Inventory models with expedited ordering : single index policies

Inventory models with expedited ordering : single index policies

Alan Scheller-Wolf, Geert-Jan van Houtum

WP 80

BETA publicatie	WP 80 (working paper)
ISBN	90-386-1837-9
ISSN	1386-9213
NUR	804
Eindhoven	June 2002
Keywords	
BETA-Research Programme	Unit Management
Te publiceren in:	.

Inventory Models with Expedited Ordering: Single Index Policies

Alan Scheller-Wolf
Graduate School of Industrial Administration, Carnegie Mellon University
Pittsburgh, PA 15213-3890

Geert-Jan van Houtum*
Faculty of Technology Management, Eindhoven University of Technology
Eindhoven, The Netherlands 5600MB

May 22, 2002

Abstract

We examine a single stage periodic-review inventory system with backorders, in which replenishment can be obtained either through a *regular* channel or, for a premium, via an *expedited* channel with a smaller lead time. The objective is to minimize the undiscounted infinite horizon average cost subject to either a minimum service level constraint or a linear penalty cost for unfilled orders. As optimal policies for these problems are complex, we propose an order-up-to policy with regular and expedited base stock levels as a heuristic, mirroring common industry behavior. For our *single index* policy, we derive simple expressions for the optimal regular level given the difference between the two levels. We present a procedure utilizing mixtures of Erlang distributions, fit to the first two moments of demand, to calculate this quantity, along with each parameter pair's cost. This enables us to efficiently find the optimal single index policy for any given grid search tolerance. We include a computational section

*This research was initiated in spring 2001 when the second author visited Carnegie Mellon. He thanks the Netherlands Organization for Scientific Research (NWO) for their financial support for this stay.

investigating the behavior of our policy as problem parameters change, and comparing our solutions with optimal single sourcing costs.

1 Introduction

Many firms are trying to construct supply chains that reduce costs while maintaining customer service, often by incorporating alternatives with respect to sourcing. This creates a need for management strategies for such chains; whereas optimal inventory policies are known for quite general single source models (see Tayur, Ganeshan and Magazine 1999), results are much more limited when there are sourcing options. Nevertheless, this is a problem which confronts industry daily. Intel faces this problem when deciding how to route chips through their supply chain, as does Caterpillar shipping construction worktools (Rao, Scheller-Wolf and Tayur 2000), and Hewlett Packard manufacturing servers (Beyer and Ward 2000). In all of these cases managers need a simple yet effective way of deciding how much to source, when, and from whom. To give the reader a fuller understanding of the issues involved in dual sourcing decisions, we briefly discuss the situation faced by Océ, a leading manufacturer of document printing systems with whom the authors have worked.

Océ produces different types of machines (office, production, and wide format printing systems) at two locations in Europe (in The Netherlands and Germany). Océ sells on the European, American and Asian market. Spare parts are kept in stock in (virtually) one central warehouse in Europe, and, in addition in a warehouse near Chicago for the American market and in a warehouse in Singapore for the Asian market. The stocks in the latter two warehouses are replenished weekly, both by ocean transport and air transport.

Replenishments for items with a high value density and/or a low demand rate travel by plane. For items with a relatively low value density and sufficiently high demand rates, in principle only ocean transport is used. For these items, the pipeline holding cost for the longer travel time by sea (three weeks vs. one week for air freight) and the costs for larger safety stocks are more than compensated for by the lower transportation costs. Occasionally air transport is used for this latter class of items, for example when peaks in the demand processes occur. Océ thus uses different transportation modes to both control their costs and meet their customer service targets.

Possibly motivated by such industrial applications, variants of the dual supplier problem have seen renewed interest in the literature lately. Zhang (1996) found complex optimal policies for systems with three delivery modes having consecutive lead times. Lawson and Porteus (2000) take a different tack, showing "top-down base stock policies" to be optimal for a special multi-echelon system. As they note, their assumption that units that have already been shipped can be expedited or delayed at will is not appropriate for many systems. Tagaras and Vlachos (2001), Teunter and Vlachos (2001), and Vlachos and Tagaras (2001) focus on the model where the review period is much greater than the expedited lead time.

We consider the infinite horizon, stochastic demand, periodic-review replenishment problem with two supply modes, with either a service level constraint or a linear penalty cost, under full back-ordering. We propose a variant of the stationary *base stock policy* for inventory control. Base stock policies, while sub-optimal for the general dual lead time case (Whittmore and Saunders 1977), are common in practice as they are simple to implement and have a reputation of performing well. Such a policy prescribes that if stocks fall below the *expedited level* an *expedited* order is placed to return them there. Then a *regular* order is placed to bring the total inventory up to the prescribed *regular level.* As the names indicate, the expedited order has a shorter lead time, and presumably a greater cost. The variant we present – the *single index policy* – bases both the regular and expedited ordering decisions on all goods on hand, owed to customers, and on order. This is the optimal policy for the dual-source penalty cost model with leadtimes that differ by one time unit (Fukuda 1964).

Under the single index policies we propose the problem costs, even for general leadtimes, decompose. Exploiting this decomposition, we derive simple expressions for the optimal regular level, z_r, given a fixed difference between the regular and expedited levels, Δ. This reduces the problem to computing the optimal $z_r(\Delta)$ for each Δ, and then finding the expected cost of each pair, $g(\Delta, z_r(\Delta))$. We compute these quickly by utilizing mixtures of Erlang random variables, fit to the mean and variance of the observed demand. As any continuous distribution on $(0, \infty)$ can be approximated arbitrarily closely by a mixture of Erlang distributions via fitting an increasing number of moments, we can thus find *optimal* single index policy parameters swiftly (in a few seconds), for any given grid search tolerance. We do so, illustrating the performance of our policy for the service level model under various problem parameters, and comparing our solutions with optimal single sourcing costs. Our use of the mixed Erlang to model leadtime demand follows in the tradition of Burgin and Wild (1967) who advocate of the use of the Gamma, a close relative of the mixed Erlang. More recently, Silver, Pyke and Peterson (1998) mention the appeal of both the Gamma and the mixed Erlang to model leadtime demand, (Section 7.7.14). For a more detailed technical discussion of mixed Erlang distributions we refer the reader to Van Houtum and Zijm (1997).

This paper makes the following contributions: We define the single index policy in Section 2, establishing properties of the optimal single index policy for the penalty cost and service level models, respectively, in Sections 3 and 4. These permit efficient calculation of optimal single index parameters – we describe this for demand modeled by a mixture of Erlang distributions analytically in Section 5 and computationally in Section 6. We discuss extensions in Section 7 and we conclude in Section 8.

2 Model Definitions and Recursions

We consider a discrete time inventory system with expedited and regular sourcing. We use a single index inventory policy; only one measure of inventory is tracked, the inventory position over the *entire* leadtime horizon. Thus when ordering, target levels are compared with inventory on hand, plus all outstanding regular or expedited orders, minus any items owed to customers. If the inventory position is below the expedited target level an expedited order is placed to bring it to this level. Then a regular order is placed to bring the final inventory position up to the regular target level. We assume excess demand is backordered. Unsatisfied demand may incur a penalty cost (Section 3) or there may be a service level constraint on the system (Section 4). We define:

n: Period index, $n \geq 0$.

d_n, F: The new customer demand in period n, $\{d_n : n \geq 0\}$, form a stationary and iid sequence. This family of random variables is generically referred to as d, having continuous distribution function F with $0 < F(x) < 1$ for all $x \in (0, \infty)$ and $F(0) = 0$. We further assume $\mathrm{E}[d] \stackrel{\text{def}}{=} \mu < \infty$ and that the standard deviation of $d \stackrel{\text{def}}{=} \sigma > 0$.

l_r, l_e, l: l_r and l_e are the nonnegative deterministic lead times for regular and expedited orders, respectively. We define $l \stackrel{\text{def}}{=} l_r - l_e \geq 0$, with the convention that vacuous summations that follow from $l = 0$ return zero (as in Lemma 3.2 and (11) for example).

c_r, c_e, c: c_r and c_e are the nonnegative unit ordering costs for regular and expedited orders, respectively. We define $c \stackrel{\text{def}}{=} c_e - c_r > 0$. (If $c_r \geq c_e$ expediting all orders is optimal.)

h, p: Strictly positive per period unit holding and backorder cost, respectively.

$B \in (0, \mu)$: Maximum permitted average backlog at the end of a period, for the service level problem. This is equivalent to a minimum γ-service level $\gamma_0 = 1 - (B/\mu) \in (0, 1)$.

z_r, z_e, Δ: z_r and z_e are the regular and expedited order-up-to levels, respectively. We let $\Delta \stackrel{\text{def}}{=} z_r - z_e \geq 0$. (A policy with $z_r < z_e$ is equivalent to policy $(\tilde{z}_r, z_e)$ with $\tilde{z}_r = z_e$.)

I_n: Inventory level at the start of period n; the amount on hand or on back-order.

IP_n: Inventory position at the start of period n; inventory level plus all goods on order.

X_n^r, X_n^e: Regular and expedited orders placed in period n, respectively.

R_n: Amount of inventory received in period n; $R_n = X_{n-l_e}^e + X_{n-l_r}^r$.

In period n orders X_n^e, X_n^r are placed, a shipment of R_n is received, demand d_n is revealed, and customers are satisfied. Costs are then assessed and any unsatisfied customers or excess inventory is carried to the next period.

Our demand assumptions, which streamline our analysis, deserve some comment:

- While we believe our analytical results remain true for discrete demands, their inclusion would entail either additional assumptions or more cumbersome proofs. If we assume continuous inventory (and continuous parameters z_r and z_e) our results extend, but this seems not reasonable. Otherwise we need to specify choosing the least integer value greater than a solution, or using a randomized base stock policy (see Van Houtum and Zijm 2000). While the former tactic yields a simpler solution, the latter will in general perform better. Yet rigorously proving our results for the class of randomized policies, while possible we believe, would not add appreciably to the insights of the paper.

- The assumption of unbounded support is not restrictive as we can limit z_r and Δ to the appropriate support as necessary. Mass at zero might modify some results, but would not change the paper significantly, as this will not necessitate randomized policies.

2.1 System Recursions

Given our definitions, the inventory level and position follow the recursions:

(1) $$I_{n+1} = I_n + R_n - d_n,$$

(2) $$IP_n = z_r - X_n^e - X_n^r,$$

(3) $$IP_{n+1} = IP_n + X_n^e + X_n^r - d_n = z_r - d_n.$$

The expedited and regular orders in period n are, respectively:

(4) $$X_n^e = (z_e - IP_n)^+,$$

(5) $$X_n^r = z_r - (IP_n + X_n^e),$$

where $x^+ \stackrel{\text{def}}{=} max(0, x)$. Similarly, $x^- \stackrel{\text{def}}{=} max(0, -x)$.

Without loss of generality, we assume $I_0 = z_r$ (obtained at purchase price c_r) and $X_n^e = X_n^r = 0$ for all $n \leq 0$. Under this assumption, substituting (3) into (4) yields:

(6) $$X_n^e = (z_e - (z_r - d_{n-1}))^+ = (d_{n-1} - (z_r - z_e))^+ = (d_{n-1} - \Delta)^+, \; n > 0.$$

Substituting (3) and (4) into (5) yields:

$$(7) \qquad X_n^r = z_r - IP_n - (z_e - IP_n)^+ = \min(z_r - z_e, z_r - IP_n) = \min(\Delta, d_{n-1}), \; n > 0.$$

Formulae (6) and (7) specify that in any period the portion of demand that exceeds Δ will be reordered by expedited delivery, and the rest – no more than Δ – will be reordered by regular delivery; large single period demands trigger expediting. For $\Delta \in [0, \infty)$ the proportion of demand filled via the expedited channel can be obtained analytically:

$$(8) \qquad \mathrm{E}[X^e]/\mathrm{E}[d] = \mathrm{E}[(d-\Delta)^+]/\mu.$$

In the extreme cases, $\Delta = 0$ or $\Delta = \infty$, the single index policy reduces to one with a single expedited or regular supplier, respectively.

3 Penalty Cost Problem

In this section we consider the problem having a per-period fixed cost for each unit of customer demand unsatisfied. For periods $0 \leq i \leq N$ an arbitrary (Δ, z_r) pair has cost:

$$(9) \qquad \sum_{i=0}^{N} g_i(\Delta, z_r) \stackrel{\text{def}}{=} Y_N + Z_N,$$

where $Y_N \stackrel{\text{def}}{=} c_r \sum_{i=0}^{N} d_i + (c_e - c_r) \sum_{i=0}^{N} X_i^e$ and $Z_N \stackrel{\text{def}}{=} h \sum_{i=1}^{N+1} I_i^+ + p \sum_{i=1}^{N+1} I_i^-$.

Thus the infinite horizon problem is:

$$\min_{\Delta, z_r} \left\{ \lim_{N\to\infty} \frac{1}{N+1} \sum_{i=0}^{N} g_i(\Delta, z_r) \right\} = \min_{\Delta, z_r} \left\{ \lim_{N\to\infty} \frac{Y_N + Z_N}{N+1} \right\}.$$

with the system evolving as in (1), (6) and (7). For simplicity, we define the time average ordering and inventory costs, $\lim_{N\to\infty} \frac{Y_N}{N+1}$ and $\lim_{N\to\infty} \frac{Z_N}{N+1}$, as $\mathrm{E}[Y]$ and $\mathrm{E}[Z]$ respectively. A coupling argument can be used to show that costs converge over the infinite horizon:

Lemma 3.1 *The infinite horizon average cost under policy* (Δ, z_r) *converges to:*

$$(10)\quad \lim_{N\to\infty} \frac{1}{N+1} \sum_{i=0}^{N} g_i(\Delta, z_r) = c_r\mu + (c_e - c_r)E[X^e] + hE[I^+] + pE[I^-] = E[Y] + E[Z]$$

where random variables without subscripts refer to stationary versions.

We next present results that pertain to the optimal parameter choices. The first of these follows trivially from $I_0 = z_r$, (1) (2), (6) and (7) and holds for every $n > l_r$:

Lemma 3.2

$$I_n = z_r - \sum_{i=n-l_r-1}^{n-1} d_i + \sum_{j=n-l_r-1}^{n-l_e-2} X_j^e = z_r - \sum_{i=n-l_e-1}^{n-1} d_i - \sum_{i=n-l_r-1}^{n-l_e-2} \min(d_i, \Delta).$$

This leads to:

Lemma 3.3 *Along any sample path, for all n:*

(i) Y_n *is determined solely by* Δ*, independent of* z_r.

(ii) For any fixed Δ, Z_n *is determined solely by* z_r.

(iii) For a fixed Δ*, the minimum value satisfying* $z_r(\Delta) \stackrel{\text{def}}{=} F^{-1}_{D(\Delta)}\left(\frac{p}{p+h}\right)$*, yields an optimal* z_r*, where* $F_{D(\Delta)}$ *is the cumulative distribution function of the random variable* $D(\Delta)$*:*

$$(11)\qquad D(\Delta) \sim \sum_{i=1}^{l_e+1} d_i + \sum_{i=l_e+2}^{l_r+1} \min(d_i, \Delta).$$

Proof :

(*i*) From the definition of Y_n, along any sample path the first term $c_r \sum_i d_i$, is fixed. This leaves $(c_e - c_r) \sum_i X_i^e$, where $X_n^e = (d_{n-1} - \Delta)^+$, due to (6).

(*ii*) Due to its definition, Z_n is solely a function of $\{I_i : 1 \le i \le n+1\}$. Given a fixed Δ and a fixed sample path of demands, Lemma 3.2 shows that I_n is a function of z_r.

(*iii*) Defining $D(\Delta)$ as in (11), Lemma 3.2 shows that $I_n \sim z_r - D(\Delta)$ for all n. Our problem thus reduces to an infinite horizon News-vendor problem, the solution to which is the critical fractile. (See Nahmias 1997).

■

Part (*i*) of Lemma 3.3 implies that for any fixed Δ, the problem of minimizing the expected cost for a single index policy reduces to minimizing $\mathrm{E}[Z]$. Parts (*ii*) and (*iii*) show that this minimization is achieved by finding the optimal $z_r(\Delta)$ via the critical fractile of the random variable $D(\Delta)$. The average cost of using such a policy is given in (10) which we rewrite using (6), the fact that $I \sim z_r - D(\Delta)$ and $\mathrm{E}[D(\Delta)] = (l_r + 1)\mu - l\mathrm{E}[(d-\Delta)^+]$:

Lemma 3.4 *The expected average cost of using policy* $(\Delta, z_r(\Delta))$ *is*

$$g_p(\Delta, z_r(\Delta)) \stackrel{\text{def}}{=} c_r\mu + (c+hl)E[(d-\Delta)^+] + hz_r(\Delta) - h(l_r+1)\mu + (p+h)E[(D(\Delta) - z_r(\Delta))^+]. \tag{12}$$

It follows that an optimal single index policy will be one that minimizes (12). In order to find such a policy, for each Δ we need expressions for $D(\Delta)$, $\mathrm{E}[(d-\Delta)^+]$, and $\mathrm{E}[(D(\Delta)-z_r(\Delta))^+]$. To obtain these efficiently, we propose (in Section 5) using a mixture of Erlang random variables with the same shape parameter to model the demand.

3.1 Properties of the Optimal Solution to the Penalty Cost Model

Recalling (11) and Lemma 3.3 it follows that:

Proposition 3.1 *For the penalty cost problem:*

(i) If $\Delta_1 > \Delta_2$*, the* $z_r(\Delta_1) \geq z_r(\Delta_2)$.

(ii) Letting $\Delta \to \infty$, $D(\Delta) \to D_r$ from below, where

$$D_r \sim \sum_{j=1}^{l_r+1} d_j. \tag{13}$$

This implies that $z_r(\Delta)$ converges from below to $z_r(\infty) = z_r^{max} \stackrel{\text{def}}{=} F_{D_r}^{-1}\left(\frac{p}{p+h}\right)$.

(iii) Setting $\Delta = 0$, $D(\Delta) \sim D_e$, where

$$D_e \sim \sum_{j=1}^{l_e+1} d_j. \tag{14}$$

This implies that $z_e(\Delta)$ converges from above to $z_r(0) = z_r^{min} = F_{D_e}^{-1}\left(\frac{p}{p+h}\right)$.

(iv) For any $\Delta \in [0, \infty)$, $0 \leq z_r^{min} \leq z_r(\Delta) \leq z_r^{max} < \infty$.

Proposition 3.1 implies that as Δ approaches its limiting values (zero or infinity), the solution of the dual supplier problem approaches that of a system with only one supplier (regular or expedited), as implied previously by (8). The costs from these single source models thus serve as upper bounds for the optimal single index policy cost. In a similar vein, (11) demonstrates that the effect of the expediting option comes through the truncation of the demand at Δ. Thus the value of this option increases as expediting becomes more powerful, i.e. as l becomes proportionally larger with respect to l_r or as demand becomes more variable.

4 Service Level Problem

In this section we consider the problem having a maximum permissible level for average customer backlog, B. By repeating the arguments of Lemma 3.1 and Lemma 3.4, the service-level problem can be formulated as the following non-linear minimization problem:

$$\min \quad g_s(\Delta, z_r) \stackrel{\text{def}}{=} c_r\mu + c\mathrm{E}[(d-\Delta)^+] + h\mathrm{E}[(z_r - D(\Delta))^+]$$

$$\mathrm{E}[(D(\Delta) - z_r)^+] \leq B\ . \tag{15}$$

We sharpen this formulation and investigate its properties below:

Lemma 4.1 *For the service level problem, given a fixed Δ:*

(i) The objective function g_s is constant for $z_r \leq 0$ and strictly increasing for $z_r > 0$.

(ii) The average backlog $E[(D(\Delta) - z_r)^+]$ is strictly decreasing in z_r.

(iii) $E[(D(\Delta) - z_r)^+] \uparrow \infty$ as $z_r \to -\infty$; $E[(D(\Delta) - z_r)^+] \downarrow 0$ as $z_r \to \infty$.

(iv) There is a unique finite positive value, $z_r(\Delta)$, for which (15) is satisfied at equality.

(v) At optimality $z_r = z_r(\Delta)$.

Proof : Parts (i) – (iii) follow from our demand assumptions. Part (iv) follows from parts (ii) and (iii), continuous demand, $B \in (0, \mu)$, and the continuity of $\mathrm{E}[(D(\Delta) - z_r)^+]$ with respect to z_r. Parts (i) and (iv) imply part (v). ■

4.1 Properties of the Solution to the Service Level Problem

Given part (v) of Lemma 4.1, the problem that remains is to minimize

$$g_s(\Delta, z_r(\Delta)) = c_r\mu + c\mathrm{E}[(d-\Delta)^+] + h\mathrm{E}[(z_r(\Delta) - D(\Delta))^+] , \qquad \Delta \geq 0$$

(16) $$\mathrm{E}[(D(\Delta) - z_r(\Delta))^+] = B .$$

Substituting (16) into $g_s(\Delta, z_r(\Delta)$ and rewriting in a manner similar to Lemma 3.4:

(17) $$g_s(\Delta, z_r(\Delta)) = c_r\mu + (c + hl)\mathrm{E}[(d-\Delta)^+] + hz_r(\Delta) - h(l_r+1)\mu + hB , \ \Delta \geq 0.$$

Proposition 4.1 *Given the service level problem defined above:*

(i) $z_r(\Delta)$ *is continuous and nondecreasing as a function of* Δ.

(ii) As $\Delta \to \infty$, $z_r(\infty) \stackrel{\text{def}}{=} z_r^{max} = \{x \in \mathbf{R} \mid E[(D_r - x)^+] = B\}$, *with* D_r *defined in (13).*

(iii) When $\Delta = 0$, $z_r(0) \stackrel{\text{def}}{=} z_r^{min} = \{x \in \mathbf{R} \mid E[(D_e - x)^+] = B\}$, *with* D_e *defined in (14).*

(iv) For any $\Delta \in [0, \infty)$, $0 \le z_r^{min} \le z_r(\Delta) \le z_r^{max} < \infty$.

Proof : Part (i) follows from the continuity of d, the definition of $D(\Delta)$, and (16). Parts (ii) and (iii) follow from the definition of $D(\Delta)$ and (16). (iv) follows from $(i) - (iii)$, $B \in (0, \mu)$, and (16). ■

We now derive a bound on the optimal value of Δ. To avoid the issue of whether derivatives exist almost everywhere we consider finite differences: For an arbitrary function $f(x)$ and positive ϵ, we define $f^\epsilon(x) \stackrel{\text{def}}{=} \frac{f(x+\epsilon)-f(x)}{\epsilon}$.

Lemma 4.2 *For all* $\Delta \ge 0$ *and* $\nu > 0$ *there exists an* $\epsilon > 0$ *such that*

$$|\{-(c+hl)\mathbf{P}(d \ge \Delta + \epsilon) + hz_r^\epsilon(\Delta)\} - g_s^\epsilon(\Delta, z_r(\Delta))| \le \nu. \tag{18}$$

Proof : Starting from (17):

$$\begin{aligned} g_s^\epsilon(\Delta, z_r(\Delta)) &= \left\{(c+hl)(\mathrm{E}[(d-\Delta-\epsilon)^+ - (d-\Delta)^+]) + h(z_r(\Delta+\epsilon) - z_r(\Delta))\right\}/\epsilon, \\ &= \{(c+hl)(-\epsilon\mathbf{P}(d \ge \Delta+\epsilon) - \mathrm{E}[(d-\Delta)I\{d \in (\Delta, \Delta+\epsilon)\}]) + h\epsilon z_r^\epsilon(\Delta)\}/\epsilon. \end{aligned}$$

Therefore

$$\begin{aligned} -(c+hl)\mathbf{P}(d \ge \Delta+\epsilon) + h(z_r^\epsilon(\Delta)) - (c+hl)\mathbf{P}(d \in (\Delta, \Delta+\epsilon)) \le \\ g_s^\epsilon(\Delta, z_r(\Delta) \le \ -(c+hl)\mathbf{P}(d \ge \Delta+\epsilon) + hz_r^\epsilon(\Delta). \end{aligned}$$

As we have a continuous demand distribution, we can choose ϵ small enough such that for any $\Delta \ge 0$ and a given $\nu > 0$, $\mathbf{P}(d \in (\Delta, \Delta+\epsilon)) \le \nu/(c+hl)$, completing the proof. ■

Lemma 4.3 *For any positive* ϵ, $D^\epsilon(\Delta) \in [0, l]$.

Proof : Recalling (11), $D(\Delta)$ is nondecreasing in Δ, but on any sample path no more than $l = (l_r + 1) - (l_e + 2) + 1$ elements of $D(\Delta)$ can increase (at rate 1) as Δ increases. ■

Lemma 4.4 *For any positive* ϵ, $z_r^\epsilon(\Delta) \in [0, l]$.

Proof : As $z_r(\Delta)$ is nondecreasing in Δ (Proposition 4.1), we know that $z_r^\epsilon(\Delta) \geq 0$.

Assume that for some Δ and ϵ, $z_r^\epsilon(\Delta) > l$. Part (v) of Lemma 4.1 establishes that for any given B and Δ, constraint (16) will hold as an equality at an optimal solution. In particular:

$$\text{(19)} \qquad \mathrm{E}[(D(\Delta) - z_r(\Delta))^+] = B = \mathrm{E}[(D(\Delta + \epsilon) - z_r(\Delta + \epsilon))^+].$$

We divide the sample paths for the demand realizations into three disjoint sets:

ω_1: Sample paths where $D(\Delta) > z_r(\Delta)$ and $D(\Delta + \epsilon) > z_r(\Delta + \epsilon)$. These will be included in both expectations in (19). From Lemma 4.3 and our assumption that $z_r^\epsilon(\Delta) > l$, we see that $D(\Delta + \epsilon) - D(\Delta) \leq l\epsilon < z_r(\Delta + \epsilon) - z_r(\Delta)$, or

$$\text{(20)} \qquad D(\Delta + \epsilon) - z_r(\Delta + \epsilon) < D(\Delta) - z_r(\Delta).$$

Relation (20) implies that the contribution of the sample paths in ω_1 to the left-hand expectation in (19) is strictly greater than their contribution on the right-hand side.

ω_2: Sample paths where $D(\Delta) > z_r(\Delta)$ and $D(\Delta + \epsilon) \leq z_r(\Delta + \epsilon)$. These make a strictly positive contribution to the left-hand side of (19) but zero on the right.

ω_3: Sample paths where $D(\Delta) \leq z_r(\Delta)$. Formula (20) implies the value of both expectations in (19) is zero.

Unless $D(\Delta) \le z_r(\Delta)$ almost surely, which violates $B < 1$, taking expectation over $\omega_1 \cup \omega_2 \cup \omega_3$ yields $\mathrm{E}[(D(\Delta) - z_r(\Delta))^+] > \mathrm{E}[(D(\Delta+\epsilon) - z_r(\Delta+\epsilon))^+]$, contradicting (19). ■

This leads to

Lemma 4.5 *If $(\Delta^*, z_r(\Delta^*))$ is an optimal solution, then for all $\nu > 0$ there exists a $\zeta > 0$ such that $\zeta \downarrow 0$ if $\nu \downarrow 0$ and $\mathbf{P}(d \ge \Delta^* + \zeta) \le \frac{hl+\nu}{c+hl}$.*

Proof : For there to be a minimum in at $(\Delta^*, z_r(\Delta^*))$, continuity dictates that for all $\epsilon > 0$ small enough, $g_s(x, z_r(x)) = g_s(x+\epsilon, z_r(x+\epsilon))$ for some $x < \Delta^* < x + \epsilon$ (equivalently $g^\epsilon(x, z_r(x)) = 0$). From (18) this implies that for all $\nu > 0$ we can find an $\epsilon > 0$ and an $x < \Delta^* < x + \epsilon$ such that: $|-(c+hl)\mathbf{P}(d \ge x+\epsilon) + hz_r^\epsilon(x)| \le \nu$.

Letting $x \stackrel{\text{def}}{=} \Delta^* - \eta$, and using $z_r^\epsilon << l$ from Lemma 4.4:

$$(c+hl)\mathbf{P}(d \ge \Delta^* - \eta + \epsilon) \le \nu + hl \quad \Rightarrow \quad \mathbf{P}(d \ge \Delta^* - \eta + \epsilon) \le \frac{hl+\nu}{c+hl}.$$

Defining $\zeta \stackrel{\text{def}}{=} \epsilon - \eta$ completes the proof. ■

Lemma 4.5 states that any $\Delta < \Delta_{min} \stackrel{\text{def}}{=} F^{-1}\left(\frac{c}{c+hl}\right)$ cannot be an optimal solution; an optimal Δ must fall within the range $[\Delta_{min}, \infty)$. Note that $l = 0$ thus implies $\Delta_{min} = \infty$, or only the regular supplier is used, which is correct. The tightness of this bound appears difficult to ascertain in general; we explore this in Section 6. Lemma 4.5 also leads to:

Corollary 4.1 *Given $c > 0$:*

(i) $\Delta^* > 0$; using the expedited supplier alone will <u>*never*</u> *be optimal.*

(ii) $c \to \infty \;\Rightarrow\; \Delta_{min} \to \infty \;\Rightarrow\; \Delta^* \to \infty$; as the relative cost of using the expedited supplier grows the proportion of expedited goods approaches zero.

Because the penalty cost problem is a Lagrangian relaxation of the service level model, we can apply Theorem 1 of Van Houtum and Zijm (2000), to extend the results listed in Lemma 4.5 and Corollary 4.1 to the penalty cost model of Section 3:

Lemma 4.6 *For the backorder cost problem, for p sufficiently large to ensure an optimal average backlog less than μ (the mean demand):*

(i): If Δ^* is an optimal solution, then for all $\nu > 0$ small enough there exists a $\zeta > 0$ such that $\zeta \downarrow 0$ if $\nu \downarrow 0$ and $\mathbf{P}(d \geq \Delta^* + \zeta) \leq \frac{hl+\nu}{c+hl}$.

(ii): $\Delta^* > 0$; using the expedited supplier alone will never be optimal.

(iii) $c \to \infty \;\Rightarrow\; \Delta_{min} \to \infty \;\Rightarrow\; \Delta^* \to \infty$; as the relative cost of using the expedited supplier grows the proportion of expedited goods approaches zero.

5 Erlang Mixtures

We now explore some properties of mixtures of Erlang distributions, which we use to model demand. Throughout we denote by $E_{k,\lambda}$ the distribution function of an Erlang random variable with $k \in \mathbf{N} := \{1, 2, \ldots\}$ phases and scale parameter $\lambda > 0$: $E_{k,\lambda}(x) = 1 - \sum_{j=0}^{k-1} \frac{(\lambda x)^j}{j!} e^{-\lambda x}$, $x \geq 0$, and $E_{k,\lambda}(x) = 0$ for all $x < 0$. The corresponding probability density function is $e_{k,\lambda}(x) = \frac{\lambda^k x^{k-1}}{(k-1)!} e^{-\lambda x}$, $x \geq 0$, and $e_{k,\lambda}(x) = 0$ for all $x < 0$. We denote by $\{q_k\}_{k \in \mathbf{N}}$ a discrete distribution on $\mathbf{N}$, for use as mixture parameters.

We make the following definitions assuming d is a mixture of Erlangs with the same shape parameter λ and mixture parameters $\{q_k\}$: for all k, $d \sim E_{k,\lambda}$ with probability q_k.

$D_m,\ F_m$: D_m is the m-fold convolution of d; $D_m = \sum_{i=1}^m d_i$ and $D_0 = 0$. Similarly, F_m is the m-fold convolution of F ($m \in \mathbf{N}_0 \stackrel{\text{def}}{=} \mathbf{N} \cup \{0\}$); F_m is the distribution function of D_m. Each $F_m,\ m \geq 1$ is a mixture of Erlang distributions with scale parameter λ and mixture parameters denoted by $\{q_k^{(m)}\}_{k\in\mathbf{N}}$.

Thus $q_k^{(1)} = q_k$ for all $k \in \mathbf{N}$, and for each $m \geq 2,\ q_k^{(m)} = \sum_{i=1}^{k-1} q_i q_{k-i}^{(m-1)}$, $k \in \mathbf{N}$, reading zero for the sum when $k = 1$. The distribution function F_0 has all its probability mass at 0: $F_0(x) = 0$ for all $x < 0$ and $F_0(x) = 1$ for all $x \geq 0$.

We add hats ($\hat{d},\ \hat{\mu},\ \hat{F},\ \hat{D}_m,$) to denote these quantities, including the mean, under truncation at Δ, e.g. $\hat{d} \stackrel{\text{def}}{=} \min\{d, \Delta\}$, and tildes ($\tilde{d},\ \tilde{\mu},\ \tilde{F},\ \tilde{D}_m$) to denote residual quantities post truncation; $\tilde{d} \stackrel{\text{def}}{=} (d - \Delta | d \geq \Delta)$. The distribution of $\tilde{d}$ is also a mixture of Erlang distributions with parameters λ and $\{\tilde{q}_k\}_{k\in\mathbf{N}}$. We can derive expressions for these $\{\tilde{q}_k\}_{k\in\mathbf{N}}$ as follows. Define for all $k \in \mathbf{N},\ j = 0, 1, \ldots, k$

$$r_{k,j} := \mathbf{P}\{j \text{ phases left after } \Delta \text{ time units} \mid d \text{ is Erlang}(k, \lambda) \text{ distributed}\}\ ,$$

Then for all $k \in \mathbf{N}$:

$$\begin{aligned}
r_{k,j} &= \frac{(\lambda\Delta)^{k-j}}{(k-j)!} e^{-\lambda\Delta}\ , \qquad j = 1, \ldots, k, \\
r_{k,0} &= \sum_{i=k}^{\infty} \frac{(\lambda\Delta)^i}{i!} e^{-\lambda\Delta} = 1 - \sum_{i=0}^{k-1} \frac{(\lambda\Delta)^i}{i!} e^{-\lambda\Delta} = E_{k,\lambda}(\Delta)\ . \\
\tilde{q}_j &= \mathbf{P}\{j \text{ phases left after } \Delta \text{ time units} \mid \text{at least 1 phase left}\} \\
&= \frac{\mathbf{P}\{j \text{ phases left after } \Delta \text{ time units}\}}{\mathbf{P}\{d > \Delta\}} \\
&= \frac{1}{p} \sum_{k=j}^{\infty} q_k r_{k,j}\ , \qquad j \in \mathbf{N}. \qquad (21)
\end{aligned}$$

Furthermore, $\tilde{\mu} = \sum_{k=1}^{\infty} \tilde{q}_k \frac{k}{\lambda}$, and $\tilde{F}(x) = \sum_{k=1}^{\infty} \tilde{q}_k E_{k,\lambda}(x)$, $x \in \mathbf{R}$. Once $\tilde{\mu}$ has been computed, via conditioning we find $\hat{\mu} = \mu - p\tilde{\mu}$. Returning to our definitions:

$\hat{Y}_{m,n}$,$\hat{G}_{m,n}$: The sum of $m \in \mathrm{N}_0$ truncated demands and $n \in \mathrm{N}_0$ non-truncated demands $\hat{Y}_{m,n} = \hat{D}_m + D_n = \sum_{i=1}^{m} \hat{d}_i + \sum_{i=m+1}^{m+n} d_i$. $\hat{G}_{m,n}$ is the m-fold convolution of $\hat{F}$ convoluted with the n-fold convolution of F; $\hat{G}_{m,n}$ is $\hat{Y}_{m,n}$'s distribution function.

We define $\tilde{Y}_{m,n}$ and $\tilde{G}_{m,n}$ analogously. Once again, each $\tilde{G}_{m,n}$, with $m \geq 1$ or $n \geq 1$, is a mixture of Erlang distributions with scale parameter λ and mixture parameters denoted by $\{\tilde{q}_k^{(m,n)}\}_{k \in \mathrm{N}}$. Then $\tilde{q}_k^{(1,0)} = \tilde{q}_k$ for all $k \in \mathrm{N}$, and for each $m \geq 2$: $\tilde{q}_k^{(m,0)} = \sum_{i=1}^{k-1} \tilde{q}_i \tilde{q}_{k-i}^{(m-1,0)}$, $k \in \mathrm{N}$, reading zero for the sum when $k = 1$. Further, for each $n \geq 1$, $\tilde{q}_k^{(0,n)} = q_k^{(n)}$ for all $k \in \mathrm{N}$, and for each $m \geq 1$ and $n \geq 1$: $\tilde{q}_k^{(m,n)} = \sum_{i=1}^{k-1} \tilde{q}_i^{(m,0)} q_{k-i}^{(n)}$, $k \in \mathrm{N}$. The distribution function $\tilde{G}_{0,0}$ has all its probability mass at 0: $\tilde{G}_{0,0}(x) = 0$ for all $x < 0$ and $\tilde{G}_{0,0}(x) = 1$ for all $x \geq 0$.

5.1 Erlang Approximations

The class of mixtures of Erlang distributions with the same scale parameter is dense in the class of all distributions on $[0, \infty)$, i.e. any distribution on $[0, \infty)$ can be approximated arbitrarily closely by such an Erlang mixture. This has been proved by Schassberger (1973) and is also described in Tijms (1986), p. 358. In general a mixture of an infinite number of Erlangs is needed for an exact approximation, but for practical purposes it is common to approximate a distribution by fitting the first two moments, as we do here. Based on the value for $c_{var} \stackrel{\text{def}}{=} \sigma/\mu$ we distinguish two cases, following Tijms (1986).

When $c_{var} \leq 1$, we fit a mixture of an Erlang$(k_0 - 1, \lambda)$ and an Erlang(k_0, λ) distribution. In that case, we choose k_0 such that $\frac{1}{k_0} < c_{var}^2 \leq \frac{1}{k_0 - 1}$, (notice that $k_0 \geq 2$). Next we choose $q_{k_0-1} = \frac{1}{1+c_{var}^2} \left[k_0 c_{var}^2 - \{k_0(1 + c_{var}^2) - k_0^2 c_{var}^2\}^{0.5} \right]$, $q_{k_0} = 1 - q_{k_0-1}$, and $q_k = 0$ for all other k. Finally, $\lambda = \frac{k_0 - q_{k_0-1}}{\mu}$. This distribution is unimodal.

When $c_{var} > 1$, we fit a mixture of an Erlang$(1, \lambda)$ and an Erlang(k_0, λ) distribution.

In that case, we choose k_0 such that $k_0 \geq 3$ and $\frac{k_0^2+4}{4k_0} \geq c_{var}^2$, taking the smallest k_0 which satisfies this inequality. Next we choose $q_1 = \frac{2k_0 c_{var}^2 + k_0 - 2 - (k_0^2 + 4 - 4kc_{var}^2)^{0.5}}{2(k_0-1)(1+c_{var}^2)}$, $q_{k_0} = 1 - q_1$, and $q_k = 0$ for all other k. Finally, $\lambda = \frac{q_1 + k_0(1-q_1)}{\mu}$. This distribution is not in general unimodal.

Given this fitting procedure, we henceforth assume d is distributed as a mixture of Erlang distributions with shape parameter λ and mixture parameters $\{q_k\}$.

5.2 Analytical Results

Below we present exact formulae for $\hat{G}_{m,n}$, which determine the distribution of $D(\Delta)$, since $D(\Delta) = \hat{Y}_{l,l_e+1}$. After that, we exploit this relation to derive exact expressions for the expectations $\mathrm{E}[(D(\Delta) - x)^+]$, needed for the computation of $z_r(\Delta)$ and $g_p(\Delta, z_r(\Delta))$. Our results are presented as a series of lemmas, with proofs in Appendix A.

Before we state our results, for expository convenience we define:

$$H_{m,n}(x) := \mathbf{P}\{\sum_{i=1}^{m+n} d_i \leq x \mid d_i < \Delta \text{ for } i = 1, \ldots, m\} \ , \qquad x \in \mathbf{R},\ m \in \mathbf{N}_0,\ n \in \mathbf{N}_0.$$

The $H_{m,n}$ are distribution functions with $H_{m,n}(x) = 0$ for all $x < 0$, $m \in \mathbf{N}_0$, $n \in \mathbf{N}_0$. If $0 \leq x < \Delta$, then $H_{m,n}(x) = \mathbf{P}(\sum_{i=1}^{m+n} d_i \leq x) = F_{m+n}(x)$, $m \in \mathbf{N}_0$, $n \in \mathbf{N}_0$. In addition, for each $m \in \mathbf{N}_0$ and $n \in \mathbf{N}_0$ with $m \geq 1$ or $n \geq 1$, $H_{m,n}$ is continuous on $\mathbf{R}$. $H_{0,0}$ is discontinuous at 0: $H_{0,0}(x) = 0$ for all $x < 0$ and $H_{0,0}(x) = 1$ for all $x \geq 0$.

Lemma 5.1 *For all $m \in \mathbf{N}_0$ and $n \in \mathbf{N}_0$:*

$$\hat{G}_{m,n}(x) = \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} H_{m-k,n}(x - k\Delta) \ , \qquad x \geq 0.$$

To further simplify the expression in Lemma 5.1, we find the following:

Lemma 5.2 *For all $m \in \mathbf{N}_0$ and $n \in \mathbf{N}_0$:*

$$H_{m,n}(x) = \frac{1}{(1-p)^m} \sum_{k=0}^{m} (-1)^k \binom{m}{k} p^k \tilde{G}_{k,m+n-k}(x - k\Delta) \ , \quad x \in \mathbf{R}.$$

By substituting the relation in Lemma 5.2 into that of Lemma 5.1 we derive:

Lemma 5.3 *For all $m \in \mathbf{N}_0$ and $n \in \mathbf{N}$:*

$$\hat{G}_{m,n}(x) = \sum_{s=0}^{m} (-1)^s \binom{m}{s} p^s \sum_{i=0}^{s} (-1)^i \binom{s}{i} \tilde{G}_{s-i,m+n-s}(x - s\Delta) \ , \qquad x \in \mathbf{R}.$$

Lemma 5.3 demonstrates that each distribution function $\hat{G}_{m,n}$ can be written as *a linear combination of shifted distribution functions* $\tilde{G}_{m,n}$. Using this result we are able to compute the functions $\hat{G}_{m,n}$, and thus the distribution of $D(\Delta)$.

For our final result we define for each $m, n \in \mathbf{N}_0$, $\hat{K}_{m,n}(x) = \mathrm{E}[(\hat{Y}_{m,n} - x)^+]$ and $\tilde{K}_{m,n}(x) = \mathrm{E}[(\tilde{Y}_{m,n} - x)^+]$, $x \in \mathbf{R}$. Then we can calculate

$$\hat{K}_{0,0}(x) = \tilde{K}_{0,0}(x) = \begin{cases} -x & \text{if } x < 0; \\ 0 & \text{if } x \geq 0. \end{cases}$$

Lemma 5.4 *For each $m, n \in \mathbf{N}_0$ with $m \geq 1$ or $n \geq 1$:*

(i): $\tilde{K}_{m,n}(x) = \sum_{k=1}^{\infty} \tilde{q}_k^{(m,n)} \left(\frac{k}{\lambda}(1 - E_{k+1,\lambda}(x)) - x(1 - E_{k,\lambda}(x)) \right) \ , \qquad x \in \mathbf{R}.$

(ii): $\hat{K}_{m,n}(x) = \sum_{s=0}^{m} (-1)^s \binom{m}{s} p^s \sum_{i=0}^{s} (-1)^i \binom{s}{i} \tilde{K}_{s-i,m+n-s}(x - s\Delta) \ , \qquad x \in \mathbf{R}.$

Lemma 5.4 enables us to carry out exact evaluations of $\mathrm{E}[(D(\Delta) - x)^+] = \hat{K}_{l,l_e+1}(x)$, $x \in \mathbf{R}$. Thus we can find $z_r(\Delta)$ using (16), and then use Lemma 5.4, the value of $z_r(\Delta)$, and the distribution of d to exactly evaluate $g_r(\Delta, z_r(\Delta))$, as given in (17).

We have derived formulae showing that the class of mixtures of Erlang distributions with common scale parameter λ is closed under convolution and conditional truncation. These formulae also hold for other classes with this property, such as *phase-type distributions*.

6 Computational Work

In the interest of brevity, we limit our computational investigation to the service level model. Further, in our results below we report average costs excluding the purchasing costs $c_r\mu$, incurred if all goods are bought via the regular channel. (Marginal costs of expediting are of course included.) Finally, in all of our experiments we model demand per period as a mixed Erlang distribution as described in Section 6. (At the end of this section we present a typical comparison of the distribution of a Mixed Erlang random variable and the random variable it approximates.) W.l.o.g., throughout our experiments we assume that $\mu = 1$. We consider three values for the standard deviation of demand, σ: $\frac{1}{3}$, 1, and 3. When $\sigma = \frac{1}{3}$, the underlying distribution is the Erlang-9 distribution with scale parameter $\lambda = \frac{1}{9}$. When $\sigma = 1$, the underlying distribution is the exponential distribution with scale parameter $\lambda = 1$. In both these cases, the probability density function is unimodal. When $\sigma = 3$, the underlying distribution is the mixture of the exponential distribution and the Erlang-36 distribution with scale parameter $\lambda = 2$ for both, with mixing probabilities $q_1 = 0.9714$ and $q_{36} = 0.0286$. Under this distribution, in each period one always has an exponential demand (with mean $\frac{1}{2}$), and with probability 0.0286 one has an additional Erlang-35 demand (with mean 17.5). The probability density function of this mixture is bimodal.

For the determination of the optimal single index policy, based on the results in Sections 4 and 5, we developed the following procedure. Let δ be a small positive number. For each $\Delta = \Delta_{min} + k\delta$, with $k = 0, 1, \ldots$, we compute $z_r(\Delta)$ and the average costs $g_s(\Delta, z_r(\Delta))$ (excluding $c_r\mu$). We track $\tilde{g}$ which denotes the lowest average cost among the points $\Delta' = \Delta_{min} + m\delta$ with $0 \leq m \leq k$. This procedure can be stopped when $hz_r(\Delta) - h(l_r + 1)\mu + hB > \tilde{g}$ (since

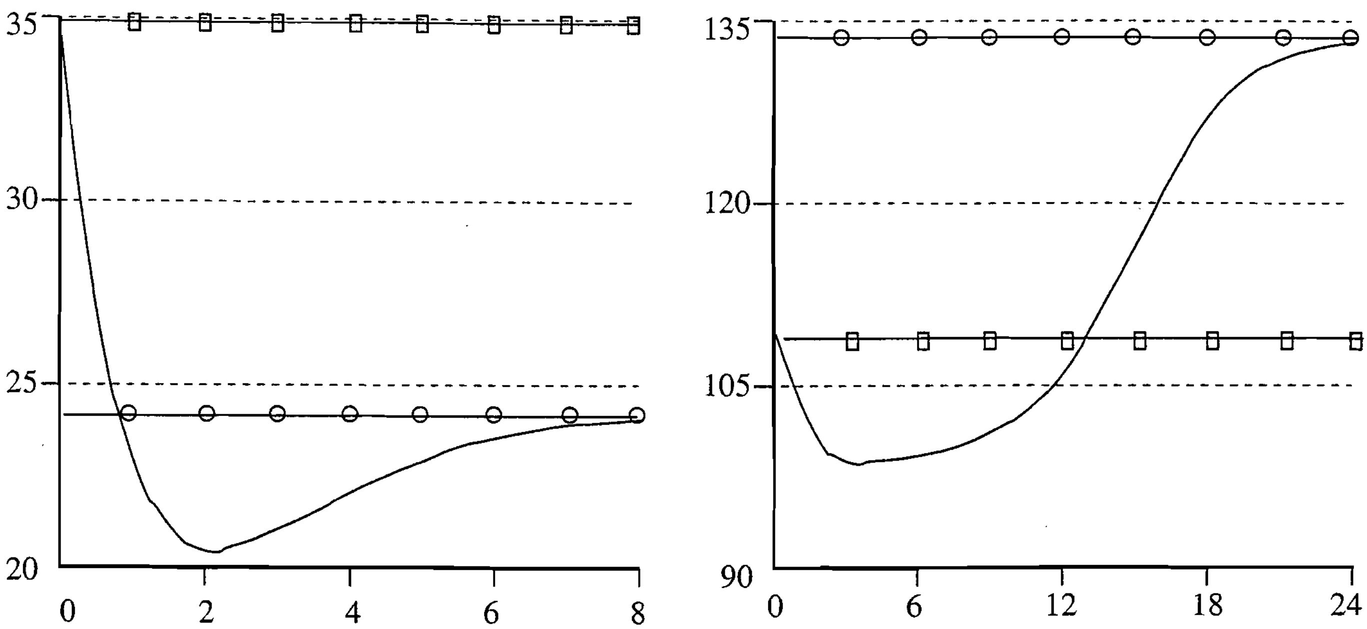

Figure 1: Average costs $g_s(\Delta, z_r(\Delta))$ (excluding $c_r\mu$) as a function of Δ for 2 instances: $\mu = 1$, $c_r = 1000$, $c_e = 1020$, $l_r = 4$, $l_e = 1$, $\gamma_0 = 0.95$, $h = 5$, and $\sigma = 1$ (lefthand side) and $\sigma = 3$ (righthand side). The horizontal lines denote the average costs when using the regular channel only (with circles) and when using the expedited channel only (with squares).

then from (17), $g_s(\tilde{\Delta}, z_r(\tilde{\Delta})) \geq hz_r(\tilde{\Delta}) - h(l_r + 1)\mu + hB > \tilde{g}$ for all $\tilde{\Delta} > \Delta$). Next we search the neighborhood of this best point to find a local optimum among all points $\Delta \geq 0$. We generated the function $g_s(\Delta, z_r(\Delta))$ analytically for many instances, and for all of these the function had a unique local minimum; thus we conclude that our procedure works well, returning the optimal single index parameters. Our optimal single index policy is *globally* optimal for the case where $l = 1$; this is a classical result for the penalty cost model (Fukuda 1964), and is an immediate consequence of Theorem 1 of Van Houtum and Zijm (2000) for the service level model. To the best of our knowledge this service level result has not appeared in the literature. In general, the function $g_s(\Delta, z_r(\Delta))$ typically behaves in one of two ways, as depicted in Figure 1, depending upon which supplier would be least expensive to use as a sole source.

We have determined the optimal single index policy for 81 instances, which are obtained by choosing the the following input parameters:

$$\mu = 1, \quad \sigma = \frac{1}{3}, \ 1, \ 3,$$

$$c_r = 1000, \quad c_e = 1020, \ 1050, \ 1100,$$

$$l_r = 2, \ 4, \ 6, \quad l_e = 1,$$

$$h = 5, \ \gamma_0 = 0.9, \ 0.95, \ 0.99.$$

The choice for these values is motivated by examples faced by Hewlett Packard and Océ. The value $l_e = 1$ represents 1 week, the time needed to transport goods by air, including preparation and receiving times. The values for l_r represent leadtimes when other transport modes are chosen (overland by truck when $l_r = 2$ or by seagoing vessel when $l_r = 4$ or $l_r = 6$). The values for μ and c_r can be set w.l.o.g., here we have chosen 1 and 1000. The values for σ have been chosen such that we have a range of coefficients of variation. The values of c_e are such that the unit price when ordering by the expedited channel is 2 %, 5 %, and 10 % more expensive than when ordering by the regular channel. As is typical, for both HP and Océ ordering by the expedited channel saves interest on the pipeline stock but incurs extra transportation costs. If the net result of this is negative (ordering by the expedited channel is cheaper) everything will be expedited. If the net result is very large everything will be ordered via the regular channel. Only when the expedited costs c_e are somewhat larger than the regular costs c_r, as is common, is it attractive to combine both modes. The inventory holding costs are fixed at $h = 5$. This corresponds to a yearly rate of 25 % for interest and storage costs ($h = 1000 \cdot 0.25/50$). Finally, the target service level γ_0 values have been chosen to mimic those typically used in industry.

$l_r \backslash \sigma$		1/3			1			3		
	$\Gamma \backslash c_e$	1020	1050	1100	1020	1050	1100	1020	1050	1100
2	0.9	∞, 3.4 2.3 1.3, 0%	∞, 3.4 2.3 1.5, 0%	∞, 3.4 2.3 1.6, 0%	3.5, 5.5 14 1.6, 3%	5.0, 5.6 14 2.4, 1%	∞, 5.7 14 3.0, 0%	17.2, 19.3 83 0.9, 5%	18.1, 19.4 84 1.4, 3%	18.9, 19.5 85 2.0, 2%
	0.95	1.9, 3.6 3.3 1.3, 0%	∞, 3.6 3.4 1.5, 0%	∞, 3.6 3.4 1.6, 0%	3.4, 6.3 18 1.6, 3%	4.6, 6.5 18 2.4, 1%	5.9, 6.6 18 3.0, 0%	17.3, 21.9 96 0.9, 4%	20.0, 22.1 97 1.4, 1%	20.7, 22.2 97 2.0, 1%
	0.99	1.7, 4.1 5.6 1.3, 1%	2.0, 4.1 5.7 1.5, 0%	2.4, 4.1 5.7 1.6, 0%	3.2, 8.1 27 1.6, 4%	4.4, 8.3 27 2.4, 1%	5.3, 8.4 28 3.0, 0%	4.7, 26.3 126 0.9, 38%	7.5, 26.9 136 1.4, 30%	13.4, 29.7 148 2.0, 13%
4	0.9	1.6, 5.5 3.3 1.0, 1%	2.4, 5.6 3.4 1.2, 0%	∞, 5.6 3.4 1.4, 0%	2.2, 7.4 16 0.8, 11%	3.6, 8.2 18 1.5, 3%	5.0, 8.5 19 2.0, 1%	8.8, 20.7 88 0.4, 26%	10.5, 21.1 95 0.8, 22%	15.9, 23.6 102 1.1, 7%
	0.95	1.5, 5.7 4.5 1.0, 2%	1.9, 5.9 4.7 1.2, 0%	2.8, 5.9 4.7 1.4, 0%	2.2, 8.2 20 0.8, 12%	3.3, 9.0 22 1.5, 4%	4.5, 9.4 23 2.0, 1%	4.0, 22.0 99 0.4, 40%	8.7, 23.4 109 0.8, 26%	11.6, 24.7 120 1.1, 18%
	0.99	1.3, 6.1 6.9 1.0, 4%	1.7, 6.4 7.4 1.2, 1%	2.0, 6.5 7.6 1.4, 0%	2.1, 10.0 30 0.8, 13%	3.1, 10.8 32 1.5, 5%	4.0, 11.2 33 2.0, 2%	2.8, 27.4 127 0.4, 44%	4.6, 27.9 140 0.8, 38%	8.2, 29.9 156 1.1, 28%
6	0.9	1.4, 7.5 4.1 0.9, 3%	1.9, 7.7 4.4 1.1, 0%	2.9, 7.8 4.4 1.3, 0%	1.7, 9.0 18 0.6, 17%	2.9, 10.3 21 1.1, 5%	4.3, 11.1 22 1.6, 1%	4.8, 21.4 89 0.3, 38%	9.2, 22.8 99 0.6, 25%	10.6, 23.6 110 0.9, 21%
	0.95	1.3, 7.6 5.3 0.9, 4%	1.7, 8.0 5.8 1.1, 1%	2.2, 8.1 5.9 1.3, 0%	1.7, 9.7 22 0.6, 18%	2.7, 11.1 25 1.1, 6%	3.9, 11.9 27 1.6, 2%	2.7, 23.0 100 0.3, 44%	5.8, 24.1 112 0.6, 35%	9.2, 26.0 127 0.9, 25%
	0.99	1.2, 8.0 7.8 0.9, 6%	1.5, 8.5 8.7 1.1, 1%	1.8, 8.7 9.1 1.3, 0%	1.6, 11.5 31 0.6, 20%	2.5, 12.8 35 1.1, 8%	3.5, 13.7 38 1.6, 3%	2.1, 28.4 128 0.3, 46%	3.6, 29.0 141 0.6, 41%	5.9, 30.4 160 0.9, 34%

Table 1: *Values for Δ^*, $z_r(\Delta^*)$, the average costs (excluding $c_r\mu$), Δ_{min}, and the percentage ordered by the expedited channel for 81 instances. The values for σ, c_e, l_r, and γ_0 vary as indicated; the other input parameters are fixed at $\mu = 1$, $c_r = 1000$, $l_e = 1$, $h = 5$.*

The results for the 81 instances are listed in Table 1. In each block we have listed the optimal single index parameters Δ^* and $z_r(\Delta^*)$ on the upper line, the optimal average cost in the middle line, and the value for Δ_{min} and the percentage ordered by the expedited policy under the optimal policy in the bottom line. The results show that:

- In keeping with our analytical results, using the regular supplier as a single source is sometimes optimal ($\Delta = \infty$), but using the expedited supplier alone never is.
- Optimal average costs are increasing as a function of σ, c_e, l_r, and γ_0.
- The value for Δ^* and the percentage ordered by the expedited channel are increasing as a function of σ and l_r, and decreasing as a function of c_e. Also, this percentage is

$l_r \backslash \sigma$		1/3			1			3		
	$\Gamma \backslash c_e$	1020	1050	1100	1020	1050	1100	1020	1050	1100
2	0.9	2.3, 22	2.3, 52	2.3, 102	14, 31	14, 61	14, 111	86, 100	86, 130	86, 180
		2.3, 0%	2.3, 0%	2.3, 0%	14, 4%	14, 0%	14, 0%	83, 4%	84, 3%	85, 1%
	0.95	3.4, 23	3.4, 53	3.4, 103	18, 35	18, 65	18, 115	99, 110	99, 140	99, 190
		3.3, 0%	3.4, 0%	3.4, 0%	18, 5%	18, 2%	18, 0%	96, 3%	97, 2%	97, 2%
	0.99	5.7, 25	5.7, 55	5.7, 105	28, 44	28, 74	28, 124	150, 137	150, 167	150, 217
		5.6, 2%	5.7, 0%	5.7, 0%	27, 6%	27, 3%	28, 2%	126, 8%	136, 9%	148, 1%
4	0.9	3.4, 22	3.4, 52	3.4, 102	19, 31	19, 61	19, 111	105, 100	105, 130	105, 180
		3.3, 2%	3.4, 0%	3.4, 0%	16, 14%	18, 5%	19, 1%	88, 12%	95, 10%	102, 4%
	0.95	4.7, 23	4.7, 53	4.7, 103	24, 35	24, 65	24, 115	134, 110	134, 140	134, 190
		4.5, 5%	4.7, 0%	4.7, 0%	20, 15%	22, 7%	23, 3%	99, 9%	109, 18%	120, 10%
	0.99	7.6, 25	7.6, 55	7.6, 105	35, 44	35, 74	35, 124	174, 137	174, 167	174, 217
		6.9, 9%	7.4, 3%	7.6, 1%	30, 16%	32, 10%	33, 6%	127, 7%	140, 17%	156, 10%
6	0.9	4.4, 22	4.4, 52	4.4, 102	23, 31	23, 61	23, 111	131, 100	131, 130	131, 180
		4.1, 7%	4.4, 0%	4.4, 0%	18, 22%	21, 9%	22, 3%	89, 10%	99, 24%	110, 16%
	0.95	5.9, 23	5.9, 53	5.9, 103	29, 35	29, 65	29, 115	152, 110	152, 140	152, 190
		5.3, 10%	5.8, 2%	5.9, 0%	22, 23%	25, 11%	27, 5%	100, 9%	112, 20%	127, 16%
	0.99	9.2, 25	9.2, 55	9.2, 105	41, 44	41, 74	41, 124	193, 137	193, 167	193, 217
		7.8, 16%	8.7, 5%	9.1, 1%	31, 24%	35, 15%	38, 9%	128, 7%	141, 16%	160, 17%

Table 2: *Optimal costs (excluding $c_r\mu$) for single sourcing by the regular channel, single sourcing by the expedited channel, and dual sourcing, respectively, and the relative savings obtained by dual sourcing in comparison with the best of the two single sourcing options. The values for σ, c_e, l_r, and γ_0 vary as indicated; the other input parameters are fixed at $\mu = 1$, $c_r = 1000$, $l_e = 1$, $h = 5$.*

usually increasing as a function of γ_0. Exceptions do occur for some instances with $\sigma = 3$, and $l_r = 2$. In these cases the increased service level is primarily achieved by increasing the regular inventory level, as the regular leadtime is relatively short.

- The lower bound Δ_{min} is best for small values of σ, and grows worse as the coefficient of variation grows.

The total computation time for all 81 was nine minutes on a regular Pentium II PC, programmed in Delphi.

In Table 2, a comparison is made between single sourcing and dual sourcing for the same 81 instances. In each block, on the upper line we have listed the average single sourcing costs when using either the regular channel or the expedited channel only, and on the lower line the average costs of our optimal single index policy and the relative savings of this policy in comparison to the best of the two single sourcing options. Under single sourcing, the optimal

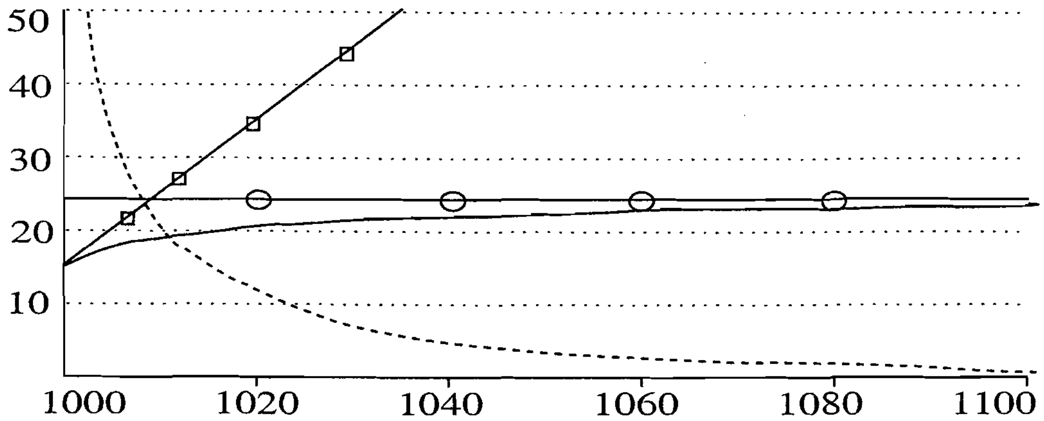

Figure 2: Average costs when using the regular channel only (with circles), when using the expedited channel only (with squares), and under dual sourcing along with the percentage ordered by the expedited channel under dual sourcing (dashed line) for varying values of c_e. Other input parameters are fixed at $\mu = 1$, $\sigma = 1$, $c_r = 1000$, $l_r = 4$, $l_e = 1$, $\gamma_0 = 0.95$, $h = 5$.

policy is a simple base stock policy and the optimal policy and costs are easily determined.

The results in Table 2 show that dual sourcing may lead to savings of 20 % or more in comparison to the best single sourcing option. We also observe that the largest savings occur when the average costs of the two single sourcing options are close to each other. This insight also follows from Figure 2, where the single sourcing costs and dual sourcing costs are depicted for one specific instance as a function of c_e, together with the percentage of demand that is ordered by the expedited channel under the optimal single index policy. For small (large) price differences $c_e - c_r$, almost everything (nothing) is ordered by the expedited channel under the optimal single index policy, and thus then the dual sourcing costs are close to the single sourcing costs when only the expedited (regular) channel is used. In between there is a region where dual sourcing leads to significant lower costs than both single sourcing options. The magnitude of the intermediate region where dual sourcing is most worthwhile to apply depends much on the coefficient of variation of the demand ($= \sigma/\mu$) and the leadtime difference $l_r - l_e$. The higher σ/μ and $l_r - l_e$, the larger this region. So, whether it is attractive for a company to consider dual sourcing depends fundamentally on whether they have many

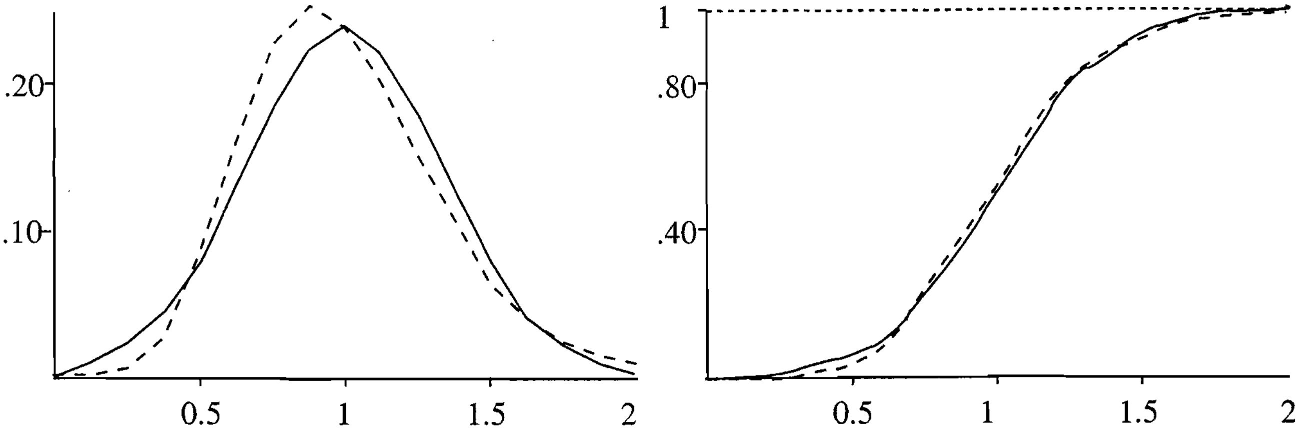

Figure 3: Comparison of the density and cumulative distribution functions of a truncated normal random variable with mean 1 and cv 1/3 (solid line) and the Erlang-9 random variable used to approximate it (dashed line), with $\lambda = 1/9$.

items in this intermediate region or not.

As we have advocated the use of the mixed Erlang as an approximation for other demand models, we illustrate the density and cumulative distribution functions of a mixed Erlang random variable and the truncated normal random variable it approximates. We have chosen the truncated normal as this is commonly used to model demand in inventory systems when the cv of demand is low (see Rao, Scheller-Wolf and Tayur, 2000 for example), but yet is very difficult to treat analytically. As can be seen in Figure 3, the approximation is quite good for the random variable, and should only improve as we convolve to arrive at the distribution of $D(\Delta^*)$. This leads us to believe that, when the true demand distribution is truncated normal, the approximate solution obtained by assuming a mixed Erlang distribution instead will be of very high quality.

7 Extensions

7.1 Capacitated Suppliers

If the expedited supplier has a capacity constraint:

Corollary 7.1 *If there is a capacity C_e on the expedited orders, Lemma 3.3 or Lemma 4.1 holds with the modification that* $D(\Delta) \sim \sum_{i=1}^{l_e+1} d_i + \sum_{j=l_e+2}^{l_r+1} M_j$, *where*

$$M_j = \begin{cases} d_j & d_j \leq \Delta, \\ \Delta & \Delta < d_j \leq C_e + \Delta, \\ d_j - C_e & C_e + \Delta < d_j. \end{cases}$$

The single index policy with a capacitated regular supplier is even simpler – placing a per-period limit of C_r on the regular order is equivalent to specifying $\Delta \leq C_r$:

Corollary 7.2 *If there is a capacity C_r on the regular orders and the additional constraint $\Delta \leq C_r$ is added, then all the results above hold.*

7.2 Multiple Suppliers

If there are more delivery options (as in Zhang 1996) a single index policy once again yields a dimensional reduction. The case of three options is illustrated below.

Corollary 7.3 *If there are three delivery modes, for each fixed Δ_1 and Δ_2, given $l_1 < l_2 < l_3$ and $z_3 = z_2 + \Delta_2$, $z_2 = z_1 + \Delta_1$, Lemma 3.3 or Lemma 4.1 holds with the modification that :*

$$D(\Delta_1, \Delta_2) \sim \sum_{i=1}^{l_1+1} d_i + \sum_{j=l_1+2}^{l_2+1} \min(d_j, \Delta_1 + \Delta_2) + \sum_{k=l_2+2}^{l_3+1} \min(d_k, \Delta_2).$$

8 Conclusion

We have defined the single index inventory replenishment policy for use when there is a choice of two supply options differing in unit cost and fixed lead time. If the lead times differ by exactly one unit our policy is known to be optimal, otherwise it serves as a heuristic. We show how to calculate optimal single index parameters for both the penalty cost and service level versions of the problem for continuous demand distributions, via a mixture of Erlang distributions, demonstrating the effectiveness of this policy via computational experiments. This line of research offers a number of avenues for future work: Further exploration of the extensions in Section 7, lost sales models, models with discrete demand, and the analysis and performance of other simple ordering policies.

A Proofs of Analytical Erlang Results

A.1 Proof of Lemma 5.1

Proof : From the definition of $\hat{G}_{m,n}$ and of $H_{m,n}$:

$$\begin{aligned}
\hat{G}_{m,n}(x) &= \mathbf{P}\{\hat{Y}_{m,n} \le x\} = \mathbf{P}\{\sum_{i=1}^{m} \hat{d}_i + \sum_{i=m+1}^{m+n} d_i \le x\} \\
&= \sum_{k=0}^{m} \binom{m}{k} \mathbf{P}\{\sum_{i=1}^{m} \hat{d}_i + \sum_{i=m+1}^{m+n} d_i \le x, \ d_i < \Delta \text{ for } i = 1, \ldots, m-k, \\
&\qquad\qquad d_i \ge \Delta \text{ for } i = m-k+1, \ldots, m\} \\
&= \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} \mathbf{P}\{\sum_{i=1}^{m} \hat{d}_i + \sum_{i=m+1}^{m+n} d_i \le x \mid d_i < \Delta \text{ for } i = 1, \ldots, m-k, \\
&\qquad\qquad d_i \ge \Delta \text{ for } i = m-k+1, \ldots, m\} \\
&= \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} \mathbf{P}\{\sum_{i=1}^{m-k} d_i + \sum_{i=m+1}^{m+n} d_i \le x - k\Delta \mid \\
&\qquad\qquad d_i < \Delta \text{ for } i = 1, \ldots, m-k\} \\
&= \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} H_{m-k,n}(x - k\Delta) \ , \qquad x \ge 0.
\end{aligned}$$

■

A.2 Proof of Lemma 5.2

Proof :

$$H_{m,n}(x) = \frac{1}{(1-p)^m} \mathbf{P}\{\sum_{i=1}^{m+n} d_i \le x, \ d_i < \Delta \text{ for } i = 1, \ldots, m\}$$

$$
\begin{aligned}
&= \frac{1}{(1-p)^m}\Bigg(\mathbf{P}\{\sum_{i=1}^{m+n} d_i \le x\} - \mathbf{P}\{\sum_{i=1}^{m+n} d_i \le x,\ d_1 \ge \Delta\} - \dots \\
&\qquad -\mathbf{P}\{\sum_{i=1}^{m+n} d_i \le x,\ d_m \ge \Delta\} + \mathbf{P}\{\sum_{i=1}^{m+n} d_i \le x,\ d_1 \ge \Delta,\ d_2 \ge \Delta\} + \dots\Bigg) \\
&= \frac{1}{(1-p)^m}\sum_{k=0}^{m}(-1)^k\binom{m}{k}\mathbf{P}\{\sum_{i=1}^{m+n} d_i \le x,\ d_i \ge \Delta \text{ for } i=1,\dots,k\} \\
&= \frac{1}{(1-p)^m}\sum_{k=0}^{m}(-1)^k\binom{m}{k}p^k\mathbf{P}\{\sum_{i=1}^{m+n} d_i \le x \mid d_i \ge \Delta \text{ for } i=1,\dots,k\} \\
&= \frac{1}{(1-p)^m}\sum_{k=0}^{m}(-1)^k\binom{m}{k}p^k\mathbf{P}\{\sum_{i=1}^{k}(d_i-\Delta) + \sum_{i=k+1}^{m+n} d_i \le x - k\Delta \mid \\
&\qquad\qquad d_i \ge \Delta \text{ for } i=1,\dots,k\} \\
&= \frac{1}{(1-p)^m}\sum_{k=0}^{m}(-1)^k\binom{m}{k}p^k\mathbf{P}\{\sum_{i=1}^{k}\tilde{d}_i + \sum_{i=k+1}^{m+n} d_i \le x - k\Delta\} \\
&= \frac{1}{(1-p)^m}\sum_{k=0}^{m}(-1)^k\binom{m}{k}p^k\tilde{G}_{k,m+n-k}(x-k\Delta)\ , \quad x \in \mathbf{R}.
\end{aligned}
$$

■

A.3 Proof of Lemma 5.3

Proof : Starting from the formula in Lemma 5.1:

$$
\begin{aligned}
\hat{G}_{m,n}(x) &= \sum_{k=0}^{m}\binom{m}{k}p^k(1-p)^{m-k} \\
&\qquad \left(\frac{1}{(1-p)^{m-k}}\sum_{j=0}^{m-k}(-1)^j\binom{m-k}{j}p^j\tilde{G}_{j,m+n-k-j}(x-j\Delta-k\Delta)\right) \\
&= \sum_{k=0}^{m}\binom{m}{k}p^k\sum_{j=0}^{m-k}(-1)^j\binom{m-k}{j}p^j\tilde{G}_{j,m+n-k-j}(x-j\Delta-k\Delta) \\
&= \sum_{k=0}^{m}\sum_{j=0}^{m-k}(-1)^j\binom{m}{k}\binom{m-k}{j}p^{k+j}\tilde{G}_{j,m+n-(k+j)}(x-(k+j)\Delta)\ , \qquad x \in \mathbf{R}.
\end{aligned}
$$

By substitution of $i = k$ and $s = j + k$ (or, equivalently, $k = i$ and $j = s - i$), we obtain

$$
\begin{aligned}
\hat{G}_{m,n}(x) &= \sum_{s=0}^{m}\sum_{i=0}^{s}(-1)^{s-i}\binom{m}{s}\binom{s}{i}p^s\tilde{G}_{s-i,m+n-s}(x-s\Delta) \\
&= \sum_{s=0}^{m}(-1)^s\binom{m}{s}p^s\sum_{i=0}^{s}(-1)^i\binom{s}{i}\tilde{G}_{s-i,m+n-s}(x-s\Delta)\ , \qquad x \in \mathbf{R}.
\end{aligned}
$$

■

A.4 Proof of Lemma 5.4

Proof : For (i):

$$\begin{aligned}\tilde{K}_{m,n}(x) &= \int_x^\infty (u-x)\mathrm{d}\tilde{G}_{m,n}(u) \\ &= \sum_{k=1}^\infty \tilde{q}_k^{(m,n)} \int_x^\infty (u-x)\mathrm{d}E_{k,\lambda}(u) \\ &= \sum_{k=1}^\infty \tilde{q}_k^{(m,n)} \left(\frac{k}{\lambda}(1-E_{k+1,\lambda}(x)) - x(1-E_{k,\lambda}(x))\right), \qquad x\in\mathbf{R},\end{aligned}$$

and for (ii):

$$\begin{aligned}\hat{K}_{m,n}(x) &= \int_x^\infty (u-x)\mathrm{d}\hat{G}_{m,n}(u) \\ &= \sum_{s=0}^m (-1)^s \binom{m}{s} p^s \sum_{i=0}^s (-1)^i \binom{s}{i} \int_x^\infty (u-x)\mathrm{d}\tilde{G}_{s-i,m+n-s}(u-s\Delta) \\ &= \sum_{s=0}^m (-1)^s \binom{m}{s} p^s \sum_{i=0}^s (-1)^i \binom{s}{i} \int_{x-s\Delta}^\infty (v-(x-s\Delta))\mathrm{d}\tilde{G}_{s-i,m+n-s}(v) \\ &= \sum_{s=0}^m (-1)^s \binom{m}{s} p^s \sum_{i=0}^s (-1)^i \binom{s}{i} \tilde{K}_{s-i,m+n-s}(x-s\Delta), \qquad x\in\mathbf{R}.\end{aligned}$$

■

References

[1] Beyer, D. and Ward, J. "Network Server Supply Chain at HP: A Case Study", Hewlett Packard Software Technology Lab Report HPL 2000-84, Palo Alto, (2000).

[2] Burgin, T.A. and Wild, A. R. "Stock Control – Experience and Usable Theory", *Operations research Quarterly*, 18 (1967) pp. 35-52.

[3] Fukuda, Y., "Optimal Policy for the Inventory Problem with Negotiable Leadtime", *Management Science*, 12 (1964) pp. 690-708.

[4] Lawson, D., and Porteus, E., "Multistage Inventory Management with Expediting", *Operations Research*, 46 6 (2000) 878-893.

[5] Nahmias, S., *Production and Operations Analysis, Third Edition*, 1997, Irwin, Chicago.

[6] RAO, U., SCHELLER-WOLF, A., AND TAYUR, S., "Development of a Rapid-Response Supply Chain at Caterpillar", *Operations Research*, 48 (2000) 189–204.

[7] SCHASSBERGER, R., *Warteschlangen*, Springer-Verlag, Berlin, 1973.

[8] SILVER, E. PYKE, D. AND PETERSON, S. *Inventory Management and Production Planning Scheduling, Third Edition*, 1998, John Wiley and Sons, New York.

[9] TAGARAS, G., AND VLACHOS, D., "An Periodic Review Inventory System with Emergency Replenishments", *Management Science*, 47 (2001) 415–429.

[10] TAYUR, S., GANESHAN, R. AND MAGAZINE, M. *Quantitative Models for Supply Chain Management*, Kluwer, Newel, Massachusetts, 1999.

[11] TEUNTER, R., AND VLACHOS, D., "An Inventory System with Periodic Regular Review and Flexible Emergency Review", *IIE Transactions* 33 (2001) 625–635.

[12] TIJMS, H.C., *Stochastic Models; An Algorithmic Approach*, Wiley, New York, 1986.

[13] VAN HOUTUM, G.J., AND ZIJM, W.H.M., "Incomplete Convolutions in Production and Inventory Models", *OR Spektrum*, 19 (1997) 97–107.

[14] VAN HOUTUM, G.J., AND ZIJM, W.H.M., "On the Relation Between Cost and Service Models for General Inventory Systems", *Statistica Neerlandica*, 54 (2000) 127–147.

[15] VLACHOS, D., AND TAGARAS, G., "An Inventory System with Two Supply Modes and Capacity Constraints", *Int. Journal of Production Economics*, 72 (2001) 41–58.

[16] WHITTMORE, A.S., AND SAUNDERS, S., "Optimal Inventory Under Stochastic Demand with Two Supply Options", *SIAM Journal of Applied Math*, 32 (1977) 293-305.

[17] V. L. ZHANG, "Ordering Policies for an Inventory System with Three Supply Modes", *Naval Research Logistics*, 43 (1996) 691-708.