UNITED ARAB REPUBLIC

THE INSTITUTE OF
NATIONAL PLANNING

Memo.No. 783

An Introduction to Dynamic Programming

by

Dr. Nadia Makary

June 1967

CAIRO
3 MOHAMED MAZHAR - ZAMALEK

"Opinions Expressed and Positions Taken by Authors are Entirely their Own and do not Necessarily Reflect the Views of the Institute of National Planning."

Table of Contents :

Preface:

Dynamic programming is an approach for analysing multistage decision processes and finding out the structure of the optimal policy. This note is a simple introduction to this approach. It first gives a general description of the situations where dynamic programming may be applied. Then, a number of examples is given to illustrate these situations and to classsify the dynamic programming technique.

An Introduction to Dynamic Programming

What is dynamic programming?

Dynamic programming is an approach for analysing multi-stage decision processes and finding out the structure of the "policy" that maximizes (or minimizes) a predefined income (or cost) function.

Historically, it was developed mainly through Bellman's papers (1950's) as a result of his trails to solve certain kinds of "programming" problems involving time as a significant element. However, the dynamic programming approach is used for analysing many static processes that can be formulated as dynamic programming processes.

In contrast to linear programming, there is no unique set up for dynamic programming problems. Yet, there are certain features common to all problems that can be solved by the dynamic programming approach.

A general description of the situation where the dynamic programming approach can be applied may be presented as follows:

A system may be found in one of a possible number of states. At each state there is a number of possible actions. By choosing any of these actions, i.e., by making a decision, the system moves from one state to another in either a deterministic or a stochastic way. Consequently, a certain income (or cost) is earned(or paid). The process continues for either finite or infinite number of times.

The sequence of decisions should be specified in such a way to maximize (or minimize) the total expected income (or cost). A discount factor may be introduced to assure that the total expected income (or cost) is finite even if the process continues infinitely. It may also be introduced in finite processes if decisions are made at successive time periods and the present value differs from the future value.

The elements of the dynamic programming problem:

S : set of states.

A : set of prossible actions,
Noticethat the action may depend on the state.

q : "the law of motion" of the system. It associates with each pair (s, a) a probability distribution on S: q(./s,a). In the deterministic case $q(s'/(s, a))$ equals one for a specified state $s' = s_o$ and equals zero for any other state $s' \neq s_o$.

i(s,a): the immediate return function. It determines the income (or cost) if the system is in state s and action a is chosen.

β : $0 \leq \beta \leq 1$; the discount factor.

The following definitions will be used:
To make a decision: is to choose one of the possible actions.
A policy : is a sequence of decisions.
An optimal policy: is a policy that maximizes (or minimizes) the total expected discounted income (or cost).

Thus the dynamic programming problem as defined above is solved if the structure of the optimal policy is known.

The solution procedure is a direct opplication to the "optimality principle". This principle, as Bellman defined it, says: "an optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision". Or, in otherwords: given the current state, an optimal policy for the remaining stages is independent of the policy adopted in previous stages. The direct result of using this principle is the development of the functional equation technque which gives the recurrence relation between the optimal value of the return function in successive stages, and thus identifies the optimal policy for each state with n stages remaining given the optimal policy for each state with (n-1) stages remaining.

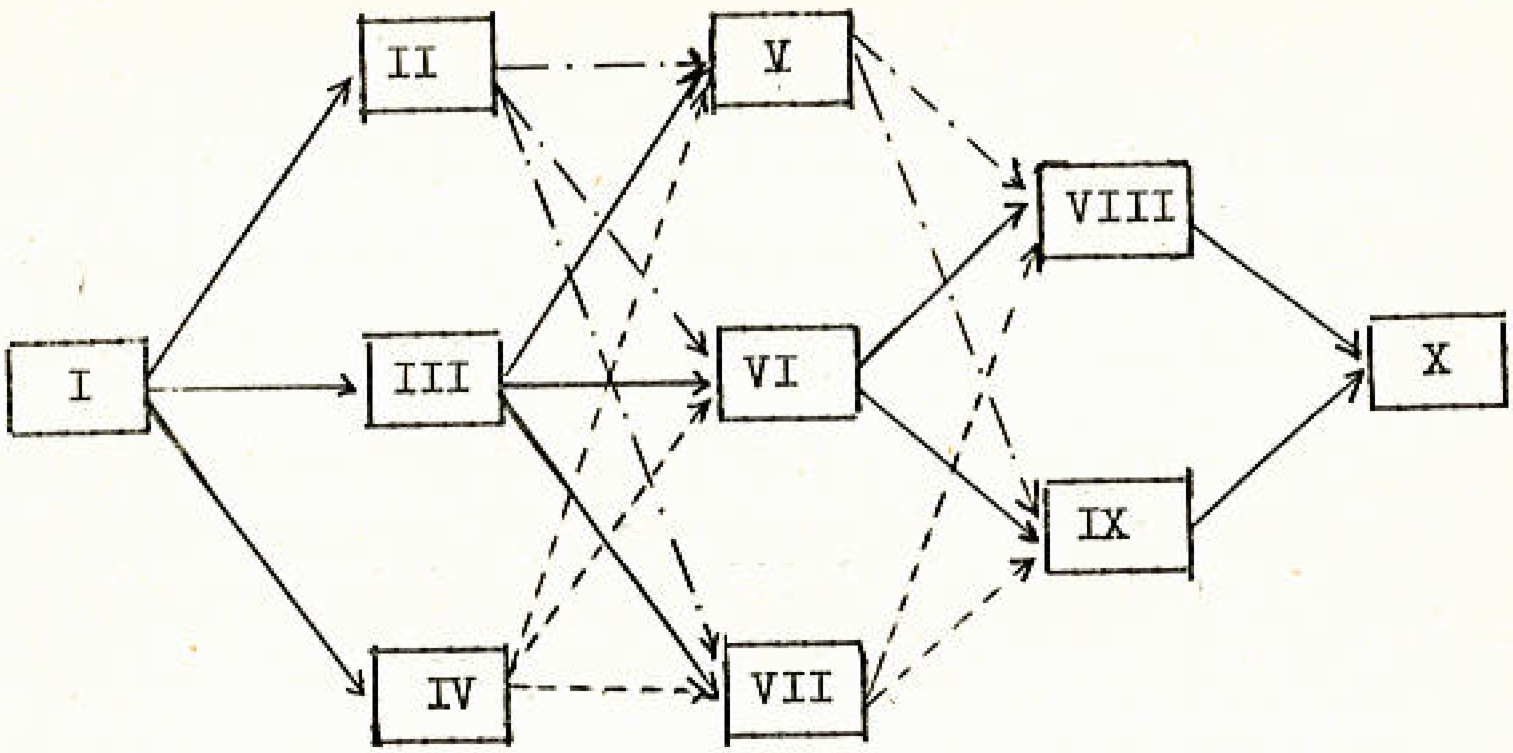
The dynamic programming approach has been successfully applied to a wide variety of problems in different fields. In what follows, a number of examples will be discussed. Some of them are given mainly to clarify the dynamic programming formulation and technique, others are presented to give an idea of some possible applications.

Example 1 (*):

(finite-stage deterministic process.)

A person wants to travel from city I to city X as fast as possible. Owing to the long distance between the two cities, he has to make several stops before reaching his destination. At each stop he can choose the route to the next stop as shown in the diagram :

(*) This example is known in the literature by: "Statecoach Problem"



The number of hours necessary to go from one stop to the next depends on the route he chooses as given in the following tables:

from \ to	II	III	IV
I	2	4	3

from \ to	V	VI	VII
II	7	4	6
III	3	2	4
IV	4	1	5

from \ to	VIII	IX
V	1	4
VI	6	3
VII	3	3

from \ to	X
VIII	3
IX	4

Which routes should he choose in order to go from I to X in the minimum number of hours?

Trail and error may be used for solving this problem, but the dynamic programming approach provides an easier and more systematic solution.

<u>The dynamic programming formulation of the problem:</u>

S = {. , II, ..., X} , i.e., each city represents a possible state,

let a(s) denote the action of going from state s to a new state a(s), then:

A = { a(s)} .

For example: a (II) = a (III) = a(IV) = V or VI or VII.

i(s,a) = the immediate "cost" of being in state s and choosing action a(s) (i.e., the member of hours needed to go from s to a(s)) .

For example : i(V,IV) = 4.

Take $\beta = 1$.

Notice that the traveller has to make four successive decisions. Each time he is confronted by a decision, will be called a "stage". So this problem is a 4-stage decision problem.

The stages will be numbered in a backward order. So at the first stage the person is either in state VIII or in state IX and he should make the decision X, while at the fourth stage, he is in state I and has to choose one of the actions II or III or IV.

Now, what is the optimal policy? In other words, what is the sequence of routes that minimizes the total "cost"?

<u>The solution</u> :

Let $f_n(s, a(s))$, n = 1, 2, 3, 4 denote the total "cost" of being in state s (at the n^{th} stage), taking action a(s) and following the optimal policy in the remaining (n-1) stages.

If $f_n(s)$ denotes the total "cost" of being in state s (at the n^{th} stage), taking the optimal action $a^*(s)$ and following the optimal policy from there on, then :

$$f_n(s, a(s)) = i\ (s, a(s)) + f_{n-1}(a(s))\ ,\ \text{and}$$

$$f_n(s) = \min_{a(s)} \left\{ i(s, a(s)) + f_{n-1}\ (a(s)) \right\} .$$

For n = 1, the only available action is X and:

f_1 (VIII) = 3,
f_1 (IX) = 4,
with a^*(VIII) = a^*(IX) = X

Suppose now that the traveller is in state VI. If he decides to go to VIII his total "cost" will be :

$$f_2\ (\text{VI, VIII}) = i\ (\text{VI, VIII}) + f_1\ (\text{VIII}) = 6 + 3 = 9.$$

But if be chooses to go to IX, his total cost will be :

$$f_2\ (\text{VI, IX}) = i(\text{VI, IX}) + f_1\ (\text{IX}) = 3 + 4 = 7.$$

Thus, f_2 (VI) = 7 and a^*(VI) = IX.

Similarly, f_2 (s) and a^*(s) can be calculated for every state in the second stage. The consequent results are:

s \ a(s)	f_2(s, a(s)) VIII	IX	f_2(s)	a^*(s)
V	4	8	4	VIII
VI	9	7	7	IX
VII	6	7	6	VIII

With this information about the optimal policy for each state in the second stage, the optimal decision in each state of the third stage can be found. For example, in state II :

f_3 (II, V) = i (II, V) + f_2 (V) = 7+4 = 11

f_3 (II, VI) = i (II, VI) + f_2 (VI) = 4+7 = 11

f (II,VII) = i (II, VII) + f_2 (VII) = 6+6 = 12

∴ f_3 (II) = 11 and a^* (II) = either V or VI.

Proceeding similarly for states III and IV yields the following results for the three-stage problem:

s \ a(s)	f_3 (s, a(s)) V	VI	VII	f_3(s)	a^*(s)
II	11	11	12	11	V or VI
III	7	9	10	7	V
IV	8	8	11	8	V or VI

In exactly the same way, the optimal decision for the only state in the fourth stage is found; from the following table; to be either III or IV:

s \ a(s)	f_4 (s,(s)) II	III	IV	f_4(s)	a^*(s)
I	13	11	11	11	III or IV

These results show that at the initial state I, the person should go to either III or IV. If he chooses III then he should go from there to V. From V he should go to VIII and from there to X. His total "cost" (i.e., the total number of hours he spends to get from I to X) if be follows this policy is 11 and it is the minimum possible cost. There are two more optimal routes that be can follow and still spends only 11 hours. These alternative routes are:

I ⟶ IV ⟶ V ⟶ VIII ⟶ X ,

and I ⟶ IV ⟶ VI ⟶ IX ⟶ X .

Example 2 : "Gold-mining"

(Finite stage stochastic process)

There are two gold mines: F and G. The amount of gold in the first is x and in the second is y. We want to get as much gold as possible from these two mines. But we have only one gold-mining machine which has the property that if used to mine gold in F, there is a probability $P_1 (0 < P_1 < 1)$ that it will mine a fraction $r_1, (0 < r_1 < 1)$ of the gold there and remain in working order, and a probability $(1-p_1)$ that it will mine no gold and be damaged beyond repair. The corresponding probabilities if it is used to mine gold in G are P_2 and $(1-P_2)$ with a fraction r_2 $(0 < P_2, r_2 < 1)$.

The process begins by using the machine in either For G. If the machine is undamaged, another choice for using the machine in either For G is made. The process continues in this way for N times if the machine is undamaged, otherwise the process terminates when the machine is damaged.

What sequence of choices maximizes the amount of gold mined before the end of the process?

The dynamic programming formulation:

$$S = \left\{ s=(\alpha, \gamma) : \alpha = (1-r_1)^k x, \quad \gamma = (1-r_2)^{\ell} y \ ; \quad k, \ell = 0, \dots, n \ ; \ k+\ell = n \ ; \ n = 0, \dots N-1 \right\}$$

$A = \{F, G\}$, where : a = F means that mine F is to be mined and a=G means that mine G should be mined.

$$q : \quad q(s'/(\alpha,\gamma), F) = \begin{cases} P_1 & \text{if } s' = ((1-r_1)\alpha, \gamma) \\ 1-P_1 & \text{if } s' = (\alpha, \gamma) \end{cases}$$

$$q(s'/(\alpha,\gamma), G) = \begin{cases} P_2 & \text{if } s' = (\alpha, (1-r_2)\gamma) \\ 1-P_2 & \text{if } s' = (\alpha, \gamma) \end{cases}$$

$$i(s,a)\ :\ i(s,F) = \begin{cases} r_1\alpha & \text{with probability } P_1 \\ 0 & \text{with probability } 1-P_1 \end{cases}$$

$$i(s,G) = \begin{cases} r_2\gamma & \text{with probability } P_2 \\ 0 & \text{with probability } 1-P_2\ . \end{cases}$$

N is the number of stages and they are numbered in a backward order.

Take $\beta = 1$.

What is the optimal policy? i.e., what is the sequence of choices that maximizes the total expected amount of gold mined?

The solution:

Let : $f_n(s,a)=$ the total expected amount of gold mined before the end of the process if the system in state s (at the n^{th} stage), action a is taken, and an optimal policy is followed in the remainign (n-1) stages.

$f_n(\ s\ )=$ the total expected amount of gold mined if the system, at the n^{th} stage, is in state s, the optimal action a^* is chosen, and an optimal policy is followed in the remaining (n-1) stages.

$$\therefore\ f_n(s,a) = E\left\{ i(s,a) + f_{n-1}(s')\right\}, \text{ and}$$

$$\begin{aligned} f_n(\ s\ ) &= \max_a E\left\{ i(s,a) + f_{n-1}(s')\right\} \\ &= \max\Big[E\left\{ i((\alpha,\gamma),F) + f_{n-1}((1-r_1)\alpha,\gamma)\right\}, \text{ and} \\ &\qquad E\left\{ i((\alpha,\gamma),G) + f_{n-1}(\alpha,(1-r_2)\gamma)\right\}\Big] \\ &= \max\Big[P_1\left\{ i((\alpha,\gamma),F) + f_{n-1}((1-r_1)\alpha,\gamma)\right\} + (1-P_1)\cdot(0) \text{ and,} \\ &\qquad P_2\left\{ i((\alpha,\gamma),G) + f_{n-1}(\alpha,(1-r_2)\gamma)\right\} + (1-P_2)(0)\Big] \\ &= \max\Big[P_1\left\{ i((\alpha,\gamma),F) + f_{n-1}((1-r_1)\alpha,\gamma)\right\}, \text{ and} \\ &\qquad P_2\left\{ i((\alpha,\gamma),G) + f_{n-1}(\alpha,(1-r_2)\gamma)\right\}\Big] \end{aligned}$$

By applying this recurrent relation we can find the optimal policy and also the expected amount of gold mined if this policy is followed.

A numerical illustration:

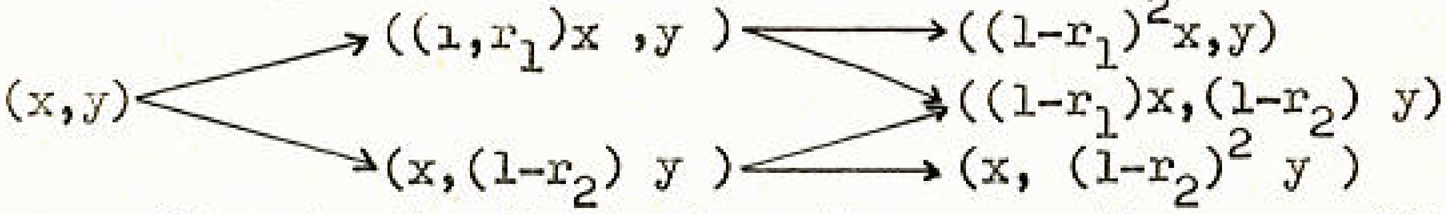
Suppose : $N = 3$

$X = 10.0$ $\qquad Y = 12.0$

$P_1 = 0.75$ $\qquad P_2 = 0.50$

$r_1 = 0.40$ $\qquad r_2 = 0.60$

The states at the three recursive stages are given in the following diagram :



Thus in the first stage there are three possible states.

Since

$$f_1(s,a) = E\, i(s,a)$$

$$= \begin{cases} P_1\ i(s,F) & \text{if } a = F \\ P_2\ i(s,G) & \text{if } a = G \end{cases}$$

$$\therefore f_1(s) = \max \{P_1\ i(s,F) \text{ and } P_2\ i(s,G)\} .$$

Therefore, we can find the optimal choice for each possible state in the first stage by calculating $f_1(s)$ and knowing the corresponding a^* :

	$f_1(s, a)$			
s \ a	F	G	$f_1(s)$	a^*
$(1-r_1)^2 x,y$	1.08	3.6	3.6	G
$(1-r_1)x, (1-r_2)y$	1.8	1.44	1.8	F
$x, (1-r_2)^2y$	3.0	0.5	3.0	F

If the optimal policy is followed in the first stage, the total expected amount of gold mined in the second stage will be given by $f_2(s,a)$. For example, if $s=((1-r_1)x,y)$ and $a=F$, then:

$$f_2\ ((1-r_1)x,y),\ F) = P_1 r_1 (1-r_1)\ x + P_1 f_1\ ((1-r_1)^2 x,\ y)$$
$$= 1.8 + (0.75)\ (3.6)$$
$$= 4.5$$

After calculating $f_2(s,a)$ for all possible combinations (s,a) the value of $f_2(s)$ can be determined. Consequently, the a^* for each possible state in the second stage can be found as shown below:

	$f_2(s,a)$			
s \ a	F	G	$f_2(s)$	a^*
$(1-r_1)x,y$	4.5	4.5	4.5	F or G
$x,(1-r_2)y$	4.25	2.95	4.25	F

Proceeding as before, we get the following table for the third stage:
(For example :

$$f_3(s,F) = P_1 r_1 x + P_1 f_2((1-r_1)\ x,y)$$
$$= 3 + (0.75)\ (4.5) = 6.375)$$

	$f_3\ (s,a)$			
s \ a	F	G	$f_3(s)$	a^*
(x,y)	6.375	5.725	6.375	F

So, the optimal policy for the given three-stage probelm starts with choosing mine F first, and if the machine is undamaged (the resulting state is $((1-r_1)x,y)$) the following choice may be For G. If F is choosen and if the machine is undamaged (this leads to the state $((1-r_1)^2x,y)$), the next choice should be G. On the other hand, if G is choosen on the second stage and if the machine is undamaged, (the resulting state is $((1-r_1)x,(1-r_2)y)$) then the following choice should be F.

Thus, there are two optimal polices:
F F G and F G F. The maximum expected amount of gold mined if any of these two polices is followed, and if the machine is undamaged, equals 6.375.

<u>Example 3</u> : "Gold-ming" - An infinite case.

Consider example 2 and suppose that the process does not terminate after N stages but continues infinitely often as long as the machine is undamaged. In this case the optimal policy will consist of an infinite number of choices and we want to use the dynamic programming opproach in order to find out the structure of the optimal policy.

Notice that the dynamic programming formulation of the problem is the same as that of example 2 except for having $N = \infty$.

Now, let the expected amount of gold mined from the two mines, if the system starts at (x,y) and an optimal policy is followed all the time, be denoted by $f(x,y)$. Then $f(x,y)$, if it exicts, should satisfy the functional equation :

$$f(x,y)=\max\left[P_1\left\{r_1x+f((1-r_1)x,y)\right\} \text{ and } P_2\left\{r_2y+f(x,(1-r_2)y)\right\}\right] \quad ;$$

where $0 < P_1, \ P_2 < 1$ and $0 < r_1 , \ r_2 < 1$.

Before using this functional equation to find the structure of the optimal policy, we should prove the existence and uniquencess of $f(x,y)$. This proof utilizes certain properties of the sequence $f_n(x,y)$ as defined in the finite-stage case (example 2) .

<u>Proof of the existance and uniquencess of $f(x,y)$:</u>

1. <u>The sequence</u> $\left\{f_n(s)\right\}_{n=o}^{n=\infty}$ <u>is monotone</u>.

Proof by induction :

$$f_1(s) \geqslant 0 = f_o(s)$$

Assume that $f_n(s) \geq f_{n-1}(s)$. Then, we want to prove that $f_{n+1}(s) \geq f_n(s)$.

But $(f_{n+1}(s) - f_n(s))$ may take any of the following values:

$$\left[P_1 r_1 \alpha + P_1 f_n((1-r_1)\alpha, \gamma)\right] - \left[P_1 r_1 \alpha + P_1 f_{n-1}((1-r_1)\alpha, \gamma)\right] \quad (1)$$

or

$$\left[P_2 r_2 \gamma + P_2 f_n(\alpha, (1-r_2)\gamma)\right] - \left\{P_2 r_2 \gamma + P_2 f_{n-1}(\alpha, (1-r_2)\gamma)\right] \quad (2)$$

or

$$\left[P_1 r_1 \alpha + P_1 f_n((1-r_1)\alpha, \gamma)\right] - \left[P_2 r_2 \gamma + P_2 f_{n-1}(\alpha, (1-r_2)\gamma)\right] \quad (3)$$

or

$$\left[P_2 r_2 \gamma + P_2 f_n(\alpha, (1-r_2)\gamma)\right] - \left[P_1 r_1 \alpha + P_1 f_{n-1}((1-r_1)\alpha, \gamma)\right] \quad (4)$$

By the induction hypothesis, it is clear that
(1) ≥ 0 and (2) $\geqq 0$. If $(f_{n+1}(s) - f_n(s))$ equals (3), this means :

$$\left[P_1 r_1 \alpha + P_1 f_n((1-r_1)\alpha, \gamma)\right] \geqq \left[P_2 r_2 \gamma + P_2 f_n(\alpha, (1-r_2)\gamma)\right].$$

Thus,

$$(3) \geq \left[P_2 r_2 \gamma + P_2 f(\alpha, (1-r_2)\gamma)\right] - \left[P_2 r_2 \gamma + P_2 f_{n-1}(\alpha, (1-r_2)\gamma)\right]$$

$\geqq 0$ by the induction hypothesis.

Similarly, if $\left[f_{n+1}(s) - f_n(s)\right]$ equals (4), it means :

$$\left[P_2 r_2 \gamma + P_2 f_n(\alpha, (1-r_2)\gamma)\right] \geq \left[P_1 r_1 \alpha + P_1 f_n((1-r_1)\alpha, \gamma)\right].$$

Consequently, by using the induction hypotheses,
(4) ≥ 0.
So, $\left[f_{n+1}(s) - f_n(s)\right] \geq 0$ for all possible cases.

Thus $\left\{f_n(s)\right\}_{n=0}^{n=\infty}$ is a monotone increasing sequence.

2. <u>The sequence</u> $\left\{f_n(s)\right\}_{n=0}^{n=\infty}$ <u>is bounded from above for all s in any finite rectangle</u>

Proof by induction:

Since $0 < p_1, p_2, r_1, r_2 < 1$ then $p_1 r_1 \alpha$ and $p_2 r_2 \gamma$ are bounded for all (α, γ) in any finite rectangle.

Let $\text{Max}\left\{p_1 r_1 \alpha,\ p_2 r_2 \gamma\right\} \leq M$.

Then $f_1(s) \leq M$.

assume that $\left[f_n(s) - f_{n-1}(s)\right] \leq p_1^{\ell} p_2^{k} M$, $\ell + k = n-1$.

Then we want to prove that :

$$\left[f_{n+1}(s) - f_n(s)\right] \leq q^n M \text{ where } q^n = p_1^{i} p_2^{j},\ i+j = n.$$

Consider the values that $\left[f_{n+1}(s) - f_n(s)\right]$ may take and notice that:

$$(1) = p_1\left[f_n((1-r_1)\alpha, \gamma) - f_{n-1}((1-r_1)\alpha, \gamma)\right] \leq p_1^{\ell+1} p_2^{k} M = q^n M.$$

$$(2) = p_2\left[f_n(\alpha, (1-r_2)\gamma) - f_{n-1}(\alpha, (1-r_2)\gamma)\right] \leq p_1^{\ell} p_2^{k+1} M = q^n M.$$

$$(3) \leq p_1\left[f_n((1-r_1)\alpha, \gamma) - f_{n-1}((1-r_1)\alpha, \gamma)\right] \leq p_1^{\ell+1} p_2^{k} M = q^n M.$$

$$(4) \leq p_2\left[f_n(\alpha, (1-r_2)\gamma) - f_{n-1}(\alpha, (1-r_2)\gamma)\right] \leq p_1^{\ell} p_2^{k+1} M = q^n M.$$

Thus, $\left[f_{n+1}(s) - f_n(s) \right] \leq q^n M$ for all s in any finite rectangle

$$\therefore \sum_{n=0}^{m} \left[f_{n+1}(s) - f_n(s) \right] \leq M \sum_{n=0}^{m} q^n$$

$$\leq \frac{M}{1-q} \quad \text{(since } 0 < q < 1 \text{)}.$$

But $\sum_{n=0}^{m} (f_{n+1}(s) - f_n(s)) = f_m(s)$.

$\therefore \; f_m(s) \leq \dfrac{M}{1-q}$ for all s in any finite rectangle.

$\therefore$ The sequence $\left\{ f_n(s) \right\}_{n=0}^{n=\infty}$ is bounded from above for all s in any finite rectangle.

3. $f_n(s)$ <u>converges uniformly to a finite f(s) which satisfies the functional equation:</u>

$$f(s) = \max_a \; E \left\{ i(s,a) + f(s') \right\}.$$

Proof:

Parts 1. and 2. prove that $\lim_{n \to \infty} f_n(s)$ exists.

In order to show that the convergence is uniform, consider $(f_{n+1}(s) - f_n(s))$:

Since $(f_{n+1}(s) - f_n(s)) \leq q^n M$ for all s in any finite rectangle.

Then $\sum_{n=0}^{\infty} (f_{n+1}(s) - f_n(s)) \leq \dfrac{M}{1-q}$ for all s in any finite rectangle.

Thus, $f_n(s)$ converges uniformly for all s in any finite rectangle, and the uniformty of convergence ensures that $f(s) = \lim_{n \to \infty} f_n(s)$ is a solution to the functional equation:

$$f(s) = \max_a E \left\{ i(s,a) + f(s') \right\}.$$

4. $f(s)$ <u>is the unique solution for this functional equation.</u>

Proof:

Let $F(s)$ be any other solution that is bounded for all s in any finite rectangle . Then:

$$\left[F(s) - f_o(s)\right] \leq M' \text{ for all s in any finite rectangle}$$

Assume that $|F(s) - f_{n-1}(s)| \leq q^{n-1} M'$ for all s in any finite rectangle . Proceeding in a way similar to that used in part 2., we can show that

$$|F(s) - f_n(s)| \leq q^n M' \text{ for all s in any finite rectangle}$$

$\therefore \sum_{n=0}^{\infty} \left[F(s) - f_n(s)\right]$ converges absolutly and uniformly .

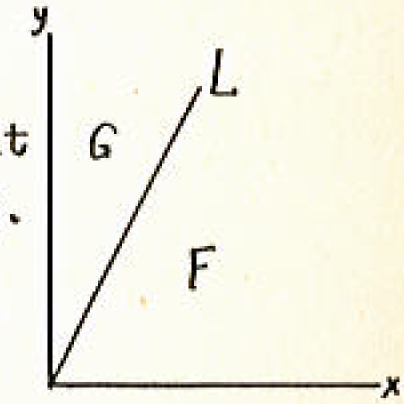
$\therefore$ $f_n(s)$ converges uniformly to $F(s)$.

$\therefore$ $F(s) \equiv f(s)$ for all s in any finite rectangle.

<u>The structure of the optimal polity:</u>

After proving the existance and uniqueness of $f(s)$ (for $o < p_1, p_2, r_1 r_2 < 1$) we will use the functional equation: $f(s) = \max_a E\left\{i(s, a) + f(s')\right\}$ to find the structure of the optimal policy in the infinite state gold-mining problem.

It is clearly noticed that if x is much bigger than y, F will be the optimal choice. But each time F is used, the amount of gold left in it decreases till it reachs a certain level where the expected amonnt of gold mined if F is chosen once more equals the expected amount of gold mined if G is chosen instead. Similarly, if y is much bigger than x , G will be the optimal choice and mining it repeatedly will decrease the amount of gold in it till it reachs a level where the choice of F is "equivelant" to the choice of G .



So, if we consider the positive (x,y) quadrent we expect it to be devided into two regions: F and G. (Region F is the region where F is the optimal choice. Simelarly for region G .)

The points (x,y) on the deviding line L should satisfy the equality:

$$p_1 r_1 x + p_1 f((1-r_1) x, y) = p_2 r_2 y + p_2 f(x, (1-r_2) y).$$

We notice also that if (x,y) is on L and action F is taken, the new state will be ($(1-r_1)x, y$) which is in region G. Consequently the optimal action G will lead the system to the state ($(1-r_1) x, (1-r_2) y$) and the expected amount of gold mined (if an optimal policy if followed from there on) will be:

$$p_1 r_1 x + p_1 (p_2 r_2 y + f((1-r_1)x, (1-r_2) y)).$$

But if at (x,y) , on L, action G is taken moving the system to state (x, $(1-r_2)$ y) which is in segion F, it should be followed by action F, the system will move to state ($(1-r_1)$ x, $(1-r_2)$ y) and the expected amount of gold mined will be:

$$p_2 r_2 y + p_2 (p_1 r_1 x + p_1 f ((1-r_1) x, (1-r_2) y)).$$

It is clear that starting at (x,y) on L, the expected amount of gold mined should be the same whether F is chosen first then followed by G , or G is chosen first and then followed by F. i.e., for every point (x,y) on L, the following equality holds:

$$p_1 r_1 x + p_1 \left[p_2 r_2 y + p_2 f ((1-r_1)x, (1-r_2) y) \right]$$

$$= p_2 r_2 y + p_2 \left[p_1 r_1 x + p_1 f ((1-r_1)x, (1-r_2)y) \right]$$

Thid yeilds: $$\frac{p_1 r_1 x}{(1-p_1)} = \frac{p_2 r_2 y}{(1-p_2)}$$

Since this equality defines L, then any point (x,y) that satisfies the inequality:

$$\frac{p_1 r_1 x}{1 - p_1} > \frac{p_2 r_2 y}{1 - p_2}$$

lies in the F region and the optimal choice there is F. If the inverse inequality holds, then the optimal choice will be G.

So, the structure of the optimal policy in the infinite-stage gold mining problem is given by:

Choose F if $\frac{p_1 r_1 x}{1 - p_1} \geqslant \frac{p_2 r_2 y}{1 - p_2}$, and

Choose G if $\frac{p_1 r_1 x}{1- p_1} \leqslant \frac{p_2 r_2 y}{1 - p_2}$.

Example 4 :[*]

A producer of a certain commodity wants to maximize his expected profit over a certain number of months. Every month he may be in either of two states. He is in the first state, s_1 , if the commodity currently produced is successful. He is in the second state, s_2, if the commodity is not successful. Each month at any state, he has to choose one of two actions. If in state s_1 he chooses action $a_1 (s_1)$ [advertising to increase the possibility of continous success] he moves, in the following month, to state s_1 with probability 0.8 and profit 4; and to state s_2 with probability 0.2 and profit 4. Choosing action $a_2 (s_1)$ [no advertising] leads to state s_1 with probability 0.5 and profit 9; and to state s_2 with probability 0.5 and profit 3. On the other hand, if in state s_2 he chooses action $a_1 (s_2)$ [more research to improve the production] he moves to state s_1 with probability 0.7 and profit 1; and to state s_2 with probability 0.3 and profit -19. Choosing action $a_2 (s_2)$ [no research] leads to state s_1 with probability 0.4 and profit 3; and to state s_2 with probability 0.6 and profit - 7.

What is the optimal policy that he should follow in order to maximize his expected profit over four months?

The dynamic programming formulation:

$S = \{ s_1 , s_2 \}$

---

[*] This example is known in the literature by: Toy Maker example.

$A = \{ a_1(s_j) , a_2(s_j) ; j=1,2 \}$

q : $q(s_1 / s_1, a_1) = 0.8$ $\quad$ $q(s_1 / s_2, a_1) = 0.7$

$q(s_2 / s_1, a_1) = 0.2$ $\quad$ $q(s_2 / s_2, a_1) = 0.3$

$q(s_1 / s_1, a_2) = 0.5$ $\quad$ $q(s_1 / s_2, a_2) = 0.4$

$q(s_2 / s_1, a_2) = 0.5$ $\quad$ $q(s_2 / s_2, a_2) = 0.6$

$i(s, a)$:

$$i(s_1, a_1) = \begin{cases} 4 & \text{with probability } 0.8 \\ 4 & \text{" " } 0.2 \end{cases}$$

$$i(s_1, a_2) = \begin{cases} 9 & \text{" " } 0.5 \\ 3 & \text{" " } 0.5 \end{cases}$$

$$i(s_2, a_1) = \begin{cases} 1 & \text{" " } 0.7 \\ -19 & \text{" " } 0.3 \end{cases}$$

$$i(s_2, a_2) = \begin{cases} 3 & \text{" " } 0.4 \\ -7 & \text{" " } 0.6 . \end{cases}$$

$\beta = 1$, $N=4$, and the months are numbered in a backward order.

<u>The solution</u> :

let $f_n(s, a)$ = the total expected profit on the n^{th} month if the producer is in state s, chooses action a , and follows an optimal policy in the remaining (n-1) months.

Let $f_n(s)$ = the total expected profit on the n^{th} month if the producer is in state s , chooses the optimal action $a^*(s)$, and follows an optimal policy in the remaining (n-1) months.

$\therefore f_n(s_j, a) = E \{ i(s_j, a) + f_{n-1}(s_k) \}$, and

$$f_n(s_j) = \max_a E \left\{ i(s_j, a) + f_{n-1}(s_k) \right\}$$

$$= \max_a \left\{ E\, i(s_j,a) + \sum_{k=1}^{2} q(s_k / s_j, a)\, f_{n-1}(s_k) \right\} .$$

Where:

$E\, i(s_1, a_1) = 4$

$E\, i(s_1, a_2) = 6$

$E\, i(s_2, a_1) = -5$

$E\, i(s_2, a_2) = -3$.

For n = 1 : $f_1(s_j, a) = E\, i(s_j, a)$.

s \ a	$f_1(s,a)$		$f_1(s)$	$a^*(s)$
	$a_1(s)$	$a_2(s)$		
s_1	4	6	6	$a_2(s_1)$
s_2	-5	-3	-3	$a_2(s_2)$

For n=2 : $f_2(s_j,a) = E\, i(s_j,a) + \sum_{k=1}^{2} q(s_k/s_j,a)\, f_1(s_k)$

$\therefore f_2(s_1,a_1) = 4 + \left[(6)(0.8) + (-3)(0.2)\right] = 8.2$

$f_2(s_1,a_2) = 6 + \left[(6)(0.5) + (-3)(0.5)\right] = 5.5$

$\therefore f_2(s_1) = 8.2$ and $a^*(s_1) = a_1(s_1)$.

Similarily for s_2 ... Thus :

s \ a	$f_2(s,a)$: $a_1(s)$	$f_2(s,a)$: $a_2(s)$	$f_2(s)$	$a^*(s)$
s_1	8.2	5.5	8.2	$a_1(s_1)$
s_2	-1.7	-2.4	-1.7	$a_1(s_2)$

For n = 3 : $f_3\ (s_j,a)+\sum_{k=1}^{2} q(s_k\ /\ s_j,a)f_2\ (s_k)$. Then

q \ a	$f_3(s,a)$: $a_1(s)$	$f_3(s,a)$: $a_2(s)$	$f_3(s)$	$a^*(s)$
s_1	10.22	9.22	10.22	$a_1(s_1)$
s_2	0.23	-0.74	0.23	$a_1(s_2)$

For n = 4 : $f_4(s_j,a) = E\ i(s_j,a) + \sum_{k=1}^{2} q(s_k/s_j,a)f_3(s_k)$. Thus:

s \ a	$f_4(s,a)$: $a_1(s)$	$f_4(s,a)$: $a_2(s)$	$f_4(s)$	$a^*(s)$
s_1	12.222	11.275	12.222	$a_1(s_1)$
s_2	2.226	1.226	2.226	$a_1(s_2)$

So, the optimal policy for the 4- stage problem is given by:

Choose action a_1 (whether in state s_1 or s_2) in all months except the last one where action a_2 should be choosen. I f this policy is followed, the expected income will be 12.222 if the producer starts at state s_1, and 2.226 if he starts at state s_2.

Example 5 : (Seasonal employment).

In seasonal industries the work load for certain jobs is subject to considerable seasonal fluctuations. Suppose that the minimum requerments for manpower in a certain job during the four reasons of the year are as follows:

Season :	Summer	Autum	Winter	Spring
Requerments :	220	240	200	255.

Suppose also that any employment above these levels is wasted at an approximate cost of $100 per man per reason. On the other hand, the estimate of hiring and firing costs is such that the total cost of changing the level of employment from one season to the other is $10 times the square of the difference in employment levels. [Fractional levels of employment , i.e., part time employments , are possible.]

What should be the employment level in each season that minimizes total costs over successive years?

Dynamic programming formulation:

It is clear that the employment level at the spring season should be 255 since this is the peak season. So, spring will be considered as stage 1, winter as stage 2, autum as stage 3, and summer as stage 4. At each season, the employment level in the previous season represents a state and the employment level chosen for the current season represents an action.

So, we have a deterministic process with an infinit number of possible states and an infinit number of possible actions :

$$S = \{ s \,;\; 0 \leq s \leq 255 \} ,$$
$$A = \{ a \,;\; 0 \leq a \leq 255 \} .$$

Take $\beta = 1$.

Let r_n denote the minimum requerments of manpower at season n.

$\therefore \; i(s, a_n) = 100\,(a_n - r_n) + 10\,(a_n - s)^2.$

<u>The solution:</u>

Let $f_n(s, a)$ = the total expected cost in season n if s is the employment level in the preceeding season, a is the employment level chosen for the current season, and an optimal policy is followed in the remaining (n-1) seasons.

$$f_n(s) = \min_{a_n \geq r_n} f_n(s,a)$$

$$\therefore \quad f_n(s,a_n) = 10(a_n-s)^2+100(a_n-r_n) + f_{n-1}(a_n) \text{ , and}$$

$$f_n(s) = \min_{a_n \geq r_n} \left\{ 10(a_n - s)^2+100(a_n-r_n)+f_{n-1}(a_n) \right\}.$$

Since the number of states and the number of possible actions are infinit , calculus will be used, in stead of direct enumeration, in order to find the value of a_n that minimizes $f_n(s, a_n)$.

<u>Note:</u> It is enough to consider one year since successive years are identical. Notice also that at the end of any year, i.e.

after the last stage of that year, the total cost of the optimal policy in the following years is a fixed constant and therefore can be omitted from consideration. Thus, $f_1\ (s,a_1)$ is given by :

$$f_1(s,a_1) = 10\ (a_1 - s)^2 + 100\ (a_1 - r_1).$$

For n = 1 :

It is clear that $a_1^* = r_1 = 255$, then :

s	$f_1(s)$	a_1^*
≤255	$10(255-s)^2$	255

For n = 2 :

$$f_2\ (s,a_2) = 10\ (a_2 - s)^2 + 100\ (a_2 - r_2) + f_1\ (a_2)$$

$$= 10\ (a_2 - s)^2 + 100\ (a_2 - 200) + 10\ (255-a_2)^2$$

$$f_2\ (s) = \min_{a_2 \geq 200} f_2\ (s,\ a_2).$$

$$\frac{\partial}{\partial a_2} f_2(s,\ a_2) = 40\ a_2 - 20\ s - 5000$$

$$\frac{\partial^2}{\partial a_2^2} f_2\ (s\ ,\ a_2) = 40 > 0\ .$$

$\therefore$ $f_2(s, a_2)$ reachs its minimum if $\frac{\partial}{\partial a_2} f_2(s, a_2) = 0$,
i.e. if $a_2 = \frac{s+250}{2}$

But, since $a_2 \geqslant 200$, then a_2^* equals $\frac{s+250}{2}$ if this value is $\geqslant 200$, i.e. if $s \geqslant 150$; and $a_2^* = 200$ otherwise.

$$\therefore \quad a_2^* = \begin{cases} \frac{s+250}{2} & \text{if} \quad s \geqslant 150 \\ 200 & \text{if} \quad s \leqslant 150. \end{cases}$$

$$\therefore \quad f_2(s) = \begin{cases} \frac{5}{2}(250-s)^2 + \frac{5}{2}(260-s)^2 + 50(s-150) & \text{if } s \geqslant 150 \\ 10(200-s)^2 + 30250 & \text{if } s \leqslant 150. \end{cases}$$

So, the results for the two-stage problem are:

s	$f_2(s)$	a_2^*
$s \leqslant 150$	$10(200-s)^2 + 30250$	200
$150 \leqslant s \leqslant 255$	$\frac{5}{2}(250-s)^2 + \frac{5}{2}(260-s)^2 + 50(s-150)$	$\frac{s+250}{2}$

For n = 3 :

$\because$ $a_3 \geqslant 240$ ($\geqslant 150$) , then

$$f_3(s, a_3) = 10(a_3 - s)^2 + 100(a_3 - 240) + f_2(a_3)$$

$$= 10(a_3-s)^2 + 100(a_3-240) + \frac{5}{2}(250-a_3)^2 + \frac{5}{2}(260-a_3)^2 + 50(a_3-150)$$

$$f_3(s) = \min_{240 \leq a_3 \leq 255} \left\{ f_3(s, a_3) \right\}$$

$$\frac{\partial}{\partial a_3} f_3(s, a_3) = 30\, a_3 - 205 - 2400$$

$$\frac{\partial^2}{\partial a_3^2} f_3(s, a_3) = 30 > 0 .$$

$\therefore$ $f_3(s, a_3)$ reachs its minimum if $\frac{\partial}{\partial a_3} f_3(s, a_3) = 0$, i.e.,

if $a_3 = \frac{25 + 240}{3}$

$$\therefore \quad a_3^* = \begin{cases} \frac{25 + 240}{3} & \text{if } s \geq 240 \left[\text{if } \frac{25+240}{3} \geq 240\right] \\ 240 & \text{if } s \leq 240 \end{cases} ,$$

and

$$f_3(s) = \begin{cases} \frac{10}{9} \quad (240-s)^2 + (255-s)^2 + (270-s)^2 \quad + 100(s-195) & \text{if } s \quad 240 \\ 10\,(240-s)^2 + 5750 & \text{if } s \leq 240. \end{cases}$$

<u>For n = 4</u> :

$\because \quad a_4 \geq 220$.

$$\therefore \quad f_4(s, a_4) = 10\,(a_4 - s)^2 + 100(a_4 - r_4) + f_3(a_4)$$

$$= \begin{cases} 10(a_4-s)^2 + 100(a_4-220) + 10(240-a_4)^2 + 5750 & \text{if } a_4 \leq 240 \\ 10(a_4-s)^2 + 100(a_4-220) + \frac{10}{9}\left[(240-a_4)^2 + (255-a_4)^2 + (270-a_4)^2 + 100(a_4-195)\right] & \text{if } a_4 \geq 240 . \end{cases}$$

In the region where $a_4 \leq 240$:

$$\frac{\partial}{\partial a_4} f_4 (s, a_4) = 20 (2 a_4 - s - 235)$$

But , it is known that s = 225 (spring employment) .

$$\therefore \frac{\partial}{\partial a_4} f_4(s,a_4) = 40 (a_4-245) < 0 \text{ for all } a_4 \leq 240.$$

$\therefore$ in this region $f_4(s,a_4)$ reachs its minimum at $a_4=240$.

In the region where $240 \leq a_4 \leq 255$:

$$\frac{\partial}{\partial a_4} f_4(s, a_4) = \frac{20}{3}\left[4 a_4 - 3s - 225\right] .$$

$$\frac{\partial^2}{\partial a_4^2} f_4(s,a_4) = \frac{80}{3} > 0.$$

$\therefore$ $f_4(s,a_4)$ reachs its minimum if $\frac{\partial}{\partial a_4} f_4(s,a_4) = 0$, i.e.

if $a_4 = \frac{3s + 225}{4}$

Since s = 255, then $f_4(s, a_4)$ reachs its minimum in this region at $a_4 = 247.5$. Since this region includes $a_4 = 240$, then $a_4 = 247.5$ minimizes $f_4(s, a_4)$ over all the region $220 \leq a_4 \leq 255$.

$\therefore$ $a_4^* = 247.5$, and

$f_4 (255) = 9250 .0$.

Therefore, the optimal policy is :

$a_4^* = 247.5$, $a_3^* = 245$, $a_2^* = 247.5$, $a_1^* = 255$, with a total estimated cost per year of $9250.

<u>Example 6</u> : Inventory problem - finite-stage case :

In spite of the storage cost and the tying up of capital, keeping inventories is a common practice in the business world for different reasons, such as : the uncertainity of future demands, the flcutuations of prices, and the economics of scale in production.

In the case considered here, we will assume that orders to increase the stock level are made at the begining of each of a finite number of equally spaced periods , at a certain cost. Orders are assumed to be fulfilled immediatly . During every period , demand decreases the inventory level. Demand is a random variable with a known density function, and demands in successive periods are independant and identially distributed . If demand happens to be greater than the available stock , it should be satisfied at the following periods, and a penalty cost should be paid. In addition, there is the cost of holding inventories,which includes the apportunity cost.

The question that needs to be answered is : " How much to order at the begining of each period in order to minimize the expected total costs ? "

Notation :

x_n : The stock level at the begining of period n , before ordering.

$y_n \geqslant x_n$: The stock level at the begining of period n , after ordering.

$\therefore\ y_n - x_n \geqslant 0$ is the quantity ordered at the begining of period n.

$h \geqslant 0$: Holding cost (per unit, per period).

$P \geqslant 0$: Penalty (or shortage) costs, (per unit, per period).

$c \geqslant 0$: Ordering costs (per units).

$z \geqslant 0$: Demand, it is a random variable with probability density $\varphi(z)$.

$L(y_n)$: The expected holding and penalty costs in period n.

$$\therefore L(y_n) = \begin{cases} \int_0^{y_n} h(y_n - z)\, \varphi(z)\, dz + \int_{y_n}^{\infty} p(z - y_n)\, \varphi(z)\, dz & \text{if } y_n > 0 \\ \int_0^{\infty} p(z - y_n)\, \varphi(z)\, dz & \text{if } y_n \leqslant 0. \end{cases}$$

So, the expected cost in the n^{th} period equals $c(y_n - x_n) + L(y_n)$.

Dynamic Programming formulation :

$S = \{ x \,,\, -\infty < x < \infty \}$, x is the stock level before ordering.

$A = \{ y \,,\, x \leqslant y < \infty \}$, y is the stock level after ordering.

$E\, i\, (s,a) = c(y-x) + L(y)$.

$$q : q(s' / s, a) = \begin{cases} \varphi(z) & \text{where } s' = y - z \\ 0 & \text{otherwise}. \end{cases}$$

$\beta = 1$, and the periods are numbered in a backword order.

The solution

Let $f_n(x,y)$ = expected total cost for n period process starting with an initial stock level x , increasing it up to y , and following an optimal ordering policy in the remaining (n-1) periods.

$f_n(x)$ = expected total cost for n- period process starting with an initial stock level x, and following an optimal ordering policy.

Then :

$$f_n(x) = \min_{y \geq x} \quad f_n\,(y,x)$$

$$= \min_{y \geq x} \left\{ c(y-x)+L(y)+ \int_0^\infty f_{n-1}(y-z)\,\varphi(z)\,dz \right. \quad ,$$

and

$$f_1(x) = \min_{y \geq x} \left\{ c(y-x) + L\,(y) \right\}$$

$$= \min_{y \geq x} \left\{ f_1(x;y) \right\} \ .$$

For n = 1 :

Since holding and shortage costs are linear, then L(y) is convex . Consequently $f_1(x,y)$ is convex and reachs its unique minimum at $\bar{y}_1$ which is defined by :

$$\frac{d}{dy}\, f_1(x,y) \;=\; o \quad \text{at} \quad y=\bar{y}_1 \ .$$

So, the optimal policy is to order up to the level $\bar{y}_1$, i.e., to order max $\left[(\bar{y}_1-x) \ , \ o \right]$.

For n = 2 :

$$f_2(x) = \min_{y \geq x} \left\{ c(y-x) + L(y) + \int_0^{\infty} f_1(y-z)\, \varphi(z)\, dz \right\}$$

$$= \min_{y \geq x} \left\{ f_2(x,y) \right\}$$

It has been proved that if a function G (x) is convex, then the function $g(y) = \min_{y \geq x} G(x)$ is also convex. Using this result shows that $f_1(x)$ is convex and consequently $f_2(x,y)$ is convex. So, it reaches its unique minimum at $\bar{y}_2$ which is defined by :

$$\frac{d}{dy} f_2(x,y) = o \quad \text{at } y = \bar{y}_2$$

Thus the optimal policy at the begining of the second period is given by : order max $[(\bar{y}_2 - x), o]$.

For n > 2 :

Repeating the same argument for any n > 2 , we reach the conclusion that the optimal policy at the begining of period n is given by: order max $[(\bar{y}_n - x), o]$, where $\bar{y}_n$ is defined by :

$$\frac{d}{dy} f_n(x,y) = o \quad \text{at } y = \bar{y}_n.$$

Example 7 : Inventory problem , Infinite - stage case.

If instead of having a finite number of periods we have an infinite horizon , we still can analyze the inventory problem under the same assumptions of the previous example except that we should have $\beta < 1$ in order to avoid infinite costs.

Let f(x) = expected total discounted costs, starting with an initial stock level x and using an optimal ordering policy.

Then , f(x) should satisfy the functional equation :

$$f(x) = \min_{y \geq x} \left\{ c(y-x)+L(y)+ \beta \int_0^\infty f(y-z)\, \varphi(z)\, dz \right\} .$$

It has been proved that if $f_n(x)$ is redefined to be :

$$f_n(x) = \min_{y \geq x} \left\{ c(y-x)+L(y)+ \beta \int_0^\infty f_{n-1}(y-z)\, \varphi(z)\, dz \right\} ,$$

then , $\left\{ f_n(x) \right\}_{n=0}^{n=\infty}$ is a monotone increasing sequence , bounded from above [because $\beta < 1$] , and converges uniformly for all x in any finite interval. The limit function f(x) is convex and is the unique solution to the functional equation :

$$f(x) = \min_{y \geq x} \left\{ c(y-x)+L(y)+ \beta \int_0^\infty f(y-z)\, \varphi(z)\, dz \right\} .$$

It has also been proved that the optimal policy for the infinite period process is given by :

order max $[(\bar{y}-x), o]$ at the begining of every period, with $\bar{y}$ defined by :

$$c(1- \beta)+ \frac{d}{dy} L(y) = o \quad \text{at } y=\bar{y} .$$

References :

Richard Bellman , Dynamic Programming, Princeton , New Jersey : Princeton University Press , 1957

David Black well : Class Notes , 1963.

Derman : Class Notes , 1965.

Howard : Dynamic Programming and Markov Processes, The Technology Press of M.I.T. , 1960, and John Wiley and Sons, Inc. , New York 1960.