J. Appl. Prob. **40,** 250–256 (2003)
Printed in Israel

SOME MEMORYLESS BANDIT POLICIES

EROL A. PEKÖZ,* *Boston University*

Abstract

We consider a multiarmed bandit problem, where each arm when pulled generates independent and identically distributed nonnegative rewards according to some unknown distribution. The goal is to maximize the long-run average reward per pull with the restriction that any previously learned information is forgotten whenever a switch between arms is made. We present several policies and a peculiarity surrounding them.

Keywords: Bandit problem; multiarmed bandit problem; average reward criteria

AMS 2000 Subject Classification: Primary 62L05; 90B36; 62L15

1. Introduction

In this paper we consider a (possibly infinite) set of arms S so that arm $i \in S$ when pulled generates independent and identically distributed (i.i.d.) nonnegative rewards according to a general unknown distribution denoted generically by the random variable X_i. The goal is to maximize the long-run average reward per pull with the restriction that whenever there is a switch from one arm to another arm, any information learned about the arms so far is forgotten. Thus, any decision on when to switch arms must be based entirely on what has occurred with the arm currently in use.

This type of problem is referred to in the literature as a 'bandit' problem, due to the analogy with playing a slot machine or 'one-arm bandit'. Initially introduced by Bellman in [1], the multiarmed bandit has received much research attention over the years. Gittins' celebrated index theorem (see [6]) gives an approach for maximizing expected discounted reward when information about previously pulled arms can be remembered. Policies for maximizing long-run average reward typically (see [2] and [8], and the references within) advocate pulling arms to learn about their corresponding reward distributions, and then returning to the best arms most often.

Our problem here without the ability to recall previously pulled arms has been studied in the special case where rewards are binary. In [7] it is shown that the policy of pulling the ith arm until there are i pulls in a row without reward is optimal. Under this policy, arms with a higher chance of reward get pulled more times on average before being discarded. In [3] policies are studied which perform well over finite time horizons, including pulling the same arm until a fixed number of failures is observed or until the failure rate exceeds a fixed value. The paper [9] studies the setting where there is a known prior distribution on the parameter for the arms. None of these policies previously studied, however, seem adaptable to the setting of general rewards. This paper considers the setting of general nonnegative rewards, and gives new optimal policies. Our interest in this problem stems from an application to maintaining a cache of Web pages on a proxy server, where there is no space to store information about pages not resident in the cache (see [10]).

Received 27 June 2002; revision received 13 November 2002.
* Postal address: School of Management, Boston University, Boston, MA 02215, USA. Email address: pekoz@bu.edu

It may be interesting to note that, in the case of binary rewards with two types of arms, the policy of switching to a new randomly chosen arm after the first pull without reward (or after some fixed number of such pulls) does not perform well. Though superior arms get pulled the majority of the time, so much time is wasted pulling the inferior arms that the average reward per pull in the long run will be below the maximum possible. In an optimal policy, it intuitively seems as though the time spent pulling inferior arms must be negligible compared with the time spent pulling superior arms.

In Section 2 we give some solutions to this problem and an associated peculiarity, and in Section 3 we give the proofs.

2. Main results

For each arm $i \in S$, let $\mu_i = \mathrm{E}[X_i]$ and $\sigma_i^2 = \mathrm{var}(X_i)$ denote the mean and variance of its corresponding reward distribution, and we assume that these are uniformly bounded. Let I be a random variable having some given probability distribution π on S (if S has a finite number of elements, then we can assume π to be a uniform distribution).

We say that any policy which achieves a long-run average reward per pull equal to the essential supremum

$$\mu^* := \inf\{t : \mathrm{P}(\mu_I \leq t) = 1\}$$

is optimal, and any policy which achieves the essential infimum

$$\mu_* := \sup\{t : \mathrm{P}(\mu_I \geq t) = 1\}$$

is the worst possible. In this paper, we consider the following two types of policies, where $\lfloor x \rfloor$ denotes the integer portion of x.

Policy A. For some given function f and a given value of t, pull a randomly selected arm (selected according to the distribution π) t times and compute $T = \prod_{j=1}^{t} f(R_j)$, where R_j is the reward obtained during the jth pull. Then pull the same arm an additional $\lfloor T \rfloor$ times, discard the arm, and start over.

Policy B. The same as Policy A except each time that an arm is discarded and another arm selected, the value of t is increased by 1.

Policy B is a natural nonstationary implementation of Policy A. Our main result shows that either one of these two policies can be be optimal for every set of arms, while at the same time the other may be the worst possible for some set of arms. But this paradoxical relation does not hold if the unknown distributions can be stochastically ordered. These possibly surprising results are summarized in the theorems below.

Let r_t^{A} be the long-run average reward per pull under Policy A using a fixed t, and let r^{B} be the corresponding long-run average reward under Policy B. These averages may not necessarily always have long-run limits, and so we define these only when the limits exist.

Theorem 2.1. (a) *When $f(x) = 1+x$, we always have $\lim_{t\to\infty} r_t^{\mathrm{A}} = \mu^*$, but it is also possible that $r^{\mathrm{B}} = \mu_*$. In other words, for any set of arms, Policy A always performs arbitrarily close to optimal for sufficiently large t, but at the same time, for some set of arms, Policy B can be the worst possible.*

(b) *When $f(x) = \mathrm{e}^x$, we always have $r^{\mathrm{B}} = \mu^*$, but it is also possible that $\lim_{t\to\infty} r_t^{\mathrm{A}} = \mu_*$. In other words, for any set of arms, Policy B is always optimal, but at the same time, for some set of arms, Policy A can perform arbitrarily close to the worst possible for sufficiently large t.*

The next theorem states that these seemingly contradictory results do not occur if the distributions can be stochastically ordered.

Theorem 2.2. *Suppose that, for all $i, j \in S$,*

$$\mu_i \geq \mu_j \implies X_i \geq_{\mathrm{st}} X_j. \tag{2.1}$$

Then, when either $f(x) = 1 + x$ or $f(x) = \mathrm{e}^x$ and, for all $i \in S$, $\mathrm{E}[f(X_i)] < \infty$, we always have $\lim_{t\to\infty} r_t^{\mathrm{A}} = r^{\mathrm{B}} = \mu^*$. In other words, Policy A always performs arbitrarily close to optimal for sufficiently large t and Policy B is always optimal.

Remark 2.1. The policies we study here may be reminiscent of solutions to the infamous 'pick the larger number' problem (see [4]). Suppose that there are two envelopes with different unknown positive numbers written in each. If allowed to open a random envelope, then, surprisingly, it is possible to have a better than 50% chance of guessing which envelope contains the larger number: pick any strictly increasing function f on $(0, \infty)$ which is bounded by 0 and 1; if the number in the opened envelope is x, then guess that it is the larger number with probability $f(x)$. This gives a strictly better than 50% chance of guessing correctly, even with no prior knowledge of the range of the numbers involved. The policies we study here use somewhat analogously an increasing function of the rewards to obtain an optimal policy even with no prior knowledge of the range of possible rewards.

3. The proofs

3.1. Proof of Theorem 2.1

First fix any b and g with $b < g < \mu^*$. We say that arm i is 'good' if $\mu_i \geq g$, 'bad' if $\mu_i < b$, and 'neutral' otherwise. We prove the first part of (a) by showing that the almost-sure long-run fraction of time that a bad arm is pulled approaches zero as $t \to \infty$.

We say that a 'cycle' starts whenever a good arm is selected. Let G, B, and N be, respectively, the number of times that a good, bad, and neutral arm is pulled during a generic cycle, and let I denote the index of an arm randomly selected according to π. When $f(x) = 1 + x$,

$$\begin{aligned}\mathrm{E}[G] &\geq t + \mathrm{E}\bigg[\prod_{j=1}^{t}(1 + X_I^j) \,\bigg|\, \mu_I \geq g\bigg] - 1\\ &\geq t + (1+g)^t - 1,\end{aligned}$$

where X_i^j are i.i.d. random variables having distribution X_i, the reward distribution for arm i. We subtract 1 at the end to account for rounding T. In each cycle, the number of distinct non-good arms pulled follows a geometric distribution having parameter $p = \mathrm{P}(\mu_I \geq b)$, and then the same reasoning gives

$$\mathrm{E}[B] \leq \frac{1}{p}(t + 1 + (1+b)^t).$$

Thus, by the renewal-reward theorem, the almost sure fraction of time that a bad arm is pulled satisfies

$$\frac{\mathrm{E}[B]}{\mathrm{E}[G] + \mathrm{E}[B] + \mathrm{E}[N]} \leq \frac{\mathrm{E}[B]}{\mathrm{E}[G]} \leq \frac{t + 1 + (1+b)^t}{p(t - 1 + (1+g)^t)} \to 0$$

as $t \to \infty$.

Since good and neutral arms give mean reward at least b when pulled, conditional on the indices of arms pulled we can use a slight generalization of the strong law of large numbers to independent but non-identically distributed random variables satisfying bounded mean and variance conditions (see for example [5, Exercise 8.4, p. 69]) to get the almost sure limit $\liminf_{t\to\infty} r_t^{\mathrm{A}} \geq b$ for any $b < \mu^*$, and hence the result.

To show the second part of (b) where $f(x) = \mathrm{e}^x$, suppose that we have only two arms with $X_2 = 5$ almost surely, $\mathrm{P}(X_1 = 0) = \mathrm{P}(X_1 = 8) = \frac{1}{2}$, and $\mathrm{P}(I = 1) = \mathrm{P}(I = 2) = \frac{1}{2}$. Then clearly $\mathrm{E}[X_1] < \mathrm{E}[X_2]$ but $\mathrm{E}[\exp\{X_1\}] > \mathrm{E}[\exp\{X_2\}]$. A similar calculation to that above shows that

$$\mathrm{E}[G] \leq t + 1 + (\mathrm{E}[\exp\{X_2\}])^t$$

and

$$\mathrm{E}[B] \geq 2(t - 1 + (\mathrm{E}[\exp\{X_1\}])^t),$$

and so this time we have

$$\frac{\mathrm{E}[G]}{\mathrm{E}[G] + \mathrm{E}[B] + \mathrm{E}[N]} \to 0 \quad \text{as } t \to \infty.$$

This gives the second part of (b).

To show the first part of (b), we again say that a cycle starts whenever a good arm is chosen and then let B_n and G_n denote, respectively, the number of times that a bad arm and a good arm is pulled under Policy B during a cycle which starts with the value $t = n$. The following lemma along with the last paragraph of the argument for the first part of (a) will establish the result.

Lemma 3.1. *For any $\varepsilon > 0$, under Policy B with $f(x) = \mathrm{e}^x$ we have*

$$\mathrm{P}(B_n \geq \varepsilon G_n \textit{ for infinitely many values } n) = 0.$$

Proof. In some cycle with $t = n$, let

$$\begin{aligned} R_{ij}^B &= \text{the reward from the } i\text{th pull of the } j\text{th bad arm selected,} \\ R_i^G &= \text{the reward from the } i\text{th pull of the good arm,} \\ Z &= \text{the number of non-good arms selected during the cycle.} \end{aligned}$$

During this cycle, the good arm is initially pulled n times, and thus

$$G_n \geq n + \exp\left\{\sum_{i=1}^{n} R_i^G\right\} - 1,$$

where we subtract 1 to account for rounding. Since there are at most Z bad arms used during the cycle and each is initially pulled at most $n + Z$ times, we also have

$$B_n \leq \sum_{j=1}^{Z}\left(n + Z + \exp\left\{\sum_{i=1}^{n+Z} R_{ij}^B\right\}\right) + 1.$$

Next choose δ sufficiently small so that $b - g + \delta b < -\delta$ and suppose that n^* is chosen sufficiently large so that $\delta n^* > \delta n^*/2 - \ln\varepsilon + \ln(\delta n^*)$. Continuing now with any $n > n^*$, define the event $A_n = \{1 + Z(n + Z) \le \delta^2 n^2/2,\ Z \le \delta n\}$. Then

$$\begin{aligned}
\mathrm{P}(B_n \ge \varepsilon G_n) &\le \mathrm{P}\biggl(1 + \sum_{j=1}^{Z}\biggl(n + Z + \exp\biggl\{\sum_{i=1}^{n+Z} R_{ij}^B\biggr\}\biggr) \ge \varepsilon \exp\biggl\{\sum_{i=1}^{n} R_i^G\biggr\}\biggr)\\
&\le \mathrm{P}\biggl(\frac{\delta^2 n^2}{2} + \sum_{j=1}^{Z}\exp\biggl\{\sum_{i=1}^{n+n\delta} R_{ij}^B\biggr\} \ge \varepsilon \exp\biggl\{\sum_{i=1}^{n} R_i^G\biggr\}, A_n\biggr) + \mathrm{P}(A_n)\\
&\le \mathrm{E}[Z]\,\mathrm{P}\biggl(\frac{\delta n}{2} + \exp\biggl\{\sum_{i=1}^{n+n\delta} R_{i1}^B\biggr\} \ge \frac{\varepsilon}{\delta n}\exp\biggl\{\sum_{i=1}^{n} R_i^G\biggr\}\biggr) + \mathrm{P}(A_n),
\end{aligned}$$

where the last inequality uses the fact that, when $X_1, \dots, X_n$ are i.i.d., $\mathrm{P}(\sum_{i=1}^n X_i \ge t) \le n\,\mathrm{P}(X_1 \ge t/n)$. Continuing by taking natural logarithms and using the fact that $\ln(x + \mathrm{e}^y) \le x + y$ for $x, y \ge 0$, we have

$$\begin{aligned}
\mathrm{P}(B_n \ge \varepsilon G_n) &\le \mathrm{E}[Z]\,\mathrm{P}\biggl(\frac{\delta n}{2} - \ln\varepsilon + \ln(\delta n) + \sum_{i=1}^{n+n\delta} R_{i1}^B \ge \sum_{i=1}^{n} R_i^G\biggr) + \mathrm{P}(A_n)\\
&\le \mathrm{E}[Z]\,\mathrm{P}\biggl(\delta n + \sum_{i=1}^{n+n\delta} R_{i1}^B \ge \sum_{i=1}^{n} R_i^G\biggr) + \mathrm{P}(A_n)\\
&= \mathrm{E}[Z]\,\mathrm{P}(x_n + y_n \ge -\delta) + \mathrm{P}(A_n),
\end{aligned}$$

where the second inequality uses the definition of n^*. Here,

$$x_n = \frac{1}{n}\sum_{i=1}^{n}(R_{i1}^B - R_i^G) \quad \text{and} \quad y_n = \frac{\delta}{\delta n}\sum_{i=n+1}^{n+\delta n} R_{i1}^B.$$

Since, by the strong law of large numbers, $\lim_{n\to\infty} x_n \le b - g$ and $\lim_{n\to\infty} y_n \le \delta b$ almost surely, we can use the definition of δ to show that

$$\lim_{n\to\infty}(x_n + y_n) < -\delta \quad \text{almost surely,}$$

and so

$$\sum_{n>n^*} \mathrm{P}(x_n + y_n \ge -\delta) < \infty,$$

and the lemma follows using the Borel–Cantelli lemma along with the fact that finiteness of the moments of the geometric random variable Z gives

$$\sum_{n>n^*} \mathrm{P}(A_n) < \infty.$$

Finally, to prove the second part of (a), suppose that we have only two arms with $X_2 = 3$ almost surely, $\mathrm{P}(X_1 = 0) = \mathrm{P}(X_1 = 8) = \frac{1}{2}$, and $\mathrm{P}(I = 1) = \mathrm{P}(I = 2) = \frac{1}{2}$. Then clearly $\mathrm{E}[X_1] > \mathrm{E}[X_2]$ but

$$b := \mathrm{E}[\ln(1 + X_1)] < g := \mathrm{E}[\ln(1 + X_2)].$$

Here, arm 1 is obviously the better arm, but we label it as the 'bad' arm for the purposes of the argument which follows. Using the same definitions as before, we will establish the second part of (a) by showing that

$$\mathrm{P}(B_n \geq \varepsilon G_n \text{ for infinitely many values } n) = 0.$$

To see why this claim is true, view the policy with $f(x) = 1 + x$ as the equivalent one where instead $f(x) = \mathrm{e}^x$ and the number of times that an arm is pulled equals $T = \prod_{j=1}^{t} f(\ln(1 + R_j))$, where R_j is the reward from the jth pull. Using the same reasoning as in Lemma 3.1 applied to the logarithm of the rewards, we have

$$\mathrm{P}(B_n \geq \varepsilon G_n) \leq \mathrm{E}[Z]\,\mathrm{P}(x_n + y_n \geq -\delta) + \mathrm{P}(A_n),$$

where

$$x_n = \frac{1}{n}\sum_{i=1}^{n}(\ln(1 + R_{i1}^{B}) - \ln(1 + R_i^{G})) \quad \text{and} \quad y_n = \frac{\delta}{\delta n}\sum_{i=n+1}^{n+\delta n}\ln(1 + R_{i1}^{B}).$$

Since, again by the strong law of large numbers,

$$\lim_{n\to\infty} x_n = b - g \quad \text{almost surely}$$

and

$$\lim_{n\to\infty} y_n = \delta b \quad \text{almost surely,}$$

we use the definition of δ to again obtain that

$$\sum_{n>n^*}\mathrm{P}(B_n \geq \varepsilon G_n) < \infty,$$

and hence the theorem is proved.

3.2. Proof of Theorem 2.2

Again fix any b and g with $b < g < \mu^*$. We say that arm i is 'good' if $\mu_i \geq g$, 'bad' if $\mu_i < b$, and 'neutral' otherwise. We say that a 'cycle' starts whenever a good arm is selected.

Let G, B, and N be, respectively, the number of times that a good, bad, and neutral arm is pulled during a generic cycle under Policy A. With $f(x) = \mathrm{e}^x$, the same argument as in the proof of Theorem 2.1 gives

$$\mathrm{E}[G] \geq t - 1 + (\mathrm{E}[\exp\{R^G\}])^t$$

and

$$\mathrm{E}[B] \leq \frac{1}{p}(t + 1 + (\mathrm{E}[\exp\{R^B\}])^t),$$

where R^G and R^B denote, respectively, the reward from a randomly selected good arm and a randomly selected bad arm. Using (2.1), we have $\mathrm{E}[\exp\{R^G\}] > \mathrm{E}[\exp\{R^B\}]$, and so $\mathrm{E}[B]/\mathrm{E}[G] \to 0$ as $t \to \infty$.

Next suppose that $f(x) = 1 + x$, and let B_n and G_n denote, respectively, the number of times that a bad arm and a good arm is pulled under Policy B during a cycle which starts with the value $t = n$. Let

$$b' = \sup_{i:\mu_i<b}\mathrm{E}[\ln(1 + X_i)] \quad \text{and} \quad g' = \inf_{i:\mu_i\geq g}\mathrm{E}[\ln(1 + X_i)].$$

The condition (2.1) ensures that $g' > b'$, and so we can choose δ sufficiently small so that $b' - g' + \delta b' < -\delta$ and pick n^* as before. Then, for $n > n^*$, we can use the same reasoning as above to show that

$$\mathrm{P}(B_n \geq \varepsilon G_n) \leq \mathrm{E}[Z]\,\mathrm{P}(x_n + y_n \geq -\delta) + \mathrm{P}(A_n),$$

where

$$x_n = \frac{1}{n}\sum_{i=1}^{n}(\ln(1 + R_{i1}^B) - \ln(1 + R_i^G)) \quad \text{and} \quad y_n = \frac{\delta}{\delta n}\sum_{i=n+1}^{n+\delta n}\ln(1 + R_{i1}^B).$$

Since, again by the strong law of large numbers,

$$\lim_{n\to\infty} x_n \leq b' - g' \quad \text{almost surely}$$

and

$$\lim_{n\to\infty} y_n \leq \delta b' \quad \text{almost surely},$$

we use the definition of δ to again obtain that

$$\sum_{n>n^*}\mathrm{P}(B_n \geq \varepsilon G_n) < \infty,$$

and hence the result follows in the same fashion as in Theorem 2.1.

Acknowledgement

Thanks are due to the referee for several helpful comments.

References

[1] Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā* **16,** 221–229.

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

[3] Berry, D. A. *et al.* (1997). Bandit problems with infinitely many arms. *Ann. Statist.* **25,** 2103–2116.

[4] Cover, T. (1987). Pick the largest number. In *Open Problems in Communication and Computation*, eds T. Cover and B. Gopinath, Springer, New York, p. 152.

[5] Durrett, R. (1996). *Probability: Theory and Examples*, 2nd edn. Wadsworth Publishing, Belmont, CA.

[6] Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In *Progress in Statistics* (Europ. Meeting Statisticians, Budapest, 1972), Vol. 1, eds J. Gani, K. Sarkadi and I. Vincze, North-Holland, Amsterdam, pp. 241–266.

[7] Herschkorn, S. J., Peköz, E. and Ross, S. M. (1996). Policies without memory for the infinite-armed Bernoulli bandit under the average-reward criterion. *Prob. Eng. Inf. Sci.* **10,** 21–28.

[8] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **6,** 4–22.

[9] Lin C.-T. and Shiau C. J. (2000). Some optimal strategies for bandit problems with beta prior distributions. *Ann. Inst. Statist. Math.* **52,** 397–405.

[10] Psounis, K. and Prabhakar, B. (2001). A randomized web-cache replacement scheme. In *Proc. IEEE Infocom 2001* (Anchorage, AK, 22–26 April 2001), pp. 1407–1415. Available at http://www.ieee-infocom.org/2001/.