The Annals of Statistics
1985, Vol. 13, No. 1, 418–422

FURTHER CONTRIBUTIONS TO THE "TWO-ARMED BANDIT" PROBLEM

BY ROBERT KEENER[1]

University of Michigan

A version of the two-armed bandit with two states of nature and two repeatable experiments is studied. With an infinite horizon and with or without discounting, an optimal procedure is to perform one experiment whenever the posterior probability of one of the states of nature exceeds a constant ξ^*, and perform the other experiment whenever the posterior is less than ξ^* with indifference when the posterior equals ξ^*. ξ^* is expressed in terms involving expectations of ladder variables and can be calculated using Spitzer series.

1. Introduction and main result. The version of the "two-armed bandit" problem that will be studied may be described as follows. At each state n, a researcher chooses one of two experiments using past information. Performing this experiment, he observes a random variable Z_n. Conditional on the value of an unknown parameter $\theta \in \{\theta_1, \theta_2\}$ and the choice of experiment, the distribution of Z_n is independent of the past. Let $\tilde{A}_i(\tilde{B}_i)$ be the distribution of Z_n given $\theta = \theta_i$ and $A(B)$ is performed. The only relevant information from an experiment is the log likelihood ratio and we will define $X_n = \log(d\tilde{A}_1/d\tilde{A}_2)(Z_n)$ when A is performed, and $X_n = \log(d\tilde{B}_1/d\tilde{B}_2)(Z_n)$ when B is performed. To allow the possibility that $\tilde{A}_1$ and $\tilde{A}_2$ are not mutually absolutely continuous, performing A, X_n will be $+\infty$ on the A_2-null set where the Radon-Nikodym derivative is undefined and $-\infty$ when $(d\tilde{A}_1/d\tilde{A}_2)(Z_n) = 0$. Similarly for B. $A_i(B_i)$ will be the distribution on $\mathbb{R} \cup \{+\infty, -\infty\}$ for X_n when $\theta = \theta_i$ and $A(B)$ is performed.

If $\theta = \theta_1$, A is the preferred experiment and a cost $c_1 > 0$ is incurred each time B is used. Conversely if $\theta = \theta_2$, a cost c_2 is incurred whenever A is used. A Bayesian approach to this problem will be pursued, and an *optimal* procedure is one which minimizes risk; i.e. expected loss. There is no horizon, i.e., the total number of experiments is infinite, and except for Corollary 1, no discounting. Let ξ denote the prior probability that $\theta = \theta_1$ and ξ_n the posterior probability that $\theta = \theta_1$ after performing n experiments. In symmetric problems where $\tilde{A}_1 = \tilde{B}_2$, $\tilde{A}_2 = \tilde{B}_1$, and $c_1 = c_2$ (equality of c_1 and c_2 can be relaxed absorbing these costs in the prior), an optimal procedure is to select A at state $n + 1$ if $\xi_n > 1/2$ and B if $\xi_n < 1/2$ with indifference for $\xi_n = 1/2$. This was established by Feldman (1962) and his methods are useful in our asymmetric problem in characterizing the form of the optimal solution. Let T_A^+ be the first strict ascending ladder epoch for a (extended) random walk with steps distributed as A_1 (see Chapter 12 of Feller, 1971, for definitions of these terms), and let T_A^- be the first weak descending ladder point for a random walk with step distribution A_2. Define $T_B^\pm$

Received February 1984; revised September 1984.

[1] Research supported by the National Science Foundation under Grant MCS-8102080.

AMS 1980 *subject classifications*. Primary 62L05, secondary 62L10.

Key words and phrases. Dynamic programming, sequential design, random walks.

similarly replacing A's with B's. Let

$$W_A = \exp - \sum_{n=1}^{\infty} (1/n)\{A_1^{*n} ([-\infty, 0)) + A_2^{*n} ([0, \infty])\}$$

where * denotes convolution. Expressed in terms of the ladder variables, $W_A = (ET_A^+ ET_A^-)^{-1}$. The mean of A_1 is a Kullback-Leibler information number and is positive (by Jensen's inequality) except in the degenerate case where $\tilde{A}_1 = \tilde{A}_2$ (i.e. experiment A is completely uninformative). Excluding this case, from the theory of random walks, $ET_A^+ < \infty$ (see Lemma 2, page 610 of Feller, 1971). Similarly $ET_A^- < \infty$ provided $\tilde{A}_1 \neq \tilde{A}_2$. Define W_B as W_A with B's replacing A's.

THEOREM 1. *Any procedure which performs A whenever*

$$W_A/(1 - \xi_n)c_2 > W_B/\xi_n c_1$$

and B whenever the reverse inequality holds is optimal (*has minimal risk*).

Bandit problems have received considerable attention in the statistical literature, in part because they address the fundamental question: When is a more costly but more informative experiment preferable to a less costly but less informative experiment? Since $(1 - \xi_n)c_2$ is the "immediate" cost for performing A, this theorem asserts that W_A is the correct measure of the information content of A in this problem. To some extent this theorem complements results of Gittins' (1979) which show that in problems with discounting and independent arms, the correct measure of the value of an experiment is a quantity called its dynamic allocation index. For k independent arms, θ should be a k-dimensional vector, with independent components, and the distribution of the observation using the jth arm would depend on θ only through its jth component. With two states of nature this possibility is precluded except in trivial cases.

When one of the experiments, B say, completely determines θ, i.e., $\tilde{B}_1$ and $\tilde{B}_2$ are singular, Theorem 1 describes the optimal solution for a power one test with two states of nature. This solution was first discovered by Lorden (1977). The numbers W_A also play an important role in his subsequent research (Lorden, 1984) leading to stopping rules asymptotically optimal to $o(c)$ as $c \to 0$ (c is the sampling cost) in a large class of composite testing problems with an indifference zone. Our theorem may be useful in an asymptotic analysis of more general bandit problems with an indifference zone.

Other articles about the two-armed bandit have appeared by Berry (1972), Fabius and van Zwet (1970) and Chernoff (1972). The books by Whittle (1972) give a good general discussion of dynamic programming. The chapter most relevant to this research, on negative programming is from Strauch (1966).

2. Proofs. Let ξ_A and ξ_B be distributed as the posterior probability that $\theta = \theta_1$ when A or B respectively are performed. For Borel f: $[0, 1] \to [0, \infty]$ define the operator T by

$$Tf(\xi) = \min\{c_2(1 - \xi) + E_\xi f(\xi_A), c_1\xi + E_\xi f(\xi_B)\}$$

where E_ξ denotes expectation when $P(\theta = \theta_1) = \xi$. From standard theorems in negative programming (Strauch, 1966) $T^{(n)}0 \uparrow R$ pointwise as $n \to \infty$ where R is the Bayes risk, $T^{(n)}$ is T composed with itself n times, and 0 denotes the zero function. Viewing R as the infimum of linear functions, it is concave and hence continuous on [0, 1] (continuity at 0 and 1 holds because there exist stationary policies with risk functions that approach 0 as $\xi \to 0$ or 1). In the proof of Theorem 2.1 of Feldman (1962) it is shown that the function

$$\Delta(\xi) = E_\xi T^{(n)}0(\xi_B) + \xi c_1 - E_\xi T^{(n)}0(\xi_A) - (1 - \xi)c_2$$

is increasing in ξ for all n. Letting $n \to \infty$, this implies using dominated convergence that

$$E_\xi R(\xi_B) + \xi c_1 - E_\xi R(\xi_A) - (1 - \xi)c_2$$

is nondecreasing. This function is continuous (dominated convergence) and varies from $-c_2$ at 0 to c_1 at 1 and has a zero at some point ξ^*. Consequently the procedure δ^* which selects B when $\xi_n \le \xi^*$ and A otherwise is optimal. To complete the proof we will show that ξ^* is given uniquely by

$$\xi^*/(1 - \xi^*) = c_2 W_B/(c_1 W_A). \tag{1}$$

Let N_A and N_B be the number of times that A and B are performed using δ^*. Let a_i be the distribution for the first (a_1 strict, a_2 weak) descending ladder height for a random walk with step distribution A_i, and b_i the distribution for the first (b_1 strict, b_2 weak) ascending ladder height for a random walk with step distribution B_i. Define the renewal measures

$$U_i^A = \textstyle\sum_{n=0}^\infty a_i^{*n} \quad \text{and} \quad U_i^B = \sum_{n=0}^\infty b_i^{*n},$$

where the initial term in these sums is a point mass at the origin. By Bayes theorem, A_i is the distribution for the change in the log odds ratio when A is performed and $\theta = \theta_i$. Using the derivation leading to equation (4.19) of Keener (1980) (being careful with the possibly discrete or extended character of the A_i's and B_i's),

$$f_2(\xi) = E_\xi(N_A \mid \theta = \theta_2) = \frac{ET_A^-}{ET_B^-} \int_{\mathbb{R}} U_2^B((s^* - s - x, \infty))\, dU_2^A(x)$$

and

$$f_1(\xi) = E_\xi(N_B \mid \theta = \theta_1) = \frac{ET_A^+}{ET_A^+} \int_{\mathbb{R}} U_1^A((-\infty, s^* - s - x))\, dU_1^B(x)$$

where $s = \log(\xi/(1 - \xi))$ and $s^* = \log(\xi^*/(1 - \xi^*))$. These functions f_1 and f_2 are discontinuous at $\xi = \xi^*$. When the A_i's and B_i's are distributions for continuous (possibly extended) random variables,

$$\lim_{\varepsilon \downarrow 0} f_2(\xi^* + \varepsilon) - f_2(\xi^* - \varepsilon) = \frac{ET_A^-}{ET_B^-},$$

and

$$\lim_{\varepsilon \downarrow 0} f_1(\xi^* + \varepsilon) - f_1(\xi^* - \varepsilon) = \frac{ET_A^+}{ET_B^+}.$$

This proves that ξ^* satisfies (1) in this case because $R(\xi) = \xi f_1(\xi)c_1 + (1 - \xi)f_2(\xi)c_2$ is continuous at ξ^*.

In the general case where some of A_i's and B_i's may have atoms in $\mathbb{R}$, it is still true that f_1 and f_2 are discontinuous at ξ^*, and the discontinuities uniquely determine ξ^* by continuity of R. Lacking a direct way of evaluating these discontinuities, we will proceed indirectly. Define $\hat{A}_1 = A_1^* N(\varepsilon, \varepsilon/2)$, $\hat{B}_1 = B_1^* N(\varepsilon, \varepsilon/2)$, $\hat{A}_2 = A_2^* N(-\varepsilon, \varepsilon/2)$ and $\hat{B}_2 = B_2^* N(-\varepsilon, \varepsilon/2)$. These are the distributions for the log likelihood ratio where, after observing the outcome associated with A or B, we observe a normal variable with variance $\varepsilon/2$ and mean $\pm\varepsilon$ according to whether $\theta = \theta_1$ or θ_2. Define $\hat{\xi}_\varepsilon^*$, $\tilde{T}_\varepsilon$, $\hat{\Delta}_\varepsilon$ and $\hat{R}_\varepsilon$ as ξ^*, T, Δ and R with $\hat{A}_i$'s and $\hat{B}_i$'s replacing A_i's and B_i's. By Jensen's inequality, if h is concave, $\hat{T}_\varepsilon h \leq Th$ and $\hat{T}_{\varepsilon_1} h \leq \hat{T}_{\varepsilon_2} h$ for $\varepsilon_1 > \varepsilon_2$ (for this assertion view $\hat{T}_{\varepsilon_2}$ as T and $\hat{T}_{\varepsilon_1}$ as $\hat{T}_{\varepsilon_1 - \varepsilon_2}$ in the first assertion). Since $\hat{R}_\varepsilon 0 = \lim_{n\to\infty} \hat{T}_\varepsilon^{(n)} 0$, $\hat{R}_\varepsilon \leq R$. If h is continuous on [0, 1] (and consequently uniformly continuous) then $\hat{T}_\varepsilon h \to Th$ as $\varepsilon \downarrow 0$ uniformly on [0, 1]. By induction, $\hat{T}_\varepsilon^{(n)} 0 \uparrow T^{(n)} 0$ as $\varepsilon \downarrow 0$ and hence $R_\varepsilon \to R$ as $\varepsilon \downarrow 0$ uniformly. This implies that $\hat{\Delta}_\varepsilon \to \Delta$ and since Δ is strictly (since ξ^* is unique) increasing at ξ^*, $\hat{\xi}_\varepsilon^* \to \xi^*$ as $\varepsilon \to 0$. The theorem now follows because

$$\hat{A}_1^{*n}([-\infty, 0)) \to A_1^{*n}([-\infty, 0)) + \tfrac{1}{2} A_1^{*n}(\{0\}) \quad \text{as} \quad \varepsilon \to 0$$

and similar statements about A_2, B_1 and B_2 show that $\hat{\xi}_\varepsilon^*/(1 - \hat{\xi}_\varepsilon^*) \to c_2 W_B/(C_1 W_A)$ (the identity $A_1^{*n}(\{0\}) = A_2^{*n}(\{0\})$ is used to eliminate the ½).

Suppose now there is a discount factor $0 \leq \beta < 1$, i.e. the goal is minimizing

$$E \sum_{n=1}^\infty \beta^{n-1}[c_1 I\{\theta = \theta_1, e_n = B\} + c_2 I\{\theta = \theta_2, e_n = A\}]$$

where e_n is the nth experiment performed. This expectation is the same as

$$E \sum_{n=1}^Q [c_1 I\{\theta = \theta_1, e_n = B\} + c_2 I\{\theta = \theta_2, e_n = A\}]$$

where Q is a geometric random variable independent of θ and all the observations. This problem is then the same as an undiscounted problem where at each stage there is a chance $1 - \beta$ of stopping the experiment. Since stopping is equivalent to learning the state of nature (i.e. no later costs are incurred) we can solve this problem by replacing A_1 with $A_1^* P$ where $P(\{0\}) = \beta = 1 - P(\{+\infty\})$, and similar substitutions for A_2, B_1 and B_2. This proves the following corollary.

COROLLARY 1. *In the discounted case, any procedure which performs A whenever*

$$V_A/(1 - \xi_n)c_2 > V_B/\xi_n c_1$$

and B whenever the reverse inequality holds is optimal. V_A *is given by*

$$V_A = \exp -\sum_{n=1}^\infty \beta^n \{A_1^{*n}([-\infty, 0)) + A_2^{*n}([0, \infty])\}/n$$

and V_B *is the same with* B_i *replacing* A_i.

REFERENCES

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

CHERNOFF, H. (1972). *Sequential Analysis and Optimal Design.* S.I.A.M. Pennsylvania.

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. (1971). *An Introduction to Probability Theory and Its Applications Vol. II.* Wiley, New York.

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

KEENER, R. W. (1980). Renewal theory and the sequential design of experiments with two states of nature. *Comm. Statist.* **A9(16)** 1699–1726.

LORDEN, G. (1977). Nearly-optimal sequential tests for finitely many parameter values. *Ann. Statist.* **5** 1–21.

LORDEN, G. (1984). Nearly optimal sequential tests for exponential families. *Ann. Statist.*, to appear.

STRAUCH, R. E. (1966). Negative dynamic programming. *Ann. Math. Statist.* **37** 871–890.

WHITTLE, P. (1982). *Optimization over Time, Vol. I and II.* Wiley, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109