The Annals of Statistics
1974, Vol. 2, No. 5, 1056–1062

A NOTE ON THE BERNOULLI TWO-ARMED BANDIT PROBLEM[1]

BY THOMAS A. KELLEY

University of Florida

Suppose the arms of a two-armed bandit generate i.i.d. Bernoulli random variables with success probabilities ρ and λ respectively. It is desired to maximize the expected sum of N trials where N is fixed. If the prior distribution of (ρ, λ) is concentrated at two points (a, b) and (c, d) in the unit square, a characterization of the optimal policy is given. In terms of a, b, c, and d, necessary and sufficient conditions are given for the optimality of the myopic policy.

1. Introduction. Suppose there are two experiments, $\mathscr{E}_1$ and $\mathscr{E}_2$, which generate independent Bernoulli random variables with expectations ρ and λ, respectively. It is desired to maximize the expected sum of N observations where N is fixed. The choice of which experiment to use may be sequential, i.e., it may be allowed to depend on previous results and prior information concerning the vector (ρ, λ). In this paper it is assumed that (ρ, λ) either is (a, b) or (c, d); for this type of prior information the optimal solution is characterized. Since the TAB has been viewed as a simplified model for the clinical trials problem of testing two drugs with unknown cure probabilities, it is of interest to know under what conditions the myopic procedure is optimal. Feldman [4] has shown that $(c, d) = (b, a)$ is a sufficient condition for this. Aside from a few obvious exceptions, it is also necessary.

2. Bernoulli two-armed bandit (TAB) model. Suppose experiment $\mathscr{E}_1$ generates i.i.d. Bernoulli random variables (generically denoted by X) with mean ρ, and suppose experiment $\mathscr{E}_2$ generates i.i.d. Bernoulli random variables (generically denoted by Y) with mean λ. Furthermore, suppose that every X is independent of every Y. If ξ denotes a prior distribution for the vector (ρ, λ) then let $\xi(X)$ denote the posterior distribution after an observation on X, $\xi(Y)$ the posterior distribution after an observation on Y, $\xi(X, Y)$ the posterior distribution after observations on X and Y in that order, and $\xi(Y, X)$ the posterior distribution after observations on Y and X in that order. Because of the independence of X and Y, $\xi(X, Y)$ and $\xi(Y, X)$ are identical.

There are a total of N trials to be sequentially allocated to $\mathscr{E}_1$ and $\mathscr{E}_2$, and the objective is to maximize the expected sum of the observations. For $n = 0, 1, \cdots, N$ let $V_n(\xi)$ denote the optimal expected gain for the remaining n trials when ξ

Received August 1973; revised October 1973.

[1] Supported by the National Science Foundation, Grant GP-33908X1.

AMS 1970 *subject classifications.* Primary 62.45; Secondary 62.35.

Key words and phrases. Bernoulli random variable, myopic, optimal, posterior distribution, relative advantage, sequential, strategy, two-armed bandit problem, two-point prior distribution.

is the current prior distribution for (ρ, λ). These functions are defined by the following recursive formulas (cf. [2], pages 395–396).

$$V_0(\xi) \equiv 0 , \tag{2.1}$$

and

$$V_n(\xi) = \max \{E[X + V_{n-1}(\xi(X))], E[Y + V_{n-1}(\xi(Y))]\} \quad \text{for } n = 1, \cdots, N . \tag{2.2}$$

It follows from (2.1) and (2.2) that there exist functions $F_n(\xi)$ and $G_n(\xi)$ such that $V_n(\xi) = \max \{F_n(\xi), G_n(\xi)\}$ for $n = 1, \cdots, N$.

These functions may be defined recursively. Let

$$F_0(\xi) \equiv 0 \quad \text{and} \quad G_0(\xi) \equiv 0. \tag{2.3}$$

Then for $n = 1, \cdots, N$,

$$F_n(\xi) = E[X + \max \{F_{n-1}(\xi(X)), G_{n-1}(\xi(X))\}] , \tag{2.4}$$

and

$$G_n(\xi) = E[Y + \max \{F_{n-1}(\xi(Y)), G_{n-1}(\xi(Y))\}] . \tag{2.5}$$

Let $D_n(\xi) = F_n(\xi) - G_n(\xi)$, the relative advantage of $\mathscr{E}_1$ over $\mathscr{E}_2$. Recursive formulas may be developed for defining $D_n(\xi)$. In fact,

$$D_1(\xi) = E(\rho) - E(\lambda) , \tag{2.6}$$

and for $n = 2, \cdots, N$,

$$D_n(\xi) = E(\rho) + E[V_{n-1}(\xi(X))] - E(\lambda) - E[V_{n-1}(\xi(Y))] , \tag{2.7}$$

$$= D_1(\xi) + E[G_{n-1}(\xi(X))] + E[D_{n-1}(\xi(X))^+] - E[F_{n-1}(\xi(Y))] + E[D_{n-1}(\xi(Y))^-] , \tag{2.8}$$

where x^+ denotes $\max \{x, 0\}$ and x^- denotes $\min \{x, 0\}$. It follows that

$$\begin{aligned} D_n(\xi) = D_1(\xi) &+ E[E(\lambda \mid X) + E(V_{n-2}(\xi(X, Y)) \mid X)] \\ &+ E[D_{n-1}(\xi(X))^+] - E[E(\rho \mid Y) + E(V_{n-2}(\xi(Y, X)) \mid Y)] \\ &+ E[D_{n-1}(\xi(Y))^-] . \end{aligned} \tag{2.9}$$

Since $\xi(X, Y)$ and $\xi(Y, X)$ are identical, it immediately follows that for $n = 2, \cdots, N$

$$D_n(\xi) = E[D_{n-1}(\xi(X))^+] + E[D_{n-1}(\xi(Y))^-] . \tag{2.10}$$

These formulas have, using different notation, been developed by Berry [1], Fabius and van Zwet [3], Quisel [5], and Zacks [6]. In terms of the functions $D_n(\xi)$ the optimal strategy may be described as follows: whenever there are n trials remaining and ξ is the current prior distribution for (ρ, λ), then the next trial should be allocated to $\mathscr{E}_1$ provided $D_n(\xi) > 0$ and to $\mathscr{E}_2$ provided $D_n(\xi) < 0$. Whenever $D_n(\xi) = 0$ the choice between $\mathscr{E}_1$ and $\mathscr{E}_2$ is arbitrary.

In this paper, only those prior distributions ξ which are concentrated at two points in the unit square will be considered. That is, it will be assumed that

$(\rho, \lambda) = (a, b)$ or $(\rho, \lambda) = (c, d)$ where $0 \leqq a, b, c, d \leqq 1$. This means that all distributions under consideration can be indexed by the real numbers in the closed unit interval; these numbers which will be denoted by ξ refer to the probability that $(\rho, \lambda) = (c, d)$. Of course, this means that each $D_n(\xi)$ is a function of a real variable. If $\bar{x}$ denotes the function $1 - x$, the following formulas for $\xi(X)$ and $\xi(Y)$ may be obtained.

$$\xi(X) = \xi c/(\xi c + \bar{\xi} a) \quad \text{if} \quad X = 1\,, \tag{2.11}$$

$$\xi(X) = \xi \bar{c}/(\xi \bar{c} + \bar{\xi} \bar{a}) \quad \text{if} \quad X = 0\,, \tag{2.12}$$

$$\xi(Y) = \xi d/(\xi d + \bar{\xi} b) \quad \text{if} \quad Y = 1\,, \tag{2.13}$$

$$\xi(Y) = \xi \bar{d}/(\xi \bar{d} + \bar{\xi} \bar{b}) \quad \text{if} \quad Y = 0\,. \tag{2.14}$$

Using these preliminary formulas and the recursive formulas for defining the functions $D_n(\xi)$, the optimal strategies will now be examined.

3. Characterization of the optimal strategies. Since $D_1(\xi) = (c - d)\xi + (a - b)\bar{\xi}$, it follows that $D_1(\xi) \geqq 0$ if $a \geqq b$ and $c \geqq d$. Then from (2.10) it follows that $D_n(\xi) \geqq 0$ for all ξ and $n = 1, \cdots, N$. Thus the optimal strategy will be to always use experiment $\mathscr{E}_1$. Likewise, if $a \leqq b$ and $c \leqq d$, then $D_n(\xi) \leqq 0$ for all $\xi \in [0, 1]$ and for $n = 1, 2, \cdots, N$. Thus the optimal strategy will be to always use experiment $\mathscr{E}_2$. So it will now be assumed that $b > a$ and $c > d$. It is now appropriate to note that for each X, $\xi(X)$ is a strictly increasing function of ξ and that for each Y, $\xi(Y)$ is a strictly increasing function of ξ. The structure of the optimal policy will be evident if it can be shown that for each $n (n = 1, \cdots, N)$, $D_n(\xi)$ is a strictly increasing function of ξ with a root in the open unit interval.

THEOREM 3.1. *For each $n = 1, 2, \cdots, N$ the following are true*:

(i) *$D_n(\xi)$ is a strictly increasing function of ξ.*
(ii) *$D_n(\xi)$ is a continuous function of ξ.*
(iii) *$D_n(0) < 0$ and $D_n(1) > 0$.*
(iv) *There exists a unique $\alpha_n \in (0, 1)$ such that $D_n(\alpha_n) = 0$.*

PROOF. Conditions (i), (ii), and (iii) are proven by induction. If they are true, condition (iv) follows from the intermediate value theorem. Since $D_1(\xi) = (a - b) + (b - a + c - d)\xi$, it follows that the theorem is true for $n = 1$. Now suppose that the theorem is true for some n between 1 and $N - 1$. Then $D_n(\xi(X))$ and $D_n(\xi(Y))$ must be strictly increasing functions of ξ. Since x^+ and x^- are non-decreasing functions of x it is certainly true that $D_{n+1}(\xi)$ is a non-decreasing function of ξ. Let

$$U_n = \sup\{\xi : E[D_n(\xi(X))^+] = 0\} \quad \text{and}$$
$$L_n = \inf\{\xi : E[D_n(\xi(Y))^-] = 0\}\,.$$

Then $E[D_n(\xi(X))^+]$ is zero on $[0, U_n]$ and strictly increasing on $(U_n, 1]$ while

$E[D_n(\xi(Y))^{-1}]$ is strictly increasing on $[0, L_n)$ and zero on $[L_n, 1]$. The only way $D_{n+1}(\xi)$ will not be strictly increasing is to have $U_n > L_n$. However,

$$U_n = \min\{a\alpha_n/(a\alpha_n + c\bar{\alpha}_n), \bar{a}\alpha_n/(\bar{a}\alpha_n + \bar{c}\bar{\alpha}_n)\}, \tag{3.1}$$

and

$$L_n = \max\{b\alpha_n/(b\alpha_n + d\bar{\alpha}_n), \bar{b}\alpha_n/(\bar{b}\alpha_n + \bar{d}\bar{\alpha}_n)\}. \tag{3.2}$$

Thus $U_n > L_n$ implies that

$$\max\{c/a, \bar{c}/\bar{a}\} < \min\{d/b, \bar{d}/\bar{b}\}. \tag{3.3}$$

And this in turn implies that $c < a$ and $b < d$. But this is impossible since $b > a$ and $c > d$. Thus $D_{n+1}(\xi)$ is a strictly increasing function of ξ. The induction hypothesis, together with the monotonicity of $D_{n+1}(\xi)$, the continuity of x^+ and x^-, and the monotone convergence theorem, guarantees the continuity of $D_{n+1}(\xi)$. Since $D_{n+1}(0) = D_n(0)$ and $D_{n+1}(1) = D_n(1)$ condition (iii) is also proven. □

This means that the optimal strategy is determined by a unique sequence of constants $\alpha_1, \alpha_2, \cdots, \alpha_N$. If there are n trials remaining to be allocated, it is optimal to allocate the next trial to experiment $\mathscr{E}_1$ provided $\xi \geqq \alpha_n$, and to experiment $\mathscr{E}_2$ otherwise.

4. Conditions for optimality of the myopic strategy. An appealing strategy to use is to allocate the next trial to experiment $\mathscr{E}_1$ whenever $E(\rho) \geqq E(\lambda)$, and to allocate the next trial to experiment $\mathscr{E}_2$ otherwise. This strategy is called the myopic strategy. It is called myopic because it allocates the next trial to $\mathscr{E}_1$ whenever $D_1(\xi) \geqq 0$; it "behaves" as if there were always just one more trial to be allocated. When the myopic strategy is optimal, it means that the optimal strategy does not depend on the number of trials remaining: it is time invariant, so to speak. It is of interest to know when the myopic strategy is optimal.

From the results of Section 3, it follows that whenever $a \leqq b$ and $c \leqq d$ the myopic strategy is optimal. Clearly, this is also true whenever $a \geqq b$ and $c \geqq d$. So now the case when $b > a$ and $c > d$ shall be considered. Let $\alpha = (b - a)/(b - a + c - d)$. Then $D_1(\xi) = (b - a)(\xi - \alpha)/\alpha$ and $D_1(\xi) \geqq 0$ iff $\xi \geqq \alpha$. From Theorem 3.1 it follows that the myopic strategy will be optimal if and only if $\alpha_1 = \alpha_2 = \cdots = \alpha_N = \alpha$ where $\alpha_1, \cdots, \alpha_N$ are those unique constants determining the optimal strategy. Obviously $\alpha_1 = \alpha$; the next theorem gives conditions under which $\alpha_2 = \alpha$.

THEOREM 4.1. $D_2(\alpha) = 0$ iff $a + b = c + d$.

PROOF. We first note that

$$(\xi c + \bar{\xi} a) D_1\left(\frac{\xi c}{\xi c + \bar{\xi} a}\right) = (b - a)(c\xi\bar{\alpha} - a\bar{\xi}\alpha)/\alpha. \tag{4.1}$$

Replacing a and c by $\bar{a}$ and $\bar{c}$, by b and d, and by $\bar{b}$ and $\bar{d}$, respectively, and using the fact that $x^+ = (x + |x|)/2$ and $x^- = (x - |x|)/2$ one obtains

$$D_2(\xi) = D_1(\xi) + \tfrac{1}{2}(b - a)(|c\xi\bar{\alpha} - a\bar{\xi}\alpha| + |\bar{c}\xi\bar{\alpha} - \bar{a}\bar{\xi}\alpha| - |d\xi\bar{\alpha} - b\bar{\xi}\alpha| - |\bar{d}\xi\bar{\alpha} - \bar{b}\bar{\xi}\alpha|)/\alpha . \tag{4.2}$$

Replacing ξ by α one immediately obtains

$$D_2(\alpha) = (b - a)\bar{\alpha}(|a - c| - |b - d|) . \tag{4.3}$$

From this, together with the assumption that (a, b) and (c, d) are on opposite sides of the main diagonal in the unit square, the conclusion follows. □

This theorem provides a necessary condition for optimality of the myopic procedure. It will now be assumed that $a + b = c + d = \gamma$. The function $D_2(\xi)$ will now be specified so that the conditions under which $D_3(\alpha) = 0$ may be examined. In order to specify $D_2(\xi)$ completely, the following four critical points must be kept in mind:

(4.4) $\xi_1 = a\alpha/(a\alpha + c\bar{\alpha})$, the point where $\xi(X) = \alpha$ for $X = 1$,

(4.5) $\xi_2 = \bar{a}\alpha/(\bar{a}\alpha + \bar{c}\bar{\alpha})$, the point where $\xi(X) = \alpha$ for $X = 0$,

(4.6) $\xi_3 = b\alpha/(b\alpha + d\bar{\alpha})$, the point where $\xi(Y) = \alpha$ for $Y = 1$,

(4.7) $\xi_4 = \bar{b}\alpha/(\bar{b}\alpha + \bar{d}\bar{\alpha})$, the point where $\xi(Y) = \alpha$ for $Y = 0$.

The following facts about these four numbers may be easily verified.

(4.8) (1) $\max\{\xi_1, \xi_4\} < \alpha < \min\{\xi_2, \xi_3\}$

(4.9) (2) $\xi_1 < \xi_4$ iff $\gamma < 1$,

(4.10) (3) $\xi_2 < \xi_3$ iff $\gamma < 1$,

(4.11) (4) $\xi_1 = \xi_4$ and $\xi_2 = \xi_3$ iff $\gamma = 1$.

From (4.2) it follows that

$$D_2(\xi) = D_1(\xi) + \tfrac{1}{2}(b - a)[(c\bar{\alpha} + a\alpha)|\xi - \xi_1| + (\bar{c}\bar{\alpha} + \bar{a}\alpha)|\xi - \xi_2| - (d\bar{\alpha} + b\alpha)|\xi - \xi_3| - (\bar{d}\bar{\alpha} + \bar{b}\alpha)|\xi - \xi_4|]/\alpha . \tag{4.12}$$

From (4.11) and (4.12) it follows that $D_2(\xi) = D_1(\xi)$ for all $\xi \in [0, 1]$ if $\gamma = 1$. Then, of course, $D_n(\xi) = D_1(\xi)$ for all $\xi \in [0, 1]$ and $n = 1, 2, \cdots, N$ and the optimal strategy is the myopic strategy. When $\gamma < 1$, one obtains

$$\begin{aligned} D_2(\xi) &= D_1(\xi) && \text{for } 0 \leqq \xi \leqq \xi_1 , \\ &= D_1(\xi) + (b - a)(c\xi\bar{\alpha} - a\bar{\xi}\alpha)/\alpha && \text{for } \xi_1 \leqq \xi \leqq \xi_4 , \\ &= \gamma D_1(\xi) && \text{for } \xi_4 \leqq \xi \leqq \xi_2 , \\ &= D_1(\xi) + (b - a)(d\xi\bar{\alpha} - b\bar{\xi}\alpha)/\alpha && \text{for } \xi_2 \leqq \xi \leqq \xi_3 , \\ &= D_1(\xi) && \text{for } \xi_3 \leqq \xi \leqq 1 . \end{aligned} \tag{4.13}$$

When $\gamma > 1$, one obtains

$$(4.14)\quad \begin{aligned} D_2(\xi) &= D_1(\xi) && \text{for } 0 \leqq \xi \leqq \xi_4 ,\\ &= D_1(\xi) + (b-a)(\bar{d}\xi\bar{\alpha} - \bar{b}\bar{\xi}\alpha)/\alpha && \text{for } \xi_4 \leqq \xi \leqq \xi_1 ,\\ &= \gamma D_1(\xi) && \text{for } \xi_1 \leqq \xi \leqq \xi_3 ,\\ &= D_1(\xi) + (b-a)(\bar{c}\xi\bar{\alpha} - \bar{a}\bar{\xi}\alpha)/\alpha && \text{for } \xi_3 \leqq \xi \leqq \xi_2 ,\\ &= D_1(\xi) && \text{for } \xi_2 \leqq \xi \leqq 1 . \end{aligned}$$

In order to compute $D_3(\alpha)$, it is necessary to know $D_2(\alpha(X))$ for $X = 1$ and $D_2(\alpha(Y))$ for $Y = 1$. First of all, for $X = 1$, $\alpha(X) = \alpha c/(\alpha c + \bar{\alpha}a)$ and for $Y = 1$, $\alpha(Y) = \alpha d/(\alpha d + \bar{\alpha}b)$. How do these two quantities compare with the four critical points ξ_1, ξ_2, ξ_3, and ξ_4? Routine manipulation with inequalities yields the following results:

$$(4.15)\quad \alpha c/(\alpha c + \bar{\alpha}a) \leqq \xi_2 \quad \text{iff} \quad a + c \geqq 1 ,$$

$$(4.16)\quad \alpha c/(\alpha c + \bar{\alpha}a) \leqq \xi_3 \quad \text{iff} \quad a + c \geqq \gamma .$$

$$(4.17)\quad \alpha d/(\alpha d + \bar{\alpha}b) \leqq \xi_1 \quad \text{iff} \quad a + c \geqq \gamma ,$$

$$(4.18)\quad \alpha d/(\alpha d + \bar{\alpha}b) \leqq \xi_4 \quad \text{iff} \quad a + c \geqq 2\gamma - 1 .$$

This means that for a fixed value of γ the value of $D_3(\alpha)$ will depend on the value of $a + c$. Suppose $\gamma < 1$, then through the use of (4.13) the following may be obtained.

$$(4.19)\quad \begin{aligned} D_3(\alpha) &= \bar{\alpha}(b-a)(c-a)(1-\gamma) && \text{for } a + c < 2\gamma - 1 , \quad \gamma > \tfrac{1}{2} ,\\ &= \bar{\alpha}(b-a)(c-a)(\gamma - a - c) && \text{for } 2\gamma - 1 \leqq a + c < 1 ,\\ &= \bar{\alpha}(b-a)(c-a)(\gamma - 1) && \text{for } 1 \leqq a + c . \end{aligned}$$

Now suppose $\gamma > 1$; then through the use of (4.14) the following may be obtained.

$$(4.20)\quad \begin{aligned} D_3(\alpha) &= \bar{\alpha}(b-a)(c-a)(1-\gamma) && \text{for } a + c < 1 ,\\ &= \bar{\alpha}(b-a)(c-a)(a + c - \gamma) && \text{for } 1 \leqq a + c < 2\gamma - 1 ,\\ &= \bar{\alpha}(b-a)(c-a)(\gamma - 1) && \text{for } 2\gamma - 1 \leqq a + c , \quad \gamma \leqq \tfrac{3}{2} . \end{aligned}$$

It is evident from (4.19) and (4.20) that as long as $\gamma \neq 1$, $D_3(\alpha) = 0$ if and only if $a + c = \gamma$. Thus we have proved the following theorem.

THEOREM 4.2. *Suppose $b > a$, $c < d$, and $a + b = c + d = \gamma$. If $\gamma \neq 1$ then* $D_3(\alpha) = 0$ iff $(c, d) = (b, a)$.

Our search for conditions under which the myopic strategy is optimal has come to an end. D. Feldman [4], in his celebrated paper, proved that if $(c, d) = (b, a)$ then the optimal strategy was the myopic rule. In a sense, Theorem 4.2 guarantees the necessity of this condition. The results of this section may be summarized in the following theorem.

THEOREM 4.3. *Suppose the prior distribution on (ρ, λ) is concentrated at two points (a, b) and (c, d) in the unit square and that $N > 2$. The myopic strategy is optimal if and only if one of the following four conditions holds.*

(i) $a \leqq b$ *and* $c \leqq d$,
(ii) $a \geqq b$ *and* $c \geqq d$,
(iii) $a + b = c + d = 1$,
(iv) $(c, d) = (b, a)$.

In all fairness it should be pointed out that even if the prior distribution is not concentrated at two points, the myopic strategy remains optimal for obvious reasons, as long as the prior is concentrated either on the line $\rho + \lambda = 1$ or on one side of the main diagonal.

REFERENCES

[1] BERRY, D. (1972). A Bernoulli two-armed bandit. *Ann. Math. Statist.* **43** 871–897.
[2] DEGROOT, M. (1970). *Optimal Statistical Decisions.* McGraw-Hill, New York.
[3] FABIUS, J. and VAN ZWET, W. R. (1970). Some remarks on the two-armed bandit. *Ann. Math. Statist.* **41** 1906–1916.
[4] FELDMAN, D. (1962). Contributions to the "two-armed bandit" problem. *Ann. Math. Statist.* **33** 847–856.
[5] QUISEL, K. (1965). Extensions of the two-armed bandit and related processes with on-line experimentation. Stanford Technical Report No. 137, Stanford Univ.
[6] ZACKS, S. (1967). Bayes sequential strategies for crossing a field containing absorption points. *Naval Res. Logist. Quart.* **14** 329–343.

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FLORIDA 32611