Rochester Center for Economic Research

A Class of Bandit Problems Yielding Myopic Optimal Strategies

Banks, Jeffrey S. and Rangarajan K. Sundaram

Working Paper No. 239
September 1990

University of Rochester

A Class of Bandit Problems Yielding Myopic Optimal Strategies

Jeffrey Banks and Rangarajan Sundaran

Rochester Center for Economic Research
Working Paper No. 239

A CLASS OF BANDIT PROBLEMS YIELDING MYOPIC OPTIMAL STRATEGIES*

Jeffrey S. Banks and Rangarajan K. Sundaram
Department of Economics, Harkness Hall
University of Rochester
Rochester, NY 14627

Working Paper No. 239

September 1990

Abstract. Consider an n-armed bandit problem in which each of the n independent arms generates rewards according to one of two known reward distributions, but the true distribution associated with some (or all) of the arms is unknown. If rewards are discounted geometrically over an infinite horizon, and the reward distributions either (a) both admit density representations, or (b) are both discrete distributions, then we show that myopic strategies are optimal in the space of all strategies. Some interesting implications of this result for a class of Bernoulli reward distributions are given.

Keywords: **n-armed bandit problems, Gittins index, myopic strategies, Bernoulli distributions, random walkc**

*Financial support from the National Science Foundation and the Sloan Foundation to the first author is gratefully acknowledged.

1. Introduction

This paper studies the class of n–armed bandit problems characterized by the property that each of the n arms generates rewards to the decision–maker according to one of two known distributions, F_1 and F_2. The arms are assumed independent, so that the probability any given arm is of type F_1 is independent of the "true" types of the remaining arms. Under some mild technical conditions on the distributions F_1 and F_2, and the assumption of geometric discounting by the decision–maker over an infinite horizon, we prove the optimality of myopic strategies, ie. strategies that specify the arm to be played in each period solely on the basis of maximizing the current expected reward. Thus, in our environment the trade–off between current rewards and information acquisition to enhance future rewards is avoided, in that the "optimal" consideration of future rewards is in essence embedded in the maximization of current rewards.

The problem we study is closely related to, but distinct from, the work of Rodman (1978), who generalizes the results of Feldman (1962). In particular, Rodman is concerned with the case where exactly one of the arms is of type F_1 and the remaining arms are of type F_2, but it is not known which arm is of type F_1.[1] Rodman shows that myopic strategies are optimal in this framework as well. In addition, our problem is related to that of Berry and Fristedt (1985, Theorem 4.3.9), where there are two arms available, both of which are Bernoulli, and where the arms have a common two–point support (thus each arm can be one of two possible types, as here). Berry and Fristedt show that as long as the discount sequence is nonincreasing, a myopic strategy is optimal. Theorem 1 below then shows that, by restricting attention to geometric discounting, their result holds for an arbitrary

[1] Rodman also allows the horizon to be finite and the discounting to be uniform.

number of non–Bernoulli arms.[2]

Section 2 below describes our framework precisely, and gathers definitions. The Dynamic Allocation Index (DAI) of Gittins and Jones (1974) is reviewed in Section 3, and then invoked to show the existence of a solution to the decision–maker's optimization problem. In Section 4 we show that the arm having the highest current expected reward is also one of the arms with the highest DAI, thus proving the optimality of myopic strategies. Finally, Section 5 provides a characterization of the optimal solution for a class of Bernoulli reward distributions.

2. The Problem

For a comprehensive and general description of bandit problems, we refer the reader to Berry and Fristedt (1985); the following suffices for our purposes.

The horizon we consider is infinite. Time is discrete, and periods are indexed by $t = 0,1,2,...$ In each period, a decision–maker, hereafter referred to as the principal, must decide which of the independent arms of the bandit to play, where $N = \{1,...,n\}$ denotes the set of available arms. Each arm yields a reward $r \in \mathbb{R}$ to the principal, where the realization of this reward is according to one of two known distributions, F_1 and F_2; however, the principal does not know the "true" type (ie. whether it is F_1 or F_2) of some or all of the arms. He begins instead with a vectors of priors $p = (p_1,...,p_n) \in [0,1]^n$, where $p_i \in [0,1]$ is the principal's prior belief that arm i is of type F_1.

We assume F_1 and F_2 have finite expectation. For technical reasons we shall also require these distributions to satisfy one of the following conditions:

[2] For further references on myopic optimal strategies, we direct the reader to the excellent monograph of Berry and Fristedt (1985). Recent additions include Fristedt and Berry (1988), and O'Flaherty (1989).

(C1) F_1 and F_2 admit densities (with respect to the Lebesgue measure), denoted f_1 and f_2, resp.; or

(C2) F_1 and F_2 are discrete distributions.

Throughout, our presentation is confined to case (C1) for notational cleanliness; however, the modifications of the proofs to cover (C2) are transparent. Let $\mathbf{R}$ = supp f_1 ∪ supp f_2, where for $k = 1,2$ supp $f_k = \{r \in \mathbb{R} : f_k(r) > 0\}$. All integrals in the material following are to be taken as being over the set $\mathbf{R}$.

Observations accumulated through the play of the various arms are used to update the vector of priors in a Bayesian manner. Since the arms are independent by assumption, information regarding a particular arm may be accumulated only when that arm is employed. Let $p^t = (p_1^t,\dots,p_n^t)$ denote the principal's beliefs at the beginning of period t, and suppose arm i is played that period and the reward $r \in \mathbf{R}$ is witnessed. The updated vector of beliefs for the principal, $p^{t+1} = (p_1^{t+1},\dots,p_n^{t+1})$, is then given by

$$p_j^{t+1} = p_j^t, \qquad \text{if } j \neq i, \text{ and} \tag{2.1a}$$

$$p_i^{t+1} = \beta(p_i^t, r) = \frac{p_i^t f_1(r)}{p_i^t f_1(r) + (1-p_i^t) f_2(r)}, \tag{2.1b}$$

where $\beta(.,.)$ represents the Bayes map. [For $r \notin \mathbf{R}$ we adopt the convention $p^{t+1} = p^t$.]

Histories and strategies are defined in the usual manner: let $H^0 = \phi$, and for integers $t \geq 1$ let H^t denote the set of all possible (partial) histories of length t of observations, with generic element h^t. A <u>strategy</u> σ for the principal is a sequence of measurable maps $\{\sigma^t\}$ such that $\sigma^0 \in N$ and for all integers $t \geq 1$, $\sigma^t: H^t \to N$.

Thus $\sigma^t(h^t)$ is the arm recommended by the strategy σ in period t if the history h^t of rewards has been observed. Let Σ denote the set of all possible strategies available to the principal.

We assume the principal discounts future rewards geometrically using the factor $\delta \in (0,1)$. The principal's objective is to maximize the (expected) discounted stream of rewards over the infinite horizon. Formally, each strategy σ defines, in the obvious manner, a t–th period expected reward for the principal based on the initial (period 0) vector of priors p, denoted $r^t(\sigma)(p)$. Thus, the total discounted reward under σ from p, or the worth of strategy σ, denoted $W(\sigma)(p)$, is given by

$$W(\sigma)(p) = \sum_{t=0}^{\infty} \delta^t r^t(\sigma)(p) \ . \tag{2.2}$$

The principal's objective is thus to find a strategy $\sigma^* \in \Sigma$ such that $W(\sigma^*) = \sup_{\sigma\in\Sigma} W(\sigma)$; if such a strategy exists, it will be called an optimal strategy.

Finally, let $\Gamma_k = \int r dF_k(r)$, k = 1,2, denote the expected reward from a type k arm. To avoid trivializing the principal's optimization problem, we assume $\Gamma_1 \neq \Gamma_2$, and, without loss of generality, that $\Gamma_1 > \Gamma_2$. For notational convenience we let $\Gamma(p_i) = p_i\Gamma_1 + (1-p_i)\Gamma_2$ be the expected one–period reward from employing arm i given belief p_i, and $f^{p_i}(r) = p_i f_1(r) + (1-p_i)f_2(r)$ be the expected density governing rewards.

3. Existence of an optimal strategy

Although the existence of an optimal strategy for the principal may be established directly by appealing to standard results in the theory of discounted dynamic programming, we do not adopt this route here. Rather, since the Dynamic Allocation Index (DAI) is an essential ingredient in the proof of our main result in the next section, we construct these indices and invoke the Theorem of Gittins and

Jones (1974) to accomplish this end.

The DAIs associated with the arms are constructed in the following manner: pick an arm $i \in N$, and suppose $p_i \in [0,1]$ is the principal's prior belief arm i is of type F_1. Consider the optimal stopping problem in which, in each period, the principal must decide whether to play arm i or stop the process and accept the terminal reward $m \in \mathbb{R}$.[3] Standard results in the bandit literature (eg. Berry and Fristedt 1985, Ross 1983) establish the existence of a unique continuous function $V_i(.,m) : [0,1] \to \mathbb{R}$ such that $V_i(p_i,m)$ is the value to the principal of this optimal stopping problem when the initial prior on arm i is p_i and the terminal reward is m. Moreover, $V_i(.,m)$ satisfies the functional equation

$$V_i(p_i,m) = \max \{m, \Gamma(p_i) + \delta \int V_i(\beta(p_i,r),m) f^{p_i}(r)dr\}. \tag{3.1}$$

It is routine to verify that if m is sufficiently large, say $m > \Gamma_1/(1-\delta)$, then $V_i(p_i,m) = m$ for all $p_i \in [0,1]$, while if m is sufficiently small, say $m < \Gamma_2/(1-\delta)$, then $V_i(p_i,m) > m$ for all $p_i \in [0,1]$. The DAI of arm i when the prior is p_i, denoted $m_i(p_i)$, is defined by

$$m_i(p_i) = \inf \{m \in \mathbb{R} : V_i(p_i,m) = m\}. \tag{3.2}$$

Since arms are identical up to the prior on their true type, it follows that $V_i(.,m) = V_j(.,m)$ for all $i,j \in N$; this implies of course that $m_i(.) = m_j(.)$ for all $i,j \in N$ as well. Henceforth these common functions will be denoted $V(.,m)$ and $m(.)$, respectively.

[3] Alternatively, one could consider the (strategically equivalent) two–armed bandit problem in which one arm is arm i, and the other generates a known constant payoff of $m(1-\delta)$.

Theorem 0 (Gittins and Jones, 1974). The optimal initial selections in the n–armed bandit problem given the priors $p = (p_1,\dots,p_n)$ are those arms i for which

$$m(p_i) = \bigvee_{j=1}^{n} m(p_j) . \tag{3.3}$$

4. The optimality of myopic rules

The myopic strategy σ^m for the principal is the strategy that in period t recommends the arms that have the highest expected current rewards based on the priors at the beginning of period t; thus, given $p^t = (p_1^t,\dots,p_n^t)$, σ^m selects any arm i for which

$$\Gamma(p_i^t) = \bigvee_{j=1}^{n} \Gamma(p_j^t) . \tag{4.1}$$

Myopic strategies are a particularly simple class of strategies in that they do not require the principal to take account of the impact of his current actions on his future rewards. In this sense, they go directly against the fundamental characteristic that bandit problems often exhibit, the trade–off between current rewards and the acquisition of information that might improve the principal's future prospects, a trade–off that Whittle (1982, p. 210) claims, "embodies in essential form the conflict evident in all human action" (emphasis added). Nonetheless, we show in this Section that in the current environment a myopic strategy by the principal is actually optimal.

Theorem 1. $W(\sigma^m) = \sup_{\sigma\in\Sigma} W(\sigma)$.

The proof of Theorem 1 follows as a consequence of several Lemmata. The underlying argument is, however, quite simple, and an outline of the proof may be useful. First, observe that since $\Gamma_1 > \Gamma_2$, so

$$\Gamma(p_i) = \bigvee_{j=1}^{n} \Gamma(p_j) \quad <=> \quad p_i = \bigvee_{j=1}^{n} p_j \ . \tag{4.2}$$

We show that whenever $p,p' \in [0,1]$ satisfy $p \geq p'$, then we must also have $m(p) \geq m(p')$, where m(.) is the DAI function of the previous Section. Since this function has the same form for all the arms, this will imply

$$p_i = \bigvee_{j=1}^{n} p_j \quad => \quad m(p_i) = \bigvee_{j=1}^{n} m(p_j) \ . \tag{4.3}$$

Theorem 1 is now an immediate consequence of (4.3) and Theorem 0.

The following Lemmata establishing (4.3) all concentrate on the optimal stopping problem described in the previous Section. For expositional ease we drop the subscript "i" and use $p \in [0,1]$ to represent the prior on a generic arm.

Lemma 1. For all $p \in [0,1]$, $V(p,.) : \mathbb{R} \to \mathbb{R}$ is continuous and non–decreasing.

Proof. Berry and Fristedt (1985, Theorem 5.0.1) prove this for the strategically equivalent case of a two–armed bandit with one known arm. □

Remark. V(p,.) continuous implies in particular that $V(p,m(p)) = m(p)$ for all $p \in [0,1]$.

Lemma 2. For all $p \in [0,1]$, $m(p) \in M \equiv [\Gamma_2/(1-\delta),\Gamma_1/(1-\delta)]$.

Proof. If $m < \Gamma_2/(1-\delta)$, then $V(p,m) > m$ for all $p \in [0,1]$; conversely, if $m > \Gamma_1/(1-\delta)$, then $V(p,m) = m$ for all $p \in [0,1]$. □

Lemma 3. For all $m \in \mathbb{R}$, $V(.,m) : [0,1] \to \mathbb{R}$ is convex on $[0,1]$.

The proof of Lemma 3 is somewhat tedious, and so is relegated to an Appendix. However the intuition behind the result is straightforward: consider $p = (p_1,\ldots,p_n)$ and $p' = (p_1',\ldots,p_n')$, and let $p_\lambda = \lambda p + (1-\lambda)p'$, where $\lambda \in (0,1)$. Then p_λ can be viewed as the expected prior before it is known whether an event has occurred, where λ is the probability of the event's occurrence. On average, knowing whether or not the event has occurred cannot lower the value of the problem to the principal, which is precisely the convexity of $V(.,m)$.

Lemma 4. For all $m \in M$, $V(.,m)$ is non-decreasing on $[0,1]$.

Proof. $V(p,m) \geq m$ for all $p \in [0,1]$ follows from (3.1). Also, if $m \geq \Gamma_2/(1-\delta)$, then $V(0,m) = m$. The result now follows from Lemma 3. □

Lemma 5. For each $m \in M$, if $V(p,m) = m$ for some $p \in (0,1]$, then $V(p',m) = m$ for all $p' \in [0,p)$.

Proof. Immediate consequence of Lemma 4 and the fact that $V(.,m) \geq m$. □

Lemma 6. $p \geq p' \Rightarrow m(p) \geq m(p')$ for all $p,p' \in [0,1]$.

Proof. If $p = p'$, then clearly $m(p) = m(p')$, so suppose $p > p'$. Then, since $V(p,m(p)) = m(p)$, so $V(p',m(p)) = m(p)$ by Lemma 5. By definition of $m(.)$, then, $m(p') \leq m(p)$. □

Proof of Theorem 1. Immediate consequence of (4.2), Lemma 6, and Theorem 0.

QED

Theorem 1 thus establishes the optimality of myopic strategies in "two–type" bandit problems regardless of the values of the remaining parameters in the model, ie. the reward distributions F_1, F_2, the principal's prior belief p, the discount factor δ, or the number of arms n.

5. A Bernoulli example

Consider the following special case of this model: the reward distributions are Bernoulli, with $q_k = \text{prob}\{r=1 : F=F_k\}$, $1-q_k = \text{prob}\{r=0 : F=F_k\}$, $k = 1,2$. We assume $q_1 = 1-q_2$, and $q_1 > q_2$, so that $q_1 > 1/2 > q_2$. Finally, we suppose that all arms are a priori identical to the principal, so that the initial prior $p = (\pi,...,\pi)$ for some $\pi \in (0,1)$. Note that under these assumptions, the posterior belief on an arm which has generated α 1's and β 0's is a function only of the difference $\alpha-\beta$; in particular, this posterior is the <u>same</u> as the one resulting from observing $\alpha-\beta$ 1's and no 0's (resp. $\beta-\alpha$ 0's and no 1's) whenever $\alpha \geq \beta$ (resp. $\beta \geq \alpha$).

We assume without loss of generality that whenever the principal is indifferent between playing any subset of arms he selects the arm with the lowest number, so let the principal begin by initially selecting arm 1. Then Bayes rule along with Theorem 1 imply that the principal will remain with arm 1 until more 0's have been observed than 1's, at which time he will begin playing arm 2. Note that this decision rule is independent of the value of π. Similarly, arm 2 will be replaced with arm 3 whenever more 0's than 1's have resulted from arm 2, and so on. Finally, the principal will return to arm 1 after the first time the n–th arm has generated more 0's than 1's, since the principal's beliefs about all arms are again identical (by the earlier observation that the posterior depends only on the difference between the number of 1's and 0's observed). And, since the above process is independent of the initial prior π, the entire procedure now repeats itself. In particular, any time a previously discarded arm is chosen the decision rule governing its replacement is exactly the same as that employed the very first time the arm was chosen, regardless of the current belief about the arm. Consequently, the "survival" probability distribution of an arm each time it is newly selected is identical to the distribution the first time it was chosen.

A second characteristic of this optimal policy is that the distribution of an arm's continued use follows a random walk. To see this, consider a newly selected arm as starting at the position 1 on the real line. If a 1 is observed, the position of the arm moves one unit to the right of its previous position, while a 0 moves it one unit to the left, where a type F_k arm moves to the right with probability q_k. From the above description of the principal's optimal policy, it follows that the arm is replaced at the first instance at which the origin is reached, ie. the first time more "left–moves" than "right–moves" occur. In random walk terminology, this is simply the first–passage to the absorbing barrier at the origin.

The following features of random walks are well–known (cf. Feller, 1968). Let q denote the probability of a right–move; then, (i) if $q < 1/2$, the probability of reaching the origin at some point in time is one, while (ii) if $q \geq 1/2$, this probability is $(1-q)/q$. Further, (iii) if $q < 1/2$, the expected first–passage time is $1/(1-2q)$, while (iv) if $q \geq 1/2$ this is evidently infinite. Therefore, in this example, all arms whose true distributions are F_2 will with probability one be replaced each time they are chosen, with an expected duration of continuous play equal to $1/(1-2q_2)$. More to the point, this expected duration is independent of the entire past use of that arm, as well as its current prior. Of course, analogous statements hold for type F_1 arms, except that with positive probability such an arm will never be replaced.

Appendix

Proof of Lemma 3. Fix $m \in M$, let $I = [0,1]$, and define $C(I,\mathbb{R})$ to be the set of all continuous functions from I to $\mathbb{R}$. Endow $C(I,\mathbb{R})$ with the topology of uniform (ie. sup–norm) convergence. It is well known that $C(I,\mathbb{R})$ is then a complete metric space. Define the operator T on $C(I,\mathbb{R})$ by

$$Tw(p) = \max \{m, \Gamma(p) + \delta\int w(\beta(p,r))f^{p}(r)dr\}. \tag{A1}$$

Routine arguments show that T maps $C(I,\mathbb{R})$ into itself, and is a contraction. Hence T has a unique fixed–point, one that is evidently $V(.,m)$ given in (3.1).

We will show[4] that there is a convex function $w^{*} \in C(I,\mathbb{R})$ such that $Tw^{*} = w^{*}$. By uniqueness of the fixed–point this establishes $V(.,m)$ is convex.

Some notational simplification will greatly aid this process. For $w \in C(I,\mathbb{R})$, let

$$Hw(p) = \int w(\beta(p,r))f^{p}(r)dr, \text{ and} \tag{A2}$$

$$Mw(p) = \Gamma(p) + \delta Hw(p). \tag{A3}$$

Then, of course,

$$Tw(p) = \max \{m, Mw(p)\}. \tag{A4}$$

As a first step in showing the existence of a convex fixed–point of T, we show that if w is convex, then Hw is convex as well. By the linearity of R, this will imply that Mw is also convex. As the max of convex functions, then, Tw will be convex.

So suppose $w \in C(I,\mathbb{R})$ is convex. Let $p^{1},p^{2} \in I$, and define $p = (1-\lambda)p^{1} + \lambda p^{2}$ for $\lambda \in (0,1)$. For all $r \in \mathbf{R}$ define $\epsilon(r)$ by

[4] These arguments largely follow McLennan (1988).

$$(1-\epsilon(r))f^{p}(r) = (1-\lambda)f^{p^{1}}(r) \tag{A5}$$

(or, equivalently, $\epsilon(r)f^{p}(r) = \lambda f^{p^{2}}(r)$),

and note that

$$\beta(p,r) = (1-\epsilon(r))\beta(p^{1},r) + \epsilon(r)\beta(p^{2},r). \tag{A6}$$

Since w is convex,

$$\begin{aligned} Hw(p) &= \int w(\beta(p,r))f^{p}(r)dr \\ &\leq \int[(1-\epsilon(r))w(\beta(p^{1},r)) + \epsilon(r)w(\beta(p^{2},r))]f^{p}(r)dr && \text{(A7)} \\ &\qquad \text{(by Jensen's inequality)} \\ &= (1-\lambda)\int w(\beta(p^{1},r))f^{p^{1}}(r)dr + \lambda\int w(\beta(p^{2},r))f^{p^{2}}(r)dr && \text{(A8)} \\ &\qquad \text{(by definition of } \epsilon(.)) \\ &= (1-\lambda)Hw(p^{1}) + \lambda Hw(p^{2}). && \text{(A9)} \end{aligned}$$

Thus, w convex implies Tw convex, completing the first step of the proof.

Now let **W** be the set of all convex $w \in C(I,\mathbb{R})$ such that $w \leq Tw$. Since $\Gamma(.)$ is bounded, **W** is nonempty and bounded above. Define w^{*} by

$$w^{*}(p) = \sup_{w\in\mathbf{W}} w(p). \tag{A10}$$

Then, clearly w^{*} is convex and $w^{*} \leq Tw^{*}$. But T is also a monotone operator (ie. $v \leq w$ implies $Tv \leq Tw$), so $Tw^{*} \leq T(Tw^{*})$. Now from the above arguments w^{*} convex implies Tw^{*} convex, so $Tw^{*} \in \mathbf{W}$ as well. Thus, by the definition of w^{*}, $Tw^{*} \leq w^{*}$, implying $Tw^{*} = w^{*}$. Since V(.,m) is the <u>unique</u> fixed-point of T, $V(.,m) = w^{*}$, so V(.,m) is convex on I. QED

References

Berry, D. and B. Fristedt (1985) Bandit Problems: Sequential Allocation of Experiments. London: Chapman and Hall.

Feldman, D. (1962) Contributions to the two–armed bandit problem, Annals of Mathematical Statistics 33:847–856.

Feller, W. (1968) An Introduction to Probability Theory and Its Applications, Vol I. New York: Wiley.

Fristedt, B. and D. Berry (1988) Optimality of myopic stopping times for geometric discounting, Journal of Applied Probability 25:437–443.

Gittins, J. and D. Jones (1974) A dynamic allocation index for the sequential design of experiments, in Progress in Statistics (eds. J. Gani et. al.), pp. 241–266. Amsterdam: North–Holland.

McLennan, A. (1988) Incomplete learning in a repeated statistical decision problem, mimeo, University of Minnesota.

O'Flaherty, B. (1989) Some results on two–armed bandits when both projects vary, Journal of Applied Probability 26:655–658.

Rodman, L. (1978) On the many–armed bandit problem, Annals of Probability 6:491–498.

Ross, S. (1983) Introduction to Dynamic Programming. New York: Academic Press.

Whittle, P. (1982) Optimization Over Time: Dynamic Programming and Stochastic Control, Vol. I. New York: Wiley.