INFORMATION AND CONTROL 1, 228–239 (1958)

Dynamic Programming and Stochastic Control Processes

RICHARD BELLMAN

The Rand Corporation, Santa Monica, California

Consider a system S specified at any time t by a finite dimensional vector $x(t)$ satisfying a vector differential equation $dx/dt = g[x, r(t), f(t)]$, $x(0) = c$, where c is the initial state, $r(t)$ is a random forcing term possessing a known distribution, and $f(t)$ is a forcing term chosen, via a feedback process, so as to minimize the expected value of a functional $J(x) = \int_0^T h(x - y, t)\, dG(t)$, where $y(t)$ is a known function, or chosen so as to minimize the functional defined by the probability that $\max_{0 \leq t \leq T} h(x - y, t)$ exceed a specified bound. It is shown how the functional equation technique of dynamic programming may be used to obtain a new computational and analytic approach to problems of this genre. The limited memory capacity of present-day digital computers limits the routine application of these techniques to first and second order systems at the moment, with limited application to higher order systems.

1. INTRODUCTION

In this paper, we wish to indicate the application of the functional equation techniques of the theory of dynamic programming to the formulation and computational solution of various types of variational problems arising in the study of control processes with stochastic elements. Although the methods displayed below are intimately related to those we have previously presented in connection with deterministic control processes (compare Bellman, 1956a, b; 1957a), the presence of stochastic effects introduces, as might be expected, new difficulties of both conceptual and analytic nature which must be carefully examined.

A fundamental problem, arising in numerous applications, is that of determining feedback control which will neutralize random disturbances. These disturbing influences are usually called "noise."

Here we shall consider the following particular version of this general

question. Let S be a physical system, specified at any time t by a finite dimensional vector $x(t)$. This vector is determined as a function of time, and the initial state of the system, by means of the differential equation

$$dx/dt = g[x, r(t)], \qquad x(0) = c. \quad (1)$$

The function $r(t)$ appearing on the right is a random function of time with known properties.

We shall not discuss here the far more difficult questions which arise from the study of processes in which $r(t)$ is only imperfectly known initially and is then determined more and more accurately as the process continues. The reader interested in these matters will find discussions of this type of problems and further references in Robbins (1952) and Bellman and Kalaba (1958).

A particularly important case, from the standpoint of both analysis and application, is that where $g[x, r(t)]$ is linear in both x and $r(t)$. The equation in (1) then has the simple form

$$dx/dt = Ax + r(t), \qquad x(0) = c. \quad (2)$$

A rigorous formulation of the theory of nonlinear differential equations with stochastic elements presents certain difficulties which we shall not enter into here for reasons we shall detail below. The linear equation, however, has been treated at great length in a number of papers in full rigor; [Compare Doob (1944); see also, Bellman, *et al.* (1954) and the recent papers of Booton (1956a, b).] Equations of the form

$$dx/dt = [A + R(t)]x,$$

where $R(t)$ is a random matrix can also be treated in some detail.

We are primarily interested here in the case where $g[x, r(t)]$ is nonlinear or where other nonlinearities arise, in a fashion we shall discuss below, to a sufficient degree to destroy any hope of using explicit analytic solutions to resolve control problems.

To counteract the influence of $r(t)$, and simultaneously to direct the unperturbed system along more desirable lines, we introduce "feedback control" in the form of a vector function $v(t)$. The defining function now has the form

$$dx/dt = g[x, r(t), v(t)], \qquad x(0) = c, \quad (3)$$

where $v(t)$ is a function of the state of the system at time t and the time t itself; that is, $v(t) \equiv v[x(t), t]$.

Let us denote by $y(t)$ the solution of the unperturbed-uncontrolled equation

$$dy/dt = g(y), \qquad y(0) = c. \quad (4)$$

In some cases, we may wish to keep x close to y over the time interval $[0,T]$. We agree then to measure the deviation from y by means of a functional of the form

$$J(v) = \int_0^T h(x - y)\, dG(t), \quad (5)$$

where $h(z)$ is a scalar function of the vector z. By introducing a step discontinuity at $t = T$, we can combine deviation over the interval with terminal control.

At other times, the function y need not be a solution of the unperturbed system but merely a desirable state of the system. In both cases, we see that we wish to determine the control vector $v(t)$ so as to minimize a prescribed functional of x and v which can be written

$$J(v) = \int_0^T h(x, v, r)\, dG(t). \quad (6)$$

Since the functional itself will be, in general, a stochastic quantity, in order to make this statement precise we must first average $J(v)$, in some suitable fashion, over the class of random functions which occur. The problem we wish to consider is that of minimizing this expected value of a function of $J(v)$, subject to constraints on $v(t)$.

A rigorous formulation of variational problems involving stochastic functions is again a matter of some difficulty. We shall avoid both this difficulty and the one mentioned concerning the meaning of stochastic differential equations by considering only *discrete* control processes. In this way, we replace differential equations by difference equations, integrals by sums, and stochastic functions by stochastic sequences. The reason for this change in format lies not so much in our desire to avoid occasionally unpleasant rigorous details, as in our desire to prepare the problem for solution by means of a digital computer.

Nothing for nothing, however! It is now a matter of some significance to study the connection between the original continuous process and the approximating discrete process. Not only is it important to know whether or not the respective minimum values are close, but it is also important to know whether the corresponding policies bear any similarity. Further-

more, the rate of convergence of the discrete process to the continuous process must be studied. This is critically dependent upon the type of discrete approximation which is employed. Some preliminary results in these directions may be found Bellman (1957a, b).

It should constantly be kept in mind that both continuous and discrete processes are approximations to the actual physical process. The important point is not so much their similarity to each other as the value of either mathematical model in treating the actual control process.

We shall first apply the functional equation technique to the general variational problem posed above. Then, as a simple example, we shall discuss its specific application to the problem of determining the scalar function $v(t)$ in such a way as to minimize the expected value of the functional

$$\int_0^T u^2\,dt + |\,u(T)\,| \tag{7}$$

where u is the solution of the Van der Pol equation with the forcing terms $r(t)$ and $v(t)$,

$$\begin{aligned} u'' + \lambda(u^2 - 1)u' + u &= r(t) + v(t), \\ u(0) = c_1\,, \qquad u'(0) &= c_2\,. \end{aligned} \tag{8}$$

To show the versatility of the method, we shall then show how to treat by means of recurrence relations the problem of minimizing the probability that $J_1(v) \geqq d$, where

$$J_1(v) = \max_{0 \leqq t \leqq T} \|\, x - y \,\|. \tag{9}$$

Here $\|\, z \,\|$ is the norm of z defined in one of the usual ways. A treatment of the deterministic version of this problem may be found in Bellman (1957c).

Finally, we shall discuss a case in which the random function $r(t)$ possesses a correlation with the value of $r(t - \Delta)$. Here t assumes only the values $\Delta, 2\Delta, \cdots$.

As a subsequent discussion of the specific equation mentioned above will show, the functional equation technique of dynamic programming furnishes a feasible computational solution for second order systems, without regard to the analytic character of either the equation or the criterion function, $J(v)$. Although equations of higher order cannot be treated at the moment by means of the same straightforward approach,

more refined analytic and computational techniques recently developed appear to offer an approach to the successful treatment of control problems for higher dimensional systems (see Bellman, 1957d, 1958).

2. FEEDBACK CONTROL AS A MULTISTAGE DECISION PROCESS

Let us now see how we can interpret feedback control as a multistage decision process. To begin with, we observe c, the initial state of the system, and make an initial choice of a control vector $v(0)$. As a result of the initial random effect $r(0)$, we find ourselves at time Δ in a new state c', determined by the equations governing the system, and we are required to make a new choice of a control vector. This situation repeats itself at times 2Δ, 3Δ, and so on.

The salient fact that enables us to break this complex process down into a sequence of simple processes is the dependence of the future upon the present, and not upon the past, or upon how the past became the present. Starting from any state at any time, say t_0, we exert control in such a way as to minimize the deviation from that time t_0 until the process ends. Whatever deviation has occurred in the past does affect the total cost of deviation of the system as measured, say, by the integral in (1.6), but does not affect the sequence of choices we make from the time t_0 on. This sequence of choices depends only upon the state of the system at this particular time t_0 and the behavior of the stochastic vector $r(t)$ from t_0 on.

This statement, which perhaps appears paradoxical at first glance and is certainly rather difficult to express verbally, is a simple consequence of the additivity of integrals, i.e.,

$$\int_0^T = \int_0^{t_0} + \int_{t_0}^T, \tag{10}$$

and the fact that the solution of a differential equation of the form given in (3) is for $t \geqq t_0$ dependent only upon its value at t_0 and the values of $r(t)$ for $t \geqq t_0$.

Let us call a *policy* any choice of $v(t)$ subject to the constraints imposed, and an *optimal policy* a policy which minimizes the prescribed criterion function. Then the remarks we have made above concerning the independence of future behavior from the past history of the process are particular consequences of what we have called the *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.

The analytic translation of this statement yields functional equations that lead to a computational solution of the control process described above. See Bellman (1957a) for further discussion and applications.

Finally, let us note in passing that, as we have discussed elsewhere (1956a, b; 1957a), not only can the variational problems derived from the study of control processes be considered to be multistage decision processes, but actually the wider discipline of the calculus of variations itself can be considered to be part of the general theory of multistage decision processes of continuous type.

3. DISCRETE VERSIONS OF CONTROL PROCESSES

Let us now prepare the way for the use of digital computers. We begin by replacing the continuous process described in the introduction by a discrete process. The interval $[0, T]$ is divided into N parts of length Δ, so that $N\Delta = T$, and t is allowed to assume only the values $0, \Delta, 2\Delta, \cdots, N$. To simplify the notation, let us write

$$x(k\Delta) = x_k, \qquad r(k\Delta) = r_k, \qquad v(k\Delta) = v_k \tag{11}$$

and replace the differential equation (3) by the difference equation

$$x_{k+1} - x_k = g(x_k, r_k, v_k)\Delta, \qquad x_0 = c. \tag{12}$$

There is now no difficulty as to what we mean by a stochastic sequence of values $\{x_k\}$ as generated by the difference equation in (12). The random sequence of vectors $\{r_k\}$ constitute a much more prosaic set than the set of values assumed by a random function $r(t)$, and one much easier to contemplate.

Instead of choosing a function $v(t)$ which minimizes the expected values of a functional, we wish to choose a sequence of vectors $\{v_k\}$ which minimize the expected value of a function,

$$J(\{v_k\}) = \sum_{k=0}^{N-1} h(x_k, r_k, v_k) + m(x_N). \tag{13}$$

This is a well-formulated problem with no conceptual loose ends.

In the next section, we shall show how the functional equation technique of dynamic programming may be applied to the problem just posed.

4. FUNCTIONAL EQUATIONS

Consider the problem of minimizing the expected value of

$$J(\{v_k\}; a) = \sum_{k=a}^{N-1} h(x_k, r_k, v_k) + m(x_N), \tag{14}$$

over sequences $\{v_k\}$, $k = a, a + 1, \cdots, N - 1$, where a is one of the quantities 0, 1, 2, $\cdots$, $N - 1$. As in Section 3, x_{k+1} is determined by the relation

$$x_{k+1} - x_k = g(x_k, r_k, v_k), \qquad k \geqq a, \qquad x_a = c. \tag{15}$$

It is clear that the minimum expected value of $J(\{v_k\}; a)$ depends upon c, the state at time a, and upon a itself. Let us then define the function

$$f_a(c) = \min_P \exp_r J(\{v_k\}; a), \tag{16}$$

where the minimum is over all policies P. The function is defined for all c, and for $a = 0, 1, \cdots, N - 1$.

We see that

$$f_{N-1}(c) = \min_{v_N} \exp_{r_N} [h(x_{N-1}, r_{N-1}, v_{N-1}) + m(x_N)], \tag{17}$$

where

$$x_N = c + g(c, r_{N-1}, v_{N-1}). \tag{18}$$

The principle of optimality, stated in Section 2, yields the recurrence relation

$$f_a(c) = \min_{v_a} \left(\exp_{r_a} \{h(c, r_a, v_a) + f_{a+1}[c + g(c, r_a, v_a)]\} \right). \tag{19}$$

Since $f_{N-1}(c)$ is determined by (4), the relation in (6) enables us to compute $f_{N-2}(c)$, and so, step-by-step, eventually $f_0(c)$.

5. AN EXAMPLE

Let us now apply these techniques to a particular example. Consider the Van der Pol equation with a forcing term,

$$\begin{aligned} u'' + \lambda(u^2 - 1)u' + u &= r(t) + v(t), \\ u(0) = c_1, \qquad u'(0) &= c_2, \end{aligned} \tag{20}$$

where the behavior of the random function $r(t)$ will be precisely speci-

fied below, and where it is desired to determine to choose $v(t)$, subject to the constraint

$$-a \leqq v(t) \leqq a, \tag{21}$$

so as to minimize the expected value of

$$J(v) = \int_0^T u^2\,dt + |\,u(T)\,|. \tag{22}$$

In place of the second order equation in (20), we consider the system

$$\begin{aligned} du/dt &= w, \; u(0) = c_1\,, \\ dw/dt &= -\lambda(u^2 - 1)w - u + r(t) + v(t), \qquad w(0) = c_2\,. \end{aligned} \tag{23}$$

This, in turn, is converted into the system of recurrence relations

$$\begin{aligned} u_{k+1} &= u_k + w_k\Delta, \qquad u_0 = c_1\,, \\ w_{k+1} &= w_k + [-\lambda(u_k^2 - 1)w_k - u_k + r_k + v_k]\Delta, \qquad w_0 = c_2\,. \end{aligned} \tag{24}$$

Let us assume that sequence $\{r_k\}$ is a sequence of independent random variables with a common distribution function $dG(r)$. We shall consider the problem of correlation below.

It is desired to choose the sequence of values $\{v_k\}$, subject to the restriction

$$-b \leqq v_k \leqq b, \tag{25}$$

so as to minimize the expected value of

$$J_a(\{v_k\}) = \Delta \sum_{k=a}^{N-1} u_k^2 + |\,u_N\,|\,, \qquad a = 0, 1, 2, \cdots, N-1. \tag{26}$$

Set

$$f_a(c_1\,, c_2) = \min_P \exp_r J_a(\{v_k\}), \tag{27}$$

for $a = 0, 1, 2, \cdots, N-1, -\infty < c_1\,, \quad c_2 < \infty$.

Then

$$f_{N-1}(c_1\,, c_2) = \min_{v_{N-1}} \exp_{r_{N-1}} (c_1^2 + |\,u_N\,|\,) \tag{28}$$

where $u_N = c_1 + \Delta c_2$. Hence

$$f_{N-1}(c_1\,, c_2) = \Delta c_1^2 + |\,c_1 + \Delta c_2\,|. \tag{29}$$

Equation (19) becomes

$$f_a(c_1, c_2) = \min_{v_a} \exp_{r_a} \{\Delta c_1^2 + f_{a+1}[c_1 + \Delta c_2, c_2 + \Delta h(c_1, c_2, r_a, v_a)]\} \quad (30)$$

or

$$f_a(c_1, c_2) = \min_{v_a} \left[\Delta c_1^2 + \int_{-\infty}^{\infty} f_{a+1}(c_1 + \Delta c_2, c_2 + \Delta h)\, dG(r) \right], \quad (21)$$

where

$$h(c_1, c_2, r_a, v_a) = r_a + v_a - c_1 - \lambda(c_1^2 - 1)c_2. \quad (32)$$

The minimization with respect to v_a is over the interval $-b \leqq v_a \leqq b$.

We have thus reduced the solution of the problem to the computation of the sequence of functions of two variables, $\{f_a(c_1, c_2)\}$.

6. DISCUSSION

Let us now discuss in more detail whether the algorithm presented in the previous section is feasible. The concept of feasibility is completely dependent upon the type of computer available. We shall think in terms of a modern high speed digital computer. As far as hand computation is concerned, the method outlined above is definitely not feasible.

To carry out the determination of $f_a(c_1, c_2)$, we must store the values of $f_{a+1}(c_1, c_2)$ in the computer, in one form or another, evaluate the integral over r appearing in (31), and then minimize over v_a.

Let us discuss these operations in turn. When we speak of storing the values of $f_{a+1}(c_1, c_2)$ in the computer, we mean that we must have a method for producing the value of $f_{a+1}(c_1, c_2)$ at any particular point (c_1, c_2) that is desired. There are two ways of accomplishing this. In the first place, we can agree that we are interested only in the points within some square $-s \leqq c_1, c_2 \leqq s$, and then only in the values of the function at a finite set of grid points $(m\delta, n\delta)$, $m, n = -M, -M + 1, \cdots, M$, where $M\delta = s$. If (c_1, c_2) is not a grid point, the value of $f_{a+1}(c_1, c_2)$ is determined by an interpolation formula.

It follows then that storing the values of the function $f_{a+1}(c_1, c_2)$ is equivalent to storing $(1 + 2M)^2$ numbers, the values at $(m\delta, n\delta)$. If $M = 50$, not a particularly fine subdivision if c_1 and c_2 are large, we require approximately 10^4 values. This is a considerable quantity, when we realize that it must be multiplied by 3, to take account of the storage

of the values of the new function $f_a(c_1, c_2)$ and the policy function $v_a = v_a(c_1, c_2)$.

Problems of this magnitude, however, can be treated with the largest of current digital computers and will be routine in a few years with the much larger machines being built at the present.

It is clear, nonetheless, that the storage of functions of many variables cannot be accomplished along the crude lines described above. Any further discussion would take us too far afield. The interested reader may consult [8] for a brief sketch of an entirely different approach.

Turn now to the problem of evaluation of the integral in Eq. (31). Since these studies are all of preliminary nature, it is wise to assume quite simple random effects. Hence if r is taken to assume the values $\pm k$ with equal probability, the expression in (31) becomes

$$\tfrac{1}{2}[f_{a+1}(c_1 + \Delta c_2, c_2 + \Delta h(c_1, c_2 k, v_a)) + f_{a+1}(c_1 + \Delta c_2, c_2 + \Delta h(c_1, c_2, -k, v_a))]. \tag{33}$$

There is thus no difficulty in this evaluation.

Finally, consider the problem of determining the minimum over v_a. For a variety of reasons, we do not wish to follow any conventional lines involving the use of derivatives. Hence, we choose a grid in the v_a-interval, say $v_a = -q\delta_1, -(q-1)\delta_1, \cdots, q\delta_1 = b$, and minimize only over the discrete set of values $\pm l\delta_1$. To do this, we need only compare numerical values at these points. If further accuracy is desired, interpolation can again be used.

A very important aspect of this direct minimization is that the presence of constraints aids rather than hurts. The more constraints, the smaller the allowable choices of v_a and the more rapid the numerical search. In particular, the simplest case is that which is occasionally called "bang-bang" control (compare Bellman *et al.*, 1956), where v_a is allowed to assume only the values $\pm b$.

7. MINIMUM OF MAXIMUM DEVIATION

So far, we have been considering variational problems of fairly conventional type. Using the same second order equation as in Section 5, let us consider the problem of determining $v(t)$ so as to minimize the probability that

$$\max_{0 \leqq t \leqq T} |u| \geqq d. \tag{34}$$

The discrete version requires us to minimize the probability that

$$\max (| u_0 |, | u_1 |, \cdots, | u_{N-1} | \geqq d). \tag{35}$$

The observation that

$$\max (| u_0 |, | u_1 |, \cdots, | u_{N-1} |) = \max [| u_0 |, \max (| u_1 |, \cdots, | u_{N-1} |)] \tag{36}$$

permits us to employ the principle of optimality in very much the same was as before.

Introduce the sequence of functions

$$f_a(c_1 , c_2) = \min_P \text{ prob } [\max (| u_a |, | u_{a+1} |, \cdots, | u_{N-1} |) \geqq d], \tag{37}$$

for $a = 0, 1, 2, \cdots, N - 1$, and $-\infty < c_1 , c_2 < \infty$.

Then

$$\begin{aligned} f_{N-1}(c_1 , c_2) &= 1, \quad | c_1 | \geqq d, \\ &= 0, \quad | c_1 | < d, \end{aligned} \tag{38}$$

and

$$\begin{aligned} f_a(c_1 , c_2) &= 1, && | c_1 | \geqq a, \\ &= \min_{v_a} \int_{-\infty}^{\infty} f_{a+1}(c_1 + c_2\Delta, c_2 + h\Delta)\, dG(r), && | c_1 | < a, \end{aligned} \tag{39}$$

$a = 0, 1, 2, \cdots, N - 2$.

8. CORRELATION

Let us now indicate how processes where the r_i are not independent random variables may be treated. The simplest of these is that the distribution of r_i depends only upon the value of r_{i-1}.

In this case, it is clear that an essential part of the information pattern at each stage is the value of r at the preceding state. Let us define

$$dG(r_i ; r_{i-1}) = \text{the distribution function of } r_i \text{ given the value of } r_{i-1}, \tag{40}$$

and returning to the model of Section 5,

$$f_a(c_1 , c_2 ; r_{a-1}) = \text{the minimum expected deviation starting at time } a \text{ in the state } (c_1 , c_2) \text{ and the information that } r \text{ at } a - 1 \text{ was } r_{a-1}. \tag{41}$$

It is easy then to see that the recurrence relation now has the form

$$f_a(c_1, c_2; r_{a-1}) = \min_{v_a} \left[\Delta c_1^2 + \int_{-\infty}^{\infty} f_{a+1}(c_1 + \Delta c_2, c_2 + \Delta h)\, dG(r_a; r_{a-1}) \right]. \tag{42}$$

RECEIVED May 5, 1958.

REFERENCES

BELLMAN, R. (1957a). "Dynamic Programming." Princeton Univ. Press, Princeton, New Jersey.

BELLMAN, R. (1956a). On the application of dynamic programming to the variational problems in mathematical economics. *In* "Proceedings of the Symposium on Calculus of Variations and Applications." McGraw-Hill, New York, 1958.

BELLMAN, R. (1956b). On the application of dynamic programming to the theory of Control Processes. *In* "Proceedings of the Symposium on Control Processes," pp. 199–213. Polytechnic Institute of Brooklyn, Brooklyn, New York.

BELLMAN, R. (1957b). Functional Equations in the theory of dynamic programming, VI: A direct convergence proof. *Ann. Math.* **65**, 215–223.

BELLMAN, R. (1957c). Notes on control processes, I: On the Minimum of maximum deviation. *Quart. Appl. Math.* **14**, 419–423.

BELLMAN, R. (1957d). Dynamic programming, Nonlinear Variational processes, and successive approximations. The RAND Corporation, Paper No. P-1133.

BELLMAN, R. (1958). Some new techniques in the dynamic programming solution of variational problems. *Quart. Appl. Math.* in press.

BELLMAN, R., AND DREYFUS, S. (1957). *Approximations and Dynamic Programming*, The RAND Corporation, Paper No. P-1176.

BELLMAN, R., AND KALABA, R., (1958). On communication processes involving learning and random duration. *1958 IRE Natl. Convention Record, Inform. Theory*, in press.

BELLMAN, R., GLICKSBERG, I., AND GROSS, O. (1954). On some variational problems occurring in the theory of dynamic programming. *Rend. Circolo Mat. Palermo* [II] **3**, 1–35.

BELLMAN, R., GLICKSBERG, I., AND GROSS, O. (1956). On the "Bang-Bang" control problem. *Quart. Appl. Math.* **14**, 11–18.

BOOTON, R. C., JR. (1956a). Optimum design of final-value control systems. *In* "Symposium on Nonlinear Circuit Analysis" (J. Fox, ed.). Polytechnic Institue of Brooklyn, Brooklyn, New York.

BOOTON, R. C., JR. (1956b). Final-value systems with Gaussian inputs. *Trans. on Inform. Theory* **IT-2**, 173–176.

DOOB, J. L. (1944). The elementary Gaussian processes. *Ann. Math. Stat.* **15**, 229–282.

ROBBINS, H. (1952). Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **58**, 527–536.