ignoring the possibility of shortages, we estimate the heuristic order quantity  $Q_r^H$  as the EOQ value. That is, we set

$$Q_r^H \equiv \sqrt{\frac{2E[D]K_r}{h_a}}. \quad (15)$$

We justify this simplification by noting that total order costs are known to frequently be robust to the choice of  $Q$  (e.g., p. 9, Lee and Nahmias, 1993). Note that this is equivalent to setting  $\partial C_Q / \partial Q_r = 0$  where

$$C_Q \equiv \frac{E[D]}{Q_r} K_r + \frac{h_a Q_r}{2}.$$

If we ignore the fixed cost of expediting (for the moment), then with the continuous time approximation described above, the system is equivalent to a lost sales model with the adjusted per-unit expediting cost,  $c_e - c_r$ , replacing the lost sales cost. Using the approximation of Hadley and Whitin (1963) (see their equation (4-20)) the long-run average cost may be estimated as

$$\frac{E[D]}{Q_r} [K_r + (c_e - c_r)E[(D - s)^+]] + h_a \left[ \frac{Q_r}{2} + E[(s - D)^+] \right].$$

We take these costs as given and now add in the additional cost for the fixed cost of expediting. Under the continuous approximation above, if demand is in single units, the fixed cost of expediting will be incurred in a cycle whenever the final demand  $D > s$ . However, this significantly overestimates the actual probability that expediting occurs because demand is not in single units but in lumps of size  $D$ . We will therefore use a renewal theory argument to more accurately reflect this cost.

In the discrete time model, consider an arbitrary cycle of periods that begins with inventory  $S$  and ends with inventory less than or equal to  $s$  (recall that no excess expediting is allowed at this time so the following period, which starts a new cycle, will again begin with inventory  $S$ ). Relabel the first period of the cycle period 1 and, for  $q \geq 0$  let  $N(q)$  be the (random) number of demands

such that  $\sum_{n=1}^{N(q)} D_n < q \leq \sum_{n=1}^{N(q)+1} D_n$ . Then  $\{N(q) : q \geq 0\}$  forms a renewal process (e.g., Karlin, 1958) and if  $Q = S - s$  then  $N(Q) + 1$  is the number of periods in the cycle. Define the residual life process (sometimes also called “excess”, although that term could cause confusion in our context),  $Y(q) = \sum_{n=1}^{N(q)+1} D_n - q$  (e.g., p. 273 Karlin, 1958); thus  $Y(Q)$  is the degree to which we “overshoot”  $s$  at the end of a cycle. Let  $I$  be a random variable that represents the inventory level at the end of the cycle immediately before expediting and/or reordering so that  $I = s - Y(Q)$ . We assume that  $Q$  is large so that this distribution of  $Y(Q)$  may be approximated by its limiting distribution which has probability mass function  $P(Y(\infty) = i) = P(D \geq i)/E[D]$  for  $i = 0, 1, 2, \dots$ . Then the probability of expediting equals

$$P(I < 0) = P(s - Y(Q) < 0) = P(Y(Q) > s) \approx \sum_{i=s+1}^{\infty} P(D \geq i)/E[D]$$

and the average cost associated with the fixed cost of expediting may be approximated by

$$\frac{E[D]}{Q_r} K_e \sum_{i=s+1}^{\infty} P(D \geq i)/E[D]$$

Putting the above together we therefore estimate the average cost per cycle,  $C_r$ , as

$$C_r \equiv \frac{E[D]}{Q_r} \left[ K_r + (c_e - c_r)E[(D - s)^+] + K_e \sum_{i=s+1}^{\infty} P(D \geq i)/E[D] \right] + h_a \left[ \frac{Q_r}{2} + E[(s - D)^+] \right].$$

Taking second differences this function is easily seen to be convex in  $s$ . Therefore, taking first differences implies that we must find the minimum  $s$  such that

$$\frac{E[D]}{Q_r} \left[ -(c_e - c_r)P(D > s) - \frac{K_e}{E[D]}P(D > s) \right] + h_a P(D \leq s) \geq 0.$$

Solving for  $F(s)$ , we define  $s^H$  as the minimum  $s$  such that:

$$F(s) \geq \left( \frac{(c_e - c_r) + \frac{K_e}{E[D]}}{(c_e - c_r) + \frac{K_e}{E[D]} + h_a \frac{Q_r^H}{E[D]}} \right). \quad (16)$$

This formula is highly intuitive. It is a newsvendor-type expression where the marginal cost of underestimating  $s$  is represented by  $c_e - c_r + K_e/E[D]$  (approximately the per unit cost of expediting rather than having produced another unit in regular production) and the marginal cost of overestimating  $s$  is given by  $h_a Q_r^H/E[D]$  (approximately the per-unit holding cost because  $Q_r^H/E[D]$  is the expected time between cycles).

To estimate  $S^*$ , the average order size in the discrete model will be larger than  $S - s$  (see, e.g., p. 139 of Porteus, 1985). In fact, it will be larger by  $Y(Q)$  (defined above). Again, assuming  $Q$

large so the limiting distribution of  $Y(Q)$  can be used, this “overshoot” has mean  $E[D^2]/(2E[D]) = E[D]/2 + Var[D]/(2E[D])$ . We therefore estimate  $S^H$  as follows:

$$S^H \equiv Q_r^H + s^H - \frac{E[D]}{2} - \frac{Var[D]}{2E[D]}. \quad (17)$$

Note that a similar approximation was made (for similar reasons) by Roberts (1962) and Schneider (1978).

Finally, relaxing our assumption of no excess expedited production above, we must determine when to use it and how much to use. We assume that demand must be met. When a shortage occurs, expedited production takes place until the shortage is filled and at that point the decision must be made whether to continue expediting or to use regular production during the next period to build up inventory. We define  $Q_e^H$  as the excess expedited production amount after the shortage is filled.

The possibility of excess expediting occurs only in those cycles where inventory has been depleted below zero (otherwise no expediting occurs). In this case, expediting must occur, and therefore the fixed cost of expediting along with the expediting costs for the backlog, may be considered “sunk”. Ignoring these sunk cost, the long run average cost over just the excess expediting cycles equals

$$C_e \equiv \frac{E[D]}{Q_e} (Q_e(c_e - c_r)) + \frac{h_a Q_e}{2}$$

Whereas, if no expediting occurs, then the long run average cost of a regular production cycle is given by

$$C_r \equiv \frac{E[D]K_r}{Q_r} + \frac{h_a Q_r}{2}.$$

Therefore, the savings associated with excess expediting may be approximated by the average savings per period ( $C_r - C_e$ ), multiplied by the average number of periods in an expedited cycle,  $Q_e/E[D]$ , which is given by:

$$\frac{K_r Q_e}{Q_r} + \frac{h_a Q_r Q_e}{2E[D]} - Q_e(c_e - c_r) - \frac{h_a Q_e^2}{2E[D]}.$$

This function is easily seen to be concave in  $Q_e$  and is maximized by

$$Q_e^H \equiv \frac{Q_r}{2} + \frac{E[D]}{h_a} \left( \frac{K_r}{Q_r} + c_r - c_e \right). \quad (18)$$

So, we define

$$\tilde{S}^H \equiv \begin{cases} S^H & \text{if } Q_e^H > S^H \\ Q_e^H & \text{if } s^H < Q_e^H \leq S^H \\ 0 & \text{if } Q_e^H \leq s^H \end{cases} \quad (19)$$

since  $\tilde{S}^* \leq S^*$ ,  $0 < Q_e^H \leq s^H$  implies the use of both excess expedited production and regular production the next period which is suboptimal, and  $Q_e^H \leq 0$  implies that optimal policy is simply to order up to 0 with expedited production.

Equations (15) - (19) define explicit heuristics  $s^H$ ,  $S^H$ , and  $\tilde{S}^H$  for  $s^*$ ,  $S^*$ , and  $\tilde{S}^*$ , respectively; in the next section we discuss their accuracy.

### 4.3 Heuristic Analysis

Here, we compare our heuristic policy developed above to the optimal policy by running computer code with the same parameters and distributions as discussed in Section 4.1. We also discuss a simple way to calculate whether excess expediting may be optimal.

For the normal distribution, we compared our heuristic to the optimal solution and it performed reasonably well. We calculated  $s^H$ ,  $S^H$  and  $\tilde{S}^H$  and the inventory/expediting costs for the same data. On average, the heuristic cost was only 0.58% higher than optimal (with median of 0.14% and a maximum of 4.73%). The heuristic appears to be unbiased in that there is no clear correlation between any of the cost parameters and its performance. [Note that for the Uniform distribution, the average heuristic cost was 0.96% higher than optimal (with a median of 0.33% and a maximum of 11.5%).]

Our estimate for the up-to point  $S^*$  appear to be good, although the estimates for the reorder point  $s^*$  and the expedited up-to point  $\tilde{S}^*$  are a little too high. For each heuristic, we determined percentage bias, the mean absolute percentage error (MAPE), and the maximum absolute percentage error (Max) when compared to the optimal solution, with the results listed in Table 4. The third column of data is for all outcomes, including 172/270 (64%) where  $\tilde{S}^* = 0$ . The last column includes only cases where  $\tilde{S}^*$  is strictly positive. During the 64% of cases when excess expediting is not optimal ( $\tilde{S}^* = 0$ ), our heuristic is correct ( $\tilde{S}^H = 0$ ) 80% of the time, but returns an incorrect positive value ( $\tilde{S}^H > 0$ ) 20% of the time. During the 36% of cases when excess expediting is optimal, our heuristic returns a positive value ( $\tilde{S}^H > 0$ ) 100% of the time. Although our values of

	$s^H$	$S^H$	$\tilde{S}^H(\tilde{S}^* \geq 0)$	$\tilde{S}^H(\tilde{S}^* > 0)$
Bias	+5.6%	+0.1%	+15.2%	+3.9%
MAPE	5.6%	1.8%	n/a	14.6%
Max	22.7%	14.6%	n/a	80%

Table 4: Heuristic Accuracy

$\tilde{S}^H$  are inexact, they typically have a limited effect on inventory/expediting costs due to the small cost savings and infrequency of excess expedited production.

Lastly, when is excess expediting worth considering? Excess expediting is optimal when  $\tilde{S}^* > 0$ , or approximately when  $Q_e^H > 0$ . Referring to Equation (18),  $Q_e^H > 0$  implies that

$$c_e - c_r < \frac{h_a Q_r^H}{2\mu} + \frac{K_r}{Q_r^H} = \sqrt{\frac{2h_a K_r}{\mu}} = h_a \frac{Q_r^H}{\mu}.$$

In other words, excess expediting may have value when the additional per-unit cost of expediting is less than what it costs to hold a unit in inventory for an average regular production cycle. For our normal distribution data, this simple estimate is accurate 87% of the time. It correctly identifies all 36% of the situations where excess expediting is optimal, it identifies 51% of the situations where excess expediting is unnecessary, but it returns false-positives ( $\tilde{S}^H > 0, \tilde{S}^* = 0$ ) 13% of the time. [For the uniform data, the estimate is accurate 79% of the time.]

It should be noted that our focus here has been on explicit equations for our heuristic. Had they not performed well, we would have revisited this decision and could likely achieve even greater accuracy by iterative techniques. For example, one could iterate between  $Q_r^H$  and  $s^H$ , as is suggested in Hadley and Whitin (1963), rather than just using the simple EOQ formula for  $Q_r^H$ . Or one could perhaps modify Zheng and Federgruen's (1991)  $(s, S)$  algorithm to account for expediting. We leave such iterative techniques as the subject of future research.

## 5 Conclusion and Extensions

In this paper we have modeled an inventory control problem where stochastic demand must always be met and shortages may be filled by expediting. Our goal was to determine optimal policies for expediting and regular production and to gain insight about this problem. We first considered

a model with one mode of expediting. We characterized the structure of the optimal expediting policy and showed that the optimal regular production policy is  $(s, S)$ . We then considered a model with two forms of expediting. Again, we explored the structure of the optimal expediting policy and showed that the optimal regular production policy is  $(s, S)$ . We presented heuristics for explicit parameter calculation and tested them using a numerical study; they were shown to perform quite well. Finally, we used numerical analysis to gain insights into the different optimal expediting policies and the frequency of excess expedited production, which was found to be moderate, and associated savings, which were found to be small.

A number of other extensions to this work are natural to consider including capacity constraints, expediting lead-time, non *i.i.d.* demand, demand forecasting, and multi-echelon supply-chains. The latter has been explored to some extent for the case of no fixed production costs in Huggins and Olsen (2003, 2007).

## Acknowledgments

We thank Brad Johnson and Michael Knox, our contacts at “a large automobile parts supplier in Michigan,” for sharing their insights about the real inventory challenges faced by their employer. We also thank Sven Axsater for introducing us to Kaj Rosling’s work prior to its publication as well as the anonymous referees and associate editors for their extremely helpful suggestions for improving this paper. We particularly note that the simple form for the proof of Theorem 2 (replacing that of Huggins, 2002) is due to a referee and much of the form of the heuristic was suggested by the AE. Finally, this research was funded in part by NSF grant DMI-9713727.

## Appendix

### Proof of Lemma 1

*Proof of the relationships between  $f^*(\cdot)$  and  $\tilde{f}^*(\cdot)$  in equations (6) and (7):*

For  $\pi \in \Pi$ , define

$$f_\pi(x_0) \equiv E_{\mathcal{D}_0} \left[ \sum_{t=0}^{\infty} \alpha^t g(x_t, y_t, \tilde{x}_t, \tilde{y}_t) \right]$$

and

$$\tilde{f}_\pi(\tilde{x}_0) \equiv E_{\mathcal{D}_1} \left[ \sum_{t=0}^{\infty} \alpha^t \tilde{g}(\tilde{x}_t, \tilde{y}_t, y_{t+1}) \right],$$

where the sums are well defined because  $g(x_t, y_t, \tilde{x}_t, \tilde{y}_t) \geq 0$  and  $\tilde{g}(\tilde{x}_t, \tilde{y}_t, y_{t+1}) \geq 0$ , respectively.

We first show that

$$f_\pi(x_0) = K_r \delta(y_0 - x_0) + E_{D_0} \left[ \tilde{f}_\pi(y_0 - D_0) \right], \quad (20)$$

where  $y_0 \geq x_0^+$ , and

$$\tilde{f}_\pi(\tilde{x}_0) = e_a(\tilde{y}_0 - \tilde{x}_0) + h_a(\tilde{y}_0) + \alpha f_\pi(\tilde{y}_0), \quad (21)$$

for  $\tilde{y}_0 \geq \tilde{x}_0^+$ .

To wit,

$$\begin{aligned} f_\pi(x_0) &= E_{\mathcal{D}_0} \left[ \sum_{t=0}^{\infty} \alpha^t g(x_t, y_t, \tilde{x}_t, \tilde{y}_t) \right] \\ &= K_r \delta(y_0 - x_0) + E_{\mathcal{D}_0} \left[ \sum_{t=0}^{\infty} \alpha^t \tilde{g}(\tilde{x}_t, \tilde{y}_t, y_{t+1}) \right] \\ &= K_r \delta(y_0 - x_0) + E_{D_0} \left[ E_{\mathcal{D}_1} \left[ \sum_{t=0}^{\infty} \alpha^t \tilde{g}(\tilde{x}_t, \tilde{y}_t, y_{t+1}) \mid D_0 \right] \right] \\ &= K_r \delta(y_0 - x_0) + E_{D_0} \left[ E_{\mathcal{D}_1} \left[ \sum_{t=0}^{\infty} \alpha^t \tilde{g}(\tilde{x}_t, \tilde{y}_t, y_{t+1}) \mid D_0 \right] \right] \\ &= K_r \delta(y_0 - x_0) + E_{D_0} \left[ \tilde{f}_\pi(\tilde{x}_0) \right] \\ &= K_r \delta(y_0 - x_0) + E_{D_0} \left[ \tilde{f}_\pi(y_0 - D_0) \right], \end{aligned}$$

where  $y_0 \geq x_0^+$ . The first equality is true by definition, the fourth equality is true by the monotone convergence theorem since  $\tilde{g}(\cdot) \geq 0$ , and the fifth equality is true by the definition of  $\tilde{f}_\pi$  and because the system is Markovian. The proof of (21) follows immediately because no expectations are needed. Minimizing these equations over  $\pi$  yields the relationships in (6) and (7). Below we show that such a minimum is finite (and hence use of the min function is justified).

*Proof that  $f^*(\cdot)$  and  $\tilde{f}^*(\cdot)$  are finite valued:*

Observe that  $f_\pi(x)$  and  $\tilde{f}_\pi(\tilde{x})$  are non-negative for all policies  $\pi \in \Pi$  and for all  $x, \tilde{x} \in \mathcal{Z}$ , because they are the sum of non-negative functions  $g$  and  $\tilde{g}$ , respectively. Thus, to show that  $f^*(x)$  and  $\tilde{f}^*(\tilde{x})$  are finite over their domains, it suffices to show that there exists a policy  $\gamma$  such that  $\tilde{f}_\gamma(\tilde{x}) < \infty$  and  $f_\gamma(x) < \infty$  for all  $x, \tilde{x} \in \mathcal{Z}$ . Let this  $\gamma$  be such that whenever the inventory level is negative

produce up to 0; otherwise, do nothing. Note that this policy applies to both expedited and regular production and that this policy is stationary.

Consider the cost per stage for expediting under this policy:

$$\tilde{g}_\gamma(\tilde{x}, \tilde{y}, y) = h_a(\tilde{x}^+) + e_a(\tilde{x}^-),$$

since  $\tilde{y} = \tilde{x}^+$ . Thus, since  $\gamma$  is stationary and  $\tilde{g}_\gamma(\tilde{x}, \tilde{y}, y) \geq 0$ , by Corollary 1.1.1 of Bertsekas (1995, p. 139) we have that

$$\tilde{f}_\gamma(\tilde{x}) = h_a(\tilde{x}^+) + e_a(\tilde{x}^-) + \alpha E[\tilde{f}_\gamma(\tilde{x}^+ - D)]$$

Now,

$$\begin{aligned} E[\tilde{f}_\gamma(0 - D)] &= E[e_a(D)] + \alpha E[\tilde{f}_\gamma(0 - D)] \\ \Rightarrow E[\tilde{f}_\gamma(0 - D)] &= \frac{E[e_a(D)]}{1 - \alpha} < \infty, \end{aligned}$$

by Assumption (A3). Thus,

$$\tilde{f}_\gamma(\tilde{x}) = \begin{cases} h\tilde{x} + \alpha E[\tilde{f}_\gamma(\tilde{x} - D)] & \text{if } \tilde{x} \geq 0 \\ e_a(\tilde{x}^-) + \alpha \frac{E[e_a(D)]}{1 - \alpha} & \text{if } \tilde{x} < 0 \end{cases}$$

and  $\tilde{f}_\gamma(\tilde{x}) < \infty$  for  $-\infty < \tilde{x} < 0$ . Now, when  $\tilde{x} \geq 0$ , the inventory level will remain non-negative for some time  $T$  and will then eventually become negative at time  $T + 1$ . Note,  $T < \infty$  almost surely (a.s.) since  $E[D] > 0$  by Assumption (A3). While the inventory level is non-negative, there will be a holding cost of at most  $h_a(\tilde{x})$  for  $T$  discounted periods. When the inventory goes negative, to some value  $N$  say, there will be an  $\alpha^{T+1}$  discounted cost of  $E[e_a(-N)] + \alpha E[e_a(D)]/(1 - \alpha)$ , where  $-\infty < N < 0$  a.s.. Now  $E[e_a(-N)] \leq E[e_a(D)]$  since  $e_a(z)$  is non-decreasing in  $z$  and, by definition,  $-N$  is stochastically smaller than  $D$ . So, for  $\tilde{x} \geq 0$ ,

$$\begin{aligned} \tilde{f}_\gamma(\tilde{x}) &\leq E \left[ \sum_{t=0}^T \alpha^t h_a(\tilde{x}) + \alpha^{T+1} \left( \frac{E[e_a(D)]}{1 - \alpha} \right) \right] \\ &\leq \frac{h_a(\tilde{x})}{1 - \alpha} + \alpha \frac{E[e_a(D)]}{1 - \alpha} < \infty. \end{aligned}$$

Thus  $\tilde{f}_\gamma(\tilde{x}) < \infty$  for  $0 \leq \tilde{x} < \infty$ . So,  $\tilde{f}^*(\tilde{x}) \leq \tilde{f}_\gamma(\tilde{x}) < \infty$  for all  $\tilde{x} \in \mathcal{Z}$ . Now,  $f^*(x)$  is also finite since

$$f_\gamma(x) = K_r \delta(x^-) + E[\tilde{f}_\gamma(x - D)] < \infty.$$

Thus,  $f^*(x) \leq f_\gamma(x) < \infty$  for all  $x \in \mathcal{Z}$ . Therefore, the optimal cost functions are both finite and satisfy the relationships in (6) and (7).

*Proof of optimality of policy that solves equations (6) and (7):*

The fact that any solution to (5) or any joint solution to (6) and (7) must be the optimal solution to (4) and to (2) and to (4), respectively, follows directly from Proposition 1.1 of Bertsekas (1995, p. 137). Further, that the corresponding policy that solves equation (5) or jointly solves (6) and (7) is the optimal stationary policy for the system follows directly from Proposition 1.3 of Bertsekas (1995, p. 143).

*Proof of (8):* Let  $\tilde{x} < 0$ . From (5),

$$\begin{aligned}\tilde{f}^*(\tilde{x}) &= \min_{\tilde{y} \geq \tilde{x}^+, y \geq \tilde{y}^+} E_D \left[ \tilde{g}(\tilde{x}, \tilde{y}, y) + \alpha \tilde{f}^*(y - D) \right] \\ &= \min_{\tilde{y} \geq 0, y \geq \tilde{y}^+} E_D \left[ \tilde{g}(\tilde{x}, \tilde{y}, y) + \alpha \tilde{f}^*(y - D) \right] \\ &\geq \min_{\tilde{y} \geq 0, y \geq \tilde{y}^+} \tilde{g}(\tilde{x}, \tilde{y}, y) + \alpha \min_{y \geq 0} E_D \left[ \tilde{f}^*(y - D) \right],\end{aligned}$$

where the inequality follows from the definition of a minimum. Assumptions (A6) and (A7) imply that  $\tilde{g}(\tilde{x}, \tilde{y}, y) \geq e_a(-\tilde{x})$ . Now, since  $\tilde{x} < 0$ , by Assumption (A5),  $e_a(-\tilde{x}) \geq \alpha K_r$ . Thus,

$$\begin{aligned}\tilde{f}^*(\tilde{x}) &\geq \alpha K_r + \alpha \min_{y \geq 0} E_D \left[ \tilde{f}^*(y - D) \right] \\ &= \alpha f_-^*.\end{aligned}$$

*Proof of (9):*

$$\begin{aligned}f_-^* &= K_r + \min_{y \geq 0} E_D \left[ \tilde{f}^*(y - D) \right] \\ &\leq K_r + \min_{y \geq 0} \{K_r \delta(y) + E_D \left[ \tilde{f}^*(y - D) \right]\} \\ &= K_r + f^*(0),\end{aligned}$$

where the inequality follows from the non-negativity of  $K_r \delta(y)$ . This completes the proof.  $\square$

### **Proof of Corollary 1**

*Proof of 1 and 2:*

Assume, for some  $\tilde{x} < 0$ ,  $\tilde{y}^*(\tilde{x}) > 0$ . We wish to show  $s^* < \tilde{y}^*(\tilde{x})$ . From equations (5) and (3)

$$\tilde{f}^*(\tilde{x}) = \min_{\tilde{y} \geq 0, y \geq \tilde{y}^+} \left\{ e_a(\tilde{y} - \tilde{x}) + h_a(\tilde{y}) + \alpha K_r \delta(y - \tilde{y}) + \alpha E_D \left[ \tilde{f}^*(y - D) \right] \right\}. \quad (22)$$

Now if  $\tilde{y} \neq y$ , then, since  $e_a(\tilde{y} - \tilde{x}) + h_a(\tilde{y})$  is non-decreasing in  $\tilde{y}$ , it is optimal to set  $\tilde{y}$  equal to zero. Thus, for any given  $y$ , there are only two possible optimal values for  $\tilde{y}$ , namely 0 or  $y$ . In other words, either expediting is used to return the system to zero or, if expediting results in

positive inventory, then regular time production is not used. This directly implies that  $s^* < \tilde{y}^*(\tilde{x})$  if  $\tilde{y}^*(\tilde{x}) > 0$ . Also, directly from (22), if for any  $\tilde{x} < 0$ ,  $e_a(y - \tilde{x}) + h_a(y) \geq e_a(-\tilde{x}) + \alpha K_r$  for all  $y > 0$ , then,  $\tilde{y}^*(\tilde{x}) = 0$ .

*Proof of 3:*

Assume  $S^* > 0$ . For  $0 \leq \tilde{x} \leq S^*$ ,  $\tilde{y}^*(\tilde{x}) = \tilde{x} \leq S^*$  (by Theorem 1). Now suppose  $\tilde{x} < 0$ . If expedited production results in positive leftover inventory, then, from part (1)  $y = \tilde{y}$  in (22), and

$$\tilde{f}^*(\tilde{x}) = \min_{y \geq 0} \left\{ e_a(y - \tilde{x}) + h_a(y) + \alpha E_D \left[ \tilde{f}^*(y - D) \right] \right\}$$

and, therefore, by definition,

$$\tilde{y}^*(\tilde{x}) = \arg \min_{y \geq 0} \left\{ e_a(y - \tilde{x}) + h_a(y) + \alpha E_D \left[ \tilde{f}^*(y - D) \right] \right\}.$$

If expedited production produces up to zero, then, from (22),

$$\tilde{f}^*(\tilde{x}) = e_a(-\tilde{x}) + \min_{y \geq 0} \left\{ \alpha K_r \delta(y) + \alpha E_D \left[ \tilde{f}^*(y - D) \right] \right\}$$

so that, by definition,

$$S^* = \arg \min_{y \geq 0} \left\{ \alpha K_r \delta(y) + \alpha E_D \left[ \tilde{f}^*(y - D) \right] \right\}.$$

By assumption,  $S^* > 0$ , so that

$$S^* = \arg \min_{y > 0} \left\{ \alpha E_D \left[ \tilde{f}^*(y - D) \right] \right\}.$$

But  $e_a(y - \tilde{x}) + h_a(y)$  is non-decreasing in  $y$ , so  $\tilde{y}^*(\tilde{x}) \leq S^*$ . □

## References

- [1] Aneja, Y., and H. Noori. 1987. The Optimality of (s,S) Policies for a Stochastic Inventory Problem with Proportional and Lump-sum Penalty Cost. *Management Science*, **33**, 750-755.
- [2] Archibald, B. 1981. Continuous Review (s,S) Policies with Lost Sales. *Management Science*, **27**, 1171-1178.
- [3] Arslan, H., H. Ayhan and T. L. Olsen. 2001. Analytic Models for When and How to Expedite in Make-to-Order Systems. *IIE Transactions*, **33**, 1019-1029.

- [4] Bertsekas, D. 1995. *Dynamic Programming and Optimal Control (Volume 2)*. Athena Scientific, Belmont, Massachusetts.
- [5] Barlow, R. E., and F. Proschan. 1965. *Mathematical Theory of Reliability*. J. Wiley and Sons, New York, NY.
- [6] Bradley, J. R. 1997. *Managing Assets and Subcontracting Policies*, Ph.D. Dissertation, Stanford University, CA.
- [7] Bradley, J. R. 2004. A Brownian Approximation of a Production-Inventory System with a Manufacturer That Subcontracts. *Operations Research*, **52**, 765 - 784.
- [8] Bradley, J. R. 2005. Optimal Stationary Control of a Dual Service Rate M/M/1 Production-Inventory Model. *European Journal of Operational Research*, **161**, 812-837.
- [9] Çetinkaya, S. and M. Parlar. 1998. Optimal Myopic Policy for a Stochastic Inventory Problem with Fixed and Proportional Backorder Costs. *European Journal of Operational Research*, **110**, 20-41.
- [10] Chiang, C. and G. Gutierrez. Optimal Control Policies for a Periodic Review Inventory System with Emergency Orders. *Naval Research Logistics*, **45**, 187-204.
- [11] Daniel, K. 1963. A Delivery-Lag Inventory Model with Emergency. In *Multistage Inventory Models and Techniques*. H. Scarf, D. Gilford, and M. Shelly (eds.). Stanford University Press, Stanford, CA.
- [12] Duenyas, I., W. J. Hopp, and Y. Bassok. 1997. Production Quotas as Bounds on Interplant JIT Contracts. *Management Science*, **43**, 1372-1386.
- [13] Duenyas, I., W. J. Hopp, and M. L. Spearman. 1993. Characterizing the Output Process of a CONWIP Line with Deterministic Processing and Random Outages. *Management Science*, **39**, 975-988.
- [14] Fukuda, Y. 1964. Optimal Policies for the Inventory Problem with Negotiable Leadtime. *Management Science*, **10**, 690-708.

- [15] Groenevelt, H., and N. Rudi. 2002. A Base Stock Inventory Model with Possibility of Rushing Part of Order. *Working Paper*, University of Rochester, Rochester NY.
- [16] Hadley, G., and T. M. Whitin. 1963. *Analysis of Inventory Systems*. Prentice-Hall, Inc., Englewood Cliffs, NJ.
- [17] Hopp, W. J., M. L. Spearman, and I. Duenyas. 1993. Economic Production Quotas for Pull Manufacturing Systems. *IIE Transactions*, **25**, 71-79.
- [18] Huggins, E. 2002. *Supply Chain Management with Overtime and Premium Freight*. Ph.D. Thesis, University of Michigan, Ann Arbor, MI.
- [19] Huggins, E. L., and T. L. Olsen. 2003. Supply Chain Management with Guaranteed Delivery. *Management Science*, **49**, 1154-1167.
- [20] Huggins, E. L., and T. L. Olsen. 2007. Supply Chain Coordination with Guaranteed Delivery. *Working Paper*, Fort Lewis College, Durango CO.
- [21] Ibragimov, I. 1956. On the Composition of Unimodal Distributions. *Theory of Probability and Its Advances*. **1** 255-260.
- [22] Ishigaki, T., and K. Sawaki. 1991. On the  $(s, S)$  Policy with Fixed Inventory Holding and Shortage Costs. *Journal of the Operations Research Society of Japan*, **34**, 47-57.
- [23] Johansen, S., and A. Thorstenson. 1998. An Inventory Model with Poisson Demands and Emergency Orders. *International Journal of Production Economics*, **56**, 275-289.
- [24] Karlin, S. 1958. The Application of Renewal Theory to the Study of Inventory Policies. In: *Studies in the Mathematical Theory of Inventory and Production*. Stanford University Press, Stanford CA.
- [25] Lee, H., and S. Nahmias. 1993. Single-Product, Single-Location Models. In *Handbooks in Operations Research and Management Science*. Nemhauser and Kan, eds., 3-50.
- [26] Lovejoy, W., and K. Sethuraman. 2000. Congestion and Complexity Costs in a Plant with Fixed Resources that Strives to Make a Schedule. *Manufacturing & Service Operations Management*, **2**, 221-239.

- [27] Mohebbi, E., and M. Posner. 1999. A Lost-Sales Continuous Review Inventory System with Emergency Ordering. *International Journal of Production Economics*, **58**, 93-112.
- [28] Moinzadeh, K. and C. Schmidt. 1991. An  $(S-1, S)$  Inventory System with Emergency Orders. *Management Science*, **39**, 308-321.
- [29] Porteus, E. L. 1971. On the Optimality of Generalized  $(s, S)$  Policies. *Management Science*, **17**, 411-426.
- [30] Porteus, E. L. 1985a. Numerical Comparisons of Inventory Policies for Periodic Review Systems. *Operations Research*, **33**, 134-152.
- [31] Porteus, E. L. 1985b. Undiscounted Approximations of Discounted Regenerative Models. *Operations Research Letters*, **3**, 293-300.
- [32] Porteus, E. L. 1990. Stochastic Inventory Theory. In *Handbooks in Operations Research and Management Science*. D.P. Heyman and M.J. Sobel, eds., 605-652.
- [33] Roberts, D. M. 1962. Approximations to Optimal Policies in a Dynamic Inventory Model. In: *Studies in Applied Probability and Management Science*. Stanford University Press, Stanford, CA.
- [34] Rosling, K. 2002. Inventory Cost Rate Functions with Nonlinear Shortage Costs. *Operations Research*, **50**, 1007-1017.
- [35] Scarf, H. 1960. The Optimality of  $(s, S)$  Policies in the Dynamic Inventory Problem. In *Mathematical Methods in the Social Sciences*. 196-202. K. Arrow, S. Karlin and P. Suppes (eds.). Stanford University Press, Stanford, CA.
- [36] Schneider, H. 1978. Methods for Determining the Re-order Point of an  $(s, S)$  Ordering Policy when a Service Level is Specified. *Journal of the Operational Research Society*, **29**, 1181-1193.
- [37] Sethi, S. P., H. Yan, and H. Zhang. 2003. Inventory Models with Fixed Costs, Forecast Updates and Two Delivery Modes. *Operations Research*, **51**, 321-328.
- [38] Smith, S. 1977. Optimal Inventories for an  $(S-1, S)$  System with No Backorders. *Management Science*, **23**, 522-533.

- [39] Song, J.-S., and P. Zipkin. Inventories with Multiple Supply Sources and Networks of Queues with Overflow Bypasses. *Working Paper*, Duke University, Durham NC.
- [40] Tagaras, G., and D. Vlachos. 2001. A Periodic Review Inventory System with Emergency Replenishments. *Management Science*, **47**, 415-429.
- [41] Van Mieghem, J. A. 1999. Coordinating Investment, Production and Subcontracting. *Management Science*, **45**, 954-971.
- [42] Veeraraghavan, S., and A. Scheller-Wolf. 2006. Now or Later: Dual Index Policies for Capacitated Dual Sourcing Systems. To appear: *Operations Research*.
- [43] Veinott, A. 1966. On the Optimality of  $(s, S)$  Inventory Policies: New Conditions and a New Proof. *SIAM Journal of Applied Mathematics*, **14**, 1067-1083.
- [44] Whittmore, A., and S. Saunders. 1977. Optimal Inventory Under Stochastic Demand with Two Supply Options. *SIAM Journal of Applied Mathematics*, **32**, 293-305.
- [45] Zheng, Y-S. 1991. A Simple Proof for Optimality of  $(s, S)$  Policies in Infinite-Horizon Inventory Systems. *Journal of Applied Probability*, **28**, 802-810.
- [46] Zheng, Y-S. 1994. Optimal Control Policy for Stochastic Inventory Systems with Markovian Discount Opportunities. *Operations Research*, **42**, 721-738.
- [47] Zheng, Y-S., and A. Federgruen. 1991. Finding Optimal  $(s, S)$  Policies is About as Simple as Evaluating a Single Policy. *Operations Research*, **39**, 654-665.