J. Soc. Indust. Appl. Math.
Vol. 13, No. 1, March, 1965
Printed in U.S.A.

OPTIMAL POLICIES FOR A SEQUENTIAL DECISION PROCESS*

LAWRENCE BROWN†

1. Introduction. In this paper we discuss a sequential decision process of a special type. For this process the decision at the nth stage of the process concerns which value among a set of possible random variables to record. The loss (or gain) function of the process is a function of the values of the random variables actually recorded.

In this broad formulation the process is related to the two-armed bandit problem posed by Robbins [5] and discussed in [2], [3], [6]. The difference lies in the amount of information given the observer about the variables he chooses not to sample. The extra information conveyed in the process considered in this paper makes possible certain kinds of conclusions which are not possible in the case of the two-armed bandit.

The process will be described in detail in §2. Briefly, it is as follows. At each stage of the process, say the nth, the two real valued random variables X_1^n, X_2^n, are observed. These variables have a fixed joint distribution (independent of n) belonging to a parametric family of distributions subject to the assumptions of §2. At each stage, *before* observing the variables, the experimenter must choose to record one of them. The choice is made on the basis of his past observations and his a priori knowledge (if any) of the distribution of x_1, x_2. After proceeding in this way for N stages (N is given in advance) the experimenter receives a payoff which is some given function of the sum of the recorded variables; the objective of the experimenter is to maximize the expectation of this payoff.

Subject to the assumptions of the next section, it is shown in Theorem 1 that an optimal decision policy exists. One of these assumptions is an assumption of symmetry on the a priori parameter distribution. Due to the finiteness of the group of symmetries the policy of Theorem 1 can easily be shown to be an optimal invariant policy, subject to the other assumptions of §2. Under these restrictions it is also the optimal maximin policy.

The method of proof of the main theorem is suggested by the dynamic programming approach used by Bellman [1]. The sequential decision process is viewed as a problem in adaptive control. As is the general case with such problems, it is then fairly easy to describe the process in terms of a recursion equation. A feature of this problem is that it is possible to

* Received by the editors October 23, 1962, and in final revised form June 26, 1964.
† Mathematics Department, Cornell University, Ithaca, New York. The author began this research under the direction of R. Bellman at the RAND Corporation.

use a modification of this equation to find an optimal policy for the process without actually computing values from the recursion equation.

2. Formal description of the process. Let $x^1, x^2, \cdots, x^n$ be a sequence of independent random variables in 2-dimensional Euclidean space E_2. Each random variable $x^k = (x_1^k, x_2^k)$ is assumed to have a probability density $p(\theta, x)$ with respect to a σ-finite measure μ on E_2. θ is a parameter, $\theta \in \Theta \subseteq E_2$. If $x = (x_1, x_2)$ is a 2-dimensional vector let the *reversal* of x, denoted by Rx, be $Rx = (x_2, x_1)$. Make the following assumption.

ASSUMPTION 1.

$$\begin{aligned} &\mu(x) = \mu(Rx), \\ &\theta \in \Theta \quad \text{if and only if} \quad R\theta \in \Theta. \end{aligned} \tag{2.1}$$

For the purposes of Theorem 1 and its corollaries we will also make the following assumption on the family of distributions $\{p(\theta, x)\}$.

ASSUMPTION 2.

$$p(\theta, x) = C(\theta) \exp(\theta_1 x_1 + \theta_2 x_2), \tag{2.2}$$

where

$$C(\theta) = C(R\theta). \tag{2.3}$$

Note that $S = \sum_{i=1}^{n-1} x^i$ is a sufficient statistic for θ on the basis of the observations $x^1, x^2, \cdots, x^{n-1}$.

The condition (2.3) is precisely what is required for (and results from) $p(\theta, x) = p(R\theta, Rx)$ and we could have alternatively formulated the restriction in that way. The condition (2.2) is not so limiting as it might at first appear. By a suitable change of variables many distributions of the exponential family can be put in the form (2.2). Also, except for a few pathological examples, the assumption (2.2) is satisfied by any family $\{p(\theta, x)\}$ for which the sum vector s is a sufficient statistic for θ [4].

We shall assume an a priori probability distribution G on E_2 for the parameter θ. We define $q(x; x^1, \cdots, x^{n-1})$ to be the conditional probability density with respect to μ of x given $x^1, \cdots, x^{n-1}$. (Statements involving q here and in the remainder of the paper are valid only almost everywhere with respect to an appropriate measure.) In view of Assumption 2 we will generally write $q(x; x^1, \cdots, x^{n-1}) = q(x; s, n-1)$.

Let ξ^n be a coordinate vector in E_2; that is, $\xi^n = \epsilon^i$ for some i, $i = 1, 2$, where ϵ^i is the ith coordinate vector in E_2. ξ^n is called the *recording choice vector* at the nth stage of the process and the value of i is called the *recording choice* at the nth stage of the process.

Let R_n, the *sum of the recorded variables after n stages*, be defined recursively by

$$R_n = R_{n-1} + \xi^n \cdot x^n, \qquad R_0 = 0, \tag{2.4}$$

where "$\cdot$" denotes the usual dot product in E_2.

Let $w(x)$ be a real valued function called the *payoff function* associated with the process.

We list formally the following assumption on w.

ASSUMPTION 3. $w(x)$ is nondecreasing in x.

An *optimal policy* (given certain information) for a process of N stages is a policy of making recording choices which maximizes the (conditional) expected value of $w(R_N)$. The proof (using (2.5) and (2.6)) that an optimal policy exists presents no special difficulties, and will not be given here.

Let $f(r, s, n)$ be the expected value of $w(R_N)$ when using an optimal policy given that $R_n = r$ and the sum vector s has been observed for the first n stages. Of course, f also depends on w, N, p, and G but these quantities are all fixed throughout a given process and need not be contained explicitly in the notation.

Using the principle of optimality [1, p. 56] we may write the following recursion relation:

$$\begin{aligned} f(r, s, n) &= \max_{i=1,2} \{f_i(r, s, n)\}, \qquad n = 0, 1, \cdots, N-1, \\ f(r, s, N) &= w(r), \end{aligned} \tag{2.5}$$

where

$$f_i(r, s, n) = \int_{E_2} f(r + x_i, s + x, n + 1) q(x, s, n)\, d\mu(x). \tag{2.6}$$

Note that f_i is the conditional expected payoff from a policy which makes the recording choice i at the nth stage and uses an optimal policy thereafter.

The sampling choice made by any optimal policy for this process at the nth stage is any one of the values i which maximize the integral on the right of expression (2.6). We shall say that we have specified the *complete* optimal policy if we know (as a function of r, s, and n) all the values of i which maximize the integral on the right of (2.6).

3. Statement and proof of the optimal policy theorem. The main optimal policy theorem is the following.

THEOREM 1. *Suppose the family $p(\theta, x)$, the measure μ, and w satisfy Assumptions 1–3 of §2. Assume $G(\theta) = G(R\theta)$. Then $s_i > s_j$ implies $f_i(r, s, n) \geqq f_j(r, s, n)$, and $s_i = s_j$ implies $f_i(r, s, n) = f_j(r, s, n)$; $i, j = 1, 2$. Thus an optimal policy at the nth stage of the process is to make any recording choice i for which $s_i \geqq s_j$, $j \neq i$.*

Proof. Define

$$A(s, \theta) = \exp\{\sum_i \theta_i s_i\}, \tag{3.1}$$

$$L(s, n) = \int_{E_2} C^n(\theta) A(s, \theta)\, dG(\theta), \tag{3.2}$$

$$\begin{aligned} g_i(r, s, n) &= L(s, n) f_i(r, s, n), \\ g(r, s, n) &= \max_i \{g_i(r, s, n)\} = L(s, n) f(r, s, n). \end{aligned} \tag{3.3}$$

Then from (2.2) the conditional density q satisfies

$$L(s, n) q(x; s, n) = L(s + x, n + 1), \tag{3.4}$$

and from (2.5), (2.6), (3.3) and (3.4) we have

$$\begin{aligned} g(r, s, N) &= L(s, N) w(r), \\ g_i(r, s, n) &= L(s, n) \int f(r + x_i, s + x, n + 1) q(x, s, n)\, d\mu(x) \\ &= \int g_i(r + x_i, s + x, n + 1)\, d\mu(x). \end{aligned} \tag{3.5}$$

We are now in a position to prove the following induction hypothesis backward from $n = N - 1$: For $s = (s_1, s_2)$ with $s_1 > s_2$ and any $x = (x_1, x_2)$ and any r, we shall show that

$$\begin{aligned} D_n(r, s, x) = {} & g_1(r, s + x, n) + g_1(r, s + Rx, n) \\ & - g_2(r, s + x, n) - g_2(r, s + Rx, n) \geqq 0. \end{aligned} \tag{3.6}$$

Note that substituting $x = (0, 0)$ in (3.5) shows the policy of the theorem to be optimal at the nth stage.

We begin the induction by proving $D \geqq 0$ for $n = N - 1$. In that case,

$$\begin{aligned} D_n = {} & \int w(r + y_1)[L(s + x + y, N) + L(s + Rx + y, N)]\, d\mu(y) \\ & - \int w(r + y_2)[L(s + x + y, N) \\ & + L(s + Rx + y, N)]\, d\mu(y) \\ = {} & \int_{y_1 > y_2} w(r + y_1)[L(s + x + y, N) + L(s + Rx + y, N) \\ & - L(s + x + Ry, N) - L(s + Rx + Ry, N)]\, d\mu(y) \\ & + \int w(r + y_2)[L(s + x + Ry, N) + L(s + Rx + Ry, N) \end{aligned} \tag{3.7}$$

$$- L(s + x + y, N) - L(s + Rx + y, N)]\, d\mu(y)$$

$$= \int_{y_1 > y_2} [w(r + y_1) - w(r + y_2)]$$

$$\cdot [L(s + x + y, N) + L(s + Rx + y, N)$$

$$- L(s + x + Ry, N) - L(s + Rx + Ry, N)]\, d\mu(y).$$

Using (3.2) we write

$$L(s + x + y, N) + L(s + Rx + y, N)$$

$$- L(s + x + Ry, N) - L(s + Rx + Ry, N)$$

$$= \int_{\theta_1 > \theta_2} C^N(\theta)[A(s + x + y, \theta) + A(s + x + y, R\theta) \tag{3.8}$$

$$+ A(s + Rx + y, \theta) + A(s + Rx + y, R\theta)$$

$$- A(s + x + Ry, \theta) - A(s + x + Ry, R\theta)$$

$$- A(s + Rx + Ry, \theta) - A(s + Rx + Ry, R\theta)]\, dG(\theta).$$

The last step of the proof for $n = N - 1$ is to show that the integrand in (3.8) is positive. To begin this step we compute

$$\frac{d^2}{da^2}[A(s + (a, -a), \theta) + A(s + (a, -a), R\theta)] \tag{3.9}$$

$$= (\theta_1 - \theta_2)^2[A(s + (a, -a), \theta) + A(s + (a, -a), R\theta)] \geqq 0$$

By Assumption 2, $A(s, \theta) = A(Rs, R\theta)$. Hence $A(s, \theta) + A(s, R\theta)$ is symmetric in s. This, together with (3.9) shows that for $P = s_1 + s_2$ fixed, $A(s, \theta) + A(s, R\theta)$ is a convex function of $|s_1 - s_2|$. Examining the arguments of the functions in the integrand of (3.8), we see that this convexity implies the integrand is always positive. Returning from (3.8) to (3.7) and using the fact that w is nondecreasing, we see that for $n = N - 1$, $D_n \geqq 0$, which is the desired result, concluding the first half of the induction.

To begin the second half of the proof we assume that $D_{n+1} \geqq 0$ (for all appropriate r, s, and x). We will then prove that $D_n \geqq 0$. As above, assume $s_1 > s_2$. Using the fact that

$$\int g_1(r + x_2, s + x, n)\, d\mu(x) = \int g_2(r + y_1, s + y, n)\, d\mu(y), \tag{3.10}$$

we may write

$$g_1(r, s, n) - g_2(r, s, n)$$

$$= \int_{p_1} (g_1(r + x_1, s + x, n + 1) - g_1(r + x_2, s + x, n + 1))\, d\mu(x)$$

$$+\int_{p_2}(g_2(r+x_1,s+x,n+1)-g_2(r+x_2,s+x,n+1))\,d\mu(x) \tag{3.11}$$

$$=\int_{p_1}(g_1(r+x_1,s+x,n+1)-g_2(r+x_1,s+x,n+1))\,d\mu(x)$$

$$+\int_{p_2}(g_1(r+x_2,s+x,n+1)-g_2(r+x_2,s+x,n+1))\,d\mu(x),$$

where $p_1=\{x: s_1+x_1\geqq s_2+x_2\}$, $p_2=\{x: x\notin p_1\}$, and $p_2'=\{x: Rx\in p_2\}$. Hence,

$$\begin{aligned}D_n(r,s,x)&=g_1(r,s+x,n)-g_2(r,s+x,n)\\&\quad+g_1(r,s+Rx,n)-g_2(r,s+Rx,n)\\&=\int_{Q_1}(g_1(r+y_1,s+x+y,n+1)\\&\qquad-g_2(r+y_1,s+x+y,n+1))\,dy\\&\quad+\int_{Q_2'}(g_1(r+y_1,s+x+Ry,n+1)\\&\qquad-g_2(r+y_1,s+x+Ry,n+1))\,d\mu(y)\\&\quad+\int_{Q_3}(g_1(r+y_1,s+Rx+y,n+1)\\&\qquad-g_2(r+y_1,s+Rx+y,n+1))\,d\mu(y)\\&\quad+\int_{Q_4'}(g_1(r+y_1,s+Rx+Ry,n+1)\\&\qquad-g_2(r+y_1,s+Rx+Ry,n+1))\,d\mu(y),\end{aligned} \tag{3.12}$$

where $Q_1=\{y: s_1+x_1+y_1\geqq s_2+x_2+y_2\}$, $Q_2'=\{y: Ry\notin Q_1\}$, $Q_3=\{y: s_1+x_2+y_1\geqq s_2+x_1+y_2\}$, $Q_4'=\{y: Ry\notin Q_3\}=\{y: s_1+x_2+y_2\leqq s_2+x_1+y_1\}$. Note that $Q_4'\subseteq Q_1$ and also $Q_3\supseteq Q_2'$. Continuing from (3.12),

$$\begin{aligned}D_n&=\int_{Q_4'}D_{n+1}(r+y_1,s,x+y)\,d\mu(y)\\&\quad+\int_{Q_4'}D_{n+1}(r+y_1,s,x+Ry)\,d\mu(y)\\&\quad+\int_{Q_1-Q_4'}(g_1(r+y_1,s+x+y,n+1)\\&\qquad-g_2(r+y_1,s+x+y,n+1))\,d\mu(y)\end{aligned} \tag{3.13}$$

$$+ \int_{Q_3 - Q_2'} (g_1(r + y_1, s + Rx + y, n + 1) - g_2(r + y_1, s + Rx + y, n + 1))\, d\mu(y).$$

Note that the integrand of the third integral is precisely $\frac{1}{2}D_{n+1}(r + y_1, s + x + y, (0, 0))$ and so is nonnegative by the definition of Q_1 and the induction hypothesis. Similarly, the other integrands in (3.13) are nonnegative. Hence $D_n \geqq 0$. As remarked before, $D_n(r, s, (0, 0)) \geqq 0$ for all n implies the policy of the theorem is optimal at every stage. The proof of the theorem is thus completed.

4. Corollaries to the theorem. The first corollary gives a necessary condition for the policy described in Theorem 1 to be the complete optimal policy in the sense of the definition of §2.

COROLLARY 1. *Assume p and μ satisfy Assumptions* 1 *and* 2 *of* §2 *and there exists a point x with $x_1 \neq x_2$ in the range of μ. Assume $G(\theta)$ is symmetric and has a point of increase at some point θ with $\theta_1 \neq \theta_2$, and suppose $w(x)$ is strictly increasing in x. Finally, assume $E(w(R_N)) < \infty$ for the optimal policy of Theorem* 1. *Then the policy of Theorem* 1 *is the complete optimal policy.*

Proof. The proof of this corollary is a minor modification of the proof of the theorem.

The left-hand side of (3.9) is positive unless $\theta_1 = \theta_2$. Using the hypothesis of this corollary concerning G, the integral on the right of (3.8) is positive. Using the hypotheses concerning w and μ, the integral on the right of (3.7) is positive. Hence for $n = N - 1$, the sign "$\geqq$" in the induction hypothesis (3.6) may be replaced by the sign "$>$". According to the reasoning following (3.13), the sign "$>$" may be carried through the induction step. Thus in (3.13), $D_n \geqq \frac{1}{2}D_{n+1} > 0$. $D_n > 0$ for all n implies that the policy of the theorem is strictly better than any other policy. This completes the proof of the corollary.

The second corollary of this section deals with what might be called a first passage process. For this process the objective is to minimize an increasing function of the stage n for which R_n first becomes larger than some fixed value R. We call this value of n, n_R.

COROLLARY 2. *Assume the hypotheses of Theorem* 1 *are satisfied. Assume also that $p(\theta, x) = 0$ for all x with $x_1 < 0$ or $x_2 < 0$. Then the policy of Theorem* 1 *is optimal if it is desired to minimize $E(f(n_s))$, where $f(j)$ is nondecreasing in $j = 1, 2, \cdots$.*

Proof. Let

$$w(r) = \begin{cases} 0 & \text{if } r \leqq R, \\ 1 & \text{if } r > R. \end{cases} \tag{4.1}$$

Let $P(n)$ be the probability that $n_R = n$ when using the policy P of the theorem. Let $P^*(n)$ be this probability when using some other recording policy, P^*. We may assume $\sum_{n=1}^{\infty} P^*(n) = 1$, for otherwise $n_R = \infty$ with positive probability for that policy and $E(n_R \mid P^*) = \infty$, so that the conclusion of the corollary is trivial. The policy of choosing at random among the possible predictions satisfies $\sum_{n=1}^{\infty} P^*(n) = 1$ (unless $\Pr\{x = (0, 0)\} = 1$), hence (except for this parenthetical case), policies giving such values of P exist.

We wish to show

$$\sum_{n=1}^{\infty} f(n)P(n) \leqq \sum_{n=1}^{\infty} f(n)P^*(n). \tag{4.2}$$

From Theorem 1 when w is given by (4.1),

$$\sum_{n=1}^{N} P(n) \geqq \sum_{n=1}^{N} P^*(n) \quad \text{for any} \quad N < \infty. \tag{4.3}$$

Hence

$$\begin{aligned} \sum_{n=1}^{\infty} f(n)P)n) &= f(1) + \sum_{k=1}^{\infty} (f(k+1) - f(k)) \left(1 - \sum_{j=1}^{k} P(j)\right) \\ &\leqq f(1) + \sum_{k=1}^{\infty} (f(k+1) - f(k)) \left(1 - \sum_{j=1}^{k} P^*(j)\right) \qquad (4.4) \\ &= \sum_{n=1}^{\infty} f(n)P^*(n), \end{aligned}$$

which proves the corollary.

This corollary can be generalized somewhat using the same methods. We shall not do this here.

5. Generalizations to m dimensions. It is very natural to consider a generalization of the previous process for which x is an m-dimensional vector $(x_1, x_2, \cdots, x_m)$ rather than a two-dimensional vector. The generalization should be evident: at each stage of the process a recording choice i is made, the objective being to maximize some function of the sum of the recorded values after N stages. There is also a natural generalization of Assumptions 1 and 2. Wherever an R operator appears, replace this by R_{ij}, $1 \leqq i < j \leqq m$, where R_{ij} induces a reversal of the ith and jth coordinates of the vector it operates on. Also replace (2.2) by

$$p(\theta, x) = C(\theta) \exp\left[\sum_{j=1}^{m} \theta_j x_j\right]. \tag{5.1}$$

It seems reasonable to expect that the natural generalization of Theorem 1 is valid for this process, i.e., if $G(\theta)$ is a symmetric a priori distribution

for θ, then the optimal policy after the sum vector denoted by s has been observed is to make the sampling choice i for which $s_i \geqq s_j$, $j = 1, 2, \cdots, m$. I have not been able to prove this result in general. For certain special cases, simple generalizations of the methods used in the preceding are sufficient to prove the generalized form of Theorem 1. The notation used in the following should be obvious. The optimal policy outlined in this paragraph is optimal in the following two cases:

Case 1. *If* $w(x) = x$.

In this case we write

$$\begin{aligned} f_i(r, s, n) &= r + f_i(s, n) \\ &= \int (r + x_i + f(s + x, n + 1))q(x, s, n)\, d\mu(x) \\ &= r + C(s, n) + \int x_i q(x, s, n)\, d\mu(x), \end{aligned} \tag{5.2}$$

where $C(s, n)$ depends only on s and n as indicated. It follows from a straightforward consideration of a posteriori probabilities that $s_i > s_j$ implies

$$\int x_i q(x, s, n)\, d\mu(x) \geqq \int x_j q(x, s, n)\, d\mu(x), \tag{5.3}$$

which proves the generalization of Theorem 1 to be true for this special m-dimensional case. (Actually, a generalization of (5.2) shows that any process with a payoff function of the form $w(x^1, \cdots, x^N) = \sum_{i=1}^{N} \omega(x^i)$ is essentially a one-stage decision process instead of a sequential decision process. This fact was used in [6] to prove a special case of Theorem 1.)

Case 2. *If* $p(\theta, x) = 0$ *for* $x_i \neq 0$ *or* 1, $i = 1, 2, \cdots, m$ (*i.e., if each* x_i *is a* 0-1 *random variable*).

In this case $s_i > s_j$ implies $s_i + x_i \geqq s_j + x_j$, so that if $s_1 > s_2$,

$$\begin{aligned} &g_1(r, s, n) - g_2(r, s, n) \\ &\quad = \int (g(r + x_1, s + x, n + 1) - g(r + x_2, s + x, n + 1))\, d\mu(x) \\ &\quad = \iint (g(r + x_1 + y_1, s + x + y, n + 2) \\ &\qquad\qquad - g(r + x_2 + y_1, s + x + y, n + 2))\, d\mu(x)\, d\mu(y) \\ &\quad = \int (g_1(r + x_1, s + x, n + 1) - g_2(r + x_1, s + x, n + 1))\, d\mu(x). \end{aligned} \tag{5.4}$$

It can be proved as in Theorem 1 that $s_i > s_j$ implies $g_i(r, s, N - 1)$

$\geqq g_j(r, s, N-1)$. Then (5.4) establishes this generalization of Theorem 1 by induction.

REFERENCES

[1] R. BELLMAN, *Adaptive Control Processes: A Guided Tour*, Princeton University Press, Princeton, 1961.

[2] R. N. BRADT, S. M. JOHNSON, AND S. KARLIN, *On sequential designs for maximizing the sum of n observations*, Ann. Math. Statist., 27 (1956), pp. 1060–1074.

[3] D. FELDMAN, *Contributions to the two armed bandit problem*, Ibid., 33 (1962), pp. 847–856.

[4] B. O. KOOPMAN, *On distributions admitting a sufficient statistic*, Trans. Amer. Math. Soc., 39 (1936), pp. 399–409.

[5] H. ROBBINS, *Some aspects of the sequential design of experiments*, Bull. Amer. Math. Soc., 58 (1952), pp. 527–535.

[6] W. VOGEL, *An asymptotic minimax theorem for the two armed bandit problem*, Ann. Math. Statist., 31 (1960), pp. 444–451.