Probability in the Engineering and Informational Sciences, **10**, 1995, 21–28. Printed in the U.S.A.

POLICIES WITHOUT MEMORY FOR THE INFINITE-ARMED BERNOULLI BANDIT UNDER THE AVERAGE-REWARD CRITERION

STEPHEN J. HERSCHKORN
School of Business and RUTCOR
Rutgers University
New Brunswick, New Jersey 08903

EROL PEKÖZ AND SHELDON M. ROSS*
Department of Industrial Engineering and Operations Research
University of California
Berkeley, California 94720

We consider a bandit problem with infinitely many Bernoulli arms whose unknown parameters are i.i.d. We present two policies that maximize the almost sure average reward over an infinite horizon. Neither policy ever returns to a previously observed arm after switching to a new one or retains information from discarded arms, and runs of failures indicate the selection of a new arm. The first policy is nonstationary and requires no information about the distribution of the Bernoulli parameter. The second is stationary and requires only partial information; its optimality is established via renewal theory. We also develop ε-optimal stationary policies that require no information about the distribution of the unknown parameter and discuss universally optimal stationary policies.

1. INTRODUCTION

Consider the following bandit problem: There are infinitely many Bernoulli arms having unknown parameters that are i.i.d. (on [0,1]). The objective is to

*Supported by the National Science Foundation under grant DMS-9401834.

select arms so that we maximize the proportion of successes over an infinite horizon. In particular, we are interested in policies without recall, that is, where each switch is to a previously unobserved arm and, furthermore, policies that require no information from the observations of discarded arms.

We establish the optimality of two such policies. In both policies, runs of successive failures indicate switching to a new arm. In the first policy, one pulls the *i*th arm until *i* failures in a row are observed; this policy is optimal regardless of the common distribution of the Bernoulli parameters. In the second, one pulls an arm until *N* failures in a row are observed, where *N* is a (possibly degenerate) random variable generated for each arm. Properties of the distribution of the Bernoulli parameter determine the distribution of *N*.

This problem is a special case of one considered by Mallows and Robbins [2]. In their work, rewards may come from general, unknown, nonidentical distributions (as opposed to Bernoulli rewards). With the assumption of uniformly bounded *r*th moments of the rewards for some $r > 1$, the authors presented a policy that obtains almost surely the supremal average reward; this policy requires no other information about the unknown distributions. Yakowitz and Lowe [4] considered the problem of general, positive-mean losses on an infinite-armed bandit. Under the assumption of a certain uniform probability bound, these authors derived a policy that guarantees zero as the expected long-run proportion of selections of suboptimal arms. Berry, Chen, Heath, Shepp, and Zame [1] consider the special case of our problem where the Bernoulli parameters are independent and uniformly distributed on the unit interval. Among other results, they show the optimality of switching to a previously unobserved arm immediately following the observation of a single failure.

The policies of Mallows and Robbins return to previously observed arms infinitely often; that of Yakowitz and Lowe pulls *each* arm infinitely often. In contradistinction, as the policy of Berry et al., each of the policies in the present paper never return to an arm once we have switched to a new one. Our restriction has the advantage that we need not retain any information about earlier arms. Furthermore, in the setting of i.i.d. Bernoulli parameters considered here, such a restriction has a certain appeal: In any realization of the problem, we have previously decided that we prefer a new unknown arm to each discarded arm already observed; because the horizon remains infinite and an infinite number of unknown arms remain, we have no reason to change our evaluation. Moreover, for each arm we discard, we will always eventually find an arm at least as good. However, the cost of the restriction of no recall lies in the rate of convergence to optimum (see, e.g., Berry et al. [1]).

We present the two optimal policies in Sections 2 and 4. In Section 3, we develop a sequence of stationary policies whose returns converge monotonically to optimum, and we discuss a necessary condition for stationary policies that are optimal for all distributions.

Throughout this article, θ represents generically the unknown Bernoulli parameter for any given arm (although the parameter for each arm is distinct and independent of the others); $v \equiv 1 - \theta$, and c denotes the essential supremum of θ, that is, $c \equiv \inf\{t : P\{\theta \le t\} = 1\}$. Thus, a policy is optimal if and only if the long-run average reward is almost surely c. (The strong law of large numbers implies that the average reward cannot exceed c.) We assume throughout that c is strictly positive.

2. A NONSTATIONARY POLICY

THEOREM 1: *The policy where we pull the ith arm until we observe i failures in a row is optimal (regardless of the distribution of θ).*

PROOF OF THEOREM 1: Fix a and r such that $1 - c < ra < a < 1$. Define an arm as *good* if $v \le ra$, *bad* if $v \ge a$, or *neutral* if $ra < v < a$. Label a trial as good, bad, or neutral in accordance with the arm pulled. We shall first show that the proportion of bad pulls is almost surely 0.

Say a *cycle* begins with each first selection of a good arm, and let G_m and B_m be the respective numbers of good and bad trials in the mth cycle ($m \ge 1$).

LEMMA: *Under the policy of Theorem* 1, *for any positive* ε,

$$P\{B_m \ge \varepsilon G_m \textit{ for infinitely many } m\} = 0.$$

PROOF OF THE LEMMA: The initial arm in a cycle is the only good arm pulled in that cycle. It is pulled until we have observed a certain number of consecutive failures thereon; call this number T_m. Heading for application of the first Borel–Cantelli lemma, we bound probabilities by conditioning:

$$\begin{aligned} P(B_m \ge \varepsilon G_m \mid T_m, T_{m+1}) &= E[P(G_m \le B_m/\varepsilon \mid B_m, T_m, T_{m+1}) \mid T_m, T_{m+1}] \\ &= E\left[\sum_{i=1}^{B_m/\varepsilon} P(G_m = i \mid T_m) \mid T_m, T_{m+1}\right] \\ &\le E[(ra)^{T_m} B_m/\varepsilon \mid T_m, T_{m+1}] \\ &\le \frac{1}{\varepsilon(1-a)} (ra)^{T_m} a^{-T_{m+1}} Z_m, \end{aligned}$$

where $Z_m \equiv T_{m+1} - T_m$. The last inequality arises from the fact that, given T_m and T_{m+1}, B_m is the sum of $Z_m - 1$ random variables, each with conditional mean no greater than $\sum_{n=1}^{T_{m+1}} a^{-n} < a^{-T_{m+1}}/(1-a)$.

Because $T_m \ge m$, it follows that

$$P(B_m \ge \varepsilon G_m \mid Z_m) \le \frac{1}{\varepsilon(1-a)} r^m Z_m a^{-Z_m}.$$

Note that, because Z_m is the number of *arms* in the mth cycle, the random variables $Z_1, Z_2, \ldots$ are i.i.d. with common geometric($P\{v \le ra\}$) distribution.

Now note that for any positive $\beta < 1$

$$\begin{aligned}\sum_{m=1}^{\infty} P\{B_m \geq \varepsilon G_m\} &= \sum_{m=1}^{\infty} [P\{B_m \geq \varepsilon G_m, Z_m > \beta m\} \\ &\qquad + P\{B_m \geq \varepsilon G_m, Z_m \leq \beta m\}] \\ &\leq \sum_{m=1}^{\infty} P\{Z_m > \beta m\} + \sum_{m=1}^{\infty} P(B_m \geq \varepsilon G_m \mid Z_m \leq \beta m) \\ &\leq \sum_{m=1}^{\infty} (P\{v > ra\})^{\beta m} + \frac{1}{\varepsilon(1-a)} \sum_{m=1}^{\infty} r^m \beta m a^{-\beta m}.\end{aligned}$$

If we choose β such that $ra^{-\beta} < 1$, then

$$\sum_{m=1}^{\infty} P\{B_m \geq \varepsilon G_m\} < \infty,$$

whence the lemma follows from the first Borel–Cantelli lemma. ■

Returning to the proof of Theorem 1, let R_n be the cumulative number of successes on good or neutral arms by the nth pull, and let M_n be the number of trials on such arms by the nth pull. Then, the total average number of successes is at least as large as

$$\liminf_{n\to\infty} \frac{R_n}{n} \geq \liminf_{n\to\infty} \frac{R_n}{M_n} \cdot \liminf_{n\to\infty} \frac{M_n}{n} \geq 1 - a.$$

The latter inequality holds because the strong law of large numbers implies, via a coupling argument, that $\liminf_{n\to\infty} (R_n/M_n) \geq 1 - a$ and because the lemma implies that the second limit is one. The optimality of the policy follows because a is arbitrary. ■

3. STATIONARY POLICIES FOR AN UNKNOWN DISTRIBUTION

The policy in the preceding section is optimal for any distribution on the unknown parameter, but it is nonstationary. That is, the switching rule changes for each arm. The next theorem yields a sequence of stationary policies that converge to optimal in a monotonic fashion.

THEOREM 2: *Let w_k be the almost sure average reward from the policy that stays with each arm until k successive failures have been observed and switches to a previously unobserved arm thereafter. Then, w_k increases to c as k grows arbitrarily large.*

PROOF: We first show that w_k is nondecreasing in k. For each k, define a renewal reward process by positing a unit reward for each success and saying a renewal occurs when we switch to a new arm. In this process, the expected cycle length is $E\sum_{i=1}^{k} v^{-i}$ (see, e.g., Ross [3, pp. 231–232]). We may condition on θ and apply Wald's equation to obtain the expected reward in a cycle as $E[\theta\sum_{i=1}^{k} v^{-i}] = E[v^{-k}] - 1$. Thus, by the renewal reward theorem,

$$w_k = (E[v^{-k}] - 1)/E\sum_{i=1}^{k} v^{-i};$$

from Theorem 3, later, we may assume without loss of generality that the numerator and denominator are finite. Thus,

$$\begin{aligned} w_k \le w_{k+1} &\Leftrightarrow \sum_{i=0}^{k-1} E[v^{-k}]E[v^{-(i+1)}] \le \sum_{i=0}^{k-1} E[v^{-(k+1)}]E[v^{-i}] \\ &\Leftarrow E[v^{-(i+1)}]^2 \le E[v^{-i}]E[v^{-(i+2)}] \qquad \text{for all } i; \end{aligned}$$

the last relation is the Cauchy–Schwarz inequality. We have thus established monotonicity.

To show convergence, fix positive a and b such that $a < b < c$. Now define a renewal reward process where we receive a unit reward for each trial on an arm with $\theta \le a$; switches remain renewals. The average reward in this case is

$$\begin{aligned} &\frac{E\left[\sum_{i=1}^{k} v^{-i} \,\middle|\, \theta \le a\right] P\{\theta \le a\}}{E\sum_{i=1}^{k} v^{-i}} \\ &\le \frac{\sum_{i=1}^{k} (1-a)^{-i} P\{\theta \le a\}}{E\left[\sum_{i=1}^{k} v^{-i} \,\middle|\, \theta \le b\right] P\{\theta \le b\} + \sum_{i=1}^{k} (1-b)^{-i} P\{\theta > b\}} \\ &\le \frac{b}{a}\,\frac{(1-a)^{-k} - 1}{(1-b)^{-k} - 1}\Bigg/ P\{\theta > b\}, \end{aligned}$$

which approaches 0 as k grows arbitrarily large.

Let r_k be the proportion of trials on arms with $\theta > a$. As at the end of the proof of Theorem 1, $w_k \ge r_k a$. Therefore, $\lim_{k\to\infty} w_k \ge a \lim_{k\to\infty} r_k = a$ by the preceding paragraph. Because a was arbitrary, w_k converges to c. ■

The existence of a stationary optimal policy remains an open question. We note, however, the following necessary condition for any such policy: The conditional expected number of trials on an arm given θ must be infinite for any positive θ. For the policy must be optimal in particular for two-point distributions, supported by a and c, say, where $0 < a < c \le 1$. As before, the fraction of trials on arms with $\theta = a$ for such a policy is

$$\frac{E[\text{number of trials} \mid \theta = a]P\{\theta = a\}}{E[\text{number of trials} \mid \theta = a]P\{\theta = a\} + E[\text{number of trials} \mid \theta = c]P\{\theta = c\}},$$

where we refer to the number of trials on a single arm. For this expression to be zero when $P\{\theta = a\} > 0$, the conditionally expected number of trials must be infinite given $\theta = c$. Thus, infinite conditional expectation is the cost of simultaneous generality and stationarity.

4. OPTIMAL POLICIES FOR A KNOWN DISTRIBUTION

When the distribution of the Bernoulli parameter is known, we may exploit this information to construct a stationary optimal policy. Let $\xi = (1 - (\theta/c))^{-1}$.

THEOREM 3: *Let ζ be the probability generating function of a distribution on the positive integers such that $\zeta(x)$ is finite for all real x and $E\zeta(\xi)$ is infinite. The following policy is optimal: For each arm, independently generate a random variable M with generating function ζ; then generate N with Pascal(M, c) distribution. Pull the arm until we observe N failures in a row; then switch to a previously unobserved arm.*

PROOF: Let a be any positive number strictly less than c. As before, we shall show that the fraction of trials on arms with a parameter no greater than a is zero.

Define a renewal reward process by positing a unit reward for each pull of such an arm; say a renewal occurs each time we start pulling a new arm. As in the proof of Theorem 2,

$$E[\text{cycle length} \mid \theta, N] = \sum_{i=1}^{N} v^{-i} \geq v^{-N},$$

whence

$$\begin{aligned} E[\text{cycle length}] &\geq EE[v^{-N} \mid \theta, M] \\ &= E\zeta\left(\frac{c/v}{1 - (1-c)/v}\right) = E\zeta(\xi) = \infty. \end{aligned}$$

Also,

$$\begin{aligned} E[\text{reward in cycle}] &= E[\text{number of pulls} \mid \theta \leq a] P\{\theta \leq a\} \\ &\leq E \sum_{i=1}^{N} (1-a)^{-i} \\ &= E\left[\frac{(1-a)^{-1}[(1-a)^{-N} - 1]}{(1-a)^{-1} - 1}\right] \\ &= \frac{\zeta(c/(c-a)) - 1}{a} < \infty. \end{aligned}$$

By the renewal reward theorem, the average reward per unit time is almost surely zero. ■

Note that if $E\xi^k$ is infinite for some integer k we may set $M = k$ degenerately. If $Ee^{s\xi}$ is infinite for some s, we may let $M - 1$ have the Poisson distribution with parameter s. More generally, we have the following.

PROPOSITION: *The generating function of each of the following distributions satisfies the hypothesis of Theorem* 3:

(a) $P\{M > m\} = \min(\alpha/E\xi^m, 1)$ *for any fixed positive* α, *and*
(b) $P\{M = m\} \propto \alpha^m/E\xi^{m-1}$ *for any fixed* $\alpha \geq 1$.

PROOF: If some moment of ξ is infinite, the statement is easy to verify; thus, assume all moments are finite. With distribution (a) and $x > 1$,

$$\begin{aligned}\zeta(x) &= \int_0^\infty P\{x^M > u\}\,du \\ &= \log x\left(\int_{-\infty}^1 x^v\,dv + \int_1^\infty P\{M > v\}x^v\,dv\right).\end{aligned}$$

The first integral is finite, and the second integral does not exceed $\alpha x \sum_{m=1}^\infty (x^m/E\xi^m)$. The ratio test confirms the convergence of the series: by Jensen's inequality,

$$\frac{E\xi^{m+1}}{E\xi^m} \geq (E\xi^m)^{1/m},$$

and the right-hand side grows arbitrarily large because ξ is essentially unbounded. Thus, ζ is finite everywhere. Also,

$$\begin{aligned}E\zeta(\xi) &\geq E\int_1^\infty P\{M > v\}\xi^v 1_{\{\xi \geq e\}}\,dv \\ &\geq \sum_{m=1}^\infty P\{M > m\}E[\xi^m 1_{\{\xi \geq e\}}] = \infty.\end{aligned}$$

Similar techniques verify that distribution (b) satisfies the desired conditions.[1] ■

Regarding this optimal policy, note that if $c = 1$, then $N = M$ and there is no need to generate a second random variable. Thus, if $c = 1$ and $E\xi^k$ is infinite for some k, we have a deterministic optimal policy, viz., to pull each arm until we observe k failures in a row. In particular, the policy of Berry et al. (see our introduction) is a special case of the one in Theorem 3. When the Bernoulli parameter is nondegenerate and bounded away from one (i.e., $c < 1$), no fixed, nonrandom number of failures will work. The renewal-theoretic approach reveals this fact: with $c < 1$ and N degenerate, the expected cycle time (number of pulls of an arm) is finite and the expected number of successes in a cycle is strictly less than c times this expected cycle time.

We can generalize the policy of Theorem 3 to bandits with general rewards. With positive probability, discussion of this topic will appear in a future article.

Note

[1] We thank A. de Acosta for pointing out that the tail $1/E\xi^m$ is a solution to this distributional problem.

References

1. Berry, D., Chen, R., Heath, D., Shepp, L., & Zame, A. (in preparation). A bandit problem with infinitely many arms.
2. Mallows, C.L. & Robbins, H. (1964). Some problems of optimal sampling strategy. *Journal of Mathematical Analysis and Applications* 8: 90–103.
3. Ross, S.M. (1983). *Stochastic processes.* New York: John Wiley.
4. Yakowitz, S. & Lowe, W. (1991). Nonparametric bandit methods. *Annals of Operations Research* 28: 297–312.