J. Appl. Prob. **29**, 625–632 (1992)
Printed in Israel

A CLASS OF BANDIT PROBLEMS YIELDING MYOPIC OPTIMAL STRATEGIES

JEFFREY S. BANKS AND
RANGARAJAN K. SUNDARAM,* *University of Rochester*

Abstract

We consider the class of bandit problems in which each of the $n \geqq 2$ independent arms generates rewards according to one of the same two reward distributions, and discounting is geometric over an infinite horizon. We show that the dynamic allocation index of Gittins and Jones (1974) in this context is strictly increasing in the probability that an arm is the better of the two distributions. It follows as an immediate consequence that myopic strategies are the uniquely optimal strategies in this class of bandit problems, regardless of the value of the discount parameter or the shape of the reward distributions. Some implications of this result for bandits with Bernoulli reward distributions are given.

MULTIARMED BANDIT PROBLEMS; DYNAMIC ALLOCATION INDEX; BERNOULLI DISTRIBUTIONS; RANDOM WALK

AMS 1991 SUBJECT CLASSIFICATION: PRIMARY 90B50

1. Introduction

We consider a class of bandit problems (cf. Berry and Fristedt (1985)) with the following structure. There are n independent arms, where $n \geqq 2$ is finite. In each period of an infinite horizon, a decision-maker (hereafter referred to as the *principal*) must decide which of the arms is to be played that period. Each arm yields rewards to the principal according to one of two known distributions F_1 and F_2, with finite expectations denoted Γ_1 and Γ_2, respectively. To avoid trivialities we suppose that $\Gamma_1 \neq \Gamma_2$, and, without loss of generality, that $\Gamma_1 > \Gamma_2$. We also assume F_1 and F_2 admit densities with respect to the Lebesgue measure, denoted f_1 and f_2, respectively. (As the reader may check, our results are also valid, with transparent modifications of the proofs, if the reward distributions are instead *discrete* (i.e. have finite or countable supports).) Let $\text{supp } f_i = \{r \mid f_i(r) > 0\}$ denote the support of f_i, and let $R = \text{supp } f_1 \cup \text{supp } f_2$.

Received 28 August 1990; revision received 8 May 1991.

* Postal address for both authors: Department of Economics, Harkness Hall, University of Rochester, Rochester, NY 14627, USA.

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

The true 'type' of some or all of the arms may be *a priori* unknown to the principal, who begins instead with a vector of priors $P = [p_1, \cdots, p_n] \in [0, 1]^n$, where $p_i \in [0, 1]$ is the prior belief that arm i is of type F_1. Observations accumulated in the course of play are then used to update this vector of prior beliefs in a Bayesian fashion: let $P^t = (p_1^t, \cdots, p_n^t)$ denote the principal's beliefs at the beginning of period t, and suppose arm i is played that period and the reward $r \in R$ is witnessed. (For $r \notin R$ we adopt the convention $P^{t+1} = P^t$.) By independence of the arms, the updated vector of beliefs for the principal, $P^{t+1} = (p_1^{t+1}, \cdots, p_n^{t+1})$, is then given by $p_j^{t+1} = p_j^t$ for $j \neq i$, and

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

Future rewards are discounted geometrically using the factor $\delta \in [0, 1)$, and the principal's objective is to maximize the discounted expected sum of rewards over the infinite horizon. Formally, a (partial) history of length t is a specification of the arms that have been chosen in each period up to t, and the consequent rewards witnessed. Let $H^0 = \varnothing$, and for integers $t \geqq 1$, let H^t denote the set of all possible histories of length t with generic element h^t. A *strategy* σ for the principal is a sequence of measurable maps $\{\sigma^t\}$ such that $\sigma^0 \in \{1, \cdots, n\}$ and for all integers $t \geqq 1$, $\sigma^t : H^t \to \{1, \cdots, n\}$. Let Σ denote the set of all possible strategies for the principal.

Each strategy σ defines, in the obvious manner, a tth period expected reward for the principal based on the initial (period 0) vector of priors P, denoted $r^t(\sigma)(P)$. The total discounted reward under σ from p, or the *worth* of strategy σ, denoted $W(\sigma)(P)$, is given by

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

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

Of special interest, from the point of view of this paper, are *myopic strategies*. A myopic strategy σ^m for the principal is a strategy that in period t recommends any of the arms having the highest expected one-period rewards based on the priors at the beginning of period t; thus, given $P^t = (p_1^t, \cdots, p_n^t)$, σ^m selects any arm i for which

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

where $\Gamma(p_i) := p_i \Gamma_1 + (1 - p_i)\Gamma_2$. It is well known that, in general, myopic strategies are suboptimal in bandit problems, since they fail to take into account the information consequences of current actions on *future* rewards (see for example Berry and Fristedt (1985)). In sharp contrast, we prove that in the current context myopic strategies uniquely identify the optimal strategies for the principal. To state the full result formally requires a few more observations.

Gittins and Jones (1974) show that for independent-armed bandit problems with geometric discounting, an index (known as the *dynamic allocation index* (DAI), or

Gittins index) can be associated with each arm, where the DAI of an arm depends solely on the distributions in the support of the arm, and the prior over these distributions. Since all arms in the family of bandit problems defined above have the same (two) distributions in their support, the dependence of the DAI on these distributions can be suppressed, and the DAI for arm i can be written as a function simply of the prior p_i, say $m(p_i)$. The DAI is defined formally in the next section. Gittins and Jones further prove that the DAIs completely characterize the principal's optimal strategy; therefore, we have the following result.

Theorem 0 (Gittins and Jones (1974)). *The optimal initial selections at the priors* $P = (p_1, \cdots, p_n)$ *in the family of bandit problems under consideration are those arms* i *for which*

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

The main result of the current paper is the following.

Theorem 1. σ^* *is an optimal strategy in the class of bandit problems under consideration if and only if the recommendations of* σ^* *are at all times myopically optimal. In particular, it is the case that, for all priors* P

$$m(p_i) = \bigvee_{j=1}^{n} m(p_j) \text{ if and only if } \Gamma(p_i) = \bigvee_{j=1}^{n} \Gamma(p_j). \tag{1.1}$$

The proof of Theorem 1 is the subject of the next section. We show there that the DAI $m(\cdot)$ is strictly increasing on [0, 1]. Evidently, so is $\Gamma(\cdot)$, since $\Gamma_1 > \Gamma_2$. Therefore, (1.1) is established, and in turn, using Theorem 0, so is Theorem 1.

Two results in the bandit literature are related to Theorem 1. Berry and Fristedt ((1985), Theorem 4.3.9) prove the optimality of myopic strategies when there are exactly two arms, each of two possible types F_1 and F_2, where F_1 and F_2 are Bernoulli distributions. Rodman (1978), who generalizes Feldman (1962), shows the optimality of myopic strategies in a model where it is known that exactly one arm is type F_1 and all others are F_2, but it is not known which is the type F_1 arm. Berry and Fristedt (1985) and Rodman (1978) allow for more general forms of discounting as well. For further references on the optimality of myopic strategies, we direct the interested reader to Berry and Fristedt (1985); recent additions include Fristedt and Berry (1988) and O'Flaherty (1989).

2. Proof of Theorem 1

We begin with a description of the DAI for a generic arm i. Consider the stopping problem in which, in each period, the principal must decide whether to play arm i for one more period, or stop the process and accept a terminal reward $m \in \mathbb{R}$. 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 $m(1-\delta)$. For notational simplicity denote the prior on arm i by just $p \in [0, 1]$.

Standard results in the bandit literature (e.g. Berry and Fristedt (1985), Ross (1983)) establish the existence of a unique continuous function $V_i(\cdot, m): [0, 1] \to \mathbb{R}$, such that $V_i(p, m)$ is the value to the principal of this optimal stopping problem when the prior on arm i is p and the terminal reward is m; additionally, V_i satisfies the functional equation

$$V_i(p, m) = \max\left\{m, \Gamma(p) + \delta \int V_i(\beta(p, r), m) f(p)(r) dr\right\},$$

where $f(p)(r) = pf_1(r) + (1 - p) f_2(r)$. The DAI of arm i when the prior is p, denoted $m_i(p)$, is then defined by

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

Since arms in our framework are identical up to the prior on their type, it follows that $V_i(\cdot, m) = V_j(\cdot, m)$ for all $i, j \in \{1, \cdots, n\}$; this implies of course that $m_i(\cdot) = m_j(\cdot)$ for all $i, j \in \{1, \cdots, n\}$ as well. These common functions are denoted $V(\cdot, m)$ and $m(\cdot)$, respectively.

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

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 2. *For all* $p \in [0, 1]$, $m(p) \geqq \Gamma(p)/(1 - \delta)$.

Proof. This follows from the observation that $V(p; m) \geqq \Gamma(p)/(1 - \delta)$ for all p. Note that if $p = 0$ or $p = 1$, then $m(p) = \Gamma(p)/(1 - \delta)$.

Lemma 3. *For all* $p \in [0, 1]$, $V(p, \cdot): \mathbb{R} \to \mathbb{R}$ *is continuous, convex, 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. Lemmata 1–3 imply $m(\cdot)$ is well defined, and $V(p, m(p)) = m(p)$ for all $p \in [0, 1]$.

Now define $MV(\cdot; m)$ by

$$MV(p; m) = \Gamma(p) + \delta \int V(\beta(p, r), m) f(p)(r) dr.$$

It is evident that for fixed p, $MV(p; \cdot)$ inherits the continuity and convexity in m of $V(p; \cdot)$.

Lemma 4. *For each* $m \in \mathbb{R}$, $MV(\cdot; m)$ *is convex on* $[0, 1]$.

Proof. The Appendix establishes that $V(\cdot; m)$ is convex in p; and that $Mw(p) := \{\Gamma(p) + \delta \int w[\beta(p, r)] f(p)(r)\}$ is convex as a function of p whenever w is.

Lemma 5. $MV(p; m) \lessgtr m$ *as* $m \lessgtr m(p)$.

Proof. If $p = 0$ or 1, this is immediate, so suppose $p \in (0, 1)$. Let m_0 and m_1 denote respectively $m(0)$ and $m(1)$, where clearly $m_1 > m_0$. Since $V(p; m) > m$ if $m \leqq m_0$, so we must have $MV(p; m) > m$ for any such m. Similarly for $m \geqq m_1$, we must have $MV(p; m) < m$. By the continuity of $MV(p; \cdot)$ it follows that there exists $m^* \in (m_0, m_1)$ such that $MV(p; m^*) = m^*$. Pick any such m^* (if there is more than one), and consider $m' \in (m^*, m_1)$; thus $m' = \lambda m^* + (1 - \lambda)m_1$ for some $\lambda \in (0, 1)$. The convexity of MV in m implies

$$\begin{aligned} MV(p; m') &\leqq \lambda MV(p; m^*) + (1 - \lambda)MV(p; m_1) \\ &< \lambda m^* + (1 - \lambda)m_1 = m'. \end{aligned}$$

But this inequality shows that m^* must be unique; i.e. there can exist only one value of m^* satisfying $MV(p; m^*) = m^*$. That $m^* = m(p)$ is immediate, completing the proof.

Proof of Theorem 1. Let $p \in (0, 1]$. We show that for all $\lambda \in [0, 1)$, we have $m(p) > m(\lambda p)$. This establishes (1.1), and hence, Theorem 1. Since $MV(\cdot; m)$ is convex as a function of p, we have

$$\begin{aligned} MV(\lambda p; m(p)) &\leqq \lambda MV(p; m(p)) + (1 - \lambda)MV(0; m(p)) \\ &< \lambda m(p) + (1 - \lambda)m(p) \\ &= m(p), \end{aligned}$$

where the strict inequality obtains by Lemma 5. But, this string of inequalities implies, again by Lemma 5, that $m(\lambda p) < m(p)$. Thus, Equation (1.1) holds.

3. A Bernoulli example

We consider in this section a special case of the n-armed bandit problem outlined above, where the space of rewards is given by $\{0, 1\}$, and the (Bernoulli) reward distributions are specified by $q_k = \Pr\{r = 1 : F = F_k\}$, $1 - q_k = \Pr\{r = 0 : F = F_k\}$, with $q_k \in [0, 1]$, $k = 1, 2$, and $q_1 > q_2$. We make the additional assumptions that $q_1 = 1 - q_2$, so $q_1 > 1/2 > q_2$; and that all arms are *a priori* identical to the principal, so the initial prior is $p = (\pi, \cdots, \pi)$ for some $\pi \in (0, 1)$.

We show that the solution to this problem given by Theorem 1 carries some strong implications. Namely, (i) any time a previously discarded arm is chosen again, 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 that (or any other) arm; (ii) the distribution of an arm's continued use follows a *random walk*, and (iii) the expected duration of an arm each time it is chosen depends *only* on its 'true' type and is independent of the entire past use of that arm, as well as its current prior.

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. As a first step, note that under our

assumptions the posterior belief on an arm which has generated α 1's and β 0's is a function only of the prior belief and the *difference* $\alpha - \beta$; in particular, this posterior is the *same* as the one resulting from observing $\alpha - \beta$ 1's and no 0's (or $\beta - \alpha$ 0's and no 1's) whenever $\alpha \geqq \beta$ (or $\beta \geqq \alpha$). Combining this observation with Theorem 1, we see 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 nth 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. Since this process is independent of the initial prior π, the entire procedure now repeats itself. This proves (i). It also implies that the 'survival' probability distribution of an arm each time it is newly selected is identical to the distribution the first time it was chosen.

To see (ii), consider a newly selected arm as starting at the position 1 on the real line. If a reward of 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, i.e. 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 1, while (ii) if $q \geqq 1/2$, this probability is $(1-q)/q$. Further, (iii) if $q < 1/2$, the expected first-passage time is $1/(1-2q)$, while (iv) if $q \geqq 1/2$ this is evidently infinite. Therefore, all arms whose true distributions are F_2 will with probability 1 be replaced each time they are chosen, with an expected duration of continuous play equal to $1/(1-2q_2)$ each time they are newly chosen. Analogous statements hold for type F_1 arms, except that with positive probability such an arm will *never* be replaced. This demonstrates (iii).

Appendix: Proof of Lemma 4

Fix $m \in [\Gamma_2/(1-\delta), \Gamma_1/(1-\delta)]$, 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 (i.e. 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

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

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 $V(\cdot, m)$ given in (2.1).

We show that there is a convex function $w^* \in C(I, \mathbb{R})$ such that $Tw^* = w^*$. By uniqueness of the fixed point this establishes $V(\cdot, m)$ is convex. The arguments here largely follow McLennan (1988).

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

$$Hw(p) = \int w(\beta(p, r)) f(p)(r) dr,$$

$$Mw(p) = \Gamma(p) + \delta Hw(p).$$

Then, of course, $Tw(p) = \max\{m, Mw(p)\}$. 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 $\Gamma(\cdot)$, 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 $r \in R$ define $\varepsilon(r)$ by $(1-\varepsilon(r)) f(p)(r) = (1-\lambda) f(p^1)(r)$ (or, equivalently, $\varepsilon(r) f(p)(r) = \lambda f(p^2)(r)$). Note that

$$\beta(p, r) = (1-\varepsilon(r))\beta(p^1, r) + \varepsilon(r)\beta(p^2, r).$$

Since w is convex,

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

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

Now let $\boldsymbol{W}$ be the set of all convex $w \in C(I, \mathbb{R})$ such that $w \leqq Tw$. Since $\Gamma(\cdot)$ is bounded, $\boldsymbol{W}$ is non-empty and bounded above. Define w^* by $w^*(p) = \sup_{w \in \boldsymbol{W}} w(p)$. Then, clearly w^* is convex and $w^* \leqq Tw^*$. But T is also a monotone operator (i.e. $v \leqq w$ implies $Tv \leqq Tw$), so $Tw^* \leqq T(Tw^*)$. Now from the above arguments w^* convex implies Tw^* convex, so $Tw^* \in \boldsymbol{W}$ as well. Thus, by the definition of w^*, $Tw^* \leqq w^*$, implying $Tw^* = w^*$. Since $V(\cdot, m)$ is the *unique* fixed point of T, $V(\cdot, m) = w^*$, so $V(\cdot, m)$ is convex on I.

References

BERRY, D. AND FRISTEDT, B. (1985) *Bandit Problems: Sequential Allocation of Experiments*. Chapman and Hall, London.

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

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

FRISTEDT, B. AND BERRY, D. (1988) Optimality of myopic stopping times for geometric discounting. *J. Appl. Prob.* **25**, 437–443.

GITTINS, J. AND JONES, D. (1974) A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics*, ed. J. Gani et al., pp. 241–266. North-Holland, Amsterdam.

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. *J. Appl. Prob.* **26**, 655–658.

RODMAN, L. (1978) On the many-armed bandit problem. *Ann. Prob.* **6**, 491–498.

ROSS, S. (1983) *Introduction to Dynamic Programming*. Academic Press, New York.