DYNAMIC PROGRAMMING AND FEEDBACK CONTROL

R. Bellman
Mathematics Division

R. Kalaba
Engineering Division
The RAND Corporation

P-1778

August 24, 1959

To be presented at The First International Congress on Automatic Control in Moscow, USSR - 1960.

ABSTRACT

By abstractly viewing feedback control processes as multi-stage decision processes, the functional equation approach of dynamic programming may be uniformly employed in the mathematical formulation and analysis of deterministic, stochastic, and adaptive control processes. Furthermore, problems in the calculus of variations may be looked upon as continuous decision problems.

Emphasis is upon the development of methods which are well-suited for high-speed digital computation.

I. INTRODUCTION

Dynamic programming, [1], the name coined for the study of multi-stage decision processes, is a field of mathematics which has been intensely cultivated during the last decade. It is possible, as will be shown, to view the operation of a feedback control system -- with its typical cycle of measurement of the state of the system, determination of a control decision, and transformation of the state of the system -- as a multi-stage decision process, the purpose of which is to optimize a system performance index or cost. As such, the functional equation technique of dynamic programming may be used as a guide in the formulation and solution of a variety of feedback control problems.

The purpose of this expository paper is to describe applications of dynamic programming methodology to three broad classes of control problems which are differentiated on the basis of the controller's knowledge of the processes under consideration. These are referred to as deterministic, [2], stochastic, [3], and adaptive, [4, 5, 6], control processes. We are interested primarily in the way in which dynamic programming can be used to construct mathematical models of control processes of the types mentioned. In some fortunate cases analytical results can be obtained, though, in general, emphasis is upon obtaining algorithms which are suitable for use with modern high-speed digital computing machines.

We first provide a treatment of deterministic feedback control processes, processes for which the controller has complete knowledge of the state of the system at each decision making opportunity, complete knowledge of the outcome of each decision made, criterion, etc. Here determinism, known cause and effect, reigns supreme.

In many realistic situations, however, unknown elements are present which are usually incorporated into mathematical models by the introduction of stochastic variables with known distribution functions. The controller must then perform under conditions of uncertainty, which leads to a discussion of stochastic control processes.

But even this is not adequate in many situations for, initially at least, the distribution functions of the stochastic variables may not be known precisely. In some instances, though, as the process unfolds, additional information concerning the unknown factors, relationships and influences may become available to the controller which may then "learn" to improve its performance based upon experience. In this case the controller adapts itself to circumstances as it finds them. Development of adequate theories of adaptive control is imperative for the understanding and development of automata and machines that learn.

The last two sections are devoted to problems in the calculus of variations as continuous decision processes and generalized trajectory guidance problems.

The aim throughout is to show how wide classes of optimal feedback control design problems, differing greatly in underlying assumptions, can be handled in unitary fashion through the functional equation technique of dynamic programming.

II. FEEDBACK CONTROL PROCESSES AS MULTI-STAGE DECISION PROCESSES

Let us focus our attention on a system S, the state of which at any time is specified by a state vector p, a point in a phase space. At certain times t_0, t_1, ... , the system S undergoes changes of state, which mathematically speaking, correspond to transformations of the state vector p. We limit ourselves to a discrete-valued time variable, rather than a continuous one, for conceptual reasons which will become apparent. In any event, the use of a digital computer automatically renders all variables discrete.

Furthermore, we shall assume that the choice of the precise transformation to be used at each time t_i is to be made by a controller, human or otherwise. At each stage the controller may then choose one of the admissible transformations T, indexed by the decision vector q. Thus, we may write

$$p_{i+1} = T(p_i, q_i), \quad i=0,1,2, \ldots , \tag{1}$$

to indicate that if the system is in state p_i at time t_i and the control decision q_i is made, then the system is transformed into the state p_{i+1} at time t_{i+1}.

Let us suppose that the control process is of finite duration and that it terminates at time t_N. We shall measure the effectiveness of the control by means of a function $\phi(p_N)$ which gives the worth of terminating the process with the system in some state p_N. The objective of the control process is to maximize $\phi(p_N)$. This is an example of a terminal control process.

Let us now review the course of the process, viewed as a sequence of transformations. Initially the system is in state $p = p_0$, and the decision q_0 is made. Then we have

$$p_1 = T(p, q_0) \tag{2}$$

for the state of the system S at time t_1.

Next, with the system in state p_1 the decision q_1 is mode, which results in the transformation of the system into the state p_2 where

$$p_2 = T(p_1,q_1). \tag{3}$$

In like manner we have

$$p_{n+1} = T(p_n,q_n),\ n = 0,1,2, \ \dots\ , N-1. \tag{4}$$

The objective of the control decisions is to attain

$$\underset{[q_0,q_1, \ \dots\ , q_{N-1}]}{\text{Max}} \phi(p_N). \tag{5}$$

We call the sequence of decisions $[q_0,q_1, \ \dots\ , q_{N-1}]$ a <u>policy</u> and a policy for which $\phi(p_N)$ attains its maximum is called an <u>optimal policy</u>.

The temptation is to view this as a straightforward maximization of a function of N variables. In general, though, it will not be possible to do this, for the methods of differential calculus leads to systems of equations with N variables, a formidable task for solution in general, even if the equations should be linear.

If each q can have any of 10 values and N is fifty, then the results of 10^{50} possible policies have to be computed which rules out a simple enumerative approach, in such cases, even where use of a digital computer is contemplated. Furthermore, for purposes of feedback control, this is not the form in which the solution is desired.

What we can do is bring the feedback aspect of the process into the foreground. Rather than seeking an optimal policy (set of q's) with respect to an initial state at time t_0, we shall determine a best first decision for the controller to make in terms of any state of the system p_k at any time t_k.

Instead of considering the isolated original process we imbed the original process within the class of all processes beginning with the system in state p_k at time t_k, k = 0,1,2, ... , N-1, the process terminating at time t_N as before. The fundamental quantity to be determined is the decision for the controller to make at time t_k with the system then in state p_k. Notice that the earlier history of the process is unimportant; only the state at time t_k is relevant in making this decision.

Next let us introduce the sequence of functions $f_k(p)$, where

$$f_k(p_k) = \max_{[q_k, q_{k+1}, \cdots, q_{N-1}]} \phi(p_N), \quad k=0,1,2, \ \cdots \ , N-1. \tag{6}$$

This function is readily determined for the one stage process beginning at time t_{N-1} with the system in state p_{N-1}. We find

$$f_{N-1}(p_{N-1}) = \max_{q_{N-1}} \phi(p_N)$$

$$= \max_{q_{N-1}} \phi(T(p_{N-1}, q_{N-1})). \tag{7}$$

To determine the functions f_{N-2}, f_{N-3}, ... , we make use of the <u>principle of optimality</u>, which states that an optimal policy has the property that whatever the initial decision is, the remaining decisions must constitute an optimal policy with respect to the state resulting from the initial decision, [1].

This yields the functional equations

$$f_{k-1}(p_{k-1}) = \max_{q_{k-1}} f_k(T(p_{k-1}, q_{k-1})), \tag{8}$$

for k = N-1, N-2, ... , 1.

Since $f_{N-1}(p_{N-1})$ can be obtained from Eq. (7), the function $f_{N-2}(p_{N-2})$ can be determined from Eq. (8), and so on. A digital computer can carry out the

indicated maximizations by a search technique in a reasonable length of time provided that the number of points at which each function is to be evaluated is not excessive (about 10^4 for current day computers).

It is important to realize that for a given state of the system p_{k-1} at time t_{k-1} a value of q_{k-1} which maximizes $f_k(T(p_{k-1}, q_{k-1}))$ is a best control decision to make under those circumstances so that q^*_{k-1}, a maximizing value of q_{k-1}, can be determined as a function of the system state p_{k-1} at time t_{k-1}

$$q^*_{k-1} = q_{k-1}(p_{k-1}, t_{k-1}), \tag{9}$$

precisely what is desired for optimal feedback control. Some details of the numerical aspects are discussed in Reference 1, and a more complete exposition will be available in Reference 7.

It is interesting to note that Eqs. (7) and (8) represent an extension of the usual causality relations of mathematical physics for purely descriptive, as opposed to control or variational processes. There at times $t_o, t_1, \ldots, t_{N-1}$ the system is transformed into new states by the transformation T, where no control decision is involved. We have

$$p_{k+1} = T(p_k), \quad k=0,1,2, \ldots, N-1, \tag{10}$$
$$p_o = p.$$

If now we let

$$g_k(p_k) = \text{the state of the system at time } t_N \text{ with the system in state } p_k \text{ at time } t_k, \tag{11}$$

we find

$$g_{N-1}(p_{N-1}) = T(p_{N-1})$$

(12)

$$g_{k-1}(p_{k-1}) = g_k(T(p_{k-1})),$$

$$k = N-1, N-2, \ldots, 0,$$

which are decision-free versions of Eqs. (7) and (8).

A general discussion of the notion of invariant imbedding -- the imbedding of a process within an appropriate class of processes for purposes of analysis is available in our paper, [8], and many additional references are given there to applications in other fields including wave propagation, neutron transport, random walk, etc.

III. STOCHASTIC CONTROL PROCESSES

In many situations in which control must be exerted to cause a system to perform satisfactorily, the result of a particular control decision cannot be specified precisely in advance. This is true in electromechanical systems subjected to random disturbing forces as well as to the control forces. Furthermore, in the control of economic and business systems, uncertainty regarding the outcome of a decision is the rule rather than the exception. Let us now recount the dynamic programming approach to such control processes.

We assume that the outcome of the choice of the decision q with the system in known state p is the transformation of the system into the random state r, where the probability distribution function of the random variable r is $G(r;\ p,q)$. (Uncertainty can enter in other ways which we shall not discuss here; for a treatment of a problem involving lack of knowledge of the precise state of the system before and after a decision is made, see Reference 1, pp. 48 - 49.) Before each control decision is made it is assumed that the controller is informed of the outcome of the previous decision; i.e., the current state of the system is made available to the controller.

Due to the presence of the random influences, it is now in general no longer possible to guarantee a particular outcome for the process. Reasonably we can only seek to maximize the expected value of $\phi(p_N)$, an expected outcome.

As before, let us consider the class of decision processes beginning at time t_k with the system in state p_k and terminating at time t_N, where

$k = 0, 1, 2, \ldots, N-1$. We define the sequence of functions

(13) $f_k(p_k)$ = the expected value of $\phi(p_N)$ for a process beginning at time t_k in state p_k and using an optimal control policy.

In the one-stage process we have

$$(14) \qquad f_{N-1}(p_{N-1}) = \max_{q_{N-1}} \int \phi(p_N)\, dG(p_N;\, p_{N-1},\, q_{N-1}),$$

and for the process beginning at time t_{k-1} the principle of optimality yields

$$(15) \qquad f_{k-1}(p_{k-1}) = \max_{q_{k-1}} \int f_k(p_k)\, dG(p_k;\, p_{k-1},\, q_{k-1})$$

for $k = N-1, N-2, \ldots, 1$.

If the system must be in one of a finite number of states z_m, $m = 1,2, \ldots, M$, then the Stieltje's integrals of the previous equations reduce to sums, and Eq. (15) becomes

$$(16) \qquad f_{k-1}(z_m) = \max_{q} \sum_{n=1}^{M} p(z_n;\, z_m,\, q)\, f_k(z_n).$$

For concrete applications of the foregoing equations to a communication problem see References 9 and 10, and for an application to a nonlinear system governed by an inhomogeneous Van der Pol equation subjected to a random disturbance see Reference 11.

It should be noted that decision-free versions of the Eqs. (15) and (16) take the form of the well-known Chapman-Kolmogoroff equations.

IV. ADAPTIVE CONTROL PROCESSES

By an adaptive feedback control process we shall mean a feedback control process in which the controller has an initial lack of information which, however, can be at least partially made up in the course of the process as the controller "learns" more and more about the unknown factors and relationships. The adaptive controller is thus confronted with the problem of simultaneously learning and deciding. In this generality it is unlikely that we shall ever have a completely satisfactory theory of adaptive control, for ignorance can manifest itself in endless ways. The controller may have incomplete (or inaccurate) information concerning the state of the system, the set of possible decisions, the outcomes of decisions, the duration of the process, and even the objective of the process. Furthermore, the methods for handling uncertainty--though quite uniformly based on the notions of random variables, probability distributions, and expected values--are limited only by the imagination of the mathematical analyst. Here we provide a treatment which is a direct extension of the theoretical framework erected in earlier sections. For a discussion along game theoretical lines see the book [5].

We assume that the outcome of each control decision is a random state of the system, the conditional probability distribution of which is not known to the controller. We do, however, wish to assume that the controller knows the physical state of the system before each decision and that it has an a priori estimate of the conditional probability distribution for the outcome of any particular decision, which, in the absence of further information is to be regarded as the actual distribution. Furthermore, we assume that the controller knows how to modify this distribution in the light of additional

information. In some statistical applications determination of the distribution function is the primary objective; in some engineering control problems this is incidental to the physical control of a system in satisfactory fashion.

Notice that now the state of the system consists of a point in phase space p and an information pattern P which summarizes the controller's knowledge of the system's response to a control decision.

As a consequence of a control decision q, p is transformed into a new point p_1, given a priori by the transformation

$$p_1 = T_1(p,P; q; r), \tag{17}$$

and P is transformed into a new information pattern given a priori by the transformation

$$P_1 = T_2(p,P; q,r). \tag{18}$$

Here r is a random variable with an a priori distribution function $G(p,P; q,r)$ which is in itself a part of the information pattern. Finally, we assume that the controller is permitted to observe the actual outcome of each decision made before making the next decision, and, in addition, knows the transformations T_1 and T_2. In many adaptive control processes, we hasten to point out, the determination of these transformations and even of the information pattern itself is an essential part of the problem.

Within this format we wish to determine an optimal decision for the controller to make at time t_k with the system in state (p,P). An optimal decision is one which leads to a minimal estimated expected value of $\phi(p_N,P_N)$. We write "estimated" because all expectations are taken with regard to

currently estimated distribution functions and not unknown actual distributions.

Following our earlier discussions we introduce the sequence of functions

$$f_k(p,P) = \text{Min Exp } \phi(p_N,P_N), \text{ for } k = N-1, N-2, \ldots, 0. \tag{19}$$

The principle of optimality then yields the functional relationships

$$f_k(p,P) = \min_q f_{k=1}(T_1(p,P;\, q;\, r),\, T_2(p,P;\, q;\, r)) \, dG(p,P;\, q;\, r) \tag{20}$$

$$K = N-1, N-2, \ldots, 0,$$

and the relationships

$$f_{N-1}(p,P) = \min_q \phi(T_1(p,P;\, q;\, r),\, T_2(p,P;\, q;\, r)) \, dG(p,P;\, q;\, r). \tag{21}$$

These formulas can be used to establish the existence of optimal policies and various structural characteristics, as in [1].

In view of its dependence upon the information pattern P, the sequence f_k may be a sequence of functionals. A number of devices, some known in mathematical statistics under the term sufficient statistics and others, permit us to reduce this sequence to a sequence of functions, a reduction that is important from the numerical and analytic viewpoint. See [12,13].

In view of the grave difficulties in treating uncertainty no pretense is made of having erected a definitive theory of adaptive control processes. What we have done, though, is create a conceptual framework which will permit many adaptive processes to be formulated in precise, if not unique, terms treated mathematically. Furthermore, in favorable circumstances this theory will lead with probability one to the same control policy as in the stochastic case, a point which will be discussed in detail in later papers.

V. DYNAMIC PROGRAMMING AND THE CALCULUS OF VARIATIONS

To apply these ideas to feedback control problems described by differential equations of the form

$$\frac{dx}{dt} = g(x,y), \quad x(0)=c, \tag{22}$$

where x is an N- dimensional vector specifying the state of the system and y is an M- dimensional control vector to be chosen so as to minimize a functional of the form

$$J(y) = \int_0^T h(x,y)\, dt + k(x(T),y(T)), \tag{23}$$

we conceive of this process as a multi-stage decision process of continuous type.

Writing

$$f(c,T) = \operatorname*{Min}_y J(y), \tag{24}$$

the principle of optimality yields the nonlinear partial differential equation

$$f_T = \operatorname*{Min}_v \left[h(c,v) + \sum_{i=1}^{N} g_i(c,v) \frac{\partial f}{\partial c_i} \right], \tag{25}$$

where $c_1, c_2, \ldots, c_N$ are the N components of c. This equation furnishes an analytic and computational approach to large classes of feedback control problems; see [1], [5],

VI. GENERALIZED TRAJECTORY PROCESSES

A class of problems that occurs throughout applied mathematics is concerned with the problem of converting a system from one state to another in minimum time, or at minimum cost of resources in general. We shall call such a problem a generalized trajectory problem.

Let p and p_T represent the initial and terminal states of a system, and suppose that t(p,q) represents the "cost" of going from the state p to the state q. Let f(p) represent the minimum cost entailed in going from a generic initial point p to the fixed terminal point p_T. Then the principle of optimality yields the basic functional equation

$$f(p) = \min_{q} \; (T(p,q) + f(q)). \tag{26}$$

This equation can be used to obtain numerical solutions to trajectory problems, [14], [15], to problems in chemical process engineering, [16], and to some problems of systems synthesis, [17], [18].

REFERENCES

1. Bellman R., Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1957.

2. Bellman, R., "On the Application of the Theory of Dynamic Programming to the Study of Control Processes," Proc. Symposium on Nonlinear Circuit Analysis, Polytechnic Institute of Brooklyn, Brooklyn, New York, 1957, pp. 199-213.

3. Bellman, R., "Dynamic Programming and Stochastic Control Processes," Information and Control, V. 1 (1958), pp. 228-239.

4. Bellman, R., and R. Kalaba, "Dynamic Programming and Adaptive Processes -- A Mathematical Foundation," IRE Transactions on Automatic Control, to appear.

5. Bellman, R., Adaptive Control; A Guided Tour, Princeton University Press, Princeton, New Jersey, 1960.

6. Bellman, R., and R. Kalaba, "On Adaptive Control Processes," 1959 IRE National Convention Record, Vol. 7, Part 4, (July, 1959), pp. 3-11.

7. Bellman, R., and S. Dreyfus, Computational Aspects of Dynamic Programming, Princeton University Press, Princeton, New Jersey, 1960.

8. Bellman, R., and R. Kalaba, "Functional Equations in Adaptive Processes and Random Transmission," IRE Transactions on Circuit Theory, v. CT-6 (Special Supplement, May, 1959), pp. 271-282.

9. Bellman, R., and R. Kalaba, "On the Role of Dynamic Programming in Statistical Communication Theory," IRE Transactions on Information Theory, v. IT-3 (1957), pp. 197-203.

10. Bellman, R., and R. Kalaba, "On Communication Processes Involving Learning and Random Duration," 1958 IRE National Convention Record, Part 4 (1958), pp. 16-21.

11. Bellman, R., and R. Kalaba, "On Adaptive Control Processes," 1959 IRE National Convention Record, Part 4 (1959), pp. 3-11.

12. Freimer, M., Ph.D. thesis, Harvard University, to appear.

13. Mood, A., Introduction to the Theory of Statistics, McGraw-Hill Book Company, Inc., New York, 1950.

14. Cartains, T., and S. Dreyfus,"Application of Dynamic Programming to the Airplane Time-to-climb Problem," Aero. Eng. Rev., 1957.

15. Kalaba, R. "On Some Communication Network Problems," a paper in the book Combinatorial Designs and Analysis, American Mathematical Society, Providence, R. I., 1960.

16. Aris, R., R. Bellman, R. Kalaba, "Some Optimization Problems in Chemical Engineering," to appear.

17. Bellman, R. J. Holland, R. Kalaba, "On the Application of Dynamic Programming to the Synthesis of Logical Systems," Journal of the Association for Computing Machinery, to appear.

18. Ash, M., R. Bellman, R. Kalaba, "On Control of Reactor Shut-down Involving Minimal Xenon Poisoning," Nuclear Science and Engineering, to appear.