OPERATIONS   RESEARCH   CENTER

Working Paper

On Production and Subcontracting Strategies for
Manufacturers with Limited Capacity and Backlog-Dependent
Demand

by

B. Tan
S. B. Gershwin

OR 354-01   May 2001

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

On Production and Subcontracting Strategies for Manufacturers with Limited Capacity and Backlog-Dependent Demand

Barış Tan
Graduate School of Business
Koç University
Rumeli Feneri Yolu, Sarıyer
Istanbul, Turkey
btan@ku.edu.tr

Stanley B. Gershwin
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139-4307
USA
gershwin@mit.edu

May 3, 2001

Abstract

We study a manufacturing firm that builds a product to stock to meet a random demand. If there is a positive surplus of finished goods, the customers make their purchases without delay and leave. If there is a backlog, the customers are sensitive to the quoted lead time and some choose not to order if they feel that the lead time is excessive. A set of subcontractors, who have different costs and capacities, are available to supplement the firm's own production capacity. We derive a feedback policy that determines the production rate and the rate at which the subcontractors are requested to deliver products. The performance of the system when it is managed according to this policy is evaluated. The subcontractors represent a set of capacity options, and we calculate the values of these options.

Contents

1   Introduction

1.1   Goal of the paper

THE purpose of this study is to extend the one-part-type, one-machine control problem of Bielecki and Kumar (1988). Bielecki and Kumar (1988) obtained an analytic solution to a special case of the hedging point problem of Kimemia and Gershwin (1983), in which a factory manager had to decide how to operate an unreliable machine to best satisfy a constant demand. We extend the model in three important directions.

1. Backlog and customer behavior   First, we treat backlog in a more fundamental way than Kimemia and Gershwin (1983), Bielecki and Kumar (1988), or any of the subsequent papers that refined or extended their models of real-time scheduling of manufacturing systems. In these papers, the difference between cumulative production and cumulative demand is called *surplus*, and is usually represented by x. When x is negative, it is *backlog*. The performance objective to be minimized is a function of x, which increases as x deviates from 0, for both x positive and x negative. In this way, the optimization tends to keep x near 0.

This makes economic sense for $x > 0$. In that case, x is *finished goods inventory*, and there are clear, tangible costs associated with inventory (including the interest cost on the raw material, the floor space devoted to storage, etc.). However, there is no such tangible cost associated with backlog. The undesirable consequence of backlog is the loss of sales, and lost sales are not related to backlog by a simple quantitative relationship.

Here, instead of including an explicit cost term for $x < 0$, we model the response of potential customers to backlogs. In this model, if there is a positive surplus of finished goods, the customers make their purchases without delay and leave. If there is a backlog, some fraction of the customers are willing to wait to make their purchases, but others depart in disgust. The greater the backlog, the more customers leave without making a purchase.

The reason for avoiding backlog comes from the fact that some potential customers choose not to place orders, and such lost sales reduce revenues. In this way, we replace an artificial, contrived cost term with a more natural model of the phenomenon that causes the cost.

2. Subcontractors   Second, we provide the factory manager with external sources for the product. In this way, if demand temporarily exceeds capacity, the manager may purchase some of the product from others to reduce backlog, improve service to customers, and reduce the number of lost sales. However, this comes at a price: the profit made from purchased finished products is less than that from items produced in-house.

The same model may be used for a different purpose. This is where there are two or more production resources available within a single factory. They have different operating costs and different maximum production rates. The manager must decide which resource to use at any time.

3. Reliable supply and variable demand   We assume a perfectly reliable factory and perfectly reliable subcontractors. Randomness in the model comes from the variability of the demand.

Customers arrive at the factory at rate d, which is either high ($d = \mu_H$) or low ($d = \mu_L$). The transitions from high to low and low to high occur at exponentially distributed time instants.

The problem The problem we solve is: How do we operate our manufacturing plant, and how do we use subcontractors to maximize profit? Profit is revenue minus cost; the revenues are diminished by customers who defect rather than wait when they see a backlog. The cost is nonzero only when the surplus is positive.

Special cases A special case of the customer behavior is the *lost sales* case, in which customers choose not to place orders whenever there is any backlog. Another special case of the model is where the firm does no manufacturing of its own, and only uses subcontractors.

Features not included We only consider the effect of backlog on present sales. We do not consider the fact that a customer who finds the backlog too great is less likely to attempt to make a purchase in the future; and we do not consider the damage to a firm's reputation when it has frequent large backlogs.

1.2   Motivation

The retail-apparel-textile channel is characterized by rapidly changing styles, uncertain customer demand, product proliferation, and long lead times. Retailers adopt lean retailing practices to place a larger fraction of their orders during the season (Abernathy, Dunlop, Hammond, and Weil 1999). This change shifts the risks associated with carrying too much or too little inventory from retailers to manufacturers. In order to respond quickly to the retailer demand, a manufacturer can produce well in advance to stock or increase its capacity to reduce the lead time.

Often, neither of these choices is desirable. Utilizing subcontractors can be an attractive option for manufacturers with limited capacity and volatile demand. Higher prices associated with subcontractors that are located near the market can be justified by reduced inventory carrying, lost sales, and markdown costs (Abernathy, Dunlop, Hammond, and Weil 2000) .

As an example, a major jean manufacturer, the VF corporation, operates two plants in the US in order to respond the replenishment orders places by Wal-Mart in four days or less. Most of the other VF plants are located outside the US. They are evidently willing to accept higher costs so they can provide short lead times. (Here the offshore plants are like our firm's own facility, and the US plants are like the high priced subcontractors.)

In the auto industry, some European companies are employing a flexible work force by paying a base salary to a group of workers who get additional hourly wage when and if they are needed. (The base hours are like the firm's facility, and the overtime is like the subcontractor.) Deciding when to use this temporary workforce or, in general, deciding when to use overtime can also be addressed by the model considered here.

1.3 Approach

We form a dynamic programming problem which is similar to that of Bielecki and Kumar (1988), except that there is no cost for backlog in the objective function. Instead, the objective function rewards revenues and penalizes cost, which is due only to inventory. The revenues are greatest for products manufactured in-house, and less for those provided by subcontractors. Revenues are also diminished by the defection of potential customers who are not willing to wait for their products. We introduce a *defection function* $B(x)$ which represents the probability that a potential customer will not complete his order when the backlog is x. Then the instantaneous demand is reduced by a factor of $1 - B(x)$.

We use the Bellman equation to determine the structure of the solution. We find that it is a generalization of the hedging point policy of Kimemia and Gershwin (1983) and Bielecki and Kumar (1988). The hedging point is a threshold indicating when there is sufficient surplus, and there are additional thresholds to indicate when to use each of the subcontractors.

To determine the optimal values of these thresholds, we derive the differential equations for the density function of the surplus. These equations can be solved analytically when $B(x)$ is piecewise constant. Since we can approximate any $B(x)$ with a piecewise constant function, we can therefore solve systems with essentially any $B(x)$. We express the solution as a function of the thresholds. Finally, we optimize over the thresholds.

1.4 Past Work

1.4.1 Dynamic programming formulations of factory scheduling and inventory control

Since the 1980s, there has been an increasing interest in devising optimal production control policies that manage production in uncertain environment. An optimal flow-rate control problem for a failure prone machine subject to a constant demand source was introduced by Olsder and Suri (1980) and Kimemia and Gershwin (1983). The single-part-type, single-machine problem was analyzed in detail by Bielecki and Kumar (1988). The optimal control is a hedging point policy where the machine operates at its maximum rate until the inventory reaches a certain level; and then it operates at a rate that keeps the inventory at this level.

Hedging point control policies are optimal or near optimal for a range of manufacturing system models. Hedging policies have been shown to be effective in a manufacturing environment by Yan, Lou, Sethi, Gardel, and Deosthali (1996). For an overview of the dynamic programming formulations of factory scheduling and inventory control and a comprehensive list of references, see Gershwin (1994).

Most of these studies assume a constant demand source. Only a few consider optimal production control problems with random demand, including Fleming, Sethi, and Soner (1987), Ghosh, Araposthathis, and Markus (1993), Tan (2000), and Perkins and Srikant (2001).

1.4.2 Subcontracting

The extension of the problem for an unreliable machine with constant demand and a single contractor whose capacity is high enough to meet the demand was introduced by Gershwin (1993). He conjectured the optimality of the hedging policy. The optimality of this policy is proven by Huang, Hu, and Vakili (1999). A simpler version of this problem where backlog is not permitted and the subcontractor with sufficient capacity is used when the inventory level is zero was analyzed by Hu (1995).

The problem of controlling a machine that can produce at a fixed rate to meet random demand is presented by Krichagina, Lou, and Taksar (1994). In this study, whenever the machine stops, a setup is performed to start production again. Moreover, backlogging is not allowed and a subcontractor can be used by paying a fixed cost and a variable cost. By using a Brownian approximation of the model, it is shown that the optimal policy is characterized by three parameters and referred as a double-band policy.

A two-period competitive stochastic investment game is presented by Van Mieghem (1999). In this model, the manufacturer and subcontractor decide on their capacity investment levels separately in the first period. After the demand uncertainty is resolved in the second period, both parties decide on their production and sales with the option to subcontract. The value of the option of subcontracting is determined is then determined by using this model.

A Brownian motion approximation for the optimal subcontracting policy for an $M/M/1$ system is given by Bradley (1999). This problem is extended by Bradley and Glynn (2000a) by optimizing the capacity, inventory, and subcontracting jointly. Competition and coordination issues in this model are addressed by Bradley and Glynn (2000b) by focusing on a multi-stage game where the equilibrium is computed by using the Brownian approximation.

For a number of discrete-time inventory models with two sources, a dual base-stock policy, with parameters s_1 and s_2, is shown to be optimal. In this policy, the first source, the manufacturer, is used when the inventory falls below s_1 and the second source, the subcontractor, is used when the inventory falls below s_2. See Fukuda (1964), Whittemore and Saunders (1977), Zhang (1995). The same policy is shown to be optimal for a continuous time $M/M/1$ system with average cost criterion by Bradley (1999).

In a different context, the problem of managing a number of power generators with different costs and capacities is studied by Schweppe, Caramanis, Tabors, and Bohn (1988). They show that the power generators are used in order of increasing per unit production cost when the marginal cost of losing the demand is higher than the marginal cost of receiving the energy from that source. Since the electrical energy cannot be stored or backlogged, this problem is a special case of the problem studied here where g^+ is very high and $B(x) = 1$ for $x \leq 0$.

1.4.3 Queues with impatient customers

The queuing literature provides models that examine the behavior of a customer who has to wait for service. In the basic models, it is assumed that the customer stays in the system until she is served. The basic queueing models are extended to include *reneging* (abandoning the queue after waiting some time) and *balking* (not joining the queue if the server is not immediately available)

(Hall 1991).

In a study motivated by telephone call centers, a method to predict waiting time in a multi-server exponential queue is proposed by Whitt (1999b). This method utilizes information about the number of customers in the system ahead of the current customer. It is stated that the waiting time prediction may be used to decide when to add additional service agents. Competition between two firms with service time sensitive customers who choose firms based on the firms' prices, their expected waiting and service times and their brands is analyzed by Cachon (1999).

In a retail queueing model (Ittig 1994), the effect of waiting time on customer demand is taken into consideration when the optimal number of clerks for the queueing system is determined.

The behavior of internet users who react to the waiting times on the web is analogous to the behavior of customers who react to the waiting times in a manufacturing environment. Dellaert and Kahn (1999) reports that the waiting times to load a web page can affect evaluation of web sites. When users experience long waits for a web site's home page to load, they either quit using the web or redirect to an alternative web page (Weinberg 2000).

Communicating waiting time information to customers is one way to improve the customer experience, according to Taylor (1994) and Hui and Tse (1996). In an $M/M/s/r$ queueing model with balking and reneging, Whitt (1999a) shows that informing customers about anticipated delays improves system performance.

In inventory management, most of the models assume that shortages are either completely lost or completely backlogged. In a recent study,partial backlogging and service-dependent sales are incorporated in a supply chain configuration study at Caterpillar Inc. (Rao, Scheller-Wolf, and Tayur 2000). Chang and Dye (1999) extended the basic economic order quantity model to include partial backlogging, where the backlog rate is inversely proportional to the waiting time for the next replenishment.

In our model, we assume that conservative estimates of the waiting time are given to the arriving customers. Having this information, the customer then decides to wait or leave the system. However, once the customer is in the queue, she does not renege, because this move is not consistent with her earlier decision to accept waiting. Our model can be described, in queueing terminology, as one with queue-length-dependent balking and no reneging.

1.4.4 Empirical work on customer behavior

A number of studies investigated the effect of waiting time on customer demand in health care. In these studies, the effect is summarized by elasticity of demand with respect to waiting time. The effect of waiting time and private insurance premiums on the demand for public and private heath care providers are estimated in the UK (McAvinchey and Yannopoulos 1993). In another study that investigates the rationing effect of waiting, elasticity of demand with respect to waiting time is estimated empirically using the waiting list for elective surgery in the British National Health Service (Martin and Smith 1999).

A model that investigates service time competition between companies is tested for two identical gasoline service stations (Mount 1994). It is found that retail demand is sensitive to service time and customers are willing to pay about 1% more for a 6% reduction in congestion.

1.5   Overview

The model description and its assumptions are given in Section 2 for the case where there are no subcontractors. In this section, backlog-dependent demand is introduced and the production control problem is stated. The optimal policy that maximizes the profit is characterized in Section 3. Subcontractors are introduced in Section 4, and the more general solution structure is derived in Section 5. In Section 6, the problem with the subcontractors is transformed so that the results are equally applicable to a system with an unreliable manufacturing facility, reliable subcontractors, and constant demand. The model is analyzed and the steady state probability distributions are formulated in Section 7. Section 8 describes the evaluations of the objective function and of other performance measures of interest. The behavior of the model is investigated in Section 9. Section 10 contains a summary of the paper and several proposed research directions, and Appendix A constructs the defection function as seen by one of the servers in a shortest-queue system.

2   Model Description — Single Producer

In this section, we describe a limited version of the problem in which there is only one production resource, and no subcontractors. We describe the structure of the solution in Section 3. We introduce subcontractors in Section 4.

2.1   Basic Model

We consider a make-to-stock system with a single manufacturing facility that produces to meet the demand for a single item. Production, demand, inventory, and backlog are all represented by continuous (real) variables. The demand rate at time t is denoted by $d(t)$. The state of the demand at time t is $D(t)$ which is either high (H) or low (L). When the demand is high, the demand rate is $d(t) = \mu_{\mathrm{H}}$ and when the demand is low, the demand rate is $d(t) = \mu_{\mathrm{L}}$. At time t, the amount of finished goods inventory or backlog is $x(t)$.

The times to switch from a high demand state to a low demand state and from a low demand state to a high demand state are assumed to be exponentially distributed random variables with rates λ_{HL} and λ_{LH}. This model is suitable for describing demand which is stationary in the long run, but whose mean shifts temporarily as a result of promotions, competitor actions, etc. The time since the last state change does not change the expected time until the next state change. The average demand rate is

$$Ed = \mu_{\mathrm{H}}e + \mu_{\mathrm{L}}(1 - e)$$

where

$$e = \frac{\lambda_{\mathrm{LH}}}{\lambda_{\mathrm{HL}} + \lambda_{\mathrm{LH}}}$$

is the percentage of the time the demand is high.

The total amount demanded during a period of length t, from this model, is asymptotically normal as $t \to \infty$ (Tan 1997). The limiting variance rate of the amount demanded during a period of length t satisfies

$$Vd = \lim_{t \to \infty} \frac{Var[N(t)]}{t} = \frac{2(\mu_{\mathrm{H}} - \mu_{\mathrm{L}})^2(1 - e)e^2}{\lambda_{\mathrm{LH}}}$$

The asymptotic coefficient of variation of the demand, defined as $cv = \sqrt{Vd}/Ed$, is used as a summary measure of the demand variability in this study. Here, it is given by

$$cv = \sqrt{\frac{\mu_{\mathrm{H}} - \mu_{\mathrm{L}}}{\lambda_{\mathrm{LH}}\mu_{\mathrm{H}} + \lambda_{\mathrm{HL}}\mu_{\mathrm{L}}} \left(\frac{1}{\lambda_{\mathrm{HL}}} + \frac{1}{\lambda_{\mathrm{LH}}}\right)^{-1}}$$

The maximum production rate of the manufacturing facility is $\underline{u}_0$. The actual production rate of the manufacturing facility at time t is a control variable which is denoted by $u_0(t)$, $0 \le u_0(t) \le \underline{u}_0$. We assume that the production capacity $\underline{u}_0$ is sufficient to meet the demand when it is low but insufficient when it is high, i.e., $\mu_{\mathrm{L}} < \underline{u}_0 < \mu_{\mathrm{H}}$. (Note that if $\underline{u}_0 > \mu_{\mathrm{H}}$, the problem is trivial: it is always possible to keep x at 0 and therefore a backlog situation never happens. Similarly, if $\underline{u}_0 < \mu_{\mathrm{L}}$, the problem is also trivial: the manufacturing facility is run with the maximum rate all the time.)

The profit coefficient (dollars per unit) for the goods produced in the factory is A_0 and the inventory carrying cost is g^+ (dollars per unit per time). As indicated earlier, we do not include the corresponding backlog cost g^-, which does appear in Bielecki and Kumar (1988) and many other papers.

2.2   Backlog-Dependent Demand

When there is backlog (i.e., when $x < 0$), a potential customer chooses not to order with probability $B(x)$ when the backlog is x. (Alternatively, $B(x)$ is the fraction of potential customers who choose not to order when the backlog is x.) $B(x)$, which is called the *defection function*, satisfies

$$\left. \begin{array}{l} 0 \le B(x) \le 1, \\[2mm] x \ge 0 \Longrightarrow B(x) = 0, \\[2mm] x < 0 \Longrightarrow B(x) > 0, \end{array} \right\} \tag{1}$$

The first condition is required by the definition of B as a probability or a fraction. The second says that no potential customers are motivated to defect when there is surplus. The third says that there are always some potential customers that refuse to wait if there is a backlog.

If $B(x)$ satisfies

$$B(x) \text{ is a non-increasing function of } x, \tag{2}$$

10
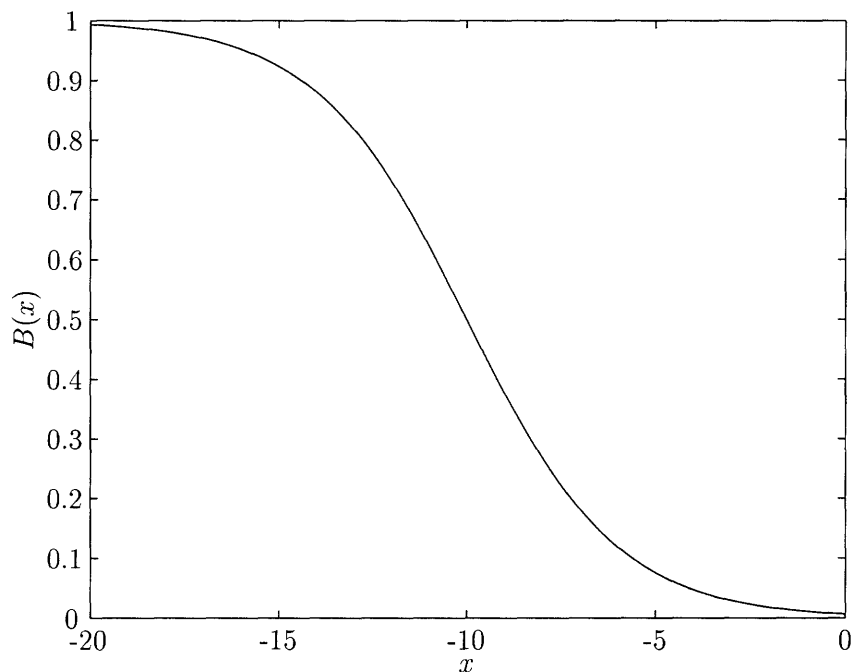
Figure 1: A sample $B(x)$ function

we say that B is a *monotonic defection function*. If B is monotonic, more customers are impatient if there is a longer wait. In the following, we restrict our attention to monotonic defection functions. Figure 1 shows an example of a $B(x)$ function that satisfies these conditions.

An additional possible condition is

$$\lim_{x \to -\infty} B(x) = 1$$

which says that nobody is infinitely patient. Because this is debatable, and not essential for the subsequent analysis, we do not require it.

In Appendix A, we derive the customer defection function that is generated by customers who choose the shortest of two queues. This preliminary analysis supports the form of $B(x)$ postulated here.Analysis of alternative functional forms of $B(x)$ resulting from different customer behavior is left for future research.

Note that $B(x) = 1$ for all $x < 0$ corresponds to the lost sales case. In this situation, no customers are willing to wait to receive their goods. All revenues are lost whenever there is any backlog.

When the surplus level is $x < 0$, the time until the next arriving customer order will be completed, i.e., the production lead time, is $-x/\underline{u}_0$. This is the time to clear the current backlog, assuming $u = \mu_0$ until the backlog is cleared. The lead-time dependent demand case can therefore be treated by using $B(x) = \tilde{B}(\underline{u}_0\tau)$ as the probability that a potential customer chooses not to order when the quoted lead time is $\tau = -x/\underline{u}_0$.

11

The dynamics of x are given by

$$\frac{dx}{dt} = u_0 - d(1 - B(x)).$$ (3)

This equation has an important property. For any given non-zero u_0, there is some sufficiently negative value of x such that $dx/dt > 0$. The profit function (4) increases with u_0 and is independent of x when x is negative, so there is no reason for u_0 to be zero when x is negative. In fact, there is no reason for u_0 to be anything but maximal when x is sufficiently negative. As a consequence, x is bounded from below. There is a value of x, called x^*, such that

$$\frac{dx}{dt} = 0 = \underline{u}_0 - \mu_{\mathrm{H}}(1 - B(x^*))$$

and $x(t) \geq x^*$ for all t. (If we had used $d = \mu_{\mathrm{L}}$ in this equation, we would have found $x^* = 0$ because $\underline{u}_0 > \mu_{\mathrm{L}}$.)

2.3   Production Control Problem

The decision variable is the rate at which the goods are produced at the plant at time t, $u_0(t)$. The profit function to be maximized is the difference between the profit generated through production and the inventory carrying costs. A linear inventory carrying cost function is assumed, i.e.,

$$g(x) = \begin{cases} g^+ x \text{ if } x \geq 0 \\ \\ 0 \text{ if } x < 0 \end{cases}$$

The production control problem is

$$V = \max_{u_0} \Pi = E \int_0^T A_0 u_0 - g(x) dt$$ (4)

subject to

$$\frac{dx}{dt} = u_0 - d(1 - B(x))$$ (5)

$$0 \leq u_0 \leq \underline{u}_0$$ (6)

$$d = \begin{cases} \mu_{\mathrm{H}} & \text{if } D = \mathrm{H} \\ \mu_{\mathrm{L}} & \text{if } D = \mathrm{L} \end{cases}$$ (7)

Markov dynamics for D with rates λ_{HL} from H to L and λ_{LH} from L to H (8)

We do not include an explicit cost for backlog in the definition of g. Instead, the firm is penalized for backlog by the defection of impatient customers and the reduced profits due to the use of subcontractors. Since the difference between the cumulative production and cumulative

demand is finite in the long run, the profit term in the objective function is written using the production rate rather than the demand rate.

We assume that T is very large, so the optimal policy discussion in Section 3 is based on the assumption that the probability distribution of (x, D) reaches a steady state. We do not need to assume that the demand is feasible. If Ed is greater than u_0, then enough impatient customers will defect to guarantee that x is bounded from below, i.e., $x \geq x^*$.

3   Characterization of the Policy

Problem (4)-(8) is a dynamic programming problem. The solution (the optimal production rate u_0 as a function of x, D, and t) satisfies the Bellman equation. In this section, we use the Bellman equation to determine the structure of the solution.

3.1   Bellman equation

Define the *value function*:

$$V(x(t), D(t), t) = \min_{u_0} E \int_t^T (-A_0 u_0 + g(x)) d\tau \tag{9}$$

V satisfies the *maximum principle*, which asserts that

$$-\frac{\partial V}{\partial t}(x, \mathrm{L}, t) = \min_{u_0} \left\{ -A_0 u_0 + g(x) + \frac{\partial V}{\partial x}(x, \mathrm{L}, t) \left(u_0 - \mu_\mathrm{L}(1 - B(x))\right) \right.$$

$$\left. + V(x, \mathrm{H}, t)\lambda_\mathrm{HL} - V(x, \mathrm{L}, t)\lambda_\mathrm{HL} \right\} \tag{10}$$

for $D = \mathrm{L}$, and

$$-\frac{\partial V}{\partial t}(x, \mathrm{H}, t) = \min_{u_0} \left\{ -A_0 u_0 + g(x) + \frac{\partial V}{\partial x}(x, \mathrm{H}, t) \left(u_0 - \mu_\mathrm{H}(1 - B(x))\right) \right.$$

$$\left. + V(x, \mathrm{L}, t)\lambda_\mathrm{LH} - V(x, \mathrm{H}, t)\lambda_\mathrm{LH} \right\} \tag{11}$$

for $D = \mathrm{H}$. The minimizations are taken over constraints (6).

It is reasonable to assume, since V is the solution of a dynamic programming problem, that V is strictly convex in x. Therefore it has a unique minimum. If that minimum were not finite, x would be increasing or decreasing without bound. If x were increasing without bound, (9) implies that V is infinite, and this cannot be optimal. It is not possible for x to decrease without bound because $B(x)$ increases with decreasing x, and there is a value of x (called x^*) below which dx/dt (3) must be zero or positive. Therefore V has a finite minimum. Figure 2 shows V as a function of x when $d = \mu_\mathrm{L}$ and we assume that V is continuously differentiable. The graph of V when $d = \mu_\mathrm{H}$ is similar.
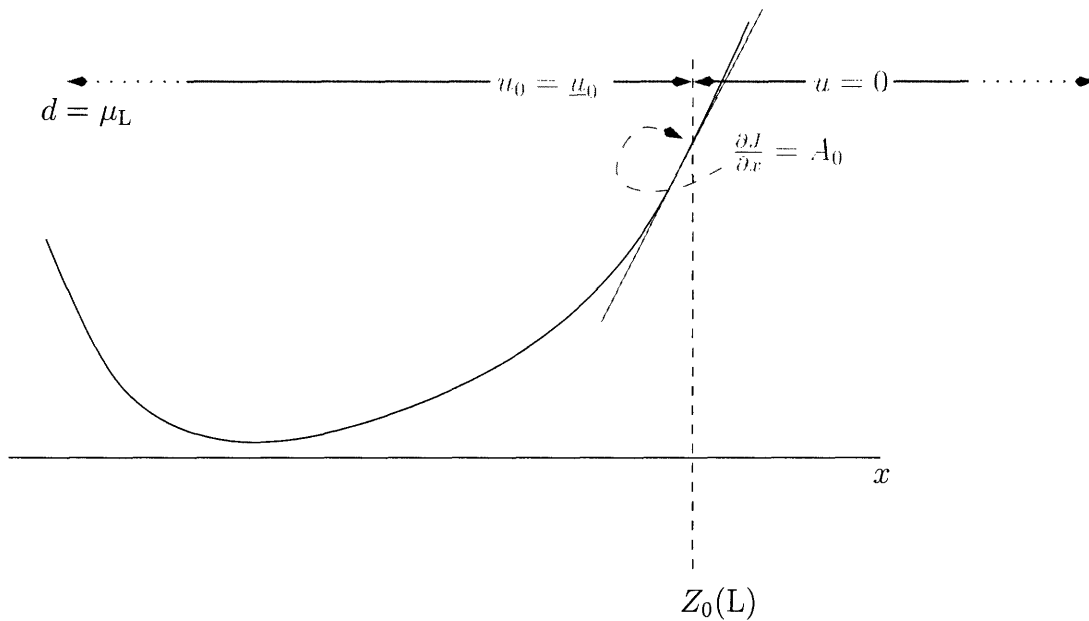
Figure 2: V vs. x, $d = \mu_{\mathrm{L}}$

¿From (10) and (11), for both values of D,

$$
u_0^*(x, D) = \begin{cases} 0 & \text{if} \quad -A_0 + \dfrac{\partial V}{\partial x}(x, D) > 0 \\[2mm] \text{undetermined} & \text{if} \quad -A_0 + \dfrac{\partial V}{\partial x}(x, D) = 0 \\[2mm] \underline{u}_0 & \text{if} \quad -A_0 + \dfrac{\partial V}{\partial x}(x, D) < 0 \end{cases} \tag{12}
$$

3.2 $D = \mathbf{L}$

(12) says that if $d = \mu_{\mathrm{L}}$, $u_0 = \underline{u}_0$ if $-A_0 + \partial V / \partial x < 0$. Since V is convex with a finite minimum, that will occur for all x less than some value which we call $Z_0(\mathrm{L})$. Conversely, $u_0 = 0$ for $x > Z_0(\mathrm{L})$.

When $x = Z_0(\mathrm{L})$ and $D = \mathrm{L}$, from (3),

$$
\frac{dx}{dt} = u_0 - \mu_{\mathrm{L}}(1 - B(Z_0(\mathrm{L}))). \tag{13}
$$

Recall that we have assumed that $\underline{u}_0 > \mu_{\mathrm{L}}$. Therefore the right side of (13) could be positive (as well as negative). If $u_0 > \mu_{\mathrm{L}}(1 - B(Z_0(\mathrm{L})))$, then (13) implies that x will increase and therefore we will soon have $x > Z_0(\mathrm{L})$. At that point, u_0 will immediately become 0 and x will decrease. After a very short time, $x < Z_0(\mathrm{L})$ so $u_0 = \underline{u}_0$ and x will increase again. Consequently, u_0 will jump infinitely rapidly (in an idealized mathematical model) between 0 and $\underline{u}_0$ and x will remain

14

very close to $Z_0(\mathrm{L})$. (Similar behavior results if we start with $u_0 < \mu_\mathrm{L}(1 - B(Z_0(\mathrm{L})))$.) To avoid this undesirable and unnecessary *chattering*, we choose $u_0 = \mu_\mathrm{L}(1 - B(Z_0(\mathrm{L}))$ when $x = Z_0(\mathrm{L})$. Note that since $dx/dt = 0$, x remains at $Z_0(\mathrm{L})$ until d changes to μ_H. This is the *hedging point* phenomenon described by Kimemia and Gershwin (1983), Bielecki and Kumar (1988), and many other papers.

Since x can remain at $Z_0(\mathrm{L})$ indefinitely with no subcontracting, it is not likely that $Z_0(\mathrm{L}) < 0$. When $x < 0$, sales are lost, and there is no offsetting benefit. We can therefore assume that $Z_0(\mathrm{L}) \geq 0$ and that $u_0(Z_0(\mathrm{L}), \mathrm{L}) = \mu_\mathrm{L}$.

To summarize, the trajectory behaves as follows when $D=\mathrm{L}$:

- If $x > Z_0(\mathrm{L})$, then $u_0 = 0$ and x decreases at rate $-\mu_\mathrm{L}$ until it reaches $Z_0(\mathrm{L})$. x remains at that value until D changes to H.

- If $x = Z_0(\mathrm{L})$, then $u_0 = d$ and x remains at that value until D changes to H.

- If $x < Z_0(\mathrm{L})$, then $u_0 = \underline{u}_0$ x increases at rate $\underline{u}_0$ until $x = Z_0(\mathrm{L})$. x remains at that value until D changes to H.

3.3 $D = \mathbf{H}$

The behavior when $d = \mu_\mathrm{H}$ is determined by the same considerations as when $d = \mu_\mathrm{L}$, but it is not the same. Equation (3) becomes, at $x = Z_0(\mathrm{H})$

$$\frac{dx}{dt} = u_0 - \mu_\mathrm{H}(1 - B(Z_0(\mathrm{H}))). \tag{14}$$

Since $\underline{u}_0 < \mu_\mathrm{H}$, the right side is guaranteed to be negative unless $Z_0(\mathrm{H}) < 0$. But $Z_0(\mathrm{H}) < 0$ cannot be optimal because if $x > Z_0(\mathrm{H})$, $u_0 = 0$. x will decrease at the maximum possible rate $(-\mu_\mathrm{H}(1 - B(Z_0(\mathrm{H}))))$, and there will be no revenues. While this may be beneficial if $x > 0$ because it reduces inventory cost, it cannot be beneficial if $0 > x > Z_0(\mathrm{H})$ because some customers (representing future revenues) are choosing not to order. As $Z_0(\mathrm{H})$ increases toward 0, fewer and fewer such future sales are lost.

Consequently, $Z_0(\mathrm{H})$ is not a hedging point (or other kind of temporary equilibrium), so if x is ever near this value, it must decrease. As x decreases, the right side of (3) increases (i.e., the rate of decrease of x diminishes), because $B(x)$ increases, until one of two events occurs. Either

1.

$$\underline{u}_0 \geq \mu_\mathrm{H}(1 - B(Z_0(\mathrm{H})))$$

Then we choose $u_0 = u_0^*$, where

$$u_0^* = \mu_\mathrm{H}(1 - B(Z_0(\mathrm{H}))); \tag{15}$$

or

2.

$$\underline{u}_0 < \mu_H (1 - B(Z_0(H)))$$

Then we choose $u_0 = \underline{u}_0$. The defection rate $B(x)$ increases enough so that the right side of (3) becomes equal to zero for some x^* which is greater than $Z_0(H)$. x^* is determined by

$$\underline{u}_0 = \mu_H (1 - B(x^*)) \tag{16}$$

In both cases, x has reached a lower limit, since dx/dt has reached 0. (In the latter case, x may approach the lower limit asymptotically, for suitable $B(x)$.) x remains at this level until the demand changes and $D = $ L. At that point, the behavior described in Section 3.2 resumes. It is convenient to define $\underline{X}$ as that lower bound. In Case 1, $\underline{X} = Z_0(H)$; in Case 2, $\underline{X} = x^*$.

To summarize, the trajectory behaves as follows when D=H:

- If $x > Z_0(H)$, then $u_0 = 0$ and x decreases.

- if $x = Z_0(H)$, then $u_0 = u_0^*$ and x remains constant.

- If $Z_0(H) > x$, then $u_0 = \underline{u}_0$, and

  - ⋆ if $x > x^*$, x decreases;
  - ⋆ if $x = x^*$, x remains constant;
  - ⋆ if $x < x^*$, x increases.

3.4   Characteristics of $Z_0(D)$

When $D = $ L, x increases until it reaches $Z_0(L)$, and it remains there until $D = $ H. After the change in demand, x decreases until it reaches $\underline{X}$. We can therefore conclude that, for a steady-state probability distribution to exist,

$$Z_0(L) \ge \underline{X}.$$

Since $\underline{X} \le x \le Z_0(L)$, the objective function cannot be minimal if $\underline{X} > 0$. In addition, since sales are lost when $x < 0$, the objective function cannot be minimal if $Z_0(L) < 0$. Therefore

$$Z_0(L) \ge 0 \ge \underline{X}.$$

16

4   Model with Subcontracting

> *Someday, and that day may never come, I will call upon you to do a service for me. But until that day, accept this ... gift ....* (Puzo 1969)

In this section, we extend the model of Section 2 by adding K additional sources of the product. Each source has a capacity — a maximum rate at which it can deliver. It also has a price that it will sell the goods at. We are not concerned here with that price; instead, we need to know the profit that the firm can earn by reselling such good. We assume that the profit from reselling is less than that from its own production facilities. In Section 5, we extend the analysis of Section 3 to cover this case.

4.1   Model

There are K subcontractors available. At time t, the plant produces finished goods at rate $u_0(t)$, and requests subcontractor i to supply materials at a rate of $u_i(t)$, $0 \leq u_i(t) \leq \underline{u}_i$. $u_i(t)$, $i = 0, 1, ..., K$ are the decision variables. The profit coefficient (dollars per unit) when subcontractor i is used is A_i. The subcontractors are indexed in decreasing A_i, i.e., $A_0 > A_1 > A_2 > A_3 > \ldots > A_K > 0$. That is, subcontractor $i + 1$ is less desirable to use than subcontractor i because it is more expensive, and therefore results in less profit. The subcontractors are perfectly reliable and they deliver instantaneously. We do not make any assumptions regarding the capacities $\underline{u}_i$ of the subcontractors.

Now, when the surplus level is $x < 0$, the time until the next arriving customer order will be completed, i.e., the production lead time, is *no greater than* $-x/\underline{u}_0$. This is the time to clear the current backlog if no subcontractors are used. The lead-time dependent demand case can therefore be treated by using $B(x) = \tilde{B}(\underline{u}_0\tau)$ as the probability that a potential customer chooses not to order when the quoted guaranteed lead time is $\tau = -x/\underline{u}_0$. A sharper bound can be developed if we assume a specific production and backlog policy, for example that of Section 5.

Combining the backlog-dependent demand rate and subcontracting, the dynamics of x are given by

$$\frac{dx}{dt} = \sum_{i=0}^{K} u_i - d(1 - B(x)). \tag{17}$$

4.2   Production Control Problem

The decision variables are the rate at which the goods are produced at the plant at time t, $u_0(t)$, and the rates at which the subcontractors are requested to supply goods at time t, $u_i(t)$ $i = 1, 2, ..., K$. The profit function to be maximized is the difference between the profit generated through production and subcontracting and the inventory carrying costs. A linear inventory carrying cost function is assumed, i.e.,

$$g(x) = \begin{cases} g^+ x \text{ if } x \geq 0 \\ \\ 0 \text{ if } x < 0 \end{cases}$$

The production control problem is

$$V = \max_{u_0, u_1, \ldots, u_K} \Pi = E \int_0^T \left(\sum_{i=0}^K A_i u_i - g(x) \right) dt \tag{18}$$

subject to

$$\frac{dx}{dt} = \sum_{i=0}^K u_i - d(1 - B(x)) \tag{19}$$

$$0 \leq u_i \leq \underline{u}_i \quad i = 0, 1, \ldots, K. \tag{20}$$

$$d = \begin{cases} \mu_{\mathrm{H}} & \text{if } D = \mathrm{H} \\ \mu_{\mathrm{L}} & \text{if } D = \mathrm{L} \end{cases} \tag{21}$$

Markov dynamics for D with rates λ_{HL} from H to L and λ_{LH} from L to H $\quad$ (22)

We do not include an explicit cost for backlog in the definition of g. Instead, we are penalized for backlog by the defection of impatient customers and the reduced profits due to the use of subcontractors.

We do not need to assume that the demand is feasible. If Ed is greater than u_0, some subcontracting will certainly be used. If Ed is greater than the total capacity $\sum_{i=0}^K \underline{u}_i$ then enough impatient customers will defect to guarantee that x is bounded from below.

We assume that T is very large, so the optimal policy discussion in Section 3 is based on the assumption that the probability distribution of (x, D) reaches a steady state.

5 Characterization of the Policy with Subcontractors

Problem (18)-(22) is a dynamic programming problem. The solution (the optimal rates of production as a function of x, D, and t) satisfies the Bellman equation. In this section, we use the Bellman equation to determine the structure of the solution.

5.1 Bellman equation

The value function is defined as:

$$V(x(t), D(t), t) = E \int_t^T \left(-\sum_{i=0}^K A_i u_i + g(x) \right) d\tau \tag{23}$$

V satisfies the *maximum principle*, which asserts that

$$-\frac{\partial V}{\partial t}(x, \mathrm{L}, t) = \min_{u_0, u_1, \ldots, u_K} \left\{ -\sum_{i=0}^K A_i u_i + g(x) + \frac{\partial V}{\partial x}(x, \mathrm{L}, t) \left(\sum_{i=0}^K u_i - \mu_{\mathrm{L}}(1 - B(x)) \right) \right.$$
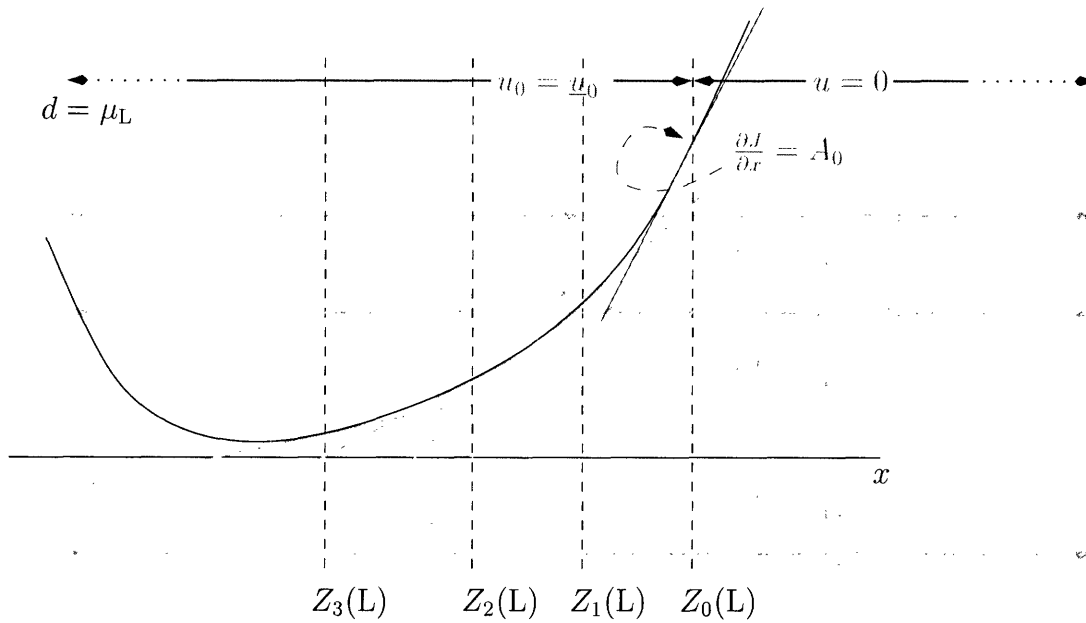
Figure 3: V vs. x, $d = \mu_L$

$$+V(x, \mathrm{H}, t)\lambda_{\mathrm{HL}} - V(x, \mathrm{L,t})\lambda_{\mathrm{HL}} \Big\} \qquad (24)$$

for $D =$L, and

$$-\frac{\partial V}{\partial t}(x, \mathrm{H}, t) = \min_{u_0, u_1, \ldots, u_K} \left\{ -\sum_{i=0}^{K} A_i u_i + g(x) + \frac{\partial V}{\partial x}(x, \mathrm{H}, t) \left(\sum_{i=0}^{K} u_i - \mu_{\mathrm{H}}(1 - B(x)) \right) \right.$$

$$\left. +V(x, \mathrm{L}, t)\lambda_{\mathrm{LH}} - V(x, \mathrm{H}, t)\lambda_{\mathrm{LH}} \right\} \qquad (25)$$

for $D =$H. The minimizations are taken over constraints (20).

It is reasonable to assume, since V is the solution of a dynamic programming problem, that V is strictly convex in x. Therefore it has a unique minimum. If that minimum were not finite, x would be increasing or decreasing without bound. If x were increasing without bound, (23) implies that V is infinite, and this cannot be optimal. It is not possible for x to decrease without bound because $B(x)$ increases with decreasing x, and there is a value of x below which dx/dt (17) must be zero or positive. Therefore V has a finite minimum. Figure 3 shows V as a function of x when $d = \mu_L$ and we assume that V is continuously differentiable. The graph of V when $d = \mu_H$ is similar.

¿From (24) and (25), for both values of D,

19

$$u_i^*(x, D) = \begin{cases} 0 & \text{if} \quad -A_i + \dfrac{\partial V}{\partial x}(x, D) > 0 \\[2ex] \text{undetermined} & \text{if} \quad -A_i + \dfrac{\partial V}{\partial x}(x, D) = 0 \\[2ex] \underline{u}_i & \text{if} \quad -A_i + \dfrac{\partial V}{\partial x}(x, D) < 0 \end{cases} \tag{26}$$

for $i = 0, 1, \ldots, K$.

5.2 $D = \mathrm{L}$

(26) says that if $d = \mu_\mathrm{L}$, $u_0 = \underline{u}_0$ if $-A_0 + \partial V / \partial x < 0$. Since V is convex with a finite minimum, that will occur for all x less than some value which we call $Z_0(\mathrm{L})$. Conversely, $u_0 = 0$ for $x > Z_0(\mathrm{L})$.

Similarly, for each i, (26) implies that there is a value of x, called $Z_i(\mathrm{L})$, such that $-A_i + \partial V / \partial x < 0$ for $x < Z_i(\mathrm{L})$ and $-A_i + \partial V / \partial x > 0$ for $x > Z_i(\mathrm{L})$. Consequently $u_i = \underline{u}_i$ for $x < Z_i(\mathrm{L})$ and $u_i = 0$ for $x > Z_i(\mathrm{L})$. Since $A_0 > A_1 > A_2 \ldots > A_K$ and since V is strictly convex, $Z_0(L) > Z_1(\mathrm{L}) > Z_2(\mathrm{L}) > \ldots > Z_K(\mathrm{L})$.

Consequently,

$$\left. \begin{array}{l} \bullet \text{ When } x > Z_0(\mathrm{L}), \qquad u_i = 0, i = 0, 1, \ldots, K. \\[2ex] \bullet \text{ When } Z_1(\mathrm{L}) < x < Z_0(\mathrm{L}), \quad u_0 = \underline{u}_0 \text{ and } u_i = 0, i = 1, \ldots, K. \\[2ex] \bullet \text{ When } Z_2(\mathrm{L}) < x < Z_1(\mathrm{L}), \quad u_0 = \underline{u}_0, u_1 = \underline{u}_1, \text{ and } u_i = 0, i = 2, \ldots, K. \\[2ex] \bullet \text{ When } Z_3(\mathrm{L}) < x < Z_2(\mathrm{L}), \quad u_0 = \underline{u}_0, u_1 = \underline{u}_1, u_2 = \underline{u}_2, \text{ and } u_i = 0, i = 3, \ldots, K. \\[2ex] \bullet \text{ etc.} \end{array} \right\} \tag{27}$$

When $x = Z_0(\mathrm{L})$ and $D = \mathrm{L}$, we already know that $u_1 = u_2 = \ldots = u_K = 0$. Therefore, from (17),

$$\frac{dx}{dt} = u_0 - \mu_\mathrm{L}(1 - B(Z_0(\mathrm{L}))). \tag{28}$$

Recall that we have assumed that $\underline{u}_0 > \mu_\mathrm{L}$. Therefore the right side of (28) could be positive (as well as negative). If $u_0 > \mu_\mathrm{L}(1 - B(Z_0(\mathrm{L})))$, then (28) implies that x will increase and therefore we will soon have $x > Z_0(\mathrm{L})$. At that point, u_0 will immediately become 0 and x will decrease. After a very short time, $x < Z_0(\mathrm{L})$ so $u_0 = \underline{u}_0$ and x will increase again. Consequently, u_0 will jump infinitely rapidly (in an idealized mathematical model) between 0 and $\underline{u}_0$ and x will remain very close to $Z_0(\mathrm{L})$. (Similar behavior results if we let $u_0 < \mu_\mathrm{L}(1 - B(Z_0(\mathrm{L})))$.) To avoid this undesirable

and unnecessary *chattering*, we choose $u_0 = \mu_L(1 - B(Z_0(L))$ when $x = Z_0(L)$. Note also that since $dx/dt = 0$, x remains at $Z_0(L)$ until d changes to μ_H. $Z_0(L)$ is therefore a hedging point.

Since x can remain at $Z_0(L)$ indefinitely with no subcontracting, it is not likely that $Z_0(L) < 0$. When $x < 0$, sales are lost, and there is no offsetting benefit. We can therefore assume that $Z_0(L) \geq 0$ and that $u_0(Z_0(L), L) = \mu_L$.

No such issue occurs for any other $Z_i(L)$. This is because the dynamics in the vicinity of $x = Z_i(L)$ are

$$\frac{dx}{dt} = \begin{cases} \sum_{j=0}^{i-1} \underline{u}_j - \mu_L(1 - B(x)) \text{ for } x > Z_i(L) \\ \\ \sum_{j=0}^{i} \underline{u}_j - \mu_L(1 - B(x)) \text{ for } x < Z_i(L) \end{cases} \tag{29}$$

Since $\underline{u}_0 > \mu_L$ and $\underline{u}_j > 0$, the right side of (29) is always positive, regardless whether $x < Z_i(L)$ or $x > Z_i(L)$. Consequently, there is no possibility of chattering at $x = Z_i(L), i > 0$.

To summarize, the trajectory behaves as follows when $D=L$:

- If $x > Z_0(L)$, all rates are 0 and x decreases at rate $-\mu_L$ until it reaches $Z_0(L)$. x remains at that value until D changes to H.

- If $x = Z_0(L)$, $u_0 = \mu_L$ and x remains constant until D changes to H.

- If $Z_i(L) < x < Z_{i-1}(L)$, $u_j = \underline{u}_j$ for $j = 0, ..., i-1$. x increases at rate $\underline{u}_0 + \underline{u}_1 + \underline{u}_2 + ... + \underline{u}_{i-1}$ until $x = Z_{i-1}(L)$. At that point, the last subcontractor is dropped; i.e., x increases at rate $\underline{u}_0 + \underline{u}_1 + \underline{u}_2 + ... + \underline{u}_{i-2}$. This continues until $x = Z_0(L)$, and x remains at that value until D changes to H.

5.3 $D = \mathbf{H}$

The behavior when $d = \mu_H$ is determined by the same considerations as when $d = \mu_L$, but it is not the same. Equation (17) becomes, at $x = Z_0(H)$,

$$\frac{dx}{dt} = u_0 - \mu_H(1 - B(Z_0(H))). \tag{30}$$

Since $\underline{u}_0 < \mu_H$, the right side is guaranteed to be negative unless $Z_0(H) < 0$. But $Z_0(H) < 0$ cannot be optimal because if $x > Z_0(H)$, $u_0 = u_1 = u_2 = ... = u_K = 0$. x will decrease at the maximum possible rate $(-\mu_H(1 - B(Z_0(H))))$, and there will be no revenues. While this may be beneficial if $x > 0$ because it reduces inventory cost, it cannot be beneficial if $0 > x > Z_0(H)$ because some customers (who would bring future revenues) choose not to order. As $Z_0(H)$ increases toward 0, fewer and fewer such future sales are lost. Therefore $Z_0(H) \geq 0$.

Consequently, since the right side of (30) is always negative, $Z_0(H)$ is not a hedging point (or any other kind of temporary equilibrium), so if x is ever near this value, it must decrease. As x

decreases past $Z_1(H)$, $Z_2(H)$, etc., the right side of (17) increases (i.e., the rate of decrease of x diminishes), because $B(x)$ increases, and because more and more subcontractors are used, until one of two events occurs. Either

1. Enough (i) subcontractors are engaged to satisfy

$$\sum_{j=0}^{i} \underline{u}_j > \mu_H(1 - B(Z_i(H)))$$

Then we choose $u_0 = \underline{u}_0$, $u_j = \underline{u}_j$ (for $j < i$), and $u_i = u_i^*$, where

$$\sum_{j=0}^{i-1} \underline{u}_j + u_i^* = \mu_H(1 - B(Z_i(H))); \tag{31}$$

and x remains constant, equal to $Z_i(H)$;

or

2. The rate of order loss $B(x)$ increases enough so that the right side of (17) becomes equal to zero for some x^* which is not equal to any $Z_j(H)$. That is,

$$\sum_{j=0}^{i} \underline{u}_j = \mu_H(1 - B(x^*)) \tag{32}$$

In both cases, x has reached a lower limit, since dx/dt has reached 0. (In the latter case, x may approach the lower limit asymptotically, for suitable $B(x)$.) x remains at this level until the demand changes and $D = L$. At that point, the behavior described in Section 5.2 resumes. It is convenient to define $\underline{X}$ as that lower bound and to define K' as the number of subcontractors used, i.e, the value of i that satisfies either (31) or (32). In Case 1, $\underline{X} = Z_{K'}(H)$; in Case 2, $\underline{X} = x^*$. Note that

$$K' = \max\{i \,|\max\{Z_i(H), Z_i(L)\} \geq \underline{X}\}$$

To summarize, the trajectory behaves as follows when $D=H$:

- If $x > Z_0(H)$, then $u_0 = \underline{u}_0$ and $u_j = 0$ for $j = 1, ..., K$. x decreases at rate $\underline{u}_0 - \mu_H$.

- If $Z_{i-1}(H) > x > Z_i(H)$ and $i < K'$, for $j = 1, ..., i \leq K$, then $u_0 = \underline{u}_0$, $u_j = \underline{u}_j$ for $j = 1, ..., i-1$, and $u_j = 0$ for $j \geq i$. x decreases at rate $\underline{u}_0 + \underline{u}_1 + ... + \underline{u}_{i-1} - \mu_H$.

- If $Z_{K'-1}(H) > x \geq Z_{K'}(H)$ and x is constant, then $u_0 = \underline{u}_0$, $u_j = \underline{u}_j$ for $j = 1, ..., K' - 1$,

  ⋆ if $x = Z_{K'}(H)$,

$$u_{K'}^* = \mu_H(1 - B(Z_{K'}(H))) - \sum_{j=0}^{K'-1} \underline{u}_j$$

  ⋆ if $x = x^*$, then $u_{K'} = \underline{u}_{K'}$

  and $u_j = 0$ for $j = K' + 1, ..., K$.

22

5.4   Characteristics of $Z_j(D)$

So far, we know that

$$Z_0(\text{L}) > Z_1(\text{L}) > Z_2(\text{L}) > ... > Z_K(\text{L}),$$

and, for the same reasons

$$Z_0(\text{H}) > Z_1(\text{H}) > Z_2(\text{H}) > ... > Z_K(\text{H}).$$

When $D = \text{L}$, x increases until it reaches $Z_0(\text{L})$, and it remains there until $D = \text{H}$. After the change in demand, x decreases until it reaches $\underline{X}$. We can therefore conclude that, for a steady-state probability distribution to exist,

$$Z_0(\text{L}) \geq \underline{X}.$$

Since $\underline{X} \leq x \leq Z_0(\text{L})$, the objective function cannot be minimal if $\underline{X} > 0$. In addition, since sales are lost when $x < 0$, the objective function cannot be minimal if $Z_0(\text{L}) < 0$. Therefore

$$Z_0(\text{L}) \geq 0 \geq \underline{X}.$$

6   Model with an Unreliable Manufacturing Facility and Constant Demand

In this section, we present a model with an unreliable manufacturing facility and constant demand. We describe the structure of the solution for this model by examining the equivalence of this model to the model with reliable manufacturing facility and uncertain demand.

6.1   Model

Consider a system where the manufacturing plant is unreliable, the subcontractors are perfectly reliable, and the demand is constant with rate $d(t) = d$. The state of the manufacturing facility at time t is $\alpha(t)$ which is either up (U) or down (D). The failure and repair times of the manufacturing plant are assumed to be exponential random variables with rates p and r respectively.

We assume that the production capacity is sufficient to meet the demand when it is up, i.e., $\underline{u}_0 > d$. All the other assumptions regarding the profits, costs, subcontracting, and the backlog-dependent demand are identical to those given in Section 4.

6.2   Production Control Problem

With these changes, the production control problem for the unreliable plant-constant demand case is

$$V = \max_{u_0, u_1, \ldots, u_K} \Pi' = E \int_0^T \left(\sum_{i=0}^K A_i u_i - g(x) \right) dt \tag{33}$$

$$\frac{dx}{dt} = \sum_{i=0}^K u_i - d'(1 - B(x)) \tag{34}$$

$$0 \le u_0 \le \alpha \underline{u}_0 \tag{35}$$

$$0 \le u_i \le \underline{u}_i \quad i = 1, \ldots, K. \tag{36}$$

$$\alpha = \begin{cases} 1 & \text{if } \alpha = \text{U} \\ 0 & \text{if } \alpha = \text{D} \end{cases} \tag{37}$$

Markov dynamics for α with rates p from U to D and r from D to U \hfill (38)

6.3 Characterization of the Policy

Following the same steps in Section 5 shows that the structure of the optimal policy for this system is identical to the case analyzed in the preceding sections.

Furthermore, setting $\underline{u}_0 = \mu_H - \mu_L$, $d' = \mu_H - \underline{u}_0$, $r = \lambda_{LH}$, and $p = \lambda_{HL}$ yields the same rates of increase and decrease in each region as those generated by the constant production–uncertain demand case. As a result, the optimal parameters for the control policy for the uncertain production–constant demand case can be determined from the solution of the reliable production–uncertain demand case.

Although the sample paths of these two cases are identical after these transformations, the effects on the customers are quite different. Even though a maximum waiting time can be guaranteed for the reliable production–uncertain demand case, an upper bound for the waiting time cannot be set if the manufacturing facility is unreliable. Since a manufacturing plant can be down for an extremely long period of time (with a very small probability), only an estimate of the waiting time can be provided. Modeling the behavior of the customer when an estimate of the waiting time is given is more challenging, since the customer may renege as well as defect.

7 Analysis of the Model

In this section, we analyze the model with subcontracting and volatile demand of Section 4. The solution of this model yields the solution for the model with a single producer of Section 2 as a special case, and also the solution for the model with an unreliable plant and uncertain demand presented in Section 6, after a transformation.

In this section, we calculate the steady-state probability distribution of x and D assuming that the system is operated under the policy of Section 5. In Section 8, we evaluate the expected profit
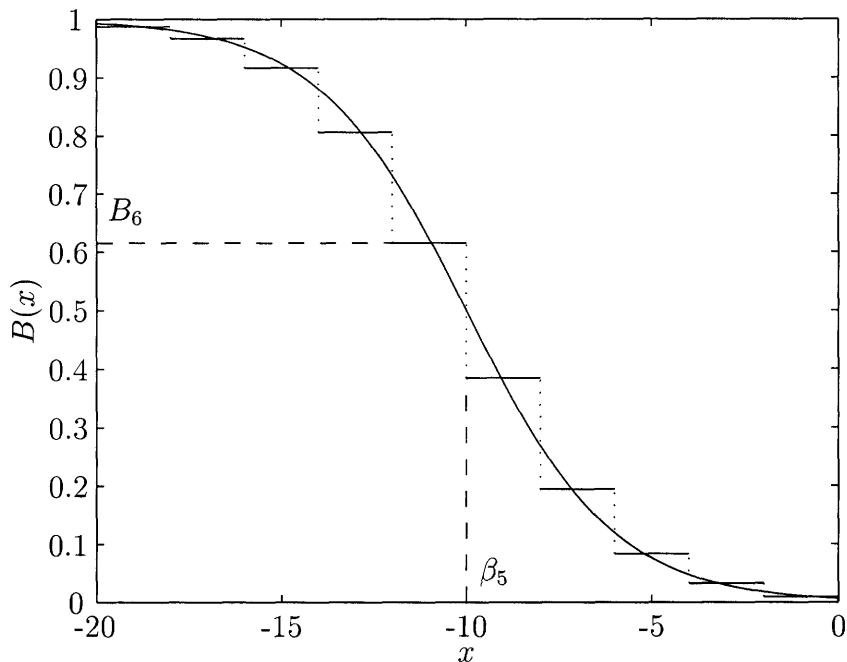
Figure 4: Step-wise constant approximation of a continuous $B(x)$ function

(as well as other performance measures). Then we find the optimal policy by finding the values of $Z_0(H)$, ..., $Z_K(H)$, $Z_0(L)$, ..., $Z_K(L)$ that maximize the expected profit.

7.1 Dynamics

The analysis of even simple systems with general non-zero $B(x)$ results in non-closed form solutions. In order to treat a wide variety of backlog-dependent demand functions, it is convenient to assume that $B(x)$ is piecewise constant. That is,

$$B(x) = \begin{cases} 0 & x \geq 0 \\ B_1 & 0 > x \geq \beta_1 \\ B_i & \beta_{i-1} > x \geq \beta_i \quad i = 2, ..., M \end{cases} \tag{39}$$

where β_i, B_i, $i = 1, ..., M$ are constants. From (1) and (2),

$$0 < B_i < B_{i+1} \leq 1,$$

$$0 > \beta_i > \beta_{i+1}.$$

By a proper choice of these constants, and for large enough M, any monotonic $B(x)$ can be arbitrarily closely approximated. Figure 4 shows a step-wise constant approximation of a continuous $B(x)$ function.

25

It is also convenient to assume that the β_i and B_i are chosen so that (32) can be satisfied exactly. That is, for each i such that $Z_i(\text{H}) < 0$, there is an $I(i)$ and a $B_{I(i)}$ such that

$$\sum_{j=0}^{i} \underline{u}_j = \mu_{\text{H}}(1 - B_{I(i)})$$

When subcontracting and backlog-dependent demand are combined, the x axis is divided into at most $2K + M + 4$ regions. Within each of these regions, the right side of the x dynamics (17) is constant and $g(x)$ is linear. Let

$$R = \{Z_i(\text{L}) | i = 0, 1, ..., K\} \cup \{Z_i(\text{H}) | i = 0, 1, ..., K\} \cup \{0\} \cup \{\beta_i | i = 1, ..., M\} \cup \{\underline{X}\}$$

and let $\|R\| + 1$ be the number of unique elements in R. Assume R is indexed in decreasing order, i.e., $R_i > R_{i+1}$, $i = 0, 1, 2, ..., \|R\| - 1 \leq 2K + M + 3$. Since $Z_0(\text{L}) > 0$ and $\beta_i < 0$ for all i,

- $R_0 = Z_0(\text{L}) > 0$,

- $R_{\|R\|} < 0$.

A sample path of a system with three subcontractors ($K = 3$) and complete backlogging ($B(x) = 0$) is depicted in Figure 5. We assume that $\underline{u}_0 + \underline{u}_1 + \underline{u}_2 < \mu_{\text{H}}$, and $\underline{u}_0 + \underline{u}_1 + \underline{u}_2 > \mu_{\text{L}}$. That is, the low demand μ_{L} can be met with the capacity of the factory and the two least expensive subcontractors, but the high demand μ_{H} cannot. When demand is high ($d = \mu_{\text{H}}$), $x(t)$ decreases and the slope, which starts steeply negative, increases to zero at $x = Z_3(\text{H})$. When demand is low ($d = \mu_{\text{L}}$), x increases and the slope, which starts steeply positive, decreases to zero at $x = Z_0(\text{L})$.

Let Δ_i^{L} be the rate of change of x in region i when the demand state D is low (L). Then

$$\Delta_i^{\text{L}} = \sum_{j=0}^{K} u_j - \mu_{\text{L}}(1 - B(x)), \ R_i < x < R_{i+1}, \ i = 0, 1, 2, ... \|R\| \tag{40}$$

where u_j, $j = 0, 1, ..., K$ are given by (27). Recall that u_j and $B(x)$ are all constant in each region, so Δ_i^{L} is constant.

Similarly, let Δ_i^{H} be the rate of change of x in region i when $D = \text{H}$. Then

$$\Delta_i^{\text{H}} = \sum_{j=0}^{K} u_j - \mu_{\text{H}}(1 - B(x)), \ R_i < x < R_{i+1}, \ i = 0, 1, 2, ... \|R\| \tag{41}$$

where u_j, $j = 0, 1, ..., K$ are described in Section 5.3.

As x decreases, more subcontractor capacity is utilized. Furthermore, some of the potential customers choose not to order and this decreases the demand rate. As a result, when $D = \text{H}$, x decreases in region $i < J$ and increases in region $i \geq J$, where J is uniquely defined as

$$J = \min\left\{j \middle| \Delta_j^{\text{H}} \geq 0\right\} \tag{42}$$

and the lower level is located at $R_J = \underline{X}$.
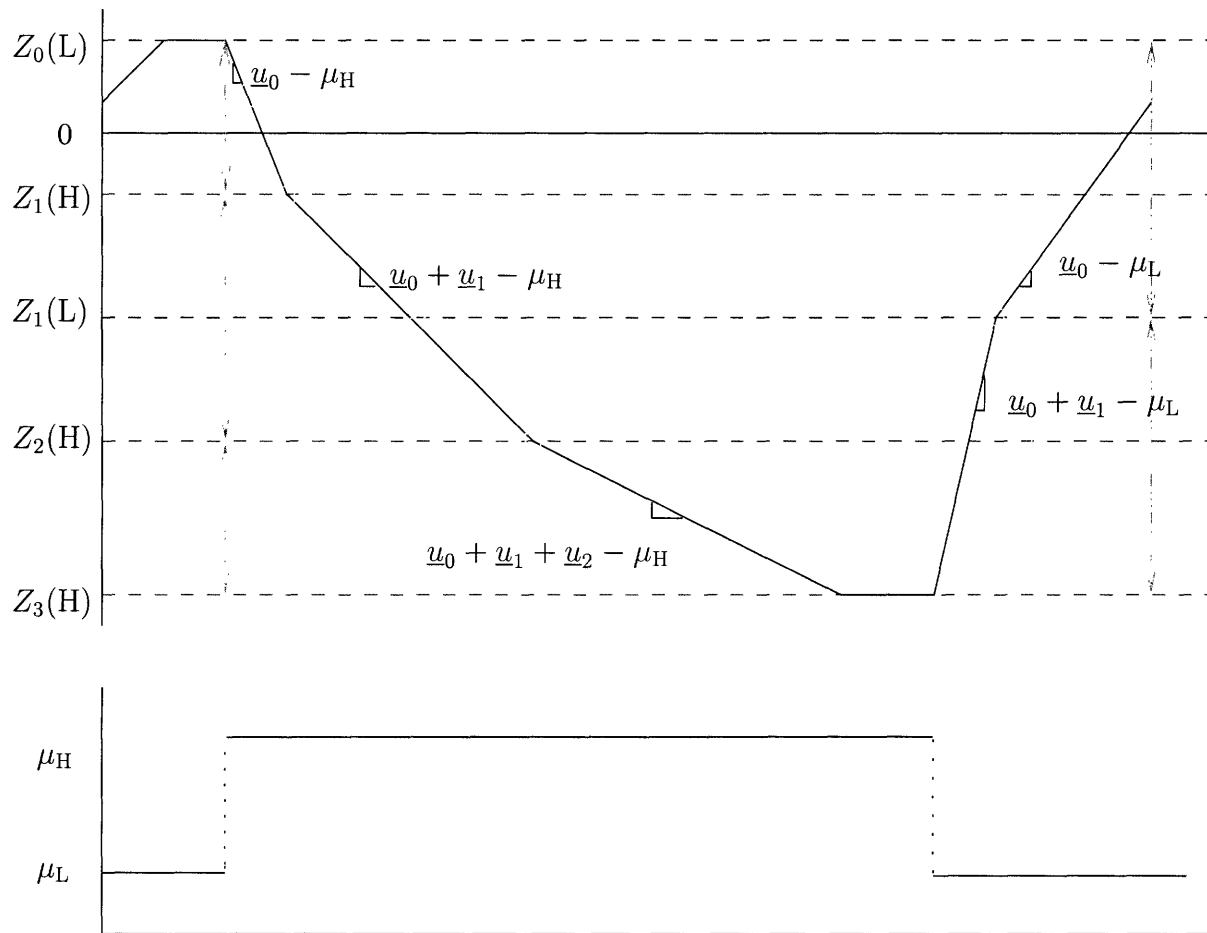
Figure 5: Sample paths for a system with three subcontractors and complete backlogging ($B(x) = 0$)
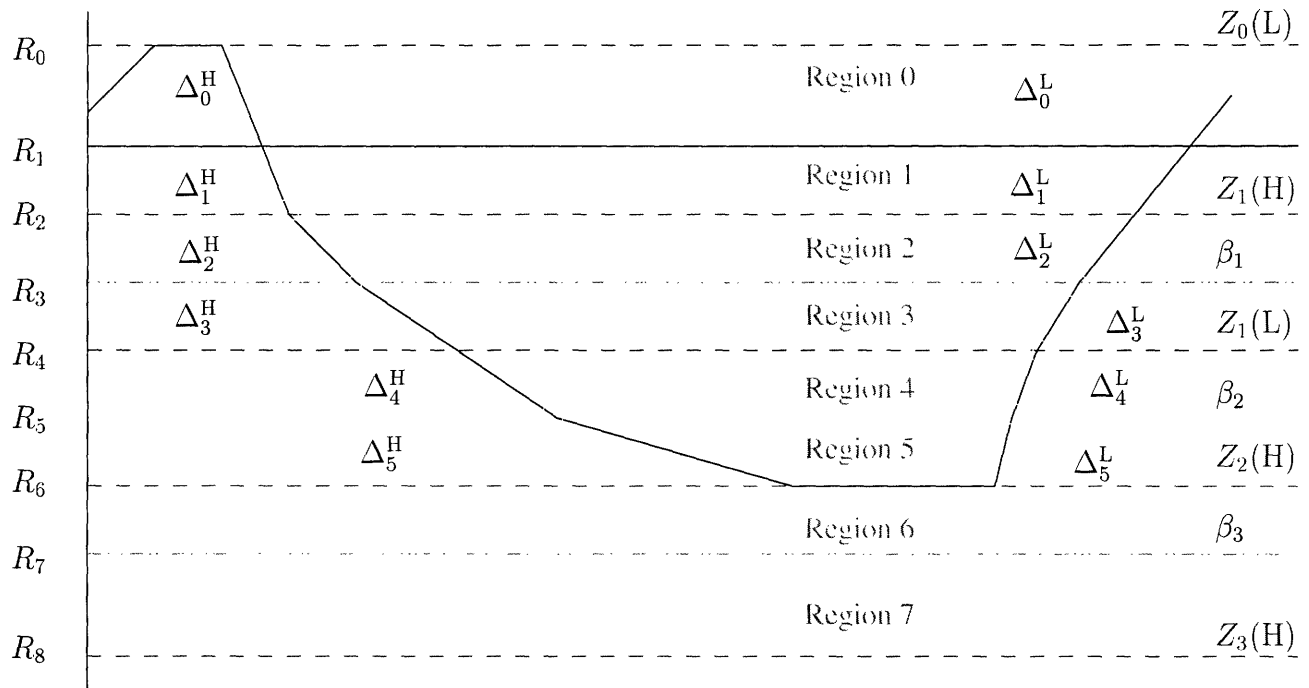
Figure 6: Sample paths for a system with three subcontractors and backorder-dependent demand

Figure 6 depicts the sample path of the system shown in Figure 5 with backorder-dependent demand described by β_i, B_i, $i = 1, 2, 3$. In this case the hedging levels are located at R_0 and R_6 and the regions 6 to 7 are transient. Furthermore, only one subcontractor is used.

Incorporating backlog-dependent demand into the model guarantees that x is always bounded between R_0 and R_J. Even if the production capacity (including the additional capacity obtained from the subcontractors) is not sufficient to meet the average demand Ed, the backlog-dependent demand of (17) yields a feasible equilibrium where a portion of the demand is matched with the available capacity and the remaining portion is lost. As a result, a steady state probability distribution always exists.

In the following sections, we describe how the optimal policy is determined. First, the system is evaluated by determining the probability density functions in the interior, and probability masses at the upper and lower levels for given values of $Z_0(L), Z_1(L), Z_2(L), \ldots, Z_{K'}(L)$ and $Z_0(H), Z_1(H), Z_2(H), \ldots, Z_{K'}(H)$. Then, the optimal values of these parameters are determined by maximizing the expected profit.

7.2 Probability distribution

When the surplus/backlog x is not equal to the upper or lower levels (R_0 or R_J), the system is said to be in the *interior*. The system state at time t is $S(t) = (x(t), D(t))$ where $R_J < x(t) < R_0$ and $D(t) \in \{H, L\}$.

The time-dependent system state probability distribution for the interior region, $F_D(t,x)$, is defined as

$$F_D(t,x) = \mathbf{prob}[D(t) = D, x(t) \le x], \quad t \ge 0, \ D \in \{\mathrm{H}, \mathrm{L}\}, \ R_J < x(t) < R_0 \tag{43}$$

The time-dependent system state density functions are defined as

$$f_D(t,x) = \frac{\partial F_D(t,x)}{\partial x} \quad t \ge 0, \ D \in \{\mathrm{H}, \mathrm{L}\}, \ R_J < x(t) < R_0 \tag{44}$$

We assume that the process is ergodic and, thus, the steady-state density functions exist. The steady-state density functions are defined as:

$$f_D(x) = \lim_{t \to \infty} f_D(t,x), \quad D \in \{\mathrm{H}, \mathrm{L}\}, \ R_J < x(t) < R_0. \tag{45}$$

It is possible to show ergodicity by observing that in the Markov process model, all of the states constitute a single communicating class. It is also possible to demonstrate aperiodicity.

7.3   Region i: $R_{i+1} < x < R_i$

Suppose $R_{i+1} < x(t + \delta t) < R_i$, $i = 1, 2, \ldots, J - 1$, and $D(t + \delta t) = \mathrm{H}$. Then, since we are modeling this system as a Markov process,

$$f_\mathrm{H}(t + \delta t, x) = f_\mathrm{H}(t, x - \Delta_i^\mathrm{H} \delta t)(1 - \lambda_{\mathrm{HL}} \delta t) + f_\mathrm{L}(t, x)(\lambda_{\mathrm{LH}} \delta t) + o(\delta t) \tag{46}$$

where $o(\delta t)$ approaches to zero faster than δt. This equation can be written in differential form, for $\delta t \to 0$, as

$$\frac{\partial f_\mathrm{H}(t,x)}{\partial t} + \Delta_i^\mathrm{H} \frac{\partial f_\mathrm{H}(t,x)}{\partial x} = -\lambda_{\mathrm{HL}} f_\mathrm{H}(t,x) + \lambda_{\mathrm{LH}} f_\mathrm{L}(t,x) \tag{47}$$

Taking the limit of (47) as $t \to \infty$ yields the following steady-state differential equation for $f_\mathrm{H}(x)$:

$$\Delta_i^\mathrm{H} \frac{df_\mathrm{H}(x)}{dx} = -\lambda_{\mathrm{HL}} f_\mathrm{H}(x) + \lambda_{\mathrm{LH}} f_\mathrm{L}(x), \quad R_{i+1} < x < R_i \tag{48}$$

Following the same steps for f_L yields

$$\Delta_i^\mathrm{L} \frac{df_\mathrm{L}(x)}{dx} = \lambda_{\mathrm{HL}} f_\mathrm{H}(x) - \lambda_{\mathrm{LH}} f_\mathrm{L}(x) \quad R_{i+1} < x < R_i \tag{49}$$

In order to solve the set of first order differential equations given in (48) and (49), two boundary conditions are needed. First, note that at any given level of the finished goods inventory, the number of upward crossings must be equal to the number of downward crossings. Let $N(D, \xi, T)$ denote the total number of level crossings in demand state D, at surplus level ξ, in the time interval $[t, t + T]$ for large T. Then

$$\lim_{T \to \infty} N(\mathrm{H}, \xi, T) = \lim_{T \to \infty} N(\mathrm{L}, \xi, T) \tag{50}$$

Renewal analysis shows that

$$\lim_{T \to \infty} \frac{N(D, \xi, T)}{T} = \Delta_i^D f_D(\xi) \tag{51}$$

where Δ_i^D is the rate of change in the buffer level when the demand state is D and ξ is in region i, and $f_D(\xi)$ is the steady-state density function. This kind of analysis was also employed by Yeralan and Tan (1997). Then, equation (50) can be written as

$$-\Delta_i^{\mathrm{H}} f_{\mathrm{H}}(x) = \Delta_i^{\mathrm{L}} f_{\mathrm{L}}(x). \tag{52}$$

Using this result in equation (48) gives the following first order differential equation

$$\frac{df_{\mathrm{H}}(x)}{dx} = \left(-\frac{\lambda_{\mathrm{HL}}}{\Delta_i^{\mathrm{H}}} - \frac{\lambda_{\mathrm{LH}}}{\Delta_i^{\mathrm{L}}} \right) f_{\mathrm{H}}(x) \tag{53}$$

whose solution is

$$f_{\mathrm{H}}(x) = c_i e^{\eta_i x}, \quad R_{i+1} < x < R_i \tag{54}$$

where

$$\eta_i = -\frac{\lambda_{\mathrm{HL}}}{\Delta_i^{\mathrm{H}}} - \frac{\lambda_{\mathrm{LH}}}{\Delta_i^{\mathrm{L}}}$$

and c_i is a constant to be determined. Following equation (52),

$$f_{\mathrm{L}}(x) = -c_i \frac{\Delta_i^{\mathrm{H}}}{\Delta_i^{\mathrm{L}}} e^{\eta_i x}, \quad R_{i+1} < x < R_i \tag{55}$$

7.4   External Boundary Conditions

The steady-state probabilities P^0 and P^J that the finished goods inventory is equal to the hedging level $Z_0(\mathrm{L})$ and the lowest level $\underline{X}$ are defined as

$$P^0 = \lim_{t \to \infty} \mathbf{prob}[x(t) = R_0], \tag{56}$$

$$P^J = \lim_{t \to \infty} \mathbf{prob}[x(t) = R_J]. \tag{57}$$

Now consider the probability P^0 that the finished goods inventory is equal to the hedging point $Z_0(\mathrm{L})$. Because $\mu_{\mathrm{L}} < \underline{u}_0 < \mu_{\mathrm{H}}$, and because we are considering an optimal policy in non-transient conditions, the inventory level can increase only when the demand is low. Therefore, if $x \to Z_0(\mathrm{L})$, $d = \mu_{\mathrm{L}}$. Each time the inventory level increases and reaches the level $R_0 = Z_0(\mathrm{L})$, it stays there until the state of the demand changes to high and the inventory level starts decreasing. As a result

of the memory-less property of the exponential distribution, the expected remaining time for the state of the demand to change from L to H is $1/\lambda_{\mathrm{HL}}$. P^0 is fraction of time that $x = Z_0(\mathrm{L})$:

$$P^0 = \lim_{T \to \infty} \frac{N(\mathrm{L}, R_0, T)}{T} \frac{1}{\lambda_{\mathrm{LH}}} = \Delta_0^{\mathrm{L}} f_{\mathrm{L}}(R_0) \frac{1}{\lambda_{\mathrm{LH}}} = -c_0 \frac{\Delta_0^{\mathrm{H}}}{\lambda_{\mathrm{LH}}} e^{\eta_0 R_0}. \tag{58}$$

Similarly

$$P^J = \lim_{T \to \infty} \frac{N(\mathrm{H}, R_J^+, T)}{T} \frac{1}{\lambda_{\mathrm{HL}}} = -\Delta_{J-1}^{\mathrm{H}} f_{\mathrm{H}}(R_J) \frac{1}{\lambda_{\mathrm{HL}}} = -c_{J-1} \frac{\Delta_{J-1}^{\mathrm{H}}}{\lambda_{\mathrm{HL}}} e^{\eta_{J-1} R_J}. \tag{59}$$

Let us also define P_i^{H} and P_i^{L} $i = 0, 1 ..., J - 1$ as the probabilities that the process is in region i in the long run when the demand is high and when it is low, respectively:

$$P_i^{\mathrm{H}} = \lim_{t \to \infty} \mathbf{prob}[R_i < x(t) < R_{i+1}, D(t) = H] \quad i = 0, ..., J - 1 \tag{60}$$

$$P_i^{\mathrm{L}} = \lim_{t \to \infty} \mathbf{prob}[R_i < x(t) < R_{i+1}, D(t) = L] \quad i = 0, ..., J - 1 \tag{61}$$

Once the density functions are available, P_i^{H} and P_i^{L} can be evaluated as

$$P_i^{\mathrm{H}} = \int_{R_{i+1}}^{R_i} f_{\mathrm{H}}(x) dx \quad i = 0, ..., J - 1 \tag{62}$$

$$P_i^{\mathrm{L}} = \int_{R_{i+1}}^{R_i} f_L(x) dx \quad i = 0, ..., J - 1 \tag{63}$$

7.5   Internal Boundary Conditions

To complete the derivation of the density functions, the coefficients c_i, $i = 0, 1, ..., J - 1$ must be determined. Since there are J unknowns, J boundary conditions are needed. The $J - 1$ internal boundary conditions come from the equality of the number of upward and downward crossings at the switching points. Let R_i^+ and R_i^- denote the points just above and just below the hedging level R_i respectively. Then for large T

$$\lim_{T \to \infty} N(j, R_i^+, T) = \lim_{T \to \infty} N(j, R_i^-, T), j \in \{\mathrm{H}, \mathrm{L}\}, i{=}1, 2, \cdots, J{-}1. \tag{64}$$

By using equation (51), this equation can be written

$$\Delta_{i-1}^j f_j(R_i^+) = \Delta_i^j f_j(R_i^-), j \in \{\mathrm{H}, \mathrm{L}\}, i{=}1, 2, \cdots, J{-}1. \tag{65}$$

8   Solution of the Model

8.1   Coefficients

Writing (65) in terms of the solution of the density function for $j = \text{H}$ given in equation (54) yields

$$\Delta_{i-1}^{\text{H}} c_{i-1} e^{\eta_{i-1} R_i} = \Delta_i^{\text{H}} c_i e^{\eta_i R_i}, \, i{=}1,2,\cdots, J{-}1, \tag{66}$$

or

$$c_i = \frac{\Delta_{i-1}^{\text{H}}}{\Delta_i^{\text{H}}} e^{(\eta_{i-1} - \eta_i) R_i} c_{i-1}, \, i = 1, 2, ..., J - 1 \tag{67}$$

Then all the constants $c_i \; i = 1, 2, ..., J - 1$ can be determined by c_0, since $c_i = \phi_i c_0$ where

$$\phi_i = \prod_{j=1}^{i-1} \frac{\Delta_{j-1}^{\text{H}}}{\Delta_j^{\text{H}}} e^{(\eta_{j-1} - \eta_j) R_j}, \, i = 1, ..., J - 1 \tag{68}$$

and $\phi_0 = 1$.

Finally, the constant c_0 is determined by using the normalizing condition. The sum of all the probabilities must add up to 1, or

$$P^0 + \sum_{i=0}^{J-1} \left(P_i^{\text{H}} + P_i^{\text{L}} \right) + P^J = 1 \tag{69}$$

Equations (54), (55), (58), (59), (62) and (63) yield

$$c_0 = \left[\frac{(\mu_{\text{H}} - u_0) e^{\eta_0 R_0}}{\lambda_{\text{LH}}} + \sum_{i=0}^{J-1} \phi_i \frac{(\Delta_i^{\text{L}} - \Delta_i^{\text{H}})}{\Delta_i^{\text{L}}} X_i - \frac{\phi_{J-1} \Delta_{J-1}^{\text{H}} e^{\eta_{J-1} R_J}}{\lambda_{\text{HL}}} \right]^{-1} \tag{70}$$

where

$$X_i = \begin{cases} (e^{\eta_i R_i} - e^{\eta_i R_{i+1}})/\eta_i & \text{if } \eta_i \neq 0, \\ \\ (R_i - R_{i+1}) & \text{if } \eta_i = 0. \end{cases}$$

8.2   Evaluation of the Objective Function

In order to determine the optimal values of the hedging levels, the profit must be evaluated. Let Π_i be the total profit rate generated through production at source i, $i = 0, 1, ..., K'$. The profit generated by the plant, Π_0 is determined by using the optimal production rate given in equation (26) as

$$\Pi_0 = \lim_{T \to \infty} E \left[\frac{1}{T} \int_0^T A_0 u_0 d\tau \right] = A_0 \left(\mathbf{prob}[x < R_0] \underline{u}_0 + \mathbf{prob}[x = R_0] \mu_L \right) \tag{71}$$

which can be simplified as

$$\Pi_0 = A_0 \left(1 - c_0 \frac{(\underline{u}_0 - \mu_{\mathrm{L}})(\mu_{\mathrm{H}} - \underline{u}_0)}{\lambda_{\mathrm{LH}}} e^{\eta_0 R_0} \right) \tag{72}$$

The fraction of time subcontractor i, $i = 1, 2, ..., K' - 1$ is used, denoted by Υ_i, can be written as

$$\Upsilon_i = \left[P^J I_{\{Z_i(\mathrm{H}) > R_J\}} + \sum_{j=1}^{J-1} \left(P_i^{\mathrm{H}} I_{\{Z_i(\mathrm{H}) > R_j\}} + P_i^{\mathrm{L}} I_{\{Z_i(\mathrm{L}) > R_j\}} \right) \right] \tag{73}$$

where $I_{\{Z_i(\mathrm{H}) \geq R_J\}} = 1$ if $Z_i(\mathrm{H}) \geq R_J$ and 0 otherwise. Note that the manufacturing plant is always used, i.e, $\Upsilon_0 = 1$.

Then, the total profit rate generated through production at subcontractor i, $i = 1, 2, ..., K' - 1$ is determined as

$$\Pi_i = A_i \underline{v}_i \Upsilon_i \tag{74}$$

The term for the profit obtained from the last subcontractor depends on whether the lower level is determined by the subcontractor hedging level or by the customer behavior. If it is determined by the subcontractor hedging level then the last subcontractor provides goods at a rate which is sufficient to keep x at this lower level as discussed in Section 5.3. Then, the profit obtained from subcontractor K' can be written as

$$\Pi'_K = A_{K'} \underline{v}_{K'} \left[P^J I_{\{Z_{K'}(\mathrm{H}) > R_J\}} + \sum_{j=1}^{J-1} \left(P_{K'}^{\mathrm{H}} I_{\{Z_{K'}(\mathrm{H}) > R_j\}} + P_{K'}^{\mathrm{L}} I_{\{Z_{K'}(\mathrm{L}) > R_j\}} \right) \right] + A_{K'} u_{K'}^* P^J I_{\{Z_{K'}(\mathrm{H}) = R_J\}} \tag{75}$$

The second term in equation (18) reflects the inventory carrying costs. Let Ψ_i be defined as

$$\Psi_i = \int_{R_i+1}^{R_i} x \left(f_H(x) + f_L(x) \right) dx = c_i \frac{\Delta_i^{\mathrm{L}} - \Delta_i^{\mathrm{H}}}{\Delta_i^{\mathrm{L}}} Q_i \tag{76}$$

where

$$Q_i = \begin{cases} \left((\eta_i R_i - 1) e^{\eta_i R_i} - (\eta_i R_{i+1} - 1) e^{\eta_i R_{i+1}} \right) / \eta_i^2 & \text{if } \eta_i \neq 0, \\[2ex] (R_i^2 - R_{i+1}^2)/2 & \text{if } \eta_i = 0. \end{cases}$$

Let i_0 be the index of the region boundary at $x = 0$:

$$i_0 = \{ j \,|\, R_j = 0 \}$$

Then the average inventory level E_{WIP} is

$$E_{\mathrm{WIP}} = \sum_{j=0}^{i_0-1} \Psi_j + R_0 P^0 \tag{77}$$

33

Finally, the average profit per unit time is

$$\Pi = \sum_{i=0}^{K'} \Pi_i - g^+ E_{\text{WIP}} \tag{78}$$

The optimal values of $Z_0(\text{L}), Z_1(\text{L}), Z_2(\text{L}), \ldots, Z_{K'}(\text{L})$ and $Z_0(\text{H}), Z_1(\text{H}), Z_2(\text{H}), \ldots, Z_{K'}(\text{H})$ are determined by maximizing Π.

8.3   Other Performance Measures

We can also evaluate other quantities of interest. The average *sales rate* or *throughput rate* is

$$TH = P^0 \mu_{\text{L}} + \sum_{i=0}^{K'} \underline{v}_i \left[P^J I_{\{Z_i(\text{H}) \geq R_J\}} + \sum_{j=1}^{J-1} \left(P_i^{\text{H}} I_{\{Z_i(\text{H}) > R_j\}} + P_i^{\text{L}} I_{\{Z_i(\text{L}) > R_j\}} \right) \right] + (u_{K'}^* - \underline{v}_{K'}) P^J I_{\{Z_{K'}(\text{H}) = R_J\}} \tag{79}$$

The *service level*, the ratio of the average sales to the average demand, is

$$S = TH/Ed \tag{80}$$

The *fill rate* is the probability that a customer receives his product as soon as he arrives:

$$FR = \mathbf{prob}[x > 0] = P^0 + \sum_{i=0}^{i_0 - 1} \left(P_i^{\text{H}} + P_i^{\text{L}} \right)$$

The *average backlog level* is E_{BL} which is evaluated as

$$E_{\text{BL}} = -\sum_{j=i_0}^{J-1} \Psi_j - R_J P^J \tag{81}$$

8.4   Waiting Time

An important performance measure of the system is W, the waiting time for a customer. In this section we derive expressions for the expected waiting time and also the minimum and maximum of the waiting time for a customer who arrives when the backlog is x.

8.4.1   Expected waiting time

In order to derive the expected waiting time, TH can be written as

$$TH = TH^+ \mathbf{prob}[x \geq 0] + TH^- \mathbf{prob}[x < 0] \tag{82}$$

where TH^+ and TH^- are the conditional throughput rates when $x \geq 0$ and $x < 0$ respectively. Since all the demand can be satisfied when $x \geq 0$, TH^+ is equal to Ed^+, the conditional average demand rate when $x \geq 0$, which can be evaluated as

$$Ed^+ = TH^+ = \frac{1}{FR}\left(\mu_L P^0 + \sum_{j=0}^{i_0-1}\left(\mu_H P_i^{\mathrm{H}} + \mu_L P_i^{\mathrm{L}}\right)\right) \tag{83}$$

Then the *expected waiting time* $E[W]$ is

$$E[W] = (1 - FR)\frac{E_{\mathrm{BL}}}{TH^-} \tag{84}$$

where TH^- is determined from equations (82) and (83).

8.4.2 Bounds on waiting time

In order to derive the bounds for the waiting time, the demand rate is set to its minimum and maximum and the resulting dynamics are solved deterministically. The *maximum waiting time* of a customer who arrives when there is a backlog of $\underline{x}$ is determined by the very conservative assumption that all new customers will arrive at rate μ_{L}. Let $y(t)$ be the amount of remaining work in front of a customer who has joined at $t = 0$ when the backlog is $y(0) = \underline{x}$. Then the following set of differential equations determine the maximum waiting time W_{MAX}:

$$\frac{dy}{dt} = u^*(x, \mathrm{L}) + \sum_{i=1}^{K} u_i^*(x, \mathrm{L}), \, y(0) = \underline{x} \tag{85}$$

$$\frac{dx}{dt} = u^*(x, \mathrm{H}) + \sum_{i=1}^{K} u_i^*(x, \mathrm{L}) - \mu_{\mathrm{L}}(1 - B(x)), \, x(0) = \underline{x} \tag{86}$$

$$W_{\mathrm{MAX}}(\underline{x}) = \min\left\{t|y(t) = 0\right\} \tag{87}$$

Similarly, the minimal waiting time is determined by the assumption that all new customers will arrive at rate μ_{H}. In this case, the *minimum waiting time* W_{MIN} is determined by the following equations:

$$\frac{dy}{dt} = u^*(x, \mathrm{H}) + \sum_{i=1}^{K} u_i^*(x, \mathrm{H}), y(0) = \underline{x} \tag{88}$$

$$\frac{dx}{dt} = u^*(x, \mathrm{H}) + \sum_{i=1}^{K} u_i^*(x, \mathrm{H}) - \mu_{\mathrm{H}}(1 - B(x)), x(0) = \underline{x} \tag{89}$$

$$W_{\mathrm{MIN}}(\underline{x}) = \min\left\{t|y(t) = 0\right\} \tag{90}$$

Figure 7 shows W_{MIN} and W_{MAX} for four different cases. In all the cases, the upper line depicts W_{MAX} and the lower one depicts W_{MIN}. In the first case (a), there is a single manufacturing facility and there are no subcontractors. In this case, the waiting time for a customer who arrives when the backlog is x is $x/\underline{u}_0$, and $W_{\mathrm{MIN}} = W_{\mathrm{MAX}} = x/\underline{u}_0$.
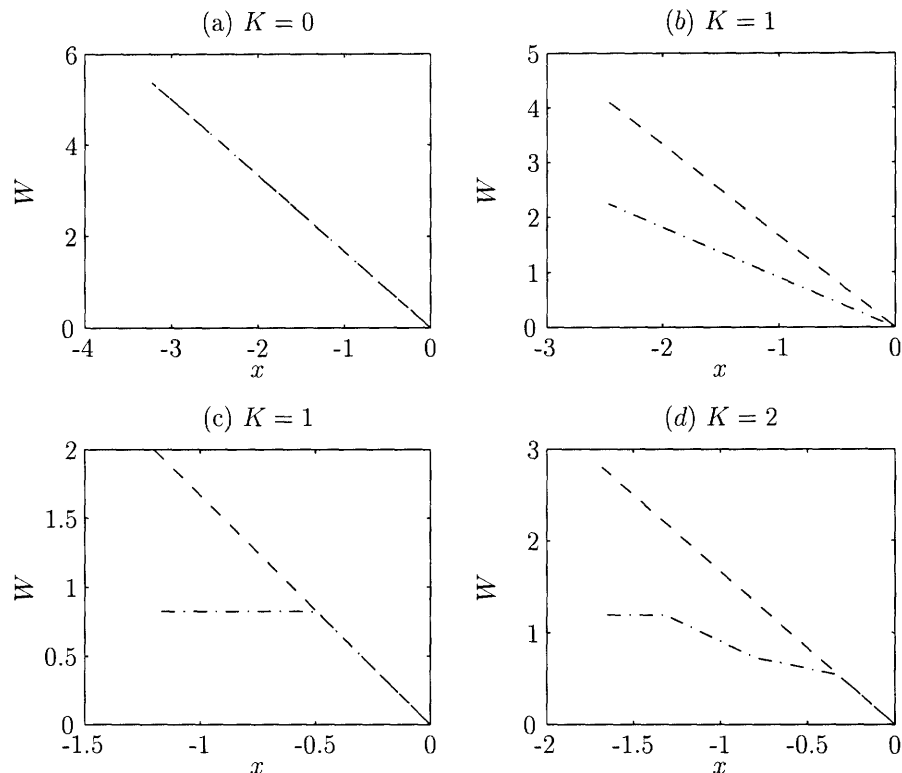
Figure 7: W_{MIN} and W_{MAX} for four different cases ($\mu_{\text{H}} = 1.5$, $\mu_{\text{L}} = 0.3$, $\lambda_{\text{HL}} = 0.05$, $\lambda_{\text{LH}} = 0.05$, $g^+ = 0.1$, $x_0 = -3$, $M = 10$, $\gamma = 0.5$, (a): $\underline{u}_0 = 0.6$, $A_0 = 3$, (b): $\underline{u}_0 = 0.6$, $\underline{u}_1 = 0.5$, $A_0 = 3$, $A_1 = 2$, (c): $\underline{u}_0 = 0.6$, $\underline{u}_1 = 1$, $A_0 = 3$, $A_1 = 1$, (d): $\underline{u}_0 = 0.6$, $\underline{u}_1 = 0.51$, $\underline{u}_2 = 1$, $A_0 = 10$, $A_1 = 6$, $A_2 = 5$.)

In the second case (b), there is one manufacturing facility and one subcontractor. The subcontractor is used when $\underline{X} \leq x < 0$ and $D = \text{H}$ and it is not used when $\underline{X} \leq x < 0$ and $D = \text{L}$. In this specific case, $W_{\text{MIN}} = x/(\underline{u}_0 + \underline{u}_1)$ and $W_{\text{MAX}} = x/\underline{u}_0$.

In the third case (c), there is also a manufacturing facility and a subcontractor. However, due to different system parameters, $Z_1(\text{H}) < 0$. In this case, when x is close to zero, $W_{\text{MIN}} = W_{\text{MAX}} = x/\underline{u}_0$. Note that W_{MIN} and W_{MAX} are not equal when x is sufficiently negative. The fourth case (d) depicts a similar picture for a system with one manufacturing plant and two subcontractors.

8.5 Performance Measures for the Subcontractors

In addition to the terms of the agreement with the manufacturing company, frequency and duration of the requested deliveries are important operational performance measures from a subcontractor's perspective.

Let Λ_i be the long run average frequency at which subcontractor i is asked to deliver. When x decreases and reaches $Z_i(\text{H})$, subcontractor i starts delivery. While $Z_i(\text{L}) < x < Z_i(\text{H})$, subcontractor i starts and stops delivery as the demand state switches from low to high and from high to low. Sub-
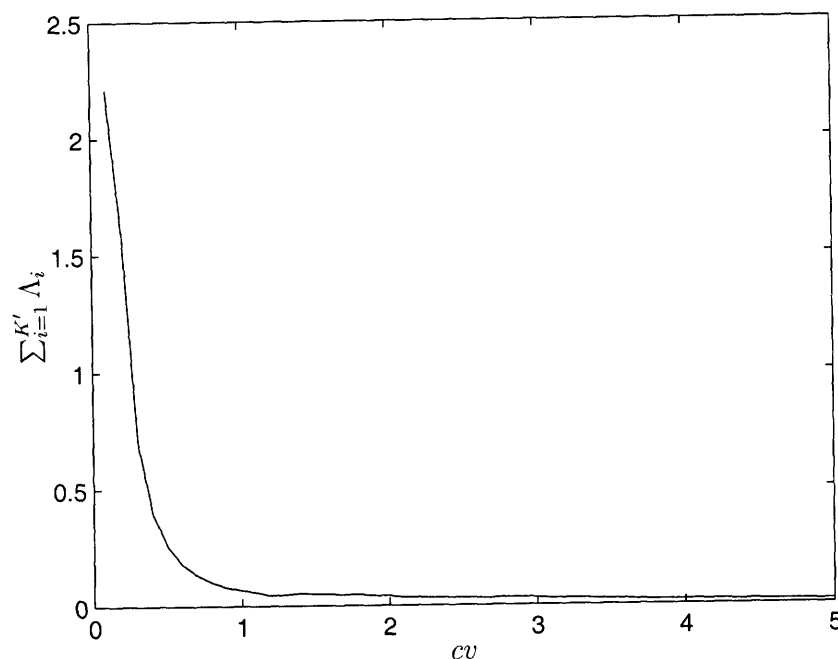
Figure 8: Effect of demand variability on the total number of managerial interventions ($\mu_H = 1.5$, $\mu_L = 0.3$, $Ed = 0.9$, $A_0 = 3$, $A_1 = 2$, $A_2 = 1$, $\underline{u}_0 = 0.5$, $\underline{u}_1 = 0.7$, $\underline{u}_2 = 1$, $g^+ = 0.1$, $x_0 = -6$, $M = 10$, $\gamma = 0.5$)

contractor i delivers all the time (regardless of the demand state) when $x < Z_i(L)$. It is reasonable to assume that $\mathbf{prob}[Z_i(L) < x < Z_i(H)]$ is much smaller than $\mathbf{prob}[x > Z_i(H)] + \mathbf{prob}[x < Z_i(L)]$ or equivalently $Z_i(H)$ is close to $Z_i(L)$. Section 9.6 justifies this approximation numerically. Under this assumption, the delivery frequency can be approximately determined by using equation (51) as

$$\Lambda_i = \lim_{T \to \infty} \frac{N(H, Z_i(H), T)}{T} = -\Delta_{j^*}^H f_H(Z_i(H)) \tag{91}$$

where $\Delta_{j^*}^H$ is the rate of decrease in the region just above $Z_i(H)$, i.e.,

$$j^* = \max \left\{ j \,|\, R_j > Z_i(H) \right\}$$

This approximation is exact if there is one subcontractor with $\underline{u}_1 \geq \mu_H - \underline{u}_0$, or $\underline{u}_1 < \mu_H - \underline{u}_0$ and $Z_1(L) < \underline{X}$. The total number of times all the subcontractors are asked to deliver per unit time, $\sum_{i=1}^{K'} \Lambda_i$ can be seen as a measure of *managerial intervention*. Figure 8 shows the effect of demand variability on the number of managerial interventions for a system with a manufacturing plant and two subcontractors. Note that as the demand variability increases, the length of a demand cycle also increases and therefore the number of managerial interventions decreases.

Let Γ_i be the expected duration of subcontract i's deliveries each time it starts a delivery. Note that the fraction of time subcontractor i is used can also be written as
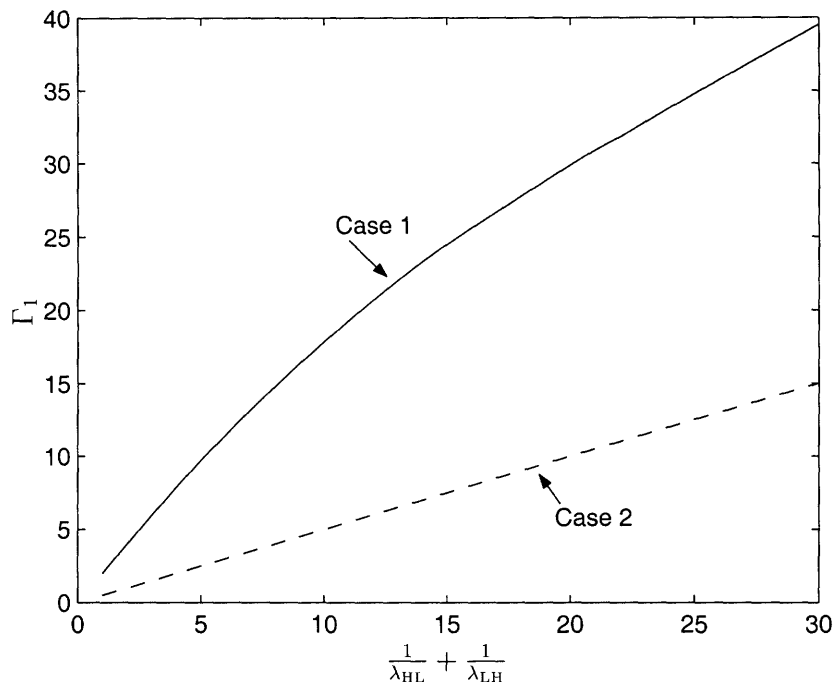
Figure 9: Demand cycle and duration of subcontractor deliveries ($\mu_H = 1.5$, $\mu_L = 0.3$, $Ed = 0.9$, $A_0 = 5$, $A_1 = 2$, $g^+ = 0.1$, $x_0 = -6$, $M = 10$, $\gamma = 0.5$, $\underline{u}_0 = 0.5$, Case 1: $\underline{u}_1 = 0.7$, Case 2: $\underline{u}_1 = 2$)

$$\Upsilon_i = \lim_{T \to \infty} \frac{N(H, Z_i(H), T)}{T} \Gamma_i = \Lambda_i \Gamma_i \qquad (92)$$

Since Υ_i is determined by equation (73), equations (91) and (92) yield the expected duration of subcontractor i's deliveries.

Figure 9 depicts Γ_1 for a system with one manufacturing plant and one subcontractor for different lengths of a demand cycle that has a duration of $1/\lambda_{HL} + 1/\lambda_{HL}$. As the figure shows, the duration of the subcontractor's deliveries increases with the duration of the demand cycle.

The frequency at which the firm starts and stops deliveries from subcontractors is comparable to the frequency at which important random events occur. In our primary model, in which demand changes, the frequency of starting and stopping subcontracting is close to the frequency with with demand changes from high to low and low to high in most cases. However, this is also influenced by other factors such as the cost of holding inventory which is also investigated in Section 9.3.

In the other version of the problem, where demand is constant and the factory is unreliable, the same statement is true: the frequency of subcontracting should be comparable to the frequency of repairs and failures. However, subcontracting is generally less frequent. In particular, if a machine typically fails for half a day, the typical worst backlog will be a day. If customers can tolerate such a backlog without defecting, there is no reason to pay for a subcontractor. Similarly, the typical inventory level (in the absence of strict control), will be about a day. If the company can afford one day's worth of inventory, again there is little reason to subcontract, and the frequency of

38

subcontracting will be low.

9   Behavior of the Model

9.1   Effect of Customer Defection Behavior

The effect of the function $B(x)$ on the performance of the system is depicted in Figure 10 for a system with a single plant. In this figure, $B(x)$ is a sigmoid, a function of the form

$$B(x) = \frac{1}{1 + e^{\gamma(x - x_0)}}.$$

Figure 1 shows a sigmoid function with $\gamma = 1/2$ and $x_0 = -10$. In the examples of Figure 10, we chose $\gamma = 1/2$ and a variety of values of x_0. In the discretization of $B(x)$, we let the step size $\delta = \frac{1}{M\gamma} \ln[\frac{\epsilon}{1-\epsilon} + x_0]$ in order to reach $1 - \epsilon$ in M steps. Then we chose $\beta_0 = 0$,

$$\left. \begin{aligned} \beta_i &= \delta i, \\[2mm] B_i &= \frac{1}{2}\left(\frac{1}{1 + e^{\gamma(\beta_{i-1} - x_0)}} + \frac{1}{1 + e^{\gamma(\beta_i - x_0)}}\right) \end{aligned} \right\} i = 1, 2, .., M.$$

and, in these cases, $\epsilon = 0.001$ and $M = 10$.

As customers become more sensitive to backlog, i.e., as x_0 increases, the profit and the service level decrease, and the expected inventory level increases. Due to the loss of more and more customers, expected backlog level also decreases. Note that the service level is quite low for this case. The upper and lower hedging levels increase as the customers become more sensitive to the waiting time.

Figure 11 shows the effect of x_0 on the hedging levels. In this case, g^+ is very high and therefore it is not desirable to hold finished goods inventory. Moreover as the customers become more impatient, that is as x_0 approaches 0, all the hedging levels approach each other and coincide at $x = 0$ when $x_0 = 0$. In this specific case, $\underline{X} = Z_{\mathrm{H}}(2) > Z_{\mathrm{L}}(2)$ and $Z_{\mathrm{H}}(0) > Z_{\mathrm{L}}(0)$ and therefore $Z_{\mathrm{L}}(2)$ and $Z_{\mathrm{H}}(0)$ do not affect the dynamics. When $x_0 = 0$, it is not possible to hold finished goods inventory or backorder. In this case, when the demand is low, only the manufacturing plant operates at the demand rate and when the demand switches to high, a number subcontractors are engaged in the order of decreasing profit until the demand is met instantaneously. When the demand switches back to low, the subcontractors stop delivery in the reverse order. Since it is not possible to store or backorder electrical power, the operation of the production/subcontracting policy for this special case is analogous to the policy used to operate different power generators with different costs and capacities described in Schweppe, Caramanis, Tabors, and Bohn (1988). Note also that when $Z_0(\mathrm{L}) = 0$, this special case is analogous to a make-to-order system where it is not possible to carry finished goods inventory.
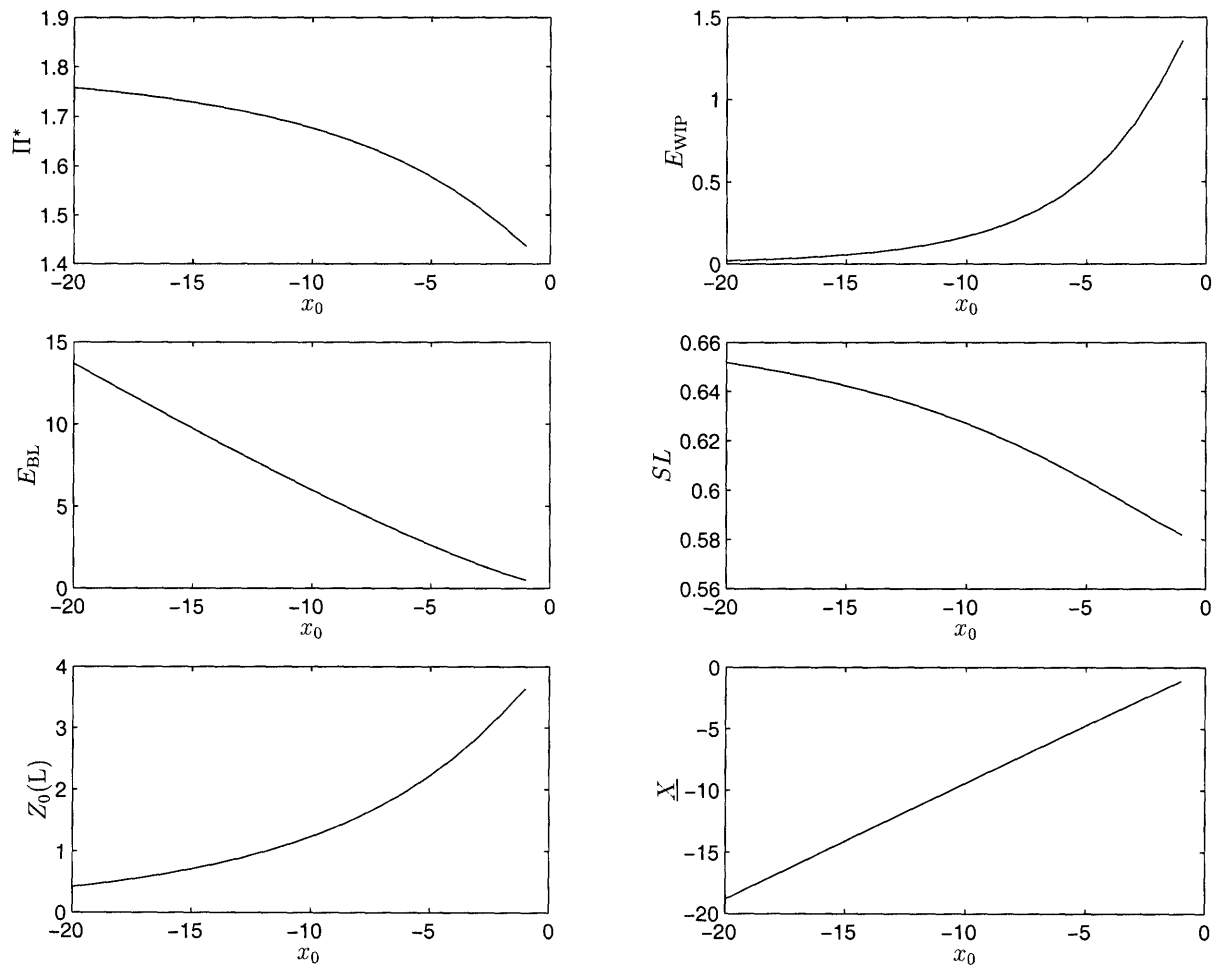
Figure 10: Effect of $B(x)$ ($\mu_{\mathrm{H}} = 1.5$, $\mu_{\mathrm{L}} = 0.3$, $\lambda_{\mathrm{HL}} = 0.05$, $\lambda_{\mathrm{LH}} = 0.05$, $\underline{u}_0 = 0.6$, $A_0 = 3$, $g^+ = 0.1$, $M = 10$, $\gamma = 0.5$)
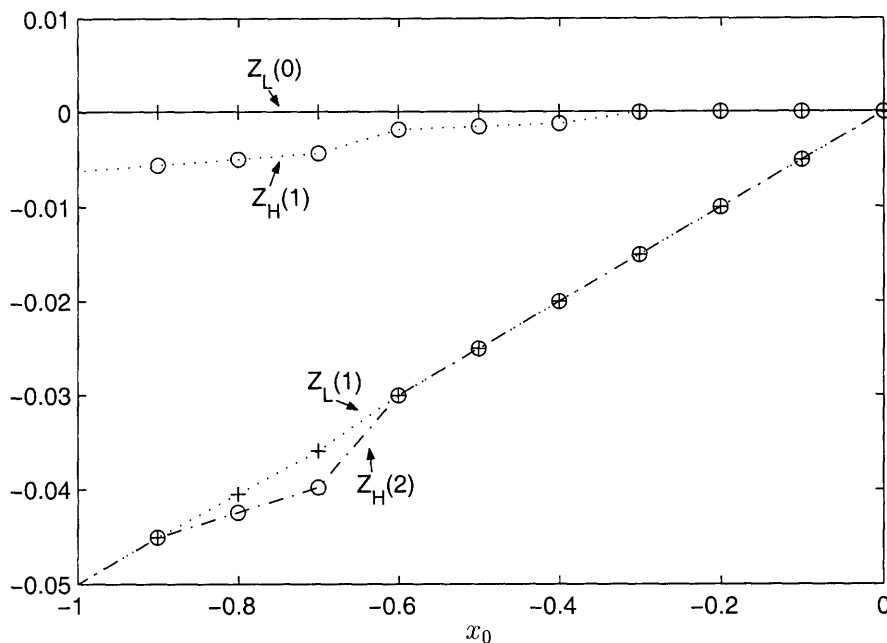
Figure 11: Effect of $B(x)$: A Special case ($\mu_H = 1.5$, $\mu_L = 0.4$, $\lambda_{HL} = 0.05$, $\lambda_{LH} = 0.05$, $\underline{u}_0 = 0.6$, $\underline{u}_1 = 0.7$, $\underline{u}_2 = 1$, $A_0 = 5$, $A_1 = 4$, $A_2 = 1$, $g^+ = 100$, $M = 10$, $\gamma = \frac{1}{x_0} \ln(\frac{\epsilon}{1-\epsilon})$)

9.2   Effect of Demand Variability

The effect of the variability of demand on the performance of the system is depicted in Figure 12 for a system with a single plant. The variability of demand is indicated by the coefficient of variation of the demand rate cv. As the demand variability increases, the profit and service level decrease and the average backlog and the average inventory increase. Increasing demand variability pushes the hedging point $Z_0(L)$ upward. Since there is a single manufacturing facility and there are no subcontractors, the lower limit is not affected by the demand uncertainty; it is determined by the customer behavior from (16).

Figure 13 shows the effect of variability on a system with one plant and one subcontractor for the lost sales case ($B(x) = 1$ for all $x < 0$). Since all the demand is lost when $x \leq 0$, the lower hedging level is located at $x = 0$ and there is no backlog. Just as in the previous case, as cv increases, the upper hedging level and the expected WIP level increases. Furthermore the profit and the service level decreases as the demand variability increases. decreases.

9.3   Effect of Inventory Carrying Cost

Figure 14 shows the effect of inventory carrying cost on the performance of a system with one manufacturing plant and one subcontractor. As g^+ increases, the upper hedging level approaches zero and consequently E_{WIP} decreases to zero. The step-wise behavior of the lower level $\underline{X}$ is caused

41

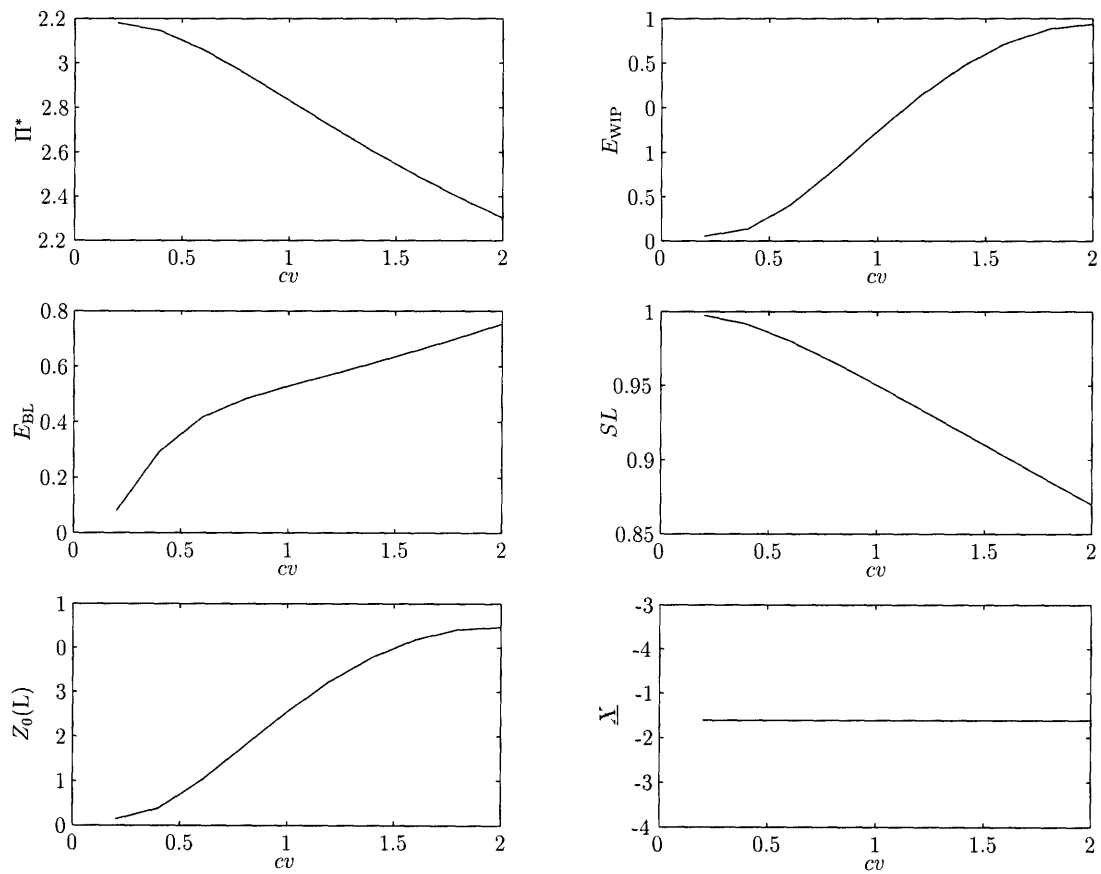Figure 12: Effect of the demand variability ($\mu_H = 1.5$, $\mu_L = 0.3$, $Ed = 0.9$, $\underline{u}_0 = 1$, $A_0 = 3$, $g^+ = 0.1$, $x_0 = -3$, $M = 10$, $\gamma = 0.5$)

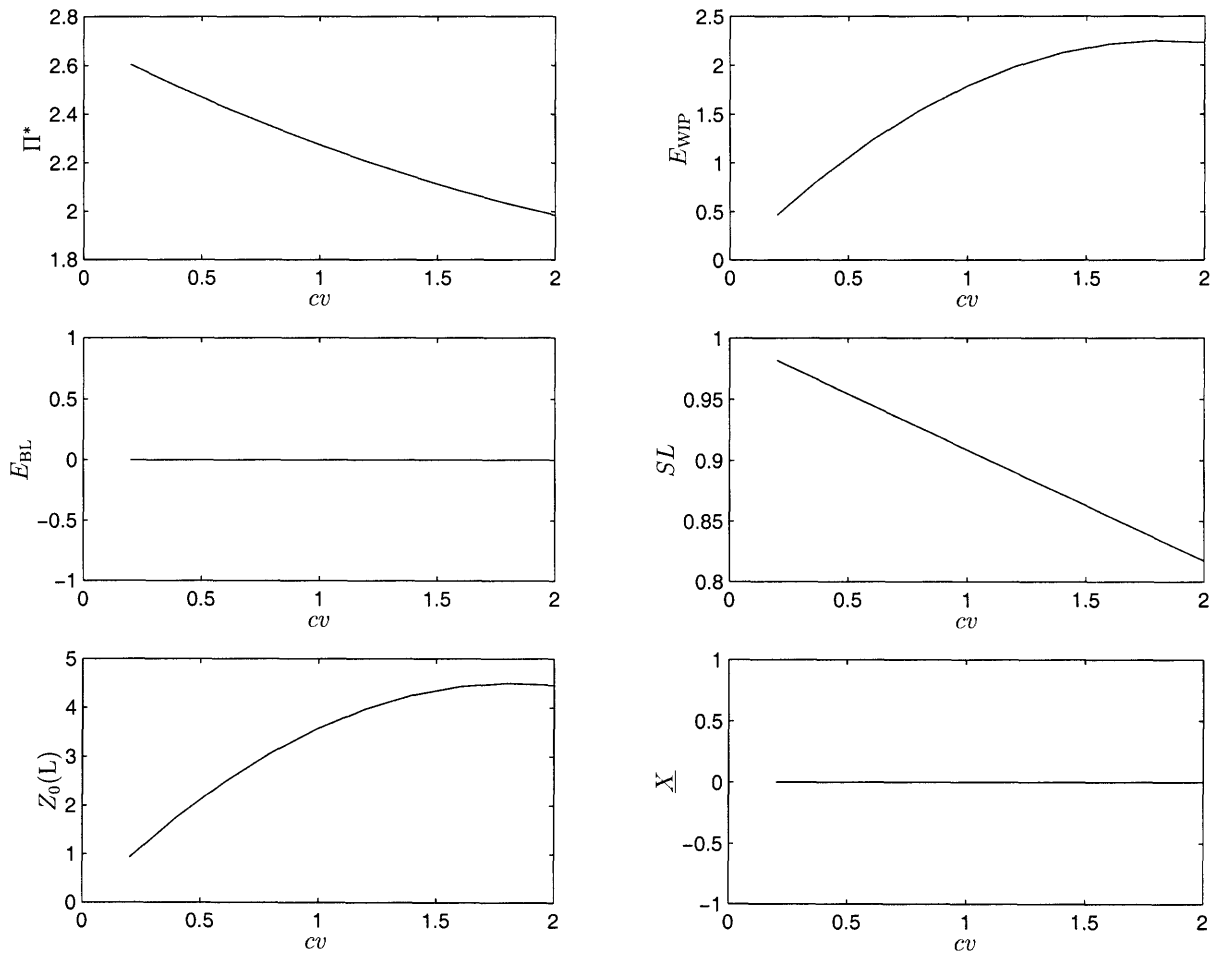Figure 13: Effect of the demand variability: lost sales case ($\mu_H = 1.5$, $\mu_L = 0.3$, $Ed = 0.9$, $\underline{u}_0 = 0.9$, $A_0 = 3$, $g^+ = 0.1$, $B(x) = 1$ for $x \leq 0$)

by the customer behavior. In this specific example, $B(x)$ takes different values at $x = 0, -0.6, -1.2$ and constant between these values. As g^+ increases, we can tolerate more customers and therefore $\underline{X}$ first switches from 0 to -0.6 after staying at 0 and then from -0.6 to -1.2. This behavior of $\underline{X}$ also drives E_{BL} and SL to behave similarly. The slight bump on SL for $g^+ = 0.27$ is caused by numerical optimization routine.

9.4   Effect of a Subcontractor's Price

Figure 15 shows the effect of the price per unit that a single subcontractor charges. If the final sales price of the item is π per unit, the cost for making it in-house is $\pi - A_0$ per unit (other than inventory cost), and the price charged by the subcontractor is $\pi - A_1$. A measure of the additional subcontractor cost relative to A_0, which is independent of final product sales price, is therefore $(A_0 - A_1)/A_0$, the horizontal axis of Figure 15. When this ratio is close to 0, the subcontractor cost is low; when it is close to 1, the price is high.

The vertical axis of Figure 15 is the fractional increase in optimal profits as a result of making use of the subcontractor. The total profit for the case where no subcontracting is used is denoted by Π^0 and the total profit for the case where the subcontractor is available is denoted by Π^1. When the profit from to the subcontractor's supplies is very small compared to the profit from using the manufacturing facility, i.e. when $1 - A_1/A_0$ is close to one, the subcontractor is not used and therefore there is no profit gain. However, as the profit due to the subcontractor approaches the profit from the manufacturing facility, i.e., as $1 - A_1/A_0$ approaches 0, the subcontracting increases the total profit substantially, as much as 63%.

9.5   Capacity Options

Having a subcontractor always available allows the manufacturer to reduce backlog and therefore customer loss. Furthermore, since the subcontractor is used when it is necessary and only paid for the volume of production received, this agreement is attractive for the manufacturer. In order to make this agreement more attractive for the subcontractor, the contractual agreement may include a fixed up-front payment for the duration of the agreement. Tan (2001) analyzes this kind of agreement as *options*. An option is the right, but not obligation, to take an action in the future and a real option is the extension of financial option theory on real (non-financial) assets (Amram and Kulatilaka 1999). Since the contractual agreement is related to increasing the capacity, we refer this option as a *capacity option*.

Similar contractual agreements are reported to be used between manufacturers and their subcontractors and analyzed in Bassok and Anupindi (1997), Eppen and Iyer (1997), Costa and Silver (1996), Jain and Silver (1995), and Tsay and Lovejoy (1999), among others. Most of these studies consider a two-period model and analyze the contracts from the buyer perspective where the buyer has an option of receiving more when the uncertainty is resolved.

Consider the following capacity option: the company pays an up-front fee of C_i to subcontractor i to receive an extra capacity of $0 \le u_i(t) \le \underline{u}_i$ at time t for a duration of T. The exercise cost of the option is $A_0 - A_i$. Due to demand volatility, the company may consider this option to decrease the
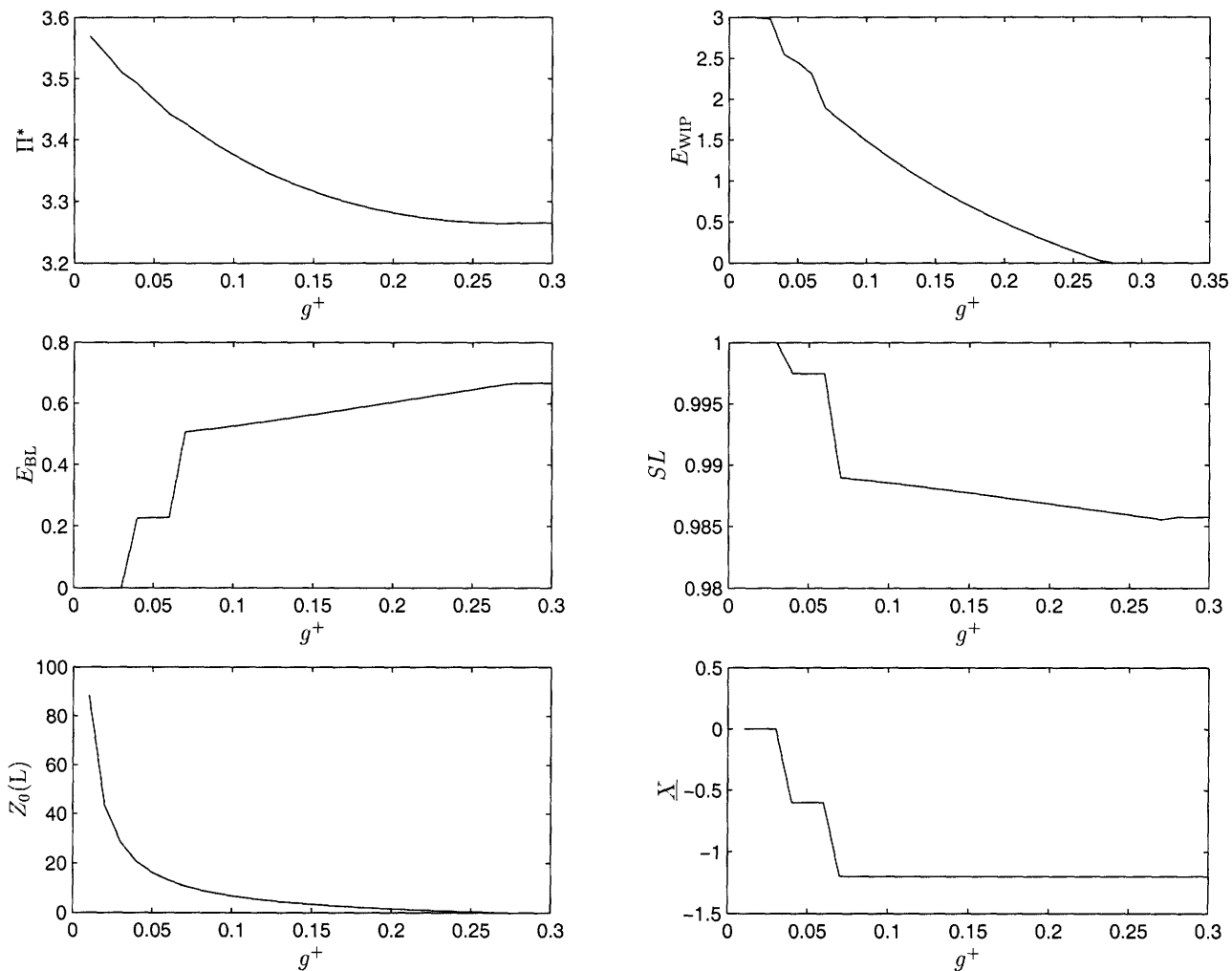
Figure 14: Effect of the inventory carrying cost ($\mu_H = 1.5$, $\mu_L = 0.3$, $Ed = 0.9$, $cv = 2.98$, $\underline{u}_0 = 0.6$, $\underline{u}_1 = 1$, $A_0 = 5$, $A_1 = 2$, $x_0 = -3$, $M = 10$, $\gamma = 0.5$)
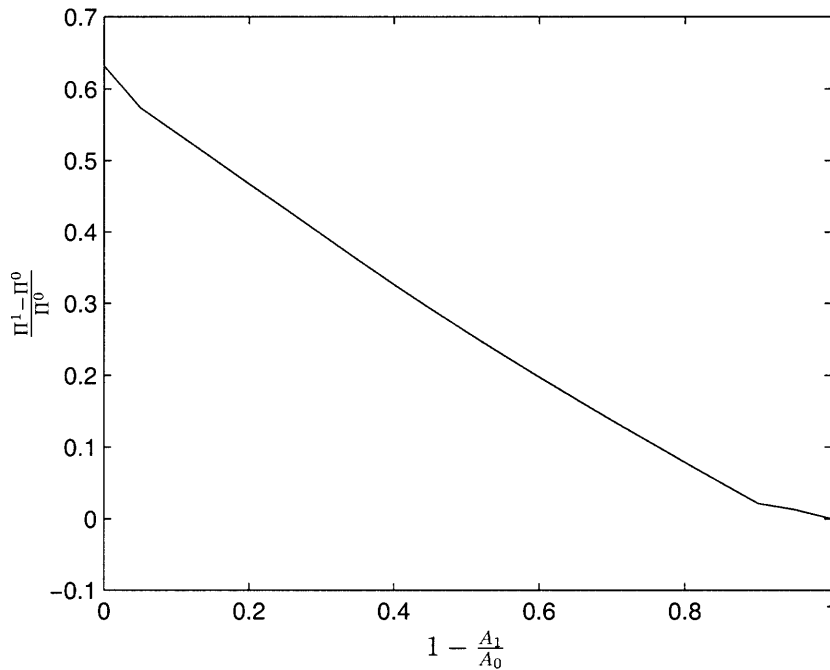
Figure 15: Effect of the subcontracting price ($\mu_H = 1.5$, $\mu_L = 0.3$, $\lambda_{HL} = 0.1$, $\lambda_{LH} = 0.1$, $\underline{u}_0 = 0.6$, $\underline{u}_1 = 1$ $A_0 = 3$, $g^+ = 0.1$, $x_0 = -3$, $M = 10$, $\gamma = 0.5$)

need of holding an excessive inventory or investing in capacity expansion. This is also advantageous for the contractor if it has extra capacity not fulfilled with its own demand. Furthermore, the up-front payment will be received regardless of whether the option is exercised or not in the specified time period.

Let Π^* be the maximum profit that the company can obtain without using subcontractor i, but possibly using the other subcontractors. Let $\Pi^{(i)}$ be the maximum profit that can be obtained by having the option of receiving additional production from subcontractor i as well, assuming a zero up-front payment. If the duration of the contract is long enough, the total profits during $[0, T)$ can be approximated by $\Pi^* T$ and $\Pi^{(i)} T$. Then the maximum amount that should be paid as an up-front payment to subcontractor i is the additional profit that is obtained by using this option, or,

$$C_i \leq (\Pi^* - \Pi^{(i)})T.$$

Figure 16 depicts how this procedure can be used to evaluate the terms of an option for a company with a single plant and only one available subcontractor. The up-front payment is set to 0, 10%, and 20% of the expected profit that the company can generate in the duration of the option without using a subcontractor. The x axis is the exercise cost of the option as a ratio of the per unit profit of the plant, i.e., $(A_1 - A_0)/A_0$. The figure suggests that this option allows the company to increase its profits even after paying an up-front payment and agreeing on a higher exercise cost. For example, if the company pays 20% of the expected profit as an up-front payment
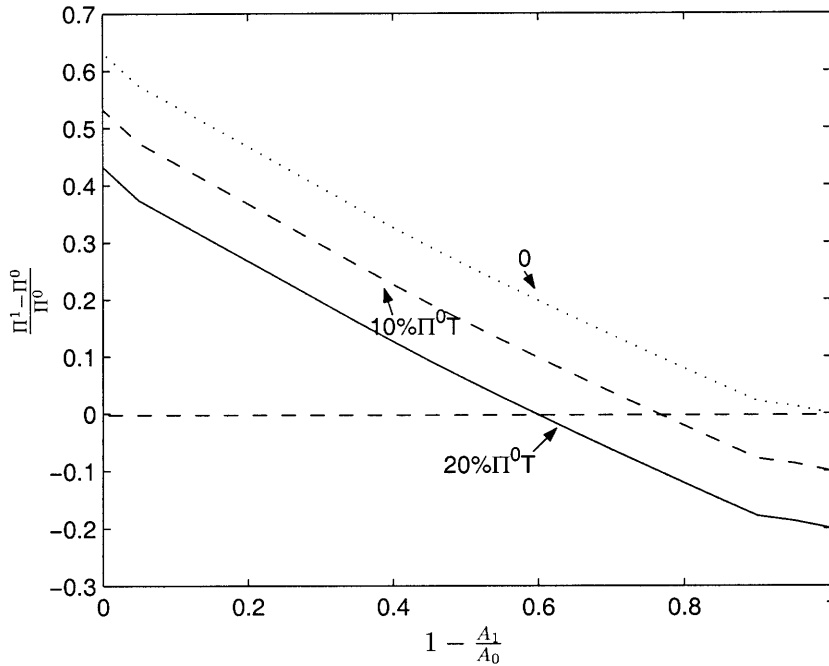
Figure 16: Evaluating the terms of a capacity option ($\mu_H = 1.5$, $\mu_L = 0.3$, $\lambda_{HL} = 0.1$, $\lambda_{LH} = 0.1$, $\underline{u}_0 = 0.6$, $\underline{u}_1 = 1$ $A_0 = 3$, $g^+ = 0.1$, $x_0 = -3$, $M = 10$, $\gamma = 0.5$, $T = 100$, $C_1 = 0$, $10\%\Pi^0$, $20\%\Pi^0$)

to the subcontractor, it can justify paying an exercise cost to the subcontractor that reduces its per unit profit to $40\%A_0$. Similarly as the up-front payment and the exercise cost decrease, this agreement with the subcontractor further increases the profit.

As the demand volatility increases, having the option of increasing the capacity temporarily becomes more important. However, if the demand volatility is low, the up-front payment and the exercise cost may not be justified. Figure 17 shows the effect of demand volatility on the profit increase for a system with a plant and a subcontractor. The up-front payment for this case is 20% of the expected profit during $T = 100$ without the subcontractor. The exercise cost of the option $2/3A_0$. In this case, the capacity option is attractive and increases the profit for $cv > 2.6$. For lower levels of demand volatility, the up-front payment and the exercise cost of the contract cannot be justified and the company satisfies the demand without the subcontractor.

9.6 An Approximate Subcontracting Policy

9.6.1 Justification

The optimal feedback policy derived in this study is a function of the inventory/surplus x and the demand state D. Although the inventory/surplus can be observed easily, the demand state may not be observable. Therefore, an approximate policy that does not depend on the demand state may be necessary. In particular, setting $Z_i(H) = Z_i(L) = Z_i$ yields an approximate policy that does
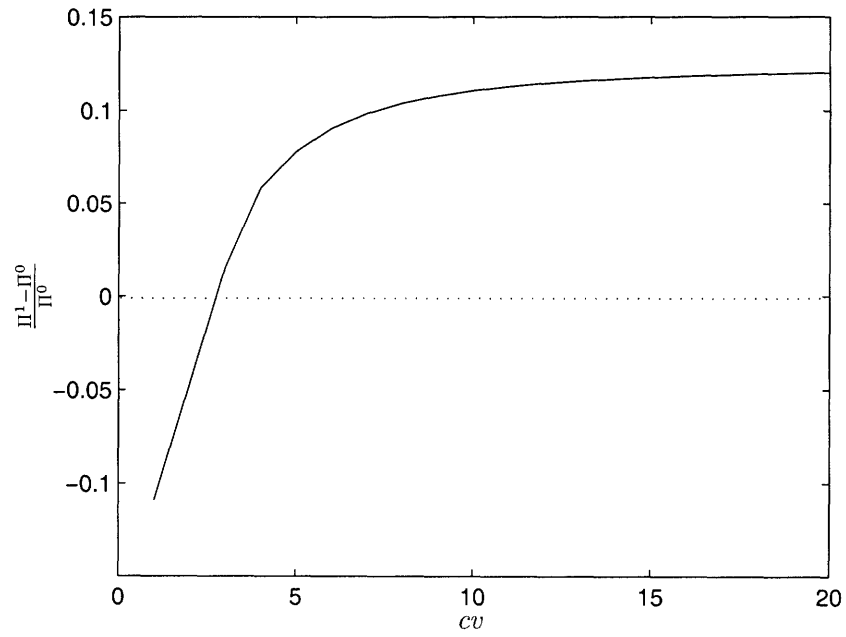
47

Figure 17: Effect of demand variability on a capacity option ($\mu_H = 1.5$, $\mu_L = 0.3$, $\lambda_{HL} = 0.1$, $\lambda_{LH} = 0.1$, $\underline{u}_0 = 0.6$, $\underline{u}_1 = 1$ $A_0 = 3$, $A_1 = 1$, $g^+ = 0.1$, $x_0 = -3$, $M = 10$, $\gamma = 0.5$, $T = 100$, $C_1 = 20\%\Pi^0$)
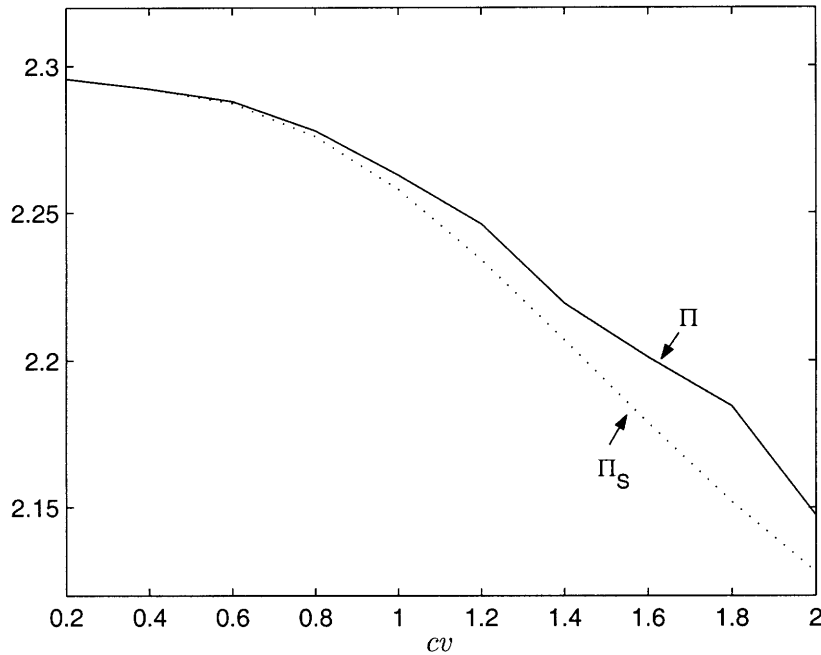
Figure 18: Value comparison for the demand-insensitive policy ($\mu_H = 1.5$, $\mu_L = 0.3$, $Ed = 0.9$, $\underline{u}_0 = 0.5$, $\underline{u}_1 = 0.8$, $\underline{u}_2 = 1$, $A_0 = 3$, $A_1 = 2$, $A_2 = 1$, $g^+ = 0.1$, $x_0 = -3$, $M = 10$, $\gamma = 0.5$)

not depend on the demand state. We call this approximation the *demand-insensitive* policy.

Moreover, as λ_{HL} and λ_{LH} approach infinity, i.e. as the demand state switches faster between high and low, incorporating the demand information in the feedback policy provides less value. Having different levels for each demand state can be beneficial only when the demand persists in a given state, that is, when the demand variability is high.

Figure 18 depicts the profits for a system with a manufacturing facility and two subcontractors for the optimal and the demand-insensitive policies denoted by Π and Π_S respectively. As the figure shows, when the demand variability is low, both policies give the same profit. Even when the demand variability is high, the difference between the optimal profit and the near-optimal profit is less than 1.5%.

9.6.2   Application

When the demand-insensitive policy is used, a system with a manufacturing facility and a number of subcontractors that is operated with this policy can be represented as a network of stations and buffers. Consider a system with a manufacturing facility and two subcontractors. Assume that a demand-insensitive policy with levels $Z_0 > 0 > Z_1 > Z_2$ is used to operate this system. A network representation of this system is given in Figure 19.
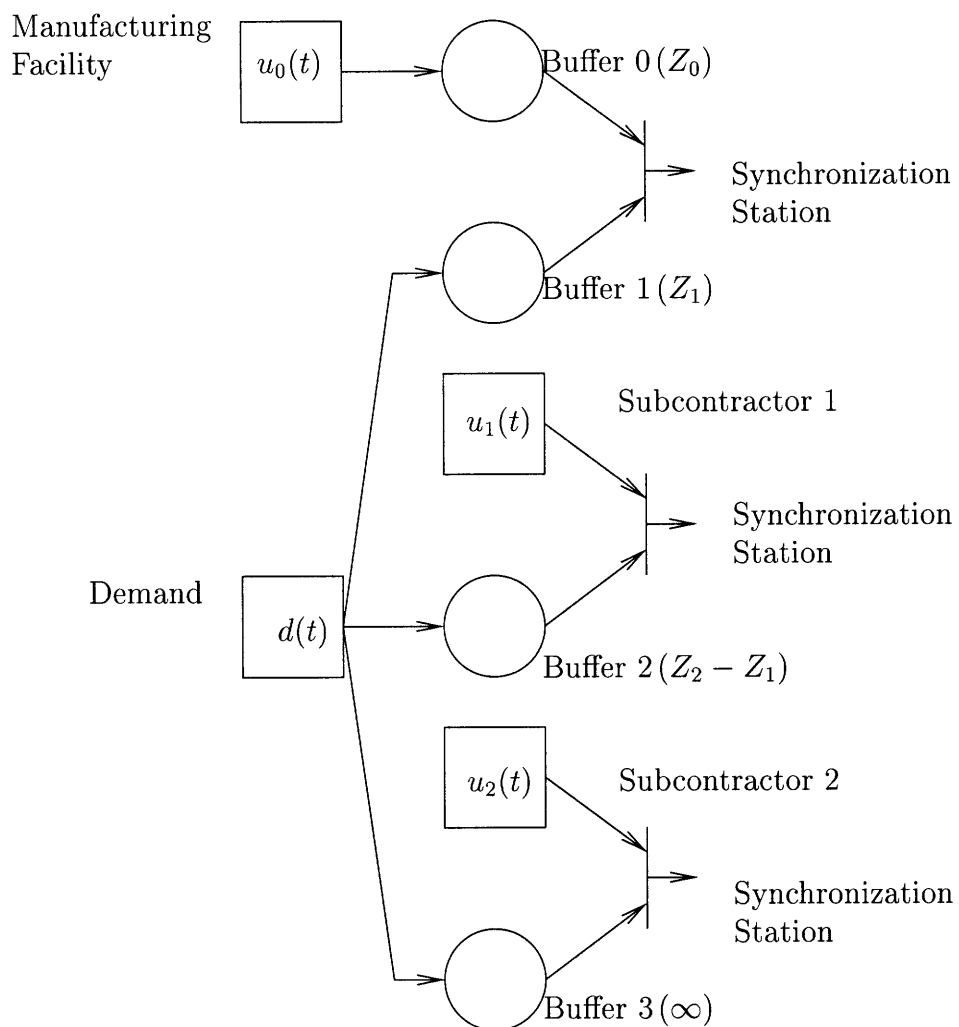
49

Figure 19: An equivalent network representation of the approximate policy

In this system, the manufacturing facility, the subcontractors, and the demand are represented by stations that are drawn as boxes. The facility and the subcontractors generate finished parts; the demand machine generates orders which are matched up with parts at the synchronization machines.

When two streams meet at an assembly station, the parts and orders are assembled instantaneously. The assembly stations are referred as synchronization stations because they are infinitely fast and perfectly reliable. If both buffers at an assembly contain material, enough material will be removed from both so that one (or possibly both) of the buffers is empty. After that, since material arrives at both at finite speeds, at least one buffer is always empty. The rate at which assembly takes place is equal to the rate at which material arrives at the empty buffer.

In particular, either Buffer 0 or Buffer 1 is empty at every time instant. Then Buffer 0 represents the finished goods inventory, x when $x \geq 0$. If Buffer 0 is empty, the total inventory level in the other buffers gives the backlog, x when $x < 0$.

The output of the manufacturing facility goes into Buffer 0, which has capacity Z_0. The output of the demand station is split into three routes where an upper route has priority over the lower one. That is, the demand first goes to Buffer 1 until it is full. When Buffer 1 becomes full, the overflow is routed to Buffer 2, which has capacity $Z_2 - Z_1$. When this buffer is also full, the remaining part of the demand goes into Buffer 3, which has infinite capacity.

The manufacturing facility works with the maximum rate $\underline{u}_0$ until Buffer 0 fills up. When that happens, the output rate from the manufacturing facility drops down to the demand rate, which is just enough to keep Buffer 0 full. This is the optimal operating policy given in (26).

Subcontractor 1 can start producing when there is demand flowing into Buffer 2, (i.e., when Buffer 1 becomes full). When Buffer 1 fills up, the amount of demand that flows into Buffer 1 is equal to the current demand minus the maximum production rate of the manufacturing facility $\underline{u}_0$. When the backlog reaches Z_1, Subcontractor 1 starts delivering goods with its maximum rate $\underline{u}_1$ and continues as long as Buffer 2 is not full, i.e., as long as $x < Z_2$. Similarly, Subcontractor 2 produces when Buffer 1 and 2 are full, or equivalently, when $x > Z_2$.

A set of control policies which can be represented as a network of stations and buffers, and whose performance can possibly be evaluated by a decomposition method, is described by Gershwin (2000). For such policies, the optimal values of the control parameters can be determined by using a buffer sizing algorithm. The policy described here is an extension of that set of policies, and decomposition methods do not currently exist for the network. We suggest that finding such a method would be valuable future research.

10   Conclusions

10.1   Summary

We have extended the widely-studied dynamic programming model of real-time scheduling control of manufacturing systems in two important ways: we model the efect of backlog on profits through an explicit representation of customer behavior; and we model the availability of subcontractors to

provide finished goods when the factory's short-term capacity is insufficient. We also model random demand.

The new model of customer behavior involves a *defection function* which indicates what fraction of the potential customers choose not to complete their orders when the backlog reaches a given level. Because of this phenomenon, the model has a novel feature: the demand need not be less than capacity (including the capacities of the subcontractors) for there to exist a steady-state probability distribution of the inventory/backlog and the demand state.

We use the Bellman equation to determine a solution structure, and we find that the solution involves a hedging point (to limit how far production should be allowed to go ahead of demand), and a set of thresholds (that indicate when to use each of the subcontractors). To calculate the hedging point and the thresholds, we find the steady-state probability distribution. We evaluate the objective function and choose values of the parameters to maximize it.

Finally, we have performed a set of numerical experiments to demonstrate the behavior of the new model and the solution.

10.2   Future Research

This research can be extended in several different directions:

- An extension of the hedging point policy to complex systems (multiple part types; multiple stages; general routing including reentrant flow) is described by Gershwin (2000). In the present system, lead time is due only to the producer falling behind demand. In the more complex system, lead time is also due to the fact that material flows from stage to stage, and may have to wait at each stage. The policy in (Gershwin 2000) is based on a dynamic programming problem that includes an explicit backlog cost. It would be of interest to replace that backlog cost with the present model of customer behavior.

- The amount of sales that a business has should be a function of its past delivery performance. However, there is no way in the present model to account for the producer's reputation for on-time delivery. One way to include such an effect might be to add an appropriate state variable. For example, consider

$$R^1(t) = \frac{\int_0^t \sum_{i=0}^{K} u_i(\tau)d\tau}{\int_0^t d(\tau)d\tau}$$

This quantity is the average amount of demand actually served as a fraction of total potential demand. We can extend the demand model so that the demand parameters (μ_{L}, μ_{H}, λ_{LH}, λ_{HL}) are functions of $R^1(t)$.

Another possible reputation variable is

$$R^2(t) = \frac{\int_0^t (1 - B(x(\tau)))d\tau}{t}$$

which might be easier to include in the system dynamics. The integrand is the fraction of customers who do not defect.

- The method to value a given option explained in Section 9.5 can be used to evaluate the terms of a given contract from a single subcontractor directly. However, if there are a number of available options from alternative subcontractors, determining the best group of subcontractors requires simultaneous evaluation of the terms of the contracts.

- An important extension would be to include competition in the formulation, and turn it into a game. Now the $B(x)$ function as seen by one firm depends on the actions taken by all competing firms. A simple example of this is described in Appendix A. A related extension would be to model the competition among the subcontractors.

- To make the model more complete, we should include possible delays in subcontractor performance. Another extension would be to guarantee lead times to customers, but this may require creating classes, one for each guaranteed lead time.

- We have postulated the existence of the defection function $B(x)$. It would be very desirable to use empirical data to confirm or refute the existence of such a function.

- In Appendix A, we analyze the effect of customer behavior on the customer defection function in a queueing model. This approach can be extended to analyze the effect of customer behavior, the capacity and the competitive position of the firm on customer defection.

Acknowledgments

References

Abernathy, F. H., J. T. Dunlop, J. H. Hammond, and D. Weil (1999). *A Stich in Time: Lean retailing and the transformation of manufacturing-lessons from the apparel and textile industries*. Oxford University Press, New York.

Abernathy, F. H., J. T. Dunlop, J. H. Hammond, and D. Weil (2000). Control your inventory in a world of lean retailing. *Harvard Business Review November-December*, 169–176.

Amram, M. and N. Kulatilaka (1999). *Real Options: Managing Strategic Investment in an Uncertain World*. Harvard Business School Press, Boston, MA.

Bassok, Y. and R. Anupindi (1997). Analysis of supply contracts with total minimum commitments. *IIE Transactions 29*(5), 373–381.

Bielecki, T. and P. R. Kumar (1988). Optimality of zero-inventory policies for unreliable manufacturing systems. *Operations Research 36*(4), 532–541.

Bradley, J. (1999). Optimal control of an M/M/1 subcontracting model. Technical report, Cornell University.

Bradley, J. and P. Glynn (2000a). Managing capacity and inventory effectively in manufacturing systems. Technical report, Cornell University.

Bradley, J. and P. Glynn (2000b). Managing the manufacturer-subcontractor relationship and the manufacturer's optimal capacity, inventory, and subcontracting policies. Technical report, Cornell University.

Cachon, G. (1999). Service competion, outsourcing and co-production in a queuing game. Technical report, University of Pennsylvania, Wharton Financial Instituions Center.

Chang, H.-J. and C.-Y. Dye (1999). An EOQ model for deteriorating items with time varying demand and partial backlogging. *The Journal of the Operational Research Society 50*(11), 1176–1182.

Costa, D. and E. Silver (1996). Exact and approximate algorithms for the multi-period procurement problem where dedicated supplier capacity can be reserved. *Operations Research Spectrum 18*(4), 197–207.

Dellaert, B. G. C. and B. E. Kahn (1999). How tolerable is delay?: Consumers' evaluations of internet web sites after waiting. *Journal of Interactive Marketing 13*(1), 41–54.

Eppen, G. and A. Iyer (1997). Backup agreements in fashion buying-the value of upstream flexibility. *Management Science 43*(11), 1469–1484.

Fleming, W., S. Sethi, and H. Soner (1987). An optimal stochastic production planning with randomly fluctuating demand. *SIAM Journal of Control Optimization 25*, 1495–1502.

Fukuda, Y. (1964). Optimal policies for the inventory problem with negotiable lead-time. *Management Science 4*, 690–708.

Gershwin, S. B. (1992-1993). Extension of FMS scheduling model. Unpublished note.

Gershwin, S. B. (1994). *Manufacturing Systems Engineering*. Prentice-Hall. See http://web.mit.edu/manuf-sys/www/gershwin.errata.html for corrections.

Gershwin, S. B. (2000). Design and operation of manufacturing systems — the control-point policy. *IIE Transactions 32*(2), 93–103.

Ghosh, M., A. Araposhathis, and S. Markus (1993). Optimal control of switching diffusions with applications to flexible manufacturing systems. *SIAM Journal of Control Optimization 31*, 1183–1204.

Hall, R. (1991). *Queueing Methods for Services and Manufacturing.* Prentice Hall, Englewood Cliffs, NJ.

Hu, J. (1995). Production rate control for failure prone production with no backlog permitted. *IEEE Transactions on Automatic Control 40*(2), 291–295.

Huang, L., J. Hu, and P. Vakili (1999, May 16-20). Optimal control of a multi-state manufacturing system: Control of production rate and temporary increase in capacity. *Second Aegean International Conference on Analysis and Modeling of Manufacturing Systems*, 191–198. http://www.samos.aegean.gr/icsd/secaic/.

Hui, M. and D. Tse (1996). What to tell customers in waits of different lengths: and integrative model of service evaluation. *Journal of Marketing 60*, 81–90.

Ittig, P. T. (1994). Planning service capacity when demand is sensitive to delay. *Decision Sciences 25*(4), 541–560.

Jain, K. and E. Silver (1995). The single-period procurement problem where dedicated supplier capacity can be reserved. *Naval Research Logistics 42*(6), 915–934.

Kimemia, J. G. and S. B. Gershwin (1983). An algorithm for the computer control of production in a flexible manufacturing systems. *IIE Transactions 15*(4), 353–362. Reprinted in *Modeling and Control of Automated Manufacturing Systems* by Alan A. Desrochers, IEEE Computer Society Press Tutorial, 1990.

Krichagina, E. V., S. X. C. Lou, and M. I. Taksar (1994). Double band policy for stochastic manufacturing systems in heavy traffic. *Mathematics of Operations Research 19*(3), 560–597.

Martin, S. and P. C. Smith (1999). Rationing by waiting lists: An empirical investigation. *Journal of Public Economics 71*(1), 141–164.

McAvinchey, I. D. and A. Yannopoulos (1993). Elasticity estimates from a dynamic model of inter-related demands for private and public acute health care. *Journal of Health Economics 12*(2), 171–186.

Mount, M. (1994). Service time competition. *The Rand Journal of Economics 79*(2), 619–634.

Olsder, G. J. and R. Suri (1980, December). Time-optimal control of flexible manufacturing systems with failure prone machines. In *Proceedings of the 19th IEEE Conference on Decision and Control*, Albuquerque, New Mexico.

Perkins, J. and R. Srikant (2001). Failure-prone production systems with uncertain demand. *IEEE Transactions on Automatic Control.* to appear.

Puzo, M. (1969). *The Godfather.* G.P. Putnam's Sons.

Rao, U., A. Scheller-Wolf, and S. Tayur (2000). Development of a rapid-response supply chain at caterpillar. *Operations Research 48*(2), 189–204.

Schweppe, F., M. Caramanis, R. Tabors, and R. Bohn (1988). *Spot Pricing of Electricity.* Kluwer Academic Publishers, Boston.

Tan, B. (1997). Variance of the throughput of an N-station production line with no intermediate buffers and time dependent failures. *European Journal of Operational Research 101* (3), 560–576.

Tan, B. (2000). Production control of a failure prone manufacturing system with variable demand. Working Paper Series 00-1, Koç University.

Tan, B. (2001). On capacity options in lean retailing. Research Paper Series February-01, Harvard University, Center for Textile and Apparel Research.

Taylor, S. (1994). Waiting for service: the relationship between delays and evaluations of service. *Journal of Marketing 58*, 56–69.

Tsay, A. and W. Lovejoy (1999). The quantity flexibility contract and supply chain performance. *Manufacturing and Service Operations Management 1* (2), 89–111.

Van Mieghem, J. A. (1999). Coordinating investment, production, and subcontracting. *Management Science 45* (7), 954–971.

Weinberg, B. D. (2000). Don't keep your internet customers waiting too long at the (virtual) front door. *Journal of Interactive Marketing 14* (1), 30–39.

Whitt, W. (1999a). Improving service by informing customers about anticipated delays. *Management Science 45* (2), 192–207.

Whitt, W. (1999b). Predicting queueing delays. *Management Science 45* (6), 870–889.

Whittemore, A. and S. Saunders (1977). Optimal inventory under stochastic demand with two supply options. *SIAM Journal of Applied Mathematics 32*, 293–305.

Yan, H., S. Lou, S. Sethi, A. Gardel, and P. Deosthali (1996, May). Testing the robustness of two-boundary control policies in semiconductor manufacturing. *IEEE Transactions on Semiconductor Manufacturing 9* (2), 285–288.

Yeralan, S. and B. Tan (1997). A station model for continuous materials flow production. *International Journal of Production Research 35* (9), 2525–2541.

Zhang, V. (1995). *Ordering Policies for an Inventory System with Supply Flexibility.* Department of industrial engineering and engineering management ph. d. thesis, Stanford University.

A   Effect of Customer Behavior on the Defection Function

In this section, we briefly discuss the effect of customer response to waiting on the customer defection realized by servers. In order to analyze this phenomenon, we employ a simple queueing model. See Figure 20.

Consider a system with two queues, two exponential servers and Poisson arrivals. Assume that when a customer arrives, she joins the shortest queue (and ties are broken with equal probability).
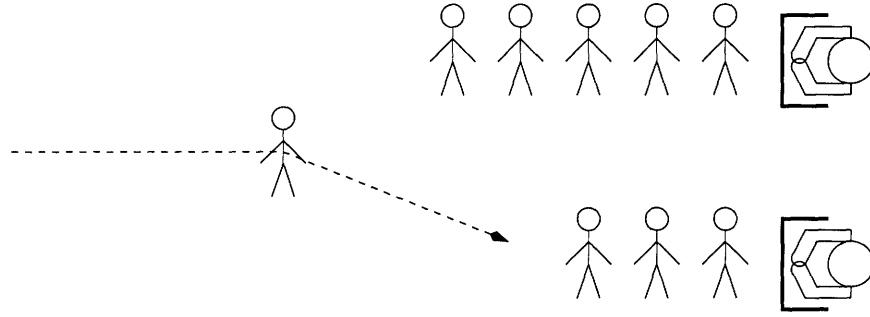
56

Figure 20: Shortest queue system

Under this customer behavior assumption, each server will see a customer defection rate that depends on the queue length. Namely, given that there are a number of customers in the first queue, the first server sees that a certain percentage of the arriving customers join the other queue.

Let N_1 and N_2 denote the number of customers in the first and the second system respectively. The steady-state joint probability function is defined as

$$p(n_1, n_2) = \mathbf{prob}[N_1 = n_1, N_2 = n_2]$$

The steady-state dynamics of the system are given by the following set of flow equations:

$$\lambda p(0, 0) = \mu_1 p(1, 0) + \mu_2 p(0, 1)$$

$$(\mu_1 + \lambda)p(1, 0) = \frac{\lambda}{2}p(0, 0) + \mu_1 p(2, 0) + \mu_2 p(1, 1)$$

$$(\mu_1 + \lambda)p(n_1, 0) = \mu_1 p(n_1 + 1, 0) + \mu_2 p(n_1, 1), \quad n_1 \geq 2$$

$$(\mu_2 + \lambda)p(0, 1) = \frac{\lambda}{2}p(0, 0) + \mu_1 p(1, 1) + \mu_2 p(0, 2)$$

$$(\mu_2 + \lambda)p(0, n_2) = \mu_1 p(1, n_2) + \mu_2 p(0, n_2 + 1), \quad n_2 \geq 2$$

$$(\mu_1 + \mu_2 + \lambda)p(n_1, n_1) = \lambda p(n_1, n_1 - 1) + \lambda p(n_1 - 1, n_1) + \mu_1 p(n_1 + 1, n_1) + \mu_2 p(n_1, n_1 + 1), \quad n_1 \geq 1$$

$$(\mu_1 + \mu_2 + \lambda)p(n_1 + 1, n_1) = \frac{\lambda}{2}p(n_1, n_1) + \mu_1 p(n_1 + 2, n_1) + \mu_2 p(n_1 + 1, n_1 + 1), \quad n_1 \geq 1$$

$$(\mu_1 + \mu_2 + \lambda)p(n_1, n_1 + 1) = \frac{\lambda}{2}p(n_1, n_1) + \mu_1 p(n_1 + 1, n_1 + 1) + \mu_2 p(n_1, n_1 + 2), \quad n_1 \geq 1$$

$$(\mu_1 + \mu_2 + \lambda)p(n_1, n_2) = \lambda p(n_1, n_2 - 1) + \mu_1 p(n_1 + 1, n_2) + \mu_2 p(n_1, n_2 + 1), \quad n_1 \geq 3, n_1 - 2 \geq n_2 \geq 1$$

$$(\mu_1 + \mu_2 + \lambda)p(n_1, n_2) = \lambda p(n_1, n_2 + 1) + \mu_1 p(n_1 + 1, n_2) + \mu_2 p(n_1, n_2 + 1), \quad n_2 - 2 \geq n_1 \geq 1, n_2 \geq 3$$

Solution of the above set of equations with $\sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} p(n_1, n_2) = 1$ yields the steady-state joint probability function.

When $n_1 > n_2$, the first server looses all the arriving customers. When $n_1 = n_2$, an arriving customer defects with 50% probability. Therefore, the conditional probability that an arriving customer defects when there are n_1 customers waiting in the system gives the *customer defection function*. By denoting the number of customers waiting as negative numbers, $B_1(x)$ and $B_2(x)$ are determined as

$$B_1(x) = \mathbf{prob}[N_1 > N_2 | N_1 = -x] + \frac{1}{2}\mathbf{prob}[N_1 = N_2 | N_1 = -x]$$

Similarly

$$B_2(x) = \mathbf{prob}[N_1 < N_2 | N_2 = -x] + \frac{1}{2}\mathbf{prob}[N_1 = N_2 | N_2 = -x]$$

Figure 21 depicts the realized $B_1(x)$ and $B_2(x)$ for four different cases. In the first case, $\mu_1 = \mu_2$ and therefore both servers see the same defection function. In the other cases, the customer defection rate is higher for the slower server. Note that the shape of the realized customer defection functions is similar to the one used in this study.

Customer behavior and the competitiveness of the company in the market determine the customer defection rate. In the previous example, the combined effect of customers choice of joining the shortest queue, availability of an alternative server, and the difference between the service rates of the servers determine the customer defection function.

Now, consider an alternative customer behavior. Assume that customers choose to join a queue with a shorter expected waiting time. In the above model, since the service time is exponential, when there are n_1 customers waiting for the first server and n_2 customers are waiting for the second server, the expected waiting time for a potential customer is n_1/μ_1 for the first server and n_2/μ_2 for the second server. Therefore, an arriving customer chooses the first server when $\frac{n_1}{\mu_1} < \frac{n_2}{\mu_2}$ and the second server when $\frac{n_1}{\mu_1} > \frac{n_2}{\mu_2}$. We assume that a tie is broken with choosing a server with equal probability. The steady-state joint density of this system can be determined by deriving the flow equations.

The *customer defection functions* for this case are defined as

$$B_1(x) = \mathbf{prob}\left[\frac{N_1}{\mu_1} > \frac{N_2}{\mu_2} | N_1 = -x\right] + \frac{1}{2}\mathbf{prob}\left[\frac{N_1}{\mu_1} = \frac{N_2}{\mu_2} | N_1 = -x\right]$$
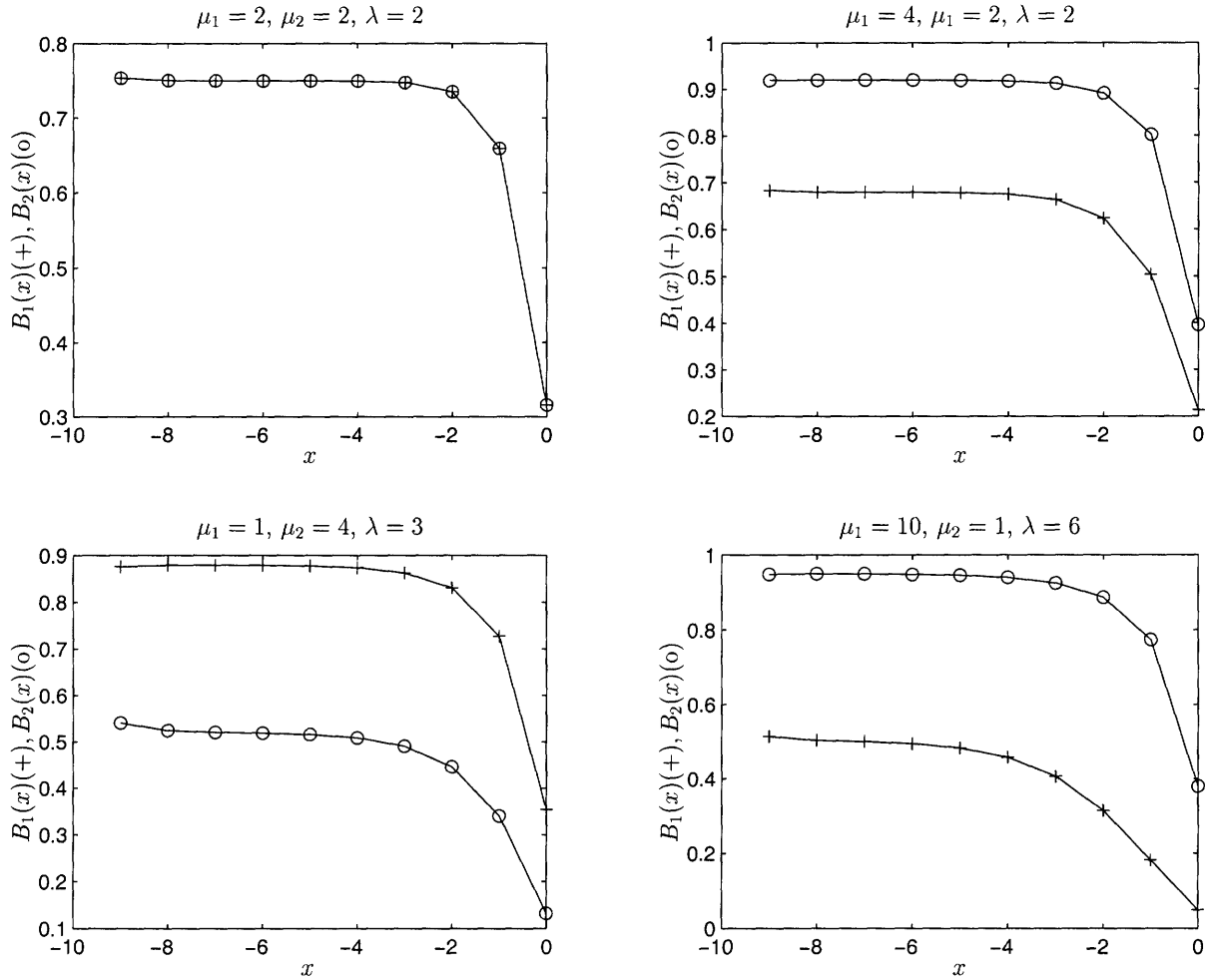
Similarly

Figure 21: Realized customer defection functions, $B_1(x)$ and $B_2(x)$ for a two-server shortest-queue system

$$B_2(x) = \mathbf{prob}\left[\frac{N_1}{\mu_1} < \frac{N_2}{\mu_2} | N_2 = -x\right] + \frac{1}{2}\mathbf{prob}\left[\frac{N_1}{\mu_1} = \frac{N_2}{\mu_2} | N_2 = -x\right]$$

It is possible to extend this approach to analyze alternative customer behaviors and also the effect of competition in the market. Furthermore, a company can obtain valuable information on customers' attitude towards waiting, the competition in the market by observing customer defections. This information can then be used to decide on capacity levels or operate the system more effectively. The investigation of this phenomenon is left for future research.