ADVANCES IN APPLIED MATHEMATICS 7, 199–204 (1986)

Multi-armed Bandits with Simple Arms*

ROBERT KEENER

University of Michigan, Ann Arbor, Michigan 48109

An exact solution to certain multi-armed bandit problems with independent and simple arms is presented. An arm is simple if the observations associated with the arm have one of two distributions conditional on the value of an unknown dichotomous parameter. This solution is obtained relating Gittins indices for the arms to ladder variables for associated random walks. © 1986 Academic Press, Inc.

1. INTRODUCTION

Bandit problems have received considerable attention in the statistical literature, in large part because the choice between taking actions for immediate reward and taking actions to acrue information must be addressed. Fundamental progress on bandit problems with independent arms was accomplished by Gittins and his co-authors in a series of papers (Gittins and Jones [4], Gittins and Galzenbrook [3] and Gittins [2]). They proved that it is possible to assign to each arm a dynamic allocation index such that the optimal strategy is to play the arm with the greatest index at each stage. In the sequel we will call this index Gittins index. It is related by their theorems to the solutions of a class of stopping problems.

The main result of this note is an explicit formula for the Gittins index of a *simple* arm. An arm is simple if the distribution of the associated observation X is governed by an unknown dichotomous parameter $\theta = 0$ or 1. Given $\theta = 0$, $X \sim P$ and given $\theta = 1$, $X \sim Q$. Theorem 2 relates the Gittins index of a simple arm to ladder variables for two associated random walks.

Bandit problems with two simple arms *both* governed by the same parameter θ (so the arms are highly dependent) have been studied by Feldman [1] and Keener [5].

*Research supported by NSF Grant MCS-8102080.

0196-8858/86 $7.50
Copyright © 1986 by Academic Press, Inc.
All rights of reproduction in any form reserved.

The specific set up is as follows: $\theta_1, \dots, \theta_k$ are independent Bernoulli variables representing the unknown parameters for the k arms. $A(n) \in \{1, \dots, k\}$ is the arm played at stage n. $X_1, X_2, \dots$ are the observations. Let $\mathscr{F}_n = \sigma\{A(1), \dots, A(n), X_1, \dots, X_n\}$. The process $\{A(n)\}_{n \geq 1}$ must satisfy the measurability constraint that $A(n+1)$ is $\mathscr{F}_n$ measurable. The distributions for the X_n are given recursively by

$$\mathscr{L}(X_{n+1} | \theta_1, \dots, \theta_k, \mathscr{F}_n) = (1 - \theta_{A(n+1)}) P_{A(n+1)} + \theta_{A(n+1)} Q_{A(n+1)}$$

for $n \geq 1$ ($\mathscr{F}_0$ is the trivial sigma field). Thus P_j and Q_j are the distributions for an observation when arm j is played and $\theta = 0$ and 1, respectively.

The control problem is to choose the process $\{A(n)\}$ to maximize

$$E \sum_{n=0}^{\infty} \beta^n R(\theta_{A(n+1)}, A(n+1))$$

where $R: \{0,1\} \times \{1, \dots, k\} \to \mathbf{R}$ is an arbitrary reward function and $\beta \in (0,1)$ is the discount factor. Let $\pi_j(n) = P(\theta_j = 1 | \mathscr{F}_n)$. The Gittins index for arm j at stage n is $G_j(\pi_j(n))$. It depends only on quantities related to arm j—the distributions P_j, Q_j, the two rewards $R(0, j)$ and $R(1, j)$, and the discount factor β. Gittins definition of G_j will be given in the next section. The following theorem is Gittins main result [4, Theorem 2] in this special case. See also Whittle [7] for a simpler proof.

THEOREM 1. *A policy which at each stage n chooses $A(n+1)$ equal to a value j which maximizes $G_j(\pi_j(n))$ is optimal.*

2. Stopping Problems for Simple Arms

From Gittins result, arms may be considered separately. For this section we will consider isolated arms. θ will be a Bernoulli parameter, and given θ, $X_1, X_2, \dots$ will be *i.i.d.* with

$$X_n | \theta = 0 \sim P$$

and

$$X_n | \theta = 1 \sim Q.$$

$\{\mathscr{F}_n\}_{n \geq 0}$ will be the filtration associated with the X_n, i.e., $\mathscr{F}_n = \sigma\{X_1, \dots, X_n\}$. Let $\pi = P(\theta = 1)$ and define

$$V(\pi, \beta, \Lambda) = \sup_{\tau} E\left\{ \Lambda \beta^{\tau} + \sum_{n=0}^{\tau-1} (r_0(1-\theta) + r_1\theta)\beta^n \right\}$$

where the supremum is over all extended stopping times with respect to the filtration $\{\mathscr{F}_n\}$. Gittins index can be defined in terms of the values for these stopping problems by

$$G(\pi) = (1 - \beta)\inf\{\Lambda : V(\pi, \beta, \Lambda) = \Lambda\}.$$

These stopping problems are closely related to a stopping problem studied by Lorden [6]. Let

$$R(\pi) = \inf_{\tau} E\left[c_0(1 - \theta)I\{\tau < \infty\} + c_1\theta\tau\right]$$

where c_0 and c_1 are positive constants and the infimum is over extended stopping times with respect to $\{\mathscr{F}_n\}$. Define

$$L = L(P, Q) = \exp - \sum_{n=1}^{\infty} \frac{1}{n}\left\{P\left(\frac{dP^n}{dQ^n} > 1|\theta = 1\right)\right.$$
$$\left. + P\left(\frac{dQ^n}{dP^n} \geq 1|\theta = 0\right)\right\}.$$

Here dP^n/dQ^n denotes the likelihood ratio,

$$\frac{dP^n}{dQ^n} = \prod_1^n \frac{dP}{dQ}(X_i).$$

If we let $\tilde{P} = \mathscr{L}\left(\log\frac{dP}{dQ}(X_1)|\theta = 0\right)$ and $\tilde{Q} = \mathscr{L}\left(\log\frac{dP}{dQ}(X_1)|\theta = 1\right)$, taking advantage of the *i.i.d.* structure we can write

$$L = \exp - \sum_{n=1}^{\infty} \frac{1}{n}\{\tilde{P}^{*n}([-\infty, 0]) + \tilde{Q}^{*n}((0, \infty])\}$$

where * denotes convolution. As noted in Lorden, $1/L$ is the product of expected ladder times for random walks generated by $\tilde{P}$ and $\tilde{Q}$. the following result is Lemma 1 of Lorden [6].

LEMMA 1. *$R(\pi) < c_0(1 - \pi)$ if and only if $(1 - \pi)c_0 L > \pi c_1$.*

A similar result holds for a discounted version of this problem. Let

$$R(\pi, \beta) = \inf_{\tau} E\left\{c_0(1 - \theta)\beta^{\tau} + c_1\theta \sum_{n=0}^{\tau-1} \beta^n\right\}$$

and

$$L(\beta) = L(\beta, P, Q) = \exp - \sum_{n=1}^{\infty} \frac{\beta^n}{n}\{\tilde{P}^{*n}([-\infty, 0]) + \tilde{Q}^{*n}((0, \infty])\}.$$

COROLLARY 1. *$R(\pi, \beta) < c_0(1 - \pi)$ if and only if $(1 - \pi)c_0 L(\beta) > \pi c_1$.*

Proof. Let $\pi_1 = P(\theta = 1 | X_1)$. From standard results in dynamic programming, $R(\pi)$ and $R(\pi, \beta)$ are unique solutions of the equations

$$R(\pi) = \min\{c_0(1 - \pi), c_1\pi + ER(\pi_1)\}$$

and

$$R(\pi, \beta) = \min\{c_0(1 - \pi), c_1\pi + \beta ER(\pi_1, \beta)\}.$$

Using Bayes theorem,

$$ER(\pi_1) = \int R\{\pi/[\pi + (1 - \pi)e^x]\}\{\pi d\tilde{Q}(x) + (1 - \pi)d\tilde{P}(x)\}.$$

Now, if we define $\hat{P} = \beta\tilde{P} + (1 - \beta)\delta_\infty$ and $\hat{Q} = \beta\tilde{Q} + (1 - \beta)\delta_{-\infty}$ where $\delta_{\pm\infty}$ are point measures on $\pm\infty$, then using $R(0, \beta) = R(1, \beta) = 0$, we can write

$$\beta ER(\pi_1, \beta) = \int R\{\pi/[\pi + (1 - \pi)e^x], \beta\}\{\pi d\hat{Q}(x) + (1 - \pi)d\hat{P}(x)\}.$$

Hence $R(\pi, \beta)$ satisfies the same defining equation as $R(\pi)$ with $\hat{P}$ replacing $\tilde{P}$ and $\hat{Q}$ replacing $\tilde{Q}$. The corollary now follows easily since $\hat{P}^{*n}([-\infty, 0]) = \beta^n\tilde{P}^{*n}([-\infty, 0])$ and $\hat{Q}^{*n}((0, \infty]) = \beta^n\tilde{Q}^{*n}((0, \infty])$. It is worth noting that $\hat{P}$ and $\hat{Q}$ are distributions for a log likelihood when at stage n, in addition to observing X_n, a variable Y_n is observed which is uninformative with probability β and is completely informative otherwise. From this observation a probabilistic proof of the corollary could be obtained.

The following result expresses Gittins index as a function of the prior π, the rewards r_1 and r_2, the discount factor β, and $L(\beta)$.

THEOREM 2. *If $r_0 \geq r_1$, the Gittins index is given by*

$$G(\pi) = \frac{r_1\pi(1 - \beta) + r_0(1 - \pi)L(\beta)}{\pi(1 - \beta) + (1 - \pi)L(\beta)}.$$

Proof. First note that if $(1 - \beta)\Lambda \geq r_0$ then $V(\pi, \beta, \Lambda) = \Lambda(\tau = 0$ is optimal) and if $(1 - \beta)\Lambda \leq r_1$, $V(\pi, \beta, \Lambda) = (\pi r_1 + (1 - \pi)r_0)/(1 - \beta) < \Lambda(\tau = \infty$ is optimal). Hence the Gittins index must lie in $[r_1, r_0]$ and we need only consider $\Lambda \in (r_1, r_0)/(1-\beta)$. After some algebra

$$E\left\{\Lambda\beta^{\tau} + \sum_{n=0}^{\tau-1}\left(r_0(1-\theta)+r_1\theta\right)\beta^n\right\}$$

$$= E\left\{\Lambda\theta + \frac{r_0(1-\theta)}{1-\beta}\right\} - E\left\{\left(\frac{r_0}{1-\beta} - \Lambda\right)(1-\theta)\beta^{\tau}\right.$$

$$\left. +((1-\beta)\Lambda - r_1)\theta\sum_{n=0}^{\tau-1}\beta^n\right\}$$

Letting $c_1 = (1-\beta)\Lambda - r_1$ and $c_0 = -\Lambda + r_0/(1-\beta)$, for $(1-\beta)\Lambda \in (r_1, r_0)$, both c_0 and c_1 are positive and

$$V(\pi, \beta, \Lambda) = \frac{r_0(1-\pi)}{1-\beta} + \pi\Lambda - R(\pi, \beta).$$

Now $V(\pi, \beta, \Lambda) = \Lambda$ if and only if $R(\pi, \beta) = c_0(1-\pi)$, and by the corollary this happens if and only if

$$(1-\pi)\left(\frac{r_0}{1-\beta} - \Lambda\right)L(\beta) \le \pi((1-\beta)\Lambda - r_1).$$

The theorem follows.

A few special cases of Theorem 2 are of interest. If the expected reward from playing an arm does not vary with time (except through the discounting) the arm is called *fixed*. This happens if $\pi = 0$ or 1, or if $r_0 = r_1 = r$. The Gittins indices in these three cases are r_0, r_1, and r, respectively. An arm is also fixed if the observations associated with the arm are completely uninformative, i.e., $P = Q$. In this case $\tilde{P}^{*n}$ and $\tilde{Q}^{*n}$ are point distributions concentrated on 0, and $L(\beta) = (1-\beta)$. In this case $G(\pi) = \pi r_1 + (1-\pi)r_0$, which remains constant over time since $\pi_n = \pi$ for all n. Finally, if P and Q are singular, so observations are completely informative, $\tilde{P}^{*n}$ and $\tilde{Q}^{*n}$ are point distributions concentrated on $+\infty$ and $-\infty$, respectively. This gives $L(\beta) = 1$ and $G(\pi) = \pi r_1 + (1-\pi)r_0 + \beta(r_0 - r_1)\pi(1-\pi)/(1-\pi\beta)$. The last term here is the value of the information from the potential observation. That Theorem 1 holds for two armed bandits with one fixed arm and one completely informative arm can be easily checked by direct computation.

REFERENCES

1. D. FELDMAN, Contributions to the "two-armed bandit" problem, *Ann. Math. Statist.* **33** (1962), 847–856.
2. J. C. GITTINS, Bandit processes and dynamic allocation indices (with discussion), *J. R. Statist. Soc. Ser. B.* **41** (1979), 148–164.

3. J. C. Gittins and K. D. Glazenbrook, On Bayesian models in stochastic scheduling, *J. Appl. Prob. ab* **14** (1977), 556–565.

4. J. C. Gittins and D. M. Jones, A dynamic allocation index for the sequential design of experiments, *in* "Progress in Statistics" (J. Gani, Ed.), pp. 241–266, North–Holland, Amsterdam 1974.

5. R. W. Keener, Further contributions to the "two-armed bandit" problem, *Ann. Statist.* **13** (1985), 418–422.

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

7. P. Whittle, Multi-armed bandits and the Gittins index, *J. R. Statist. Soc. Ser. B.* **42** (1980), 143–149.