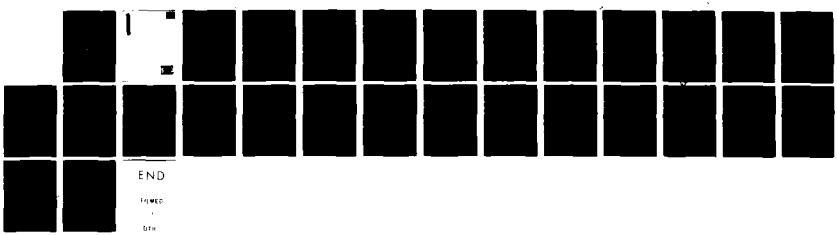
AD-A120 645 A DIFFUSION MODEL OF INVENTORY AND PRODUCTION CONTROL 1/1
(U) STANFORD UNIV CA J M HARRISON AUG 82 TR-205
N00014-75-C-0561

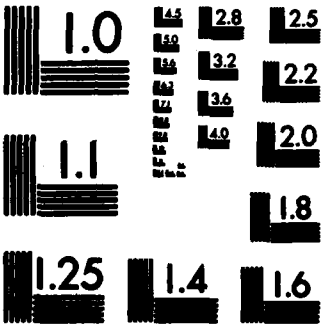
UNCLASSIFIED F/G 13/8 NL



END

FILMED

DTIC



A DIFFUSION MODEL OF INVENTORY AND PRODUCTION CONTROL

by

J. MICHAEL HARRISON

TECHNICAL REPORT NO. 205

August 1982

SUPPORTED UNDER CONTRACT N00014-75-C-0561 (NR-047-200)
WITH THE OFFICE OF NAVAL RESEARCH

Gerald J. Lieberman, Project Director

DEPARTMENT OF OPERATIONS RESEARCH
AND
DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

A DIFFUSION MODEL OF INVENTORY AND PRODUCTION CONTROL

J. Michael Harrison
Graduate School of Business
Stanford University

1. Introduction

This paper is concerned with the diffusion limits of several closely related production planning problems. Each problem involves a firm producing a single durable commodity, which it sells for π dollars per unit. Production flows into a finished goods inventory, and demand which cannot be met from stock on hand is simply lost, with no adverse effect on future demand. With the price π fixed, product demand is viewed as an exogenous source of uncertainty, and very specific assumptions will be made about its stochastic structure.

Given a fixed investment in plant and equipment, we consider the problem of adjusting production, inventory and workforce levels as demand quantities are observed. This is a production smoothing problem of the general type studied by Holt, Modigliani, Muth and Simon (1960). Simplest of the several formulations to be considered here is the following two-stage problem with lost sales.

At time zero the firm must select a work force size, or equivalently a regular-time production capacity. For simplicity, assume that the work-force size cannot be varied thereafter, the firm being obliged to pay workers at a stated wage rate regardless of whether they are productively employed. Let k be the capacity level selected, in production units per unit time. The firm then incurs a labor cost of ℓk dollars per unit time ever afterward, where ℓ is a specified constant, even if it occasionally

chooses to operate below full capacity. In this first formulation, overtime production is assumed to be impossible or forbidden. In addition to its labor costs, the firm incurs a materials cost of m dollars for each unit of actual production. Given the initial capacity decision (workforce level), labor costs are fixed, and thus the marginal cost of production is m dollars per unit. Next, a physical holding cost of h dollars per unit time is incurred for each unit of production held in inventory. (This does not include the financial cost of holding inventory.) Finally, it is assumed that the firm earns interest at rate $\alpha > 0$, compounded continuously, on funds which are not required for production operations. The firm must choose a capacity level k and then at each time $t \geq 0$ select a production rate from the interval $[0,k]$. Its objective is to minimize the expected present value of sales revenues received minus operating expenses incurred over an infinite planning horizon, where discounting is continuous at interest rate α. When a production rate below k is selected, we shall say that undertime is being employed.

To specify the stochastic character of demand, let us consider the sequence of discrete time periods ending at epochs $t = 1, 2, \ldots$. Let ξ_t denote the total demand experienced during period t. It is assumed that $\{\xi_1, \xi_2, \ldots\}$ form a sequence of independent and identically distributed (IID) positive random variables with

$$(1) \qquad E(\xi_t) = \lambda > 0 \qquad \text{and} \qquad \text{Var}(\xi_t) = \sigma^2 > 0 .$$

Furthermore, as a convenient idealization, assume that demand arrives at a uniform rate during each individual period. Thus, defining the partial

sums $S_n = \xi_1 + \cdots + \xi_n$ (with $S_0 = 0$ by convention), the cumulative demand up to time t is

$$(2) \qquad D(t) = (t-n)S_{n+1} + (n+1-t)S_n \quad \text{if} \quad n \leq t \leq n+1 .$$

A picture of the interpolated random walk $D = \{D(t),\ t \geq 0\}$ is given in Figure 1.

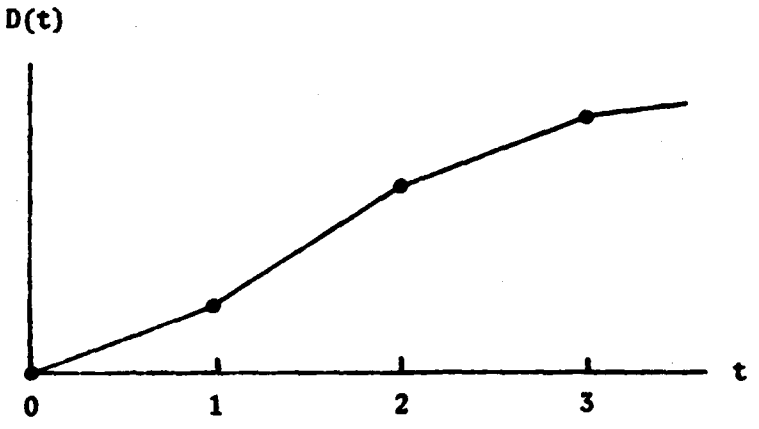


Figure 1. A Typical Sample Path of the Cumulative Demand Process D

These assumptions rule out seasonality, growth, and every other sort of secular trend in demand. It might be argued that this makes the production planning problem uninteresting, since optimal response to such secular trends is the dominant issue in real life. My objective here, however, is to explore the potential role of diffusion models in production and inventory theory, starting with the simplest possible situation. When

certain basic principles have been established, more realistic generalizations will naturally suggest themselves.

The two-stage problem with lost sales will be stated in precise mathematical terms in Section 2. Fixing the demand process D, we then consider a sequence of cost structures in which

(3) $$\varepsilon \equiv h + m\alpha \downarrow 0 ,$$

with the wage rate ℓ and the <u>contribution margin</u> $c = \pi - m$ remaining constant. Under these conditions, it will be shown that our production smoothing problem approaches a certain two-stage optimal control problem for Brownian motion, which can be solved explicitly. The first stage of this limiting problem involves selecting a drift rate for an underlying Brownian motion. Its second stage is the instantaneous control problem formulated and solved by Harrison-Taylor (1977). When the solution of this limiting problem is interpreted in terms of the original production planning problem, we arrive at the following conclusion. If ε is small, then a nearly optimal policy is to

(4) select capacity level $k = \lambda + \varepsilon^{1/2}\mu^*$, where μ^* is an easily computed constant,

(5) use only as much undertime as required to keep the inventory level below $b^*\varepsilon^{-1/2}$ where b^* is another easily computed constant,

and (of course) to forego potential sales only when obliged to do so by lack of stock. The capacity choice (4) creates a condition of <u>balanced</u>

loading, meaning that the full-capacity production rate and the average demand rate are very nearly equal. (In queueing theory this would be called a heavy traffic condition, but the term balanced loading seems more appropriate for production systems.) A novel and important feature of the treatment given here is that no such assumption is imposed a priori. From the economic assumption (3), it is shown that optimal system design creates a balanced loading condition, and this in turn justifies a diffusion approximation for the subsequent problem of inventory control through manipulation of the production rate.

Readers will doubtless recognize $m\alpha$ as the financial cost of carrying inventory, in dollars per unit of inventory per unit time, with inventory valued at its marginal production cost of m dollars per unit. Thus our sole assumption (3) is that the total cost of carrying inventory (both explicit and implicit) is vanishingly small compared with the opportunity loss on foregone sales and the cost of labor. This leads to the high tolerance for inventory manifested in (5) and the disinterest in excess capacity manifested in (4).

It was stated earlier that (4) and (5) constitute a nearly optimal policy when ε is small, but this statement will not in fact be justified by a formal proof of convergence. By carefully developing a sequence of equivalent transformed problems, I hope to make it clear how the diffusion model emerges as ε becomes small, without the technical complexity required for a rigorous proof of convergence. It seems likely that this line of argument can be extended to a formal limit theorem, and others may have some interest in that task.

Section 3 is devoted to a two-stage planning problem with overtime. This is identical to the problem discussed above except that, after the regular-time production capacity (workforce level) has been set, overtime production is also available at the cost of a premium labor rate. We find that this problem is mathematically equivalent to its predecessor, and thus their diffusion limits coincide.

Section 4 treats a more complex and realistic problem where the workforce level can be dynamically adjusted, and all else is as before. (Overtime production may or may not be available.) Each decrease in the workforce or capacity level entails a proportional transaction cost, and the same is true for increases. Again we consider a sequence of cost structures where $\varepsilon \equiv h + m\alpha \downarrow 0$ and all else remains constant. The production planning problem approaches a two-dimensional stochastic control problem where one simultaneously controls the state and the drift rate of a Brownian motion, subject to proportional transaction costs for each type of control. This limiting problem has not been studied as yet. The prospects for explicit solution seem bright, however, since a very similar stochastic control problem was solved completely by Benes-Shepp-Witsenhausen (1980).

In the development of Sections 2-4, we consistently speak in terms of a fixed demand process D and a family of cost structures such that $\varepsilon \downarrow 0$. There is an alternataive interpretation for each of our problems, however. This involves a fixed cost structure and a sequence of demand processes characterized by increasing average demand rates. This alternative interpretation in terms of high volume production systems will be developed in Section 5 for the two-stage problem with lost sales.

It is assumed throughout that $\ell > 0$, $m > 0$, $h \geq 0$, and $\pi > \ell + m$. Relaxing any of these conditions yields a degenerate and uninteresting problem. When we say that a given stochastic process is a (μ,σ) Brownian motion, this means that the drift rate is μ and the variance parameter is σ^2.

2. The Two-Stage Problem with Lost Sales

To state the problem in mathematical terms, we denote by $U(t)$ the cumulative amount of undertime used over the interval $[0,t]$. Thus actual production up to time t is $kt - U(t)$. Let $L(t)$ denote the cumulative amount of potential sales foregone, or cumulative lost sales, over the interval $[0,t]$. Denoting by $Z(t)$ the inventory level at time t, and assuming for simplicity that $Z(0) = 0$, we then have the tautological relationship (inventory = cumulative input - cumulative output)

$$Z(t) = [kt - U(t)] - [D(t) - L(t)], \qquad t > 0 . \tag{6}$$

The undertime process $U = \{U(t), t \geq 0\}$ is of course a manifestation of managerial policy, and we shall shortly specify the set of processes U that constitute legal policy choices. To symmeterize the statement of the control problem, one can also treat the lost sales process $L = \{L(t), t \geq 0\}$ as a policy component, meaning that management may specify the amount of potential sales to be sacrificed, provided that this specification meets certain constraints. Given a capacity choice k, the pair (L,U) is said to be admissible if

(7) L and U are right continuous with left limits (RCLL),

(8) L and U are non-anticipating with respect to D,

(9) L and U are non-decreasing with $L(0) = U(0) = 0$,

(10) $Z(t) \geq 0$ for all $t \geq 0$,

where Z is defined in terms of k and (L,U) by (6). Condition (7) is purely technical in nature, while (9) and (10) are obviously essential physical restrictions. The key condition (8) requires that any decision to work below full capacity or forego potential sales must be based solely on demand information available at the time of the decision. Physical considerations also suggest the restrictions $L(b) - L(a) \leq D(b) - D(a)$ and $U(b) - U(a) \leq k(b-a)$ whenever $0 \leq a < b < \infty$. But one finds that these restrictions are always satisfied by an optimal policy even if omitted from the formal problem statement. Thus they can safely be ignored.

In order to state precisely the objective for our control problem, note first the following. Over any time interval $(a,b]$ the total revenue earned is $\pi\{[D(b) - L(b)] - [D(a) - L(a)]\}$, the total labor cost incurred is $\ell k(b-a)$, the total materials cost incurred is $m\{[kb-U(b)] - [ka-U(a)]\}$, and the total inventory holding cost incurred is

$$h \int_a^b Z(t)dt \ .$$

Thus the expected present value of revenues received minus operating expenses incurred over an infinite planning horizon is

$$(11) \quad E \int_0^\infty e^{-\alpha t}\{\pi[dD(t) - dL(t)] - \ell k dt - m[k dt - dU(t)] - hZ(t)dt\} \equiv V \ ,$$

where the integrals involving $dD(t)$, $dL(t)$ and $dU(t)$ are defined path-by-path in the Riemann-Stieltjes sense. Our objective is to choose a capacity level k and then an admissible pair (L,U) so as to minimize (11). Given k and U, it will of course turn out that the optimal choice for L elects to forego potential sales only when this is necessary to satisfy $Z(t) \geq 0$.

It should be emphasized that the opportunity loss on capital tied up in inventory is fully accounted for by the discounting in (11), so h should include only the direct or out-of-pocket expenses associated with holding inventories. To put it another way, no explicit financial cost of inventory appears in (11), and including any such financial cost in h would be double counting, cf. Problem 2-69 of Hadley-Whitin (1963). It will now be helpful to derive an equivalent objective function in which a financial cost of inventory <u>does</u> appear explicitly. Let

$$I = E\int_0^\infty e^{-\alpha t}(\pi - i - m)dD(t) \quad , \tag{12}$$

this representing the firm's expected present value in the <u>ideal</u> situation where units are produced precisely as demanded, with labor and materials paid for only when needed for such production and no inventories held. Using integration by parts and the fact that $E[D(t)] = \lambda t$ (this is exact for all t), we find that

$$E\int_0^\infty e^{-\alpha t}dD(t) = \alpha E\int_0^\infty e^{-\alpha t}D(t)dt$$

$$= \alpha\int_0^\infty e^{-\alpha t}\lambda t dt = \int_0^\infty e^{-\alpha t}\lambda dt \ ,$$

and thus (12) can be rewritten as

$$I = E\int_0^\infty e^{-\alpha t}\{\pi dD(t) - \ell\lambda dt - m dD(t)\} \ . \tag{13}$$

Now defining the excess capacity

$$\mu = k - \lambda \ , \tag{14}$$

we see from (13) that (11) can be rewritten as

$$V = I - \Delta \tag{15}$$

where

$$\Delta = E\int_0^\infty e^{-\alpha t}\{\pi dL(t) + \ell\mu dt + hZ(t)dt + m[kdt - dU(t) - dD(t)]\} \ . \tag{16}$$

Using (6) and (Riemann-Stieltjes) integration by parts, we have

$$(17)\qquad \int_0^\infty e^{-\alpha t}[k\,dt - dU(t) - dD(t)] = \int_0^\infty e^{-\alpha t}[dZ(t) - dL(t)]$$

$$= \int_0^\infty e^{-\alpha t}[\alpha Z(t)dt - dL(t)] \ .$$

Thus, defining the <u>contribution margin</u>

$$(18)\qquad c = \pi - m \ ,$$

we substitute (17) into (16) to obtain

$$(19)\qquad \Delta = E\int_0^\infty e^{-\alpha t}\{c\,dL(t) + \ell\mu dt + (h+\alpha m)Z(t)dt\} \ ,$$

Obviously Δ represents the amount by which our plan falls short, in expected present value terms, of the ideal profit level I. Equation (19) expresses this shortfall in terms of three distinct effects. First, the excess capacity μ costs us $\mu\ell$ dollars per unit time more than the ideal capacity level λ. (Note that this component of Δ may actually be negative if k is taken smaller than λ.) Second, the contribution margin of c dollars is lost each time a unit of potential sales is foregone. Finally, for each unit of inventory we continuously incur an out-of-pocket expense of h plus an opportunity loss of α times the marginal production cost m.

Since the demand process D is uncontrollable, I is a constant, and thus our original objective (maximizing V) is equivalent to minimizing Δ. To further transform the problem, let us define

(20) $$\varepsilon = h + m\alpha \quad \text{and} \quad \beta = \alpha/\varepsilon \ ,$$

(21) $$\mu^* = \mu\varepsilon^{-1/2} \ ,$$

(22) $$U^*(t) = \varepsilon^{1/2}U(t/\varepsilon),\ L^*(t) = \varepsilon^{1/2}L(t/\varepsilon) \quad \text{and}$$

$$Z^*(t) = \varepsilon^{1/2}Z(t/\varepsilon) \quad \text{for} \quad t \geq 0 \ .$$

Making the change of variable $s = \varepsilon t$ in (19), we find that

(23) $$\beta\varepsilon^{1/2}\Delta = E\int_0^\infty \beta e^{-\beta s}[\ell\mu^* ds + Z^*(s)ds + cdL^*(s)] \equiv \Delta^* \ .$$

Obviously, our original objective is equivalent to minimizing Δ^*, and we may view μ^* and (L^*,U^*) as the objects of choice rather than k and (L,U). To re-express the constraints on (L,U) in convenient form, let us define

(24) $$X(t) = \lambda t - Y(t) \quad \text{and} \quad X^*(t) = \varepsilon^{1/2}X(t/\varepsilon)$$

for $t \geq 0$. The inventory equation (6) is equivalent to $Z^*(t) = X^*(t) + \mu^* t + L^*(t) - U^*(t)$, $t \leq 0$, and thus our constraints (7)-(10) may be equivalently expressed as

(25) L^* and U^* are RCLL

(26) L^* and U^* are non-anticipating with respect to X^*,

(27) L^* and U^* are non-decreasing with $L^*(0) = U^*(0) = 0$,

(28) $Z^*(t) \equiv X^*(t) + \mu^* t + L^*(t) - U^*(t) \geq 0$ for all $t \geq 0$.

We summarize the development up to here as follows.

(29) Proposition. The original two-stage problem with lost sales is equivalent to choosing first μ^* and then (L^*,U^*) so as to minimize Δ^*, defined by (23), subject to the constraints (25)-(28).

Note that the uncontrollable process X^* drives this stochastic control problem through (26) and (28). If demand is deterministic, meaning that $\sigma = 0$ in (1), then the optimal policy is obviously to take $k = \lambda$ (the average demand rate) and $U(t) = L(t) = 0$ for $t \geq 0$. This achieves the ideal profit level of I. Thus one may interpret the minimal value of Δ or Δ^* as a cost of stochastic variability, in expected present value terms.

As in Section 1, we now consider a sequence of cost structures in which

(30) $\alpha \downarrow 0$ and $h \downarrow 0$ but all other cost date remain constant.

This of course implies $\varepsilon \downarrow 0$, and for the moment let us further assume that α and h vanish in such a way that

(31) h/α remains constant, and thus $\beta \equiv (h/\alpha + m)^{-1}$ does as well.

Our several problem transformations have of course been chosen with an eye to the fact that

(32) X^* converges weakly to a $(0,\sigma)$ Brownian motion as $\varepsilon \downarrow 0$ (Donsker's Theorem),

cf. Billingsley (1968), page 68. Thus, the natural diffusion approximation for our production smoothing problem is found by taking X^* to be a $(0,\sigma)$ Brownian motion in (23) and (25)-(28). Taking β to be a fixed positive constant in accordance with (31), let us now consider that limiting problem. After the excess capacity level (or drift rate) μ^* has been chosen, one is left with the instantaneous control problem solved by Harrison-Taylor (1977) and later generalized by Harrison-Taksar (1982). (The term instantaneous control is introduced in the latter paper.) Given a fixed but arbitrary drift rate, the optimal policy (L^*,U^*) enforces a lower reflecting barrier at zero and an upper reflecting barrier at b^*, where b^* is the unique solution of a certain transcendental equation. This means that L^* increases in the minimal amounts necessary to insure $Z^* \geq 0$, while U^* increases in the minimal amounts necessary to insure $Z^* \leq b^*$. In terms of our original problem, the latter constraint is expressed as $\varepsilon^{1/2}Z \leq b^*$, which is in accordance with (5). The optimal barrier height b^*, and thence the minimal objective value Δ^*, are functions of the drift rate μ^* selected initially; one of course wishes to choose μ^* so that Δ^* is minimized. The algebraic expression for Δ^* in terms of μ^* is complicated enough to make this a difficult calculus problem, but numerical solution for the minimizing μ^* value is of course trivial. In terms of the original problem statement, one wants to choose excess capacity level $\mu = \varepsilon^{1/2}\mu^*$, or capacity level $k = \lambda + \varepsilon^{1/2}\mu^*$, as stated earlier in (4).

The question naturally arises whether (31) is the only interesting condition regarding the relative magnitude of physical versus financial costs of holding inventory. Since β is bounded above by $1/m$, it seems

to me that the only other interesting condition is

$$h/\alpha \to \infty \quad \text{and hence} \quad \beta \to 0 \quad \text{as} \quad \varepsilon \downarrow 0 \ ,$$

meaning that α vanishes faster than h and hence the physical costs of holding inventory eventually dominate the financial costs. This causes the interest rate in our transformed objective (23) to vanish, which presumably leads to the instantaneous control problem of Harrison-Taylor (1977) with a minimum long-run average cost criterion. This problem has not been analyzed in the literature, but it is obviously tractable and may be substantially simpler than its discounted analog.

3. The Two-Stage Problem with Overtime

Suppose now that the firm can obtain unlimited amounts of instantaneous overtime production at a premium labor cost of p dollars per unit of production. In order to get an interesting problem, we assume $\ell < p < \pi - m$. The firm then prefers to use overtime rather than lose sales, but it has no motivation to use overtime except as required to avoid lost sales. Having selected a capacity level k, we are then left with a problem of inventory control through usage of overtime and undertime production. Formally, this is identical to the problem treated in Section 2 except that now $L(t)$ is interpreted as the cumulative amount of overtime production used up to time t, and c (the lost contribution margin on sales foregone) must be replaced by p.

Complicating things a bit, suppose that overtime production can only be achieved at a finite rate, and that this rate may not exceed δk

$(0 < \delta < 1)$. If $\delta = 0.2$, for example, this would mean that overtime production during any period cannot exceed 20% of regular-time capacity. When using both regular-time and overtime production in the maximum allowable amounts, the firm's inventory dynamics, ignoring lost sales and undertime, and then given by

$$(34) \qquad \begin{aligned} dZ(t) &= (1+\delta)k\,dt - dD(t) \\ &= \lambda\,dt + \mu\,dt + k\delta\,dt - dD(t) \\ &= dX(t) + \mu\,dt + k\delta\,dt \ . \end{aligned}$$

In terms of the scaled quantities $Z^*(t) = \varepsilon^{1/2}Z(t/\varepsilon)$, $X^*(t) = \varepsilon^{1/2}X(t/\varepsilon)$ and $\mu^* = \mu\varepsilon^{-1/2}$ introduced in Section 2, we can equivalently express (34) as

$$(35) \qquad dZ^*(t) = dX^*(t) + \mu^*dt + (k\delta\varepsilon^{-1/2})dt \ .$$

For small values of ε, this means that by using overtime at the maximum permissible rate, one can effectively achieve instantaneous upward displacements in the scaled inventory process Z^*. As ε vanishes, the rate restriction on overtime production ceases to be a significant features of the problem, and we are reduced to the situation described in the first paragraph of this section.

4. Dynamic Adjustment of Capacity

Let us return to the problem of Section 2 (no overtime production allowed), altered by the assumption that an initial capacity level $k(0)$ is specified, and that this can be varied over time at the expense of

certain workforce smoothing costs. For simplicity, assume that $k(0) = \lambda$ and that a smoothing cost of $q\delta$ is incurred for either an increase or a decrease of size δ in the capacity. (This assumption of symmetric costs is made only to simplify notation. The general case will be discussed shortly.) Denoting by $k(t)$ the capacity level at time t, the expected present value of labor costs plus smoothing costs is then given by

$$E\int_0^\infty e^{-\alpha t}\{\ell k(t)dt + q\left|dk(t)\right|\}\ , \tag{36}$$

where $\int\left|dk(t)\right|$ represents the total variation of $k(\cdot)$. Setting $\mu(t) = k(t) - \lambda$ in the obvious way, we observe that $\mu(0) = 0$ and $\left|dk(t)\right| = \left|d\mu(t)\right|$, so (36) can be rewritten as

$$E\int_0^\infty e^{-\alpha t}\{\ell\lambda dt + \ell\mu(t)dt + q\left|d\mu(t)\right|\}\ . \tag{37}$$

Let $U(t)$, $L(t)$ and $Z(t)$ be defined as in Section 2. Then generalize (21) to

$$\mu^*(t) = \mu(t)\varepsilon^{-1/2}, \qquad t \geq 0\ , \tag{38}$$

and let the scaled processes U^*, L^* and Z^* be defined by (22) again. Both the original statement and the appropriate transformations of our current problem are precisely analogous to those developed in Section 2, so the obvious steps will be skipped. In the end, we arrive at the problem of choosing a scaled excess capacity process $\mu^* = \{\mu^*(t),\ t \geq 0\}$ and a pair (L^*,U^*) so as to minimize

$$(39) \qquad E\int_0^\infty \beta e^{-\beta s}\{\ell\mu^*(s)ds + Z^*(s)ds + cdL^*(s) + q\left|d\mu^*(s)\right|\}$$

subject to the constraints

(40) μ^*, L^* and U^* are RCLL,

(41) μ^*, L^* and U^* are non-anticipating with respect to X^*,

(42) L^* and U^* are non-decreasing with $L^*(0) = U^*(0) = 0$,

$$(43) \quad Z^*(t) \equiv X^*(t) + \int_0^t \mu^*(s)ds + L^*(t) - U^*(t) \geq 0 \quad \text{for all} \quad t \geq 0 .$$

In (39)-(43) we are of course defining $\beta = \alpha/\varepsilon$ and $X^*(t) = \varepsilon^{1/2}X(t/\varepsilon)$ $= \varepsilon^{1/2}[\lambda t/\varepsilon - Y(t/\varepsilon)]$ exactly as in Section 2. Again we imagine a sequence of cost structures with $\varepsilon \downarrow 0$ and all else remaining constant. Invoking Donsker's Theorem, we conclude that the natural diffusion approxmation for (39)-(43) is obtained by replacing X^* with a $(0,\sigma)$ Brownian motion. This will hereafter be called the <u>limiting problem</u>, and $\mu^*(t)$ will be referred to as a <u>drift rate</u> rather than an excess capacity level.

For the limiting problem, the state of the system at time $t \geq 0$ is adequately summarized by the current drift rate $\mu^*(t)$ and the current inventory level $Z^*(t)$. Thus the relevant state space for this stochastic control problem is the half plane

$$S = \{(\mu^*,Z^*) : Z^* \geq 0\}$$

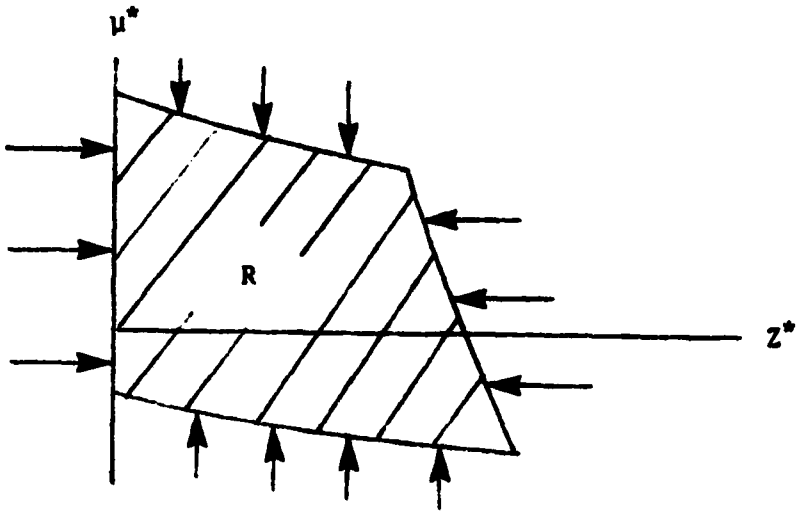


Figure 2. Conjectured Form of Optimal Policy for the Limiting Problem

pictured in Figure 2. In the absence of external control, μ^* remains constant and Z^* evolves as a (μ^*,σ) Brownian motion, meaning that movement is constrained to a horizontal line in Figure 2. The controller can effect instantaneous changes in the value of μ^* at a cost of q per unit of change (either upward or downward). Also, he can instantaneously increase Z^* at a cost of c per unit of increase, and can instantaneously decrease Z^* without cost. Finally, runnings costs are continuously incurred at rate $Z^*(t) + \ell\mu^*(t)$. From the development in Section 2 of Harrison-Taylor (1977), readers will see that this problem is equivalent to one having no running costs (or holding costs) associated with Z^* but a positive reward associated with decreases in Z^* enforced by the controller. In the discussion to follow, however, we shall continue to speak in terms of the cost structure embodied in the minimand (39).

Because its state descriptor is two-dimensional, our limiting stochastic control problem looks very difficult, but the potential saving grace is that changes in μ^* (movement in the vertical dimension of Figure 2) occur only as a result of exogenously imposed controls. Put another way, random disturbances only affect the Z^* component in Figure 2. In this regard, our problem is like the finite-fuel follower problem solved explicitly in the beautiful paper by Benes-Shepp-Witsenhausen (1980). Looking at the results of that analysis and of Harrison-Taylor (1977), one is led to conjecture an optimal policy of the form pictured in Figure 2. Here we have a four-sided control region R (cross-hatched in the figure), and no action is taken so long as the state of the system remains within R. Rightward displacement of Z^* (lost sales) is used to create a reflecting barrier on the left side of R, downward displacement of μ^* (capacity reduction) is used to create a reflecting barrier on the upper side of R, leftward displacement of Z^* (undertime) is used to create a reflecting barrier on the right side of R, and upward displacement of μ^* (capacity increase) is used to create a reflecting barrier on the lower side of R. In each case, the term reflecting barrier means that the control in question is employed in the minimal amounts necessary to insure that the process $[Z^*(t),\mu^*(t)]$ does not cross over its corresponding boundary section. (The paper by Benes-Shepp-Witsenhausen contains a lengthy discussion of such reflecting barriers.) If one starts at the origin ($\mu^* = Z^* = 0$) and uses the particular control region pictured in Figure 2, then the controlled process (Z^*,μ^*) will never leave the axis $\mu^* = 0$. To prove that an optimal policy has the form pictured in Figure 2, and to determine the control surfaces explicitly, is certainly not an easy problem, but it may just be within the reach of currently available techniques.

If overtime production is allowed at any given capacity level, one arrives at the same formal problem, with $L(t)$ now interpreted as cumulative overtime used up to time t, just as in Section 2. If there are different proportional costs associated with increasing and decreasing the capacity level, then notation gets more complicated, but all the ideas are the same. One must represent $k(t)$ as the difference of two non-decreasing processes, say $k(t) = A(t) - B(t)$, representing cumulative capacity increases and cumulative capacity decreases respectively. The last term in (36) is then replaced by two terms, one involving $dA(t)$ and the other involving $dB(t)$. In the limit problem, this of course results in different costs associated with upward and downward displacement of μ^*.

5. High Volume Systems

For ease of exposition, let us assume throughout this section that the physical holding cost h is zero. Ostensibly, the analysis of Section 2 involves a fixed demand process and a vanishing interest rate, but another important interpretation is available. After redefining the units in which production, time and cost were to be measured, we arrived at the equivalent minimand

$$(44) \qquad E\int_0^\infty \beta e^{-\beta s}[\lambda\mu^* ds + Z^*(s)ds + c dL^*(s)] \equiv \Delta^* .$$

In terms of these units, the interest rate is β, which we assume constant, and the cumulative demand process $D^*(s) \equiv c^{1/2}D(s/c)$ has the form

$$(45) \qquad D^*(s) = \lambda^* s + X^*(s), \qquad s \geq 0 ,$$

there $\lambda^* \equiv \lambda\varepsilon^{-1/2}$ and X^* is defined as before. As a consequence of our assumptions, we have

(46) $\lambda^* \to \infty$ and X^* converges weakly to a $(0,\sigma)$ Brownian motion.

For an alternate interpretation of the entire development, one may accept as natural the units of measurement in (44), fix the interest rate β, and directly hypothesize a family of demand processes of the form (45)-(46). Such a family displays a constant degree of stochastic variability around an ever increasing average demand rate λ^*. With the linear cost structure hypothesized in Section 2, we conclude that for large values of λ^*, a nearly optimal policy is to

(47) set the production capacity at $\lambda^* + \mu^*$, and

(48) use undertime as necessary to keep the inventory level below b^*,

where μ^* and b^* are computed as in Section 2. A key point is that μ^* and b^* are constants (do not depend on λ^*). Thus as λ^* grows, (47) leads to an increasingly well-balanced system, and average inventory is a smaller and smaller fraction of sales volume by (48).

For all intents and purposes, the policy recommendation (47)-(48), with μ^* and b^* computed as in Section 2, amounts to a proposal that cumulative demand be modeled as a Brownian motion with drift λ^* and variance σ. It is important to recognize that this recommendation hinges critically on the assumption that λ^* is large. If λ^* is moderate relative to σ, there is no reason to believe that a (λ^*,σ) Brownian motion can provide a good approximation for the non-decreasing cumulative demand process.

References

[1] V. Benes, L. A. Shepp and H. M. Witsenhausen (1982), "Some Solvable Stochastic Control Problems," Stochastics, Vol. 4, 39-83.

[2] P. Billingsley (1968), Convergence of Probability Measures, Wiley, New York.

[3] G. Hadley and T. M. Whitin (1963), Analysis of Inventory Systems, Prentice-Hall, Englewood Cliffs, N.J.

[4] J. M. Harrison and A. J. Taylor (1977), "Optimal Control of a Brownian Storage System," Stoch. Proc. Appl., Vol. 6, 179-194.

[5] J. M. Harrison and M. I. Taksar (1982), "Instantaneous Control of Brownian Motion," Math. Ops. Rsch,. to appear.

[6] C. C. Holt, F. Modigliani, J. F. Muth and H. A. Simon (1960), Planning Production, Inventories and Work Force, Prentice-Hall, Englewood Cliffs, N.J.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 205	2. GOVT ACCESSION NO.: AD-A120645	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): A DIFFUSION MODEL OF INVENTORY AND PRODUCTION CONTROL		5. TYPE OF REPORT & PERIOD COVERED: Technical Report
		6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): J. Michael Harrison		8. CONTRACT OR GRANT NUMBER(s): N00014-75-C-0561
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Department of Operations Research and Department of Statistics, Stanford University Stanford, California 94305		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (NR-047-200)
11. CONTROLLING OFFICE NAME AND ADDRESS: Operations Research, Code 434 Office of Naval Research Arlington, Virginia 22217		12. REPORT DATE: August 1982
		13. NUMBER OF PAGES: 23
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report): Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION IS UNLIMITED

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Also supported in part by National Science Foundation Grant ECS 80-17867 Department of Operations Research, Stanford University and issued as Technical Report No. 67.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

INVENTORY THEORY
PRODUCTION PLANNING
DIFFUSION APPROXIMATIONS
BROWNIAN MOTION
STOCHASTIC CONTROL

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(See next page)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

A DIFFUSION MODEL OF INVENTORY AND PRODUCTION CONTROL

by J. Michael Harrison

Report No. 205

Abstract. We consider the diffusion limits of several closely related production planning problems. Each involves a make-to-stock producer who faces IID demands over a sequence of future periods. In the simplest formulation, a production capacity or workforce level must be fixed at time zero, and thereafter the actual production rate is adjusted dynamically in response to inventory fluctuations. The capacity decision fixes certain operating costs and constrains subsequent decisions regarding production rates. Fixing the demand process, we consider a sequence of cost structures in which the total inventory carrying cost approaches zero. (This requires that both the physical cost of carrying inventory and the interest rate earned on external investment vanish.) Under this condition, the production planning problem approaches a two-stage optimal control problem for Brownian motion. The first stage of the limiting problem involves drift rate selection for a Brownian motion, and its second stage is the instantaneous control problem formulated and solved by Harrison-Taylor (1977). With small holding costs, we find that an optimal capacity decision calls for near equality of the average production and demand rates, which in turn justifies a diffusion approximation for the subsequent problem of inventory control through production rate adjustment. A novel and important feature of this analysis is that no assumption of system balance, or heavy traffic, is imposed a priori. Instead, it is shown that optimal system design will create such balanced loading under certain economic conditions. Under the same assumption of small holding costs, a production planning problem with dynamic capacity adjustment is shown to yield a more complex diffusion limit, which has not been studied as yet. Finally, it is shown that our asymptotic analyses apply equally well to a sequence of production planning problems with fixed cost structure and increasing average demand rate.