The Annals of Statistics
1983, Vol. 11, No. 2, 707–711

A NOTE ON DISCOUNTED FUTURE TWO-ARMED BANDITS

BY RICHARD KAKIGI

California State University at Hayward

This paper is concerned with the problem of finding Bayes sequential designs for successively choosing between two given Bernoulli variables so as to maximize the total discounted expected sum. Simple hypotheses concerning the success probabilities are assumed and dynamic programming methods are used to characterize optimal designs. Explicit solutions are described for certain special cases.

1. Introduction. An important prototype for sequential design problems is the two-armed bandit: Let X and Y be two Bernoulli variables with success probabilities $x = P[X = 1] = 1 - P[X = 0]$ and $y = P[Y = 1] = 1 - P[Y = 0]$. Periodically, say once a day, you must choose either X or Y and then are paid the outcome of the chosen variable. X and Y can be viewed as the left and right arms of a two-armed slot machine and the problem of how to choose between X and Y so as to maximize our total expected payoff is known as the two-armed bandit problem. It is assumed that the daily outcomes are independent conditional on the unknown probabilities x and y, so that the choices of X and Y are exchangeable rather than independent.

The two-armed bandit problem has received considerable attention in the literature. Robbins (1952) considers the problem in the absence of prior information and finds designs which maximize, with probability one, the relative frequency of positive payoffs when x and y are unknown. In most of the literature concerning the two-armed bandit a prior distribution for the values of x and y is assumed. Bradt, Johnson and Karlin (1956) consider the problem of maximizing the total payoff for a fixed number of days of play, and characterize the optimal design for the case (known as the one-armed bandit) when the value of y is known and x has a prior density. Feldman (1962) assumes a prior distribution for x and y which concentrates mass on two pairs of probabilities: $x = x_0, y = y_0$ versus $x = y_0, y = x_0$, where $x_0 > y_0$ are known probabilities. Feldman shows that the design which chooses X whenever $P[x = x_0, y = y_0] > \frac{1}{2}$ is optimal. Fabius and van Zwet (1970) characterize optimal designs for general prior distributions for x and y, derive expressions for the resulting Bayes risk, and show there is a symmetric, minimax-risk, admissible design. Berry (1972) assumes that a prior distribution under which x and y are independent, and derives sufficient conditions for choosing X. Also the "stay on a winner" rule is shown to be a property of an optimal design. Berry and Fristedt (1979) have considered arbitrary rather than geometric discounting and assume that one probability of positive payoff is known with probability 1. Kelly (1974) finds sufficient conditions for the optimality of the myopic design: choose between X and Y as if only one play were remaining. Gittins (1979) discusses discounted future Markov decision processes with important results which apply to multiarmed bandit problems. A distinction which should be noted is that in Gittins (1979) the distributions for success probabilties are assumed to be independent, whereas the distribution for x and y are dependent.

In this paper we consider an unlimited number of days of play and introduce a discount factor, β, $0 < \beta < 1$, so that a payoff of 1 on the nth day is worth only β^{n-1}. This approach may be appropriate when the future is indefinite and the number of days of play cannot be fixed. Another interpretation of discounting is the following: the number of days of play is a random variable with a geometric distribution where $(1 - \beta)$ is the probability of

Received January 1982; revised November 1982.
AMS 1980 *subject classifications.* 62L05, 90C50, 62F15.
Key words and phrases. Bayes sequential design, discounted dynamic programming, two-armed bandit.

stopping on any given day of play, and the total payoff is proportional to the sum of the payoffs up until stopping. We investigate Bayes designs for prior distributions which concentrate on two simple hypotheses for x and y, namely H_0: $x = x_0 > y = y_0$ and H_1: $x = x_1 < y = y_1$, where x_0, y_0, x_1, y_1 are known probabilities.

In this setting a prior distribution is specified by $s = P[H_0 \text{ is true}] = 1 - P[H_1 \text{ is true}]$, and given s, the aim is to find a design which maximizes the total discounted expected payoff. Let $D(s^*)$ denote the design which selects Y when and only when the current posterior probability that H_0 is true falls below s^*. $D(s^*)$ is shown to be optimal for some s^* in Section 3 and the s^* is calculated for certain special cases in Section 4.

2. Preliminaries. The discounted future two-armed bandit problem is a discounted dynamic programming problem as described in Blackwell (1965). The set of actions, $A = \{X, Y\}$, consist of the available daily choices; the set of states, $S = \{s : 0 \le s \le 1\}$, consist of the possible (posterior) probabilities that H_0 is true; and the law of motion, q, is the conditional distribution, given by Bayes rule, of the new posterior probability of H_0 given the action selected and the current posterior probability. Given a reward function r, the objective is to find designs (plans) with returns (incomes) which are optimal. For functions f on S, let $T^X f (T^Y f)$ denote the expected payoff from choosing $X(Y)$ and receiving the sum of the outcome for the first day and βf as a function of the new posterior probability. That is, for $s \in S$

$$(T^X f)(s) = (sx_0 + \bar{s}x_1) + \textstyle\sum_{s'} \beta f(s') q(s' \mid X, s);$$

and

$$(T^Y f)(s) = (sy_0 + \bar{s}y_1) + \textstyle\sum_{s'} \beta f(s') q(s' \mid Y, s).$$

Finally, let $Uf = \max\{T^X f, T^Y f\}$. Some results of Blackwell (1965) applied to the present problem yield the following theorem.

THEOREM 1. (a) *If the function f on S satisfies $f \ge T^a f$ for all actions a, then f is an upper bound on returns.* (*Blackwell*, 1965, *Theorem* 6(d)). (b) *U is a contraction with modulus β and has a unique fixed point: $v = Uv$.* (*Blackwell*, 1965, *Theorem* 5). (c) *A design is optimal if and only if its return is the fixed point of U.* (*Blackwell*, 1965, *Theorem* 6(f)). (d) *The return of the design $D(s^*)$ is the unique solution to the equation*:

$$g = \begin{cases} T^X g & \text{if} \quad s \ge s^*, \\ T^Y g & \text{if} \quad s < s^*, \end{cases}$$

(*Blackwell*, 1965, *Theorem* 3(c)).

The following notation will be used. Let v denote the fixed point of U, and for any real number t, let $\bar{t} = 1 - t$. For a real-valued function f defined on S, let $E^a[f]$ be the expected value of f with respect to the distribution $q(\cdot \mid a, s)$ where $a \in A$. Note the dependence of $E^a[f]$ on the initial state s. Let r^X denote the expected value of $r(s, X, .)$ with respect to $q(\cdot \mid X, s)$, where

$$r(s, X, s') = \begin{cases} 1 & \text{if} \quad s' = sx_0/(sx_0 + \bar{s}x_1) \quad \text{and} \quad 0 < s < 1, \\ x_0 & \text{if} \quad s' = s = 1, \\ x_1 & \text{if} \quad s' = s = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Define r^Y similarly. Note that with this notation

$$r^X(s) = sx_0 + \bar{s}x_1, \quad T^X f = r^X + \beta E^X[f].$$

Finally, note that $D(s^*)$ is, in the language of Blackwell (1965), the stationary plan $d^{(\infty)}$, where

$$d(s) = \begin{cases} X & \text{if} \quad s \ge s^* \\ Y & \text{if} \quad s < s^*. \end{cases}$$

Letting $\Pi = (d, d, d, \cdots)$, Theorem 1(d) is a restatement of Theorem 3(c), Blackwell (1965). Moreover, uniqueness follows since the operator T associated with d is a contraction with modulus β.

3. General form of optimal design. Intuitively, if an optimal design chooses X at the posterior probability s, then X should be chosen at any posterior $s' > s$. This property holds for the designs $D(s)$ and the following theorem states that such a design is optimal.

THEOREM 2. *There exists an $s^* \in S$ such that $D(s^*)$ is an optimal design. Moreover, s^* is any solution of the equation $T^X v = T^Y v$.*

For the proof of Theorem 2 we will need:

LEMMA 1. *For any real-valued function f on S, if $T^X f - T^Y f$ is a non-decreasing function of s then so is $T^X(Uf) - T^Y(Uf)$.*

PROOF. Since

$$T^X(Uf) - T^Y(Uf) = [T^X(Uf) - T^X(T^Y f)] + [T^Y(T^X f) - T^Y(Uf)] + [T^X(T^Y f) - T^Y(T^X f)] \tag{3.1}$$

it is enough to show that each of the summands in (3.1) is non-decreasing.

Since $T^X f - T^Y f$ is non-decreasing, there is an s' such that

$$(Uf)(s) = \begin{cases} (T^X f)(s) & \text{if } s \geq s', \\ (T^Y f)(s) & \text{if } s < s'. \end{cases}$$

Define g by

$$g(s) = \begin{cases} (T^X f)(s) - (T^Y f)(s) & \text{if } s \geq s', \\ 0 & \text{if } s < s'. \end{cases}$$

Then the first summand in (3.1) can be expressed as $E^X[g]$. To see that $E^X[g]$ is non-decreasing, it is enough to notice, g being non-decreasing, that a variable with distribution $q(. \mid X, s)$ is stochastically larger than a variable with distribution $q(. \mid X, s'')$ whenever $s > s''$. A similar argument shows that the second summand in (3.1) is non-decreasing. Given the initial state s, the distribution of the new state after observing both X and Y is independent of the order in which they are observed by exchangeability. Therefore, $E^X(E^Y[f]) = E^Y(E^X[f])$ and the third summand in (3.1) depends only on the first two plays. Easy calculation shows that the third summand equals $(r^X - r^Y)\bar{\beta}$ which is monotone increasing since $x_0 > y_0$ and $y_1 > x_0$. □

PROOF OF THEOREM 2. Let $v_0 = 0$ and for $n = 0, 1, 2, \cdots$, define $v_{n+1} = Uv_n$. Since $(T^X v_0) - (T^Y v_0) = r^X - r^Y$ is non-decreasing, Lemma 1 implies that for $n = 0, 1, 2, \cdots$, $T^X v_n - T^Y v_n$ is non-decreasing. Since v, the fixed point of U, is the limit of the v_n's with respect to supremum norm, it follows that $T^X v - T^Y v$ is non-decreasing. If s' is any solution of $T^X v = T^Y v$, then

$$v(s) = (Uv)(s) = \begin{cases} (T^X v)(s) & \text{if } s \geq s^*, \\ (T^Y v)(s) & \text{if } s < s^*, \end{cases}$$

which by Theorem 1(d) is the return of $D(s^*)$. The optimality of $D(s^*)$ follows from Theorem 1(c). □

4. Applications. This section gives conditions on the break-even value s^* for some special cases. When $x_0 = y_1 > x_1 = y_0$, it is immediate from the symmetry of the roles of X and Y that the optimality of $D(s^*)$ implies the optimality of $D(\bar{s}^*)$. The following theorem is the result of Feldman (1962) for the discounted future setting.

THEOREM 3. *If $x_0 = y_1 > x_1 = y_0$ then the design $D(\frac{1}{2})$ is optimal.*

Fix $s' \in S$ and let v' be the return of the design $D(s')$. If $D(s')$ is optimal then

$$T^X v'(s') = T^Y v'(s'). \tag{4.1}$$

If there is only one such s' which satisfies (4.1) then by theorem 2 $D(s')$ must be an optimal design. If $x_1 = y_0 = 0$ then there is a unique s' satisfying (4.1).

THEOREM 4. *If $x_1 = y_0 = 0$ then the design $D(s^*)$ is optimal with*

$$s^* = \frac{y_1(1 - \bar{y}_1\beta)}{y_1(1 - \bar{y}_1\beta) + x_0(1 - \bar{x}_0\beta)}.$$

PROOF. Fix s' and let v' denote the return of $D(s')$. Noting that whenever a payoff of 1 is received the true hypothesis is known with certainty. It follows that

$$\begin{aligned} T^X v'(s') = {} & s'x_0\{1 + x_0\beta)/\bar{\beta}\} + \bar{s}'y_1\{\beta + (y_1\beta^2)/\bar{\beta}\} \\ & + (s'\bar{x}_0 + \bar{s}'\bar{y}_1)[\beta^2 v'\{(s'\bar{x}_0)/(s'\bar{x}_0 + \bar{s}'\bar{y}_1)\}], \end{aligned}$$

and

$$\begin{aligned} T^Y v'(s') = {} & \bar{s}'y_1\{1 + (y_1\beta)/\bar{\beta}\} + s'x_0\{\beta + (x_0\beta^2)/\bar{\beta}\} \\ & + (s'\bar{x}_0 + \bar{s}'\bar{y}_1)[\beta^2 v'\{(s'\bar{x}_0)/(s'\bar{x}_0 + \bar{s}'\bar{y}_1)\}]. \end{aligned}$$

Thus, (4.1) holds if and only if $s' = s^*$. □

If $x_0 = y_1 = 1$ then whenever the payoff is 0 the true hypothesis is known with probability one.

THEOREM 5. *If $x_0 = y_1 = 1$ then the design $D(s^*)$ is optimal with*

$$s^* = \frac{\bar{x}_1(1 - y_0\beta)}{\bar{x}_1(1 - y_0\beta) + \bar{y}_0(1 - x_1\beta)}.$$

The proof is similar to that for Theorem 4 and is omitted.

For the remainder of this paper the cases $y_1 = \bar{y}_0 = 1$ and $y_1 = y_0$ will be considered. The latter case is discussed in Berry and Fristedt (1979). In these cases v is partially linear.

LEMMA 2. *Assume $y_1 = \bar{y}_0 = 1$ and define the linear function f on S by $f(1) = \beta x_0/\bar{\beta}$ and $f(0) = 1/\bar{\beta}$. Then $v = f$ on the interval $[0, s^*]$ where $D(s^*)$ is an optimal design.*

PROOF. It is enough to note that selecting Y determines the true hypothesis with probability one.

THEOREM 6. *Assume $y_1 = \bar{y}_0 = 1$. Then the design $D(s^*)$ is optimal for some s^* such that $s_L \le s^* \le s_R$ where $s_L = \bar{x}_1/(\bar{x}_1 + x_0)$ and $s_R = \bar{x}_1/(\bar{x}_1 + \bar{\beta}x_0)$.*

PROOF. Let u be the return of the design which always selects X. Define the linear function g on S by $g(1) = x_0/\bar{\beta}$ and $g(0) = x_1 + \beta/\bar{\beta}$. Note that g is the return from initially selecting X and then being told the true hypothesis. Let s_L and s_R be the solutions to the equations $f = g$ and $f = u$, respectively, where f is given in Lemma 2. Let $D(s^*)$ be an optimal design. In the interval $(s_R, 1]$, $f < u$ and so Y cannot be initially optimal. Therefore, $s^* \le s_R$. On the other hand, in the interval $[0, s_L)$, $f > g$ and so X cannot be initially optimal. Therefore, $s^* \ge s_L$. □

LEMMA 3. *Assume $y_0 = y_1 = y$ and define the linear function f on S by $f(1) = f(0) = y/\bar{\beta}$. Then $v = f$ on the interval $[0, s^*]$ where $D(s^*)$ is an optimal design.*

PROOF. If $y_0 = y_1 = y$ then selecting Y does not change the prior probability that H_0 is true. Therefore, on $[0, s^*]$ $v = T^Y v = f$. □

THEOREM 7. *Assume $y_0 = y_1 = y$. Then the design $D(s^*)$ is optimal for some s^* such that $s_L \leq s^* \leq s_R$ where $s_L = (y - x_1)/[(x_0 - x_1) + (\beta/\bar{\beta})(x_0 - y)]$ and $s_R = (y - x_1)/(x_0 - x_1)$.*

The proof is similar to that of Theorem 6 utilizing Lemma 3 instead of Lemma 2, and is omitted.

REFERENCES

BERRY, D. (1972). A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43** 871–897.

BERRY, D. A. and FRISTEDT, B. (1979). Bernoulli one-armed bandits-arbitrary discount sequences. *Ann. Math. Statist.* **7** 1086–1105.

BLACKWELL, DAVID (1965). Discounted dynamic programming. *Ann. Math. Statist.* **36** 226–235.

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

FABIUS J. and VAN ZWET, W. R. (1970). Some remarks on the two-armed bandit. *Ann. Math. Statist.* **41** 1906–1916.

FELDMAN, D. (1962). Contributions to the "two-armed bandit" problem. *Ann. Math. Statist.* **33** 847–856.

FELLER, W. (1966). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

GITTINS, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. B.* **41** 148–177.

ROBBINS, H. (1952). Some aspects of the sequential designs of experiments. *Bull. Amer. Math. Soc.* **58** 527–535.

DEPARTMENT OF STATISTICS
CALIFORNIA STATE UNIVERSITY
HAYWARD, CALIFORNIA 94542