(1)

DYNAMIC PROGRAMMING AND MULTI-STAGE DECISION PROCESSES OF STOCHASTIC TYPE

Richard Bellman

P-589

2 November 1954

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA

SUMMARY

This paper is a summary of some applications of the theory of dynamic programming to various classes of multi-stage decision problems of stochastic type.

DYNAMIC PROGRAMMING AND MULTI-STAGE DECISION PROCESSES OF STOCHASTIC TYPE

Richard Bellman

Table of Contents

§1. Introduction

§2. Some Representative Multi-Stage Decision Processes of Stochastic Type

2.1 Allocation

2.2 Optimal Inventory

2.3 Gold-Mining

2.4 Learning

2.5 A Deterministic Maximization Problem

§3. A Modicum of Nomenclature

§4. The Principle of Optimality

§5. Functional Equations

5.1 Allocation

5.2 Optimal Inventory

5.3 Gold-Mining

5.4 Learning

5.5 A Deterministic Maximization Problem

§6. Successive Approximations

§7. Approximation in Policy Space - Monotone Convergence

Bibliography

§1. Introduction

A large number of problems of theoretical and practical importance reduce to the computation of the maximum of a function of the form

$$(1)\quad F(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} a_i \phi_i(x_1, x_2, \ldots, x_n),$$

subject to a series of constraints of the form

$$(2)\quad R_j(x_1, x_2, \ldots, x_n) \leq 0, \qquad j = 1, 2, \ldots, m.$$

If n is a number of even moderate size, prosaic computational techniques are utterly unavailing, and the use of unadorned calculus is equally fruitless, and sometimes even illegitimate.

If the functions ϕ_i and R_j are linear, numerical results, and sometimes even analytic results, may be obtained using various versions of the elegant "simplex" technique devised by G. Dantzig. If some of the functions are non-linear, the theory of non-linear programming, as conceived by Kuhn and Tucker and others, must be invoked. Naturally, the computational road is not as smooth as in the linear case.

If the functions above possess certain features of symmetry, or if, in particular, $F(x_1, x_2, \ldots, x_n)$ represents the "return" of a multi-stage process, the theory of dynamic programming is often useful as we have shown in a number of papers, cf. [1], [2], [3], [4], [5], [6], [7], [8], [9]. In general, the more non-linear the problem, the more useful the techniques of this theory.

For problems of deterministic type, such as those posed above, we have at least a choice of the techniques we may wish to employ. However, if we consider multi-stage decision processes of stochastic type, where the functions and coefficients may be stochastic, then, in general, there seems to be little alternative to some variation of the functional equation approach.

Use of the functional equation technique is, as pointed out above, only profitable when the problem possesses certain symmetrical features. The general linear programming problem of determining the expected value of the maximum of

$$(3) \quad F(x) = \sum_{i=1}^{n} a_i x_i$$

where the x_i are subject to constraints of the form

$$(4) \quad (a) \quad x_i \geq 0$$

$$(b) \quad \sum_{j=1}^{n} c_{ij} x_j \leq e_i, \qquad i = 1, 2, \ldots, m,$$

where the c_{ij} and e_i are stochastic parameters subject to given probability distributions, seems at the moment still to be outside mathematical ken.

In this paper we shall consider some multi-stage decision processes of stochastic type which are particularly suited to the techniques of the theory of dynamic programming. Five processes of stochastic type, concerning a wide range of topics, will be

treated in outline, and then a deterministic maximization problem, inserted to show how a problem of this type can profitably be considered as a multi-stage decision problem.

Some applications of these ideas to the calculus of variations and theory of integral equations may be found in [10], [11], [12] and [13].

In the section following the presentation of the problems we shall present some basic terminology of the theory of dynamic programming. Following that we shall discuss the "Principle of Optimality", which we utilize to obtain functional equations for the determination of optimal policies in the above decision processes.

We shall then indicate the use of successive approximations in determining numerical and analytic solutions and the application of the concept of "approximation in policy space" to obtain monotone convergence.

§2. Representative Multi-Stage Decision Processes of Stochastic Type.

Let us now consider some representative processes which we shall show below may be treated by the methods of dynamic programming.

§2.1 Optimal Allocation

Over a period of n years, it is necessary, at the beginning of each year, to order some equipment to perform certain tasks. We possess an initial amount of money x which is to be divided into two parts, y and x-y. The first part, y, is to be used to purchase equipment of type A, and the remaining amount, x-y, is to be used to purchase equipment of type B.

Let us assume, for simplicity, that if we spend y dollars to purchase A-equipment there are two possibilities.

(1) (a) with probability p_1 we obtain $g_1(y)$ man-hours from the equipment and retain a salvage value of $a_1 y$ dollars, where $0 < a_1 < 1$.

(b) with probability $1-p_1$ we obtain $g_2(y)$ man-hours from the equipment and retain a salvage value of $a_2 y$ dollars, where $0 < a_2 < 1$.

Similarly, if (x-y) dollars are spent for B-equipment, we have corresponding probabilities q_1 and $1-q_1$, functions $h_1(x-y)$ and $h_2(x-y)$ and parameters b_1 and b_2.

At the start of each year we repeat the process with the new initial amount equal to the sum of the salvage values.

The problem is to determine the allocation policy which maximizes the total expected man-hours obtained over the n year period.

§2.2 Optimal Inventory

At various specified times we have an opportunity to order supplies of a certain set of items, where the cost of ordering is some function of the amount ordered which may or may not include fixed administrative or "red tape" costs. At various other times, demands are made upon the stocks of these items. These demands are stochastic and their joint distribution function is known. The incentive for ordering lies in a penalty which is assessed whenever the demand of an item exceeds the supply.

We wish to determine the ordering policy which minimizes some average cost.

§2.3 Stochastic Gold-Mining

We are fortunate enough to possess two gold mines, Anaconda and Bonanza, the first of which contains an amount, x, of gold, while the second possesses an amount, y. In addition, we have a rather delicate gold-mining machine which has the property that if used to mine gold in Anaconda there is a probability p_1 that it will mine a fraction r_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. Similarly, Bonanza has associated the probabilities p_2 and $(1-p_2)$ and the fraction r_2.

We begin by using the machine in either the Anaconda or Bonanza mine. If the machine is undamaged, we again make a choice of using the machine in either of the two mines, and continue in this way, making a choice before each mining operation, until the machine is damaged.

What sequence of choices maximizes the amount of gold mined before the machine is damaged?

§2.4 Learning

Let us assume that we have two machines, I and II, with the following properties. If machine I is used, there is a probability r of receiving a gain of one unit and a probability (1-r) of receiving nothing. If machine II is used, there is a corresponding

probability s. These probabilities are not known. We do, however, possess an à priori probability distribution for their values.

Given a fixed number of trials, the problem is to determine the sequence of choices which maximizes the total expected return.

§2.5 A Deterministic Maximization Problem

We wish to determine the maximum of $\sum_{i=1}^{n} x_i$, subject to the constraints

(1) (a) $x_i \geq 0$

(b) $\sum_{i=1}^{n} F(x_i) \leq a$,

(c) $\sum_{i=1}^{n} G(x_i) \leq b$

§3. A Modicum of Nomenclature

Let us now define some useful terms. We are considering processes, finite or infinite, discrete or continuous, which require a sequence of decisions. We consider only feasible sequences, which is to say those which are consistent with various limitations and constraints which may be imposed. Every feasible sequence of decisions is called a policy, and a policy which maximizes the "return" of the process is called an optimal policy. By the solution of a problem, we mean the determination of all optimal policies.

In order to define precisely the "return" of a process, we must introduce the concept of a state variable or state parameter. We consider a system S of economic, industrial, engineering or other source, whose physical state at any time is specified by a set of

quantities, P. In many cases, P is an N-dimensional vector, while in more complicated situations, P may consist of a set of points and functions, and involve functionals as well. The components of P are called the state variables.

The effect of a decision is to transform P into another vector P^1. Hence a decision is equivalent to a transformation of the state variables. Every policy then yields a sequence of states

(1) $P_1, P_2, \ldots, P_n, \ldots$

We now introduce a criterion function, $\phi(p)$, measuring the value of a state P, and postulate that the purpose in carrying out the process is to maximize this function of the final state. The function, $\phi(P_F)$, of the final state P_F is called the return of the process, and write $\phi_D(P)$ to denote the value, $\phi(P_F)$, obtained starting in a state P and using a policy D.

Finally, we define

(2) $f(P) = \underset{D}{\text{Max}}\ \phi_D(P)$,

where we maximize over all feasible policies. The policies, D, which yield the return f(P) are the optimal policies.

§4. The Principle of Optimality

We now characterize optimal policies by means of the following intuitive

PRINCIPLE OF OPTIMALITY: 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.

This principle immediately yields the functional equation

$$(1) \quad f(P) = \max_{D} f(P(D)),$$

where P(D) is the vector resulting from a choice of an initial decision D.

If we can separate P into a "space" vector, π, and a "time" vector T, then (1) takes the form

$$(2) \quad f(\pi, S+T) = \max_{D[0,S]} f(\pi_D(S), T),$$

where $D[0,S]$ represents a sequence of decisions over the time-interval $[0,S]$.

Observe that the form of the equation is the same regardless of whether we are dealing with a deterministic or stochastic situation; in one case we have a function of the final state, in the other case an average of functional values. Further discussion will be found in [2], [5] and [11].

§5. Functional Equations

Let us now use the method expounded above to convert the problems discussed previously in §2.1 – §2.5 involving policies into problems concerning the solution of functional equations.

5.1 Allocation

Let us define

(1) $f_n(x)$ = expected return in manhours obtained over an N-stage period starting with x-dollars and using an optimal policy

Then

$$(2)\quad f_1(x) = \max_{0\le y\le x} \left[p_1 g_1(y)+(1-p_1)g_2(y)+q_1 h_1(x-y)+(1-q_1)h_2(x-y) \right],$$

and

$$\begin{aligned}(3)\quad f_{n+1}(x) = \max_{0\le y\le x} \Big[& p_1q_1\left[g_1(y)+h_1(x-y)+f_n(a_1y+b_1(x-y)) \right] \\ & +p_2q_1\left[g_2(y)+h_2(x-y)+f_n(a_2y+b_1(x-y)) \right] \\ & +p_1q_2\left[g_1(y)+h_2(x-y)+f_n(a_1y+b_2(x-y)) \right] \\ & +p_2q_2\left[g_2(y)+h_2(x-y)+f_n(a_2y+b_2(x-y)) \right].\end{aligned}$$

for $n \ge 1$; see [1], [2], [3], [4].

§5.2 Optimal Inventory

Let us consider for the sake of simplicity a process involving the stocking of one item where we may order at each of a finite number of equally spaced times, and we must fulfill the demand at these same times. We assume that there is no delay in filling an order or a demand.

We assume that we know completely the following functions:

(1) (a) $\phi(s)ds$ = the probability that the demand will lie between s and s+ds.

(b) $k(z)$ = the cost of ordering z items initially to increase the stock level

(c) $p(z)$ = the cost of ordering z items to meet an excess, z, of demand over supply, the penalty cost

Let us define

(2) $f_n(x)$ = expected total cost for an n-stage process starting with an initial supply x and using an optimal ordering policy

Then

$$(3)\quad f_1(x) = \min_{y \geq x} \left[k(y-x) + \int_y^\infty p(s-y)\phi(s)ds \right],$$

$$f_{n+1}(x) = \min_{y \geq x} \left[k(y-x) + \int_y^\infty p(s-y)\phi(s)ds + f_n(0)\int_y^\infty \phi(s)ds + \int_0^y f_n(y-s)\phi(s)ds \right].$$

See [2], [7].

§5.3 Stochastic Gold-Mining

Let us define

(1) $f(x,y)$ = expected amount of gold mined before the machine is damaged when A has x, B has y and an optimal policy is employed

Then we see that $f(x,y)$ satisfies the functional equation

$$(2)\quad f(x,y) = \text{Max}\left[\begin{array}{ll} \text{A:} & p_1\left[\, r_1x+f((1-r_1)x,y)\,\right], \\ \text{B:} & q_1\left[\, r_2y+f(x,(1-r_2)y)\,\right] \end{array}\right]$$

If we wish to maximize the expected value of some function of the total return, R, say $\phi(R)$, then we must, in general, introduce another state variable, a, the amount of gold already mined. Let us define

(3) $f(x,y,a)$ = expected value of $\phi(R)$ when A has x, B has y, a has already been mined and an optimal policy is employed.

Then $f(x,y,a)$ satisfies the equation

$$(4)\quad f(x,y,a) = \text{Max}\left[\begin{array}{ll} \text{A:} & p_1f((1-r_1)x,y,a+r_1x) + (1-p_1)\phi(a), \\ \text{B:} & q_1f(x,(1-r_2)y,a+r_2y) + (1-q_1)\phi(a) \end{array}\right]$$

See [2], [4], [11].

§5.4 Learning

Let us consider the case where we have an unbounded set of trials and the return on the $k^{\underline{th}}$ trial is discounted by a factor a^{k-1}, where $0 < a < 1$. Furthermore, to simplify the notation, let us treat only the case where one machine, II, has a known probability of success, s, and the other machine, I, has an à priori distribution of probabilities of success, dF(r).

Our fundamental assumption is that the new à priori distribution function after m successes and n failures on the first machine is

(1) $$dF_{mn}(r) = \frac{r^m(1-r)^n dF(r)}{\int_0^1 r^m(1-r)^n dF(r)}$$

On the basis of this assumption, we obtain for the function

(2) $f_{m,n}(s)$ = expected total return obtained using an optimal policy after m successes and n failures on the first machine,

the functional equation

(3) $$f_{m,n}(s) = \text{Max} \begin{bmatrix} \text{I:} & \left[\int_0^1 r dF_{mn}(r)\right]\left[1 + a f_{m+1,n}(s)\right] \\ & + \left[\int_0^1 (1-r) dF_{mn}(r)\right]\left[a f_{m,n+1}(s)\right] \\ \text{II:} & s/(1-a) \end{bmatrix}$$

See [9].

§5.5 A Deterministic Maximization Problem

Let us write

(1) $f_n(a,b)$ = the maximum value of $\sum_{i=1}^{n} x_i$ subject to the constraints of (2.5.1).

Then clearly for $n \geq 2$

(2) $$f_n(a,b) = \underset{(x_1)}{\text{Max}} \left[x_1 + f_{n-1}(a-F(x_1), b-G(x_1))\right],$$

where x_1 is bound by the constraints

(3) $$0 \leq x_1 \leq \text{Min}\left[F^{-1}(a), G^{-1}(b)\right],$$

and

(4) $$f_1(a,b) = \text{Min}\left[F^{-1}(a), G^{-1}(b)\right].$$

Regardless of the value of N each maximization in (2) is a two-dimensional problem and can be programmed for computing machines quite easily.

§6. Successive Approximations

Generally speaking, the functional equations of the above sections are as intractable as far as exact solutions are concerned as the differential equations of mathematical physics, engineering or mathematical economics are. They serve, however, two useful purposes. In the first place, a great many structural properties of optimal policies can be deduced from simple properties of the coefficient functions which appear. Secondly, the method of successive approximations can be used to good effect to compute the maximum return, and in that way the optimal policies.

In the allocation and optimal inventory problems discussed above, the use of successive approximations is quite natural. We are merely computing the N-stage return for $n = 1, 2, \ldots$. If n is large, and we are not particularly interested in the results for small n, this method of approximation is not as efficient as others we can devise.

One particularly useful approximation is that of an unbounded process. Thus, for example, in (3) of §5.2, we may, if n is large replace $f_n(x)$ by $f(x) = \lim_{n\to\infty} f_n(x)$. This function satisfies the equation

$$(1)\quad f(x) = \max_{0\le y\le x} \left[p_1 q_1 \left[g_1(y) + h_1(x-y) + f(a_1 y + b_1(x-y)) \right] \right.$$

$$+p_2q_1\left[g_2(y)+h_1(x-y)+f(a_2y+b_1(x-y))\right]$$
$$+p_1q_2\left[g_1(y)+h_2(x-y)+f(a_1y+b_2(x-y))\right]$$
$$+p_2q_2\left[g_2(y)+h_2(x-y)+f(a_2y+b_2(x-y))\right]\Big], \quad x > 0,$$

with $f(0) = 0$.

This equation may now be solved by the method of successive approximations, using not the original sequence $\{f_N(x)\}$, the sequence of N-stage returns, but a sequence chosen to approximate $f_\infty(x)$ more closely from the very beginning.

We shall discuss a simple way of doing this in the next section.

§7. Approximation in Policy Space

A characteristic of great importance possessed by these dynamic programming processes is the duality between the return function and the policy which yields this function. This duality is actually inherent in many other functional equations, but not as obviously, cf. [10], [11], [12], [13].

Let us consider equation (6.1) to illustrate our remarks. We write this equation in the form

$$(1) \quad f(x) = \max_{0 \le y \le x} T(f,y).$$

The quantity $y = y(x)$, the allocation when we have an amount x of resources, we call the policy function or policy.

Observe that the choice of a policy yields a return function, and conversely if we have obtained f(x), explicitly or by some iterative

technique, then the maximization of $T(f,y)$ yields all optimal value of y, and hence all optimal policies.

It follows that we have the privilege of solving the problem by determining optimal policies or maximum returns. The method we presented above was approximation in function space. It turns out that the alternative method of approximation in policy space possesses very important theoretical and computational advantages.

Let us discuss the theoretical advantages first. We obtain a first approximation by choosing an initial policy $y_0 = y_0(x)$. Using this policy, we compute $f_0(x)$ by iteration, using the formula

$$(2) \quad f_0(x) = T(f_0, y_0).$$

We now define

$$(3) \quad f_1(x) = \max_{0 \le y \le x} T(f_0, y).$$

It is clear that $f_1(x) \ge f_0(x)$. Continuing we define

$$(4) \quad f_{n+1}(x) = \max_{0 \le y \le x} T(f_n, y), \quad n \ge 1.$$

Since $T(f,y)$ is a positive operator, $f_1 \ge f_0$ yields $f_2 \ge f_1$ and hence, inductively $f_{n+1} \ge f_n$. We thus have monotone convergence. The applications of this concept to other fields such as the calculus of variations have been discussed in [10], [11], [12], [13].

Let us now discuss the practical aspects. Many of the processes we consider have been carried out for some time in the real world.

In the course of these years, a great deal of experience has been gained and a number of techniques have been devised which yield results considerably better than what might be obtained by an inexperienced person.

Consequently, in all these cases we have fair approximations in policy space, even though the corresponding approximations in function space may be completely lacking.

BIBLIOGRAPHY

1. Bellman, R., "The Theory of Dynamic Programming", Proc. Nat. Acad. Sci., Vol. 38 (1952), p. 716-719.

2. ———, "An Introduction to the Theory of Dynamic Programming", RAND Report No. R-245, 1953.

3. ———, "Some Problems in the Theory of Dynamic Programming", Econometrica, January 1954, pp. 37-48.

4. ———, "The Theory of Dynamic Programming - A Review", Operations Research Quarterly, August 1954.

5. ———, "The Theory of Dynamic Programming", Bull. Amer. Math. Soc. (to appear).

6. ———, "Bottleneck Problems, Functional Equations and Dynamic Programming", Econometrica (to appear).

7. ———, "On Some Mathematical Problems Arising in the Theory of Optimal Inventory and Stock Control", RAND Paper No. P-580.

8. ———, "Decision-Making in the Face of Uncertainty I, II", Navy Quarterly of Logistics (to appear).

9. ———, "A Problem in the Sequential Design of Experiments", (to appear).

10. ———, "Dynamic Programming and a New Formalism in the Calculus of Variations", Proc. Nat. Acad. Sci., Vol. 40 (1954), pp. 231-5.

11. ———, "Dynamic Programming and Continuous Processes", RAND Report No. R-271, November 1954.

12. ———, "Monotone Convergence in Dynamic Programming and the Calculus of Variations", Proc. Nat. Acad. Sci., November 1954.

13. ———, "Dynamic Programming and a New Formalism in the Theory of Integral Equations", Proc. Nat. Acad. Sci., (to appear).