

Note 16: Time-varying, Stochastic Demand: Policy Optimization

We now envision a world of discrete time. There is a sequence of *time points* $t = 0, 1, \dots, T$. The *horizon* T may be finite or infinite. *Time period* t is the interval from point t until just before point $t + 1$. Denote

$z(t) =$ order size at time t ;

$d(t) =$ demand at time t , $t = 0, 1, \dots, T - 1$.

The demand $d(t)$ is now a nonnegative *random variable* with finite mean. Also, the $d(t)$ are *independent* over t . For convenience, $d(t)$ is *continuous*, as is the control variable $z(t)$. Because demand is random, we must allow backorders, so the state variable is

$x(t) =$ net inventory at time t , $t = 0, 1, \dots, T$.

The initial net inventory is the constant x_0 .

The sequence of events at each time $t < T$ is as follows:

1. We observe the net inventory $x(t)$.
2. We decide the order size $z(t)$.

Then, sometime during period t , the demand $d(t)$ occurs and the order $z(t)$ arrives. The system dynamics are described by the following equations:

$$x(t+1) = x(t) + z(t) - d(t), \quad t = 0, 1, \dots, T. \quad (1)$$

At the horizon T we observe $x(T)$, but this is all; there is no order $z(T)$ or demand $d(T)$.

What we want is not a set of fixed decisions, but rather an order *policy*, a rule for making decisions based on current information. One of our primary goals is *to determine which type of policy is best*, given the economics of the situation.

The cost parameters are

$k(t) =$ fixed order cost at time t ;

$c(t) =$ unit variable order cost at time t ;

$h(t) =$ inventory-holding cost rate at time t ;

$b(t) =$ backorder-penalty cost rate at time t ;

$\gamma =$ discount factor, $0 < \gamma \leq 1$.

Future costs are discounted at rate γ .

Letting $\delta(z)$ denote the Heavyside function (1 when $z > 0$, 0 otherwise), the total order cost at time t is $k(t)\delta(z(t)) + c(t)z(t)$. The inventory-holding or backorder-penalty cost is assessed on $x(t)$ at step 1 of time t . This cost can be written compactly as $\hat{C}(t, x(t))$, where

$$\hat{C}(t, x) = h(t)[x]^+ + b(t)[x]^-.$$

Denote

$$y(t) = x(t) + z(t).$$

This is the inventory position just after ordering at step 2. Thus, (1) is equivalent to

$$x(t+1) = y(t) - d(t), \quad t = 0, 1, \dots, T-1.$$

Define the function

$$C(t, y) = E[\hat{C}(t+1, y - d(t))].$$

Then, $C(t, y(t))$ measures the *expected* inventory-backorder cost, as viewed from step 2. We call $C(t, \cdot)$ the *one-period cost function*. When T is finite, we also adopt the terminal cost $-c(T)x(T)$.

Now consider the case with $k(t) = 0$. Thus, the cost to order z at time t is the *linear* function $c(t)z$, $z \geq 0$. There are several interrelated decisions to make. In choosing $y(0)$ we need to consider the costs in future periods, not just the current one. Worse, when we reach point $t = 1$, we shall choose $y(1)$, which also affects future costs. But now, at $t = 0$, we haven't yet made that choice. So, how can we intelligently select $y(0)$? And then there is $y(2)$, which should probably depend on $y(1)$ and $y(0)$, or maybe it's the other way around... We seem to be trapped!

Fortunately, there is a way out. This is precisely the sort of dilemma addressed by *dynamic programming*. The key idea is to think *backward in time*. The fact that this approach works is sometimes called the *principle of optimality*. It should be intuitively clear, and it can be proven rigorously.

Here is a precise statement of the approach: Define the following functions for $t = 0, 1, \dots, T$:

$$\begin{aligned} V(t, x) &= \text{minimal expected discounted cost in periods } t, t+1, \dots, T, \\ &\text{assuming period } t \text{ begins with } x(t) = x. \end{aligned}$$

We compute these functions recursively. In the process, we obtain the optimal policy. First,

$$V(T, x) = -c(T)x. \tag{2}$$

Suppose we have determined $V(t+1, x)$, along with the optimal policies for points $t+1$ through $T-1$. To address the problem at time t , given $x(t) = x$, define

$$H(t, y) = c(t)y + C(t, y) + \gamma E[V(t+1, y - d(t))]. \tag{3}$$

This quantity measures all the relevant costs if we choose $y(t) = y$. The first two terms (minus the constant $c(t)x$) represent the costs at time t itself. The last term is the expected value of all future costs, assuming we act optimally in the future, since $x(t+1) = y - d(t)$.

The problem at point t , then, is to choose y to minimize $H(t, y)$, subject to $y \geq x$. The solution to this problem for each x gives the optimal policy at time t . Then set

$$V(t, x) = -c(t)x + \min\{H(t, y) : y \geq x\}. \quad (4)$$

Now, using $V(t, \cdot)$, we can proceed to compute $V(t-1, \cdot)$ in the same way, then $V(t-2, \cdot)$, and so on.

Equations (3) and (4), together with the boundary or terminal condition (2), describe a recursive scheme to compute the optimal-cost functions for all time points, as well as the optimal policy. The full optimal policy has a simple, appealing structure.

Theorem 1 *For all t :*

- (a) $H(t, y)$ is a convex function of y .
- (b) A base-stock policy is optimal. The optimal base-stock level $s^*(t)$ is the smallest value of y minimizing $H(t, y)$.
- (c) $V(t, x)$ is a convex function of x .

We can prove this by induction on t .

We can obtain useful and interesting information about the $s^*(t)$ by means of a slight transformation of the problem: Define

$$\begin{aligned} V^+(t, x) &= c(t)x + V(t, x), & t \leq T; \\ c^+(t) &= c(t) - \gamma c(t+1); \\ C^+(t, y) &= \gamma c(t+1)E[d(t)] + c^+(t)y + C(t, y), & t < T. \end{aligned}$$

The following recursion is equivalent to (2) to (4):

$$\begin{aligned} V^+(T, x) &= 0; \\ H(t, y) &= C^+(t, y) + \gamma E[V^+(t+1, y - d(t))]; \\ V^+(t, x) &= \min\{H(t, y) : y \geq x\}. \end{aligned}$$

We can write

$$V^+(t, x) = H(t, \max\{s^*(t), x\}).$$

Now, C^+ plays the role of the current period's cost in this transformed model. Let $s^+(t)$ be the smallest value of y that minimizes $C^+(t, y)$. The corresponding base-stock policy minimizes the current cost while ignoring the future, so we call it the *myopic policy*. The myopic and optimal policies are closely related. But we omit the details here.

Suppose the horizon T is infinite. Assume that the problem data are stationary in time. So, $c(t) = c$, $C(t, y) = C(y)$, etc., and the $d(t)$ all have the same distribution. Let d denote a generic demand random variable. Define

$$c^+ = (1 - \gamma)c;$$

$$C^+(y) = \gamma c E[d] + c^+ y + C(y).$$

Let s^* be the smallest value of y minimizing $C^+(y)$, that is, s^* solves

$$F_d^0(y) = \frac{c^+ + h}{b + h}.$$

Assume that the cost ratio lies strictly between 0 and 1, so s^* is finite. Thus, s^* is the myopic base-stock policy.

Theorem 2 *The stationary base-stock policy with base-stock level s^* is optimal.*

Now we allow positive fixed costs $k(t)$. It turns out that an (r, s) (or (s, S)) policy is optimal at each time point. To prepare for later developments, we introduce a property, called k -convexity. Let $f(x)$ be a function and m a nonnegative number. We say that f is m -convex if, for all x and all positive numbers ξ and ν ,

$$f(x) + \xi \left[\frac{f(x) - f(x - \nu)}{\nu} \right] \leq f(x + \xi) + m. \quad (5)$$

If f is differentiable, the following condition is equivalent: For all x and all positive ξ ,

$$f(x) + \xi f'(x) \leq f(x + \xi) + m. \quad (6)$$

Let us try to understand this property intuitively: If f is actually convex, then (5) holds for $m = 0$; 0-convexity is equivalent to convexity itself. Also, the condition becomes weaker as m grows; that is, if f is m_1 -convex and $m_2 > m_1$, then f is also m_2 -convex. The expression on the left of (5) or (6) can be viewed as a linear approximation of $f(x + \xi)$ for fixed x . When $m = 0$ and f is convex, we recover the familiar fact that the linear approximation *underestimates* the true value $f(x + \xi)$. When $m > 0$, the approximation may overestimate $f(x + \xi)$, but *the error is bounded* above by m .

We state some additional facts:

- (a) If $f(x)$ is m -convex, then so is the shifted function $f(x + \Psi)$, for any fixed Ψ .
- (b) If f_1 is m_1 -convex, f_2 is m_2 -convex, and α_1 and α_2 are positive numbers, then the function $f = \alpha_1 f_1 + \alpha_2 f_2$ is m -convex, where $m = \alpha_1 m_1 + \alpha_2 m_2$.
- (c) If \tilde{f} is m -convex, and $f(x) = E[\tilde{f}(x - d)]$, then f is m -convex.

Now we consider the following problem:

$$\begin{aligned} &\text{Minimize} && k\delta(y - x) + H(y) \\ &\text{subject to} && y \geq x. \end{aligned} \quad (7)$$

Theorem 3 *Let H be any continuous, k -convex function in problem (7). The optimal policy is an (r, s) policy with parameters (r^*, s^*) , where s^* is the smallest y minimizing $H(y)$, and r^* is the largest $x \leq s^*$ satisfying*

$$H(x) = H(s^*) + k.$$

Proof: Clearly, it is optimal not to order for $x \geq s^*$. Also, for $r^* < x < s^*$, we have, by the definition of r^* ,

$$H(x) < H(s^*) + k,$$

so again it is optimal not to order. So, we need only show that it is optimal to order up to s^* , i.e.,

$$H(x) \geq H(s^*) + k,$$

for all $x \leq r^*$. But, suppose this inequality is violated at $x = r^* - \nu$ for some $\nu > 0$, and set $\xi = s^* - r^*$. Then, $H(r^*) > H(r^* - \nu)$, so

$$H(r^*) + \xi \left[\frac{H(r^*) - H(r^* - \nu)}{\nu} \right] > H(r^*) = H(r^* + \xi) + k.$$

This violates the k -convexity of H at r^* . ■

The dynamic-programming formulation for the current problem is:

$$V(T, x) = -c(T)x;$$

$$H(t, x) = c(t)y + C(t, y) + \gamma E[V(t+1, y - d(t))];$$

$$V(t, x) = -c(t)x + \min\{k\delta(y - x) + H(t, y) : y \geq x\}.$$

Theorem 4 *For all t ,*

(a) *$H(t, y)$ is a continuous, k -convex function of y .*

(b) *The optimal policy for point t is an (r, s) policy; the target stock level $s^*(t)$ is the smallest value of y minimizing $H(t, y)$, and the reorder point $r^*(t)$ is the largest value of $x \leq s^*(t)$ satisfying*

$$H(t, x) = k + H(t, s^*(t)).$$

(c) *$V(t, x)$ is a continuous, k -convex function of x .*

We have myopic-policy and infinite-horizon results that are comparable to those for the linear-cost case.