STATIONARY MULTI-CHOICE BANDIT PROBLEMS

BY

DIRK BERGEMANN
AND
JUUSO VÄLIMÄKI

COWLES FOUNDATION PAPER NO. 1022

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
AT YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281
2001

ELSEVIER

Journal of Economic Dynamics & Control
25 (2001) 1585–1594

JOURNAL OF
Economic
Dynamics
& Control

www.elsevier.nl/locate/econbase

Stationary multi-choice bandit problems☆

Dirk Bergemann[a,*], Juuso Välimäki[b,c]

[a]*Department of Economics, Yale University, New Haven, CT 06520-8268, USA*
[b]*Department of Economics, Northwestern University, Evanston, IL 60208-2009, USA*
[c]*Department of Economics, University of Southampton, Southampton, SO17 1BJ, UK*

Received 16 June 1999; accepted 18 October 1999

Abstract

This note shows that the optimal choice of k simultaneous experiments in a stationary multi-armed bandit problem can be characterized in terms of the Gittins index of each arm. The index characterization remains equally valid after the introduction of switching costs. © 2001 Elsevier Science B.V. All rights reserved.

JEL classification: D81; D83

Keywords: Multi-armed bandits; Gittins index; Stationary bandits; Job search

1. Introduction

It is well known that in a multi-armed bandit situation, the optimality of the Gittins index policy is sensitive to changes in the number of experiments performed in each period.[1] A different generalization to which the index policy is not robust is the introduction of switching costs between the arms as shown in

☆The authors thank the referees and the editor for comments which improved the exposition of the paper. Financial support from NSF Grant SBR 9709887 and 9709340, respectively, is gratefully acknowledged.

* Corresponding author.
E-mail address: dirk.bergemann@yale.edu (D. Bergemann).

[1] See the extensive discussion in Gittins (1989, Section 2.11) on multiple processors.

PII: S0165-1889(99)00064-0

Banks and Sundaram (1994). The purpose of this note is to show that with a countable infinity of ex ante identical arms, the problem is much better behaved. This setup was called a stationary bandit problem in Banks and Sundaram (1992), who analyzed the optimality of the index policy with denumerably many arms. More precisely, we find that the main problem in the case with finitely many arms is the possible need to revert back to arms that were once abandoned. In a stationary bandit problem, once an arm is abandoned, it will never be revisited in either of the two environments just mentioned.

The basic setup of a stationary bandit has been used in the theory of job search as a description of a labor market. Hence it is of interest to note that the simple index rules characterizing optimal search can be generalized for stationary bandits. Within the context of a job market, the generalization to multiple searches in each time period seems relevant to models where a firm is searching for the most productive workers to fill k vacancies or, alternatively, a model of job search among family members when the resources within a family are pooled. In the context of allocating research and development expenditure, one can imagine a model where k separate research teams direct their efforts among a countable set of ex ante equally profitable projects. In each of these cases, our results demonstrate that selecting the alternatives with the highest individual indices is an optimal strategy. We also show that the use of a generalized index rule with a large but finite set of arms is approximately optimal.

2. Basic model

The general problem in this note is to operate k arms in a multi-armed bandit simultaneously so as to maximize the expected discounted sum of returns from all operated arms in all future time periods. Time is discrete and indexed by $t \in \{0, 1, \dots\}$. The set of available arms is given by $\mathbb{N} = \{1, 2, \dots\}$ and a particular element of the set is indexed by i. All arms are assumed to be ex ante identical and statistically independent. In period t, arm i yields a reward x_t^i if operated and zero otherwise. In the most general setup, the reward from arm i follows a stochastic process, not necessarily Markov, that depends only on past realizations from arm i.[2] For ease of notation, we focus on the case where each arm is a sampling process. In other words, $x_t^i \sim G(\cdot \mid \theta)$, where $\theta \in \Theta \subset \mathbb{R}^n$ is an unknown parameter. Assume also that $\inf_\theta \mathbb{E}[x_t^i \mid \theta] > 0$ so that selecting an arm is always strictly better than not selecting one and also that $\sup_\theta \mathbb{E}[x_t^i \mid \theta] < \infty$. Let π_0^i be the prior on θ (common to all i). The posterior belief π_t^i is updated using Bayes' rule in the periods when arm i is operated. Let $\pi_t = (\pi_t^1, \pi_t^2, \pi_t^3, \dots)$ and observe that for all t, $\pi_t^i = \pi_0^i$ for all but finitely many i. Let $\delta \in (0,1)$ be the

[2] See Varaiya et al. (1985) for the extension to the non-Markovian case.

discount factor between periods. The problem is then to choose for each t and π_t a subset $A_t(\pi_t)$ of $\mathbb{N}$ to solve the following:

$$\max_{A_t(\pi_t)\subset\mathbb{N}} \mathbb{E}_{\pi_0}\left[\sum_{t=0}^{\infty}\delta^t \sum_{i\in A_t(\pi_t)} x_t^i\right]$$

subject to

$$|A_t(\pi_t)| \leq k,$$

where $|A_t(\pi_t)|$ denotes the cardinality of the set $A_t(\pi_t)$.

We briefly recall the definition of the Gittins index in its flow characterization. For arm i at posterior π_t^i, the Gittins index is given by $m^i(\pi_t^i)$ if a decision maker is indifferent in period t between the following two payoff flows: (i) she can either receive a fixed reward of $m^i(\pi_t^i)$ in the current period as well as in all future periods or (ii) she can receive the random payoff x_t^i in the current period and maintain the option to continue sampling from x^i or receive $m^i(\pi_t^i)$ in all future periods.[3] An alternative characterization of the index $m^i(\pi_t^i)$ is given by

$$m^i(\pi_t^i) = \max_{\tau} \frac{\mathbb{E}_{\pi_t^i}\sum_{s=t}^{\tau}\delta^{s-t}x_s^i}{\mathbb{E}_{\pi_t^i}\sum_{s=t}^{\tau}\delta^{s-t}}, \tag{1}$$

where τ is a stopping time. In words, the index $m^i(\pi_t^i)$ is the maximal expected average discounted payoff per unit of expected discounted time. We denote the Gittins index of any arm in period 0 by

$$m_0 \triangleq m^i(\pi_0^i).$$

The following recursive definition extends the Gittins index rule to the case where k arms are selected simultaneously. Let

$$A_0(\pi_0) = \{1,\ldots,k\},$$

and associated with A_0, let

$$B_1(\pi_1) = \{i\in A_0(\pi_0)\,|\,m^i(\pi_1^i)\geq m_0\}.$$

The sequence $\{A_t(\pi_t), B_{t+1}(\pi_{t+1})\}_{t=0}^{\infty}$ is then extended recursively by

$$A_1(\pi_1) = B_1(\pi_1)\cup\{k+1,\ldots,2k-|B_1(\pi_1)|\}.$$

and

$$B_2(\pi_2) = \{i\in A_1(\pi_1)\,|\,m^i(\pi_2^i)\geq m_0\},$$

[3] For a more complete exposition of the theory of multi-armed bandit processes and the optimality of index policies, see Gittins (1989), Whittle (1982) or Banks and Sundaram (1992).

and for any arbitrary t by

$$A_t(\pi_t) = B_t(\pi_t) \cup \left\{ tk + 1 - \sum_{j=1}^{t-1} |B_j(\pi_j)|, \ldots, (t+1)k - \sum_{j=1}^{t} |B_j(\pi_j)| \right\},$$

and

$$B_{t+1}(\pi_{t+1}) = \{ i \in A_t(\pi_t) \mid m^i(\pi^i_{t+1}) \geq m_0 \}.$$

We call the sequence $\{A_t(\pi_t)\}_{t=0}^{\infty}$ the *Gittins index k-rule*. We may omit the obvious dependence on the sample path and posterior belief and simply write $\{A_t\}_{t=0}^{\infty}$. The sequence A_t operates all those arms whose Gittins indices are above the common index of the untried arms and abandons those arms with indices below that value. New arms are tried in the order of their labels. In the next section, we prove that this rule is optimal and we also establish the limit result that a finite version of this rule achieves approximately the optimal payoff in an economy with N arms as N gets large.

3. Optimality of the index rule

3.1. An example with finitely many arms

We start by giving an example of the failure of the index rule when k arms are to be selected in every period and there are only finitely many arms. For simplicity, we assume that the arms are not identical at the beginning of the problem. The reader may want to think of this situation as one arising as a continuation problem in an initially symmetric situation.

Consider three arms, of which at most two can be employed in any given period. The rewards of the arms are given by

$$x^1 = 1 \text{ with probability } p^1 = 1 \text{ for all } t.$$

$$x^2 = \begin{cases} 2 & \text{with } p^2 = \frac{1}{3}, \\ 0 & \text{with } 1 - p^2 = \frac{2}{3}, \end{cases} \text{ for all } t.$$

$$x^3 = \begin{cases} 3 & \text{with } p^3 = \frac{1}{3}, \\ 0 & \text{with } 1 - p^3 = \frac{4}{3}, \end{cases} \text{ for all } t.$$

Future payoffs are discounted with a discount factor $\delta = \frac{1}{2}$. Due to the simple structure of the uncertain arms 2 and 3, all uncertainty is resolved in a single trial.[4]

[4] This is assumed only to make the example computationally simple. The qualitative nature of the example does not depend on this.

It is straightforward to verify that $m^1 = m^2 = m^3 = 1$, as the Gittins index of each arm i can be computed by

$$m^i = \max\left\{m \left| p^i x^i + \frac{\delta p^i x^i}{1-\delta} + \frac{\delta(1-p^i)m}{1-\delta} \geq \frac{m}{1-\delta}\right.\right\}.$$

The Gittins index rule is hence completely indifferent about the temporal order in which the arms are selected. However a simple calculation shows that the values to the decision maker are different depending on the order by which she chooses among the arms. In particular, setting $A_0 = \{1,2\}$ or $A_0 = \{1,3\}$ yields an overall value of $\frac{56}{15}$ while starting with the two uncertain arms, $A_0 = \{2,3\}$ yields only $\frac{52}{15}$. If we parametrize the value, V, of the entire program by the reward, and thus the Gittins index, of the certain alternative, x_1, it is easy to show that $V(x_1)$ is a continuous function of x_1. But this tells us immediately that for small enough ε and $x_1 = m_1 = 1 - \varepsilon$, the optimal choice of A_0 includes alternative 1 even though it has a lower Gittins index than the other alternatives.

The loss resulting from the choice of $\{2,3\}$ rather than $\{1,2\}$ or $\{1,3\}$ can be understood as follows. If an uncertain arm fails, i.e. generates a reward of 0 in the first period, then it is optimal to abandon that arm and choose the safe arm in the following period. The option value of an uncertain arm is defined as the (expected) gain resulting from a switch to the safe arm. If both uncertain arms are chosen in the initial period, then the option value of one of the arms is lost immediately. Since, if both arms fail in the initial period, only one of the arms can have a positive option value. The probability of failing is $\frac{2}{5}$ for arm 2 and $\frac{4}{5}$ for arm 3. If $\{1,2\}$ are chosen in the initial period, then this loss in option value is delayed by one period and hence the difference in the losses is given by $(1-\delta)\frac{2}{5}\cdot\frac{4}{5} = \frac{4}{15}$. If the decision maker had two certain arms available, each yielding a reward of 1 per period, then all policies not using dominated arms would yield a payoff of 4. The simultaneous use of the uncertain arms would not result in any losses in terms of the option value as there would be sufficiently many safe arms available after all histories. In a stationary bandit problem, the unlimited supply of untried arms guarantees that the option value of any given arm is not dependent on the past choices of arms.

3.2. *Optimality with infinitely many identical arms*

The crucial point in the following proof is the observation that, with infinitely many identical arms, an arm that is abandoned once, will never be employed again. This form of independence from the state of the other arms is sufficient to show that the index policy is an optimal policy.

Theorem 1. The Gittins index k-rule is optimal. The initial value of the stationary k-choice bandit is given by

$$V(\pi_0) \triangleq \frac{km_0}{1-\delta}. \tag{2}$$

Proof. The proof proceeds in two steps. In Step 1, we show that the value $V(\pi_0)$ can be achieved by the Gittins index k-rule and then, in Step 2, we show that no other policy can achieve a larger expected payoff.

Step 1: Consider any of the available k slots. The following policy achieves $m_0/(1-\delta)$ for every slot. Start with an untried arm i and continue with it until its Gittins index $m^i(\pi_t^i)$ drops below m_0. At the first occurrence of this event switch to an untried arm and play according to the same strategy. We claim that this strategy achieves $m_0/(1-\delta)$. The claim is proved if we show that the stopping time:

$$\tau^i = \min\{t \mid m^i(\pi_t^i) < m_0\}$$

solves the optimization problem on the right-hand side of (1). But this follows immediately from the optimality of the index policy for $k = 1$. As the same argument can be given for all slots, we conclude that this policy results in a payoff of $km_0/(1-\delta)$.

Step 2: The argument is by contradiction. Suppose that there is another policy which achieves a strictly higher payoff than $km_0/(1-\delta)$. Then there must exist at least one arm i for which the expected average payoff strictly exceeds m_0. Denote by the indicator function $I^i(\pi_t)$ the employment of arm i under this policy:

$$I^i(\pi_t) = \begin{cases} 1 & \text{if } i \in A_t(\pi_t), \\ 0 & \text{if } i \notin A_t(\pi_t). \end{cases}$$

Restating the claim just made, it must be that there exists at least one i such that for some $m^i \in \mathbb{R}_+$ with $m^i > m_0$:

$$\frac{\mathbb{E}_{\pi_0} \sum_{t=0}^{\infty} I^i(\pi_t)\delta^t x_t^i}{\mathbb{E}_{\pi_0} \sum_{t=0}^{\infty} I^i(\pi_t)\delta^t} = m^i > m_0. \tag{3}$$

Notice that the indicator function depends on π_t, and not only on π_t^i, and that the indicator may switch arbitrarily often between 0 and 1. For every policy and its representation through the indicator function $I^i(\pi_t)$ we can construct another policy and associated indicator function $\hat{I}^i(\pi_t^i, t)$, depending on π_t^i and time t only, such that

$$\Pr(I^i(\pi_t) = 1 \mid \pi_t^i, t) = \Pr(\hat{I}^i(\pi_t^i, t) = 1 \mid \pi_t^i, t)$$

for every π_t^i and t.[5] It follows that the ratio formed by the new indicator function still satisfies the (in-)equalities in (3):

$$\frac{\mathbb{E}_{\pi_0^i}\sum_{t=0}^{\infty}\hat{I}^i(\pi_t^i,t)\delta^t x_t^i}{\mathbb{E}_{\pi_0^i}\sum_{t=0}^{\infty}\hat{I}^i(\pi_t^i,t)\delta^t} = m^i > m_0. \tag{4}$$

Consider next a modified stream of payoffs, $\{u_t\}_{t=0}^{\infty}$, based on $\hat{I}^i(\pi_t^i,t)$ and defined as follows:

$$\hat{I}^i(\pi_t^i,t) = 1 \Leftrightarrow u_t = x_t^i,$$

$$\hat{I}^i(\pi_t^i,t) = 0 \Leftrightarrow u_t = m^i.$$

Notice that $\{u_t\}_{t=0}^{\infty}$ can be thought of as the realized payoffs in a two-armed bandit problem with one certain and one uncertain arm. Denote the discounted expected payoff from the policy $\hat{I}^i(\pi_t^i,t)$ in this modified problem by $V(\hat{I}^i,m^i)$. From (4), we have

$$V(\hat{I}^i,m^i) = \mathbb{E}_{\pi_0^i}\sum_{t=0}^{\infty}\delta^t u_t = \frac{m^i}{1-\delta}$$

and for all $m < m^i$, we have

$$V(\hat{I}^i,m) > \frac{m}{1-\delta}.$$

Instead of the allocation policy $\hat{I}^i(\pi_t^i,t)$ consider now the *optimal* allocation policy between the uncertain arm x^i and the certain arm m^i. Finding the optimal allocation policy in this case is a standard two-armed bandit problem. A stationary solution to this problem exists. Denote this policy by $I^{i*}(\pi_t^i)$. By definition

$$V(I^{i*},m^i) \geq V(\hat{I}^i,m^i) = \frac{m^i}{1-\delta}. \tag{5}$$

As the optimal policy is stationary, there exists a stopping policy τ^* based only on π_t^i that achieves the same payoff as the optimal policy. But by inequality (5), this implies that

$$\frac{\mathbb{E}_{\pi_0^i}\sum_{t=0}^{\tau^*}\delta^t x_t^i}{\mathbb{E}_{\pi_0^i}\sum_{t=0}^{\tau^*}\delta^t} \geq m^i,$$

[5] At this point, we are using the stochastic independence of the arms, i.e. the fact that

$$\mathbb{E}[x_t^i \mid \pi_t] = \mathbb{E}[x_t^i|\pi_t^i].$$

which yields the desired contradiction as the initial hypothesis stated that

$$\max_{\tau} \frac{\mathbb{E}_{\pi_0^i} \sum_{t=0}^{\tau} \delta^t x_t^i}{\mathbb{E}_{\pi_0^i} \sum_{t=0}^{\tau} \delta^t} = m_0 < m^i,$$

and this concludes the proof. □

An immediate consequence of Theorem 1 is the following:

Corollary 1. The Gittins index k-rule is optimal in any bandit problem with infinitely many arms where all but $k' < k$ arms have initially the same Gittins index.

We conclude this section by providing a limit result for approximate optimality of the Gittins index k-rule when there are N arms and N goes to infinity. Let $V(k, N)$ denote the optimal value in a problem with N arms. Denote the value in the case with an infinite number of identical arms by $V(k)$.

Lemma 1. $V(k, N)$ is increasing in N.

Proof. Let $N' < N$. Then $V(k, N') \leq V(k, N)$ as, with N arms, it is always possible to restrict all choices to $\{1, \ldots, N'\}$. □

In view of this lemma, the value of the problem with a countable infinity of arms exceeds (weakly) the value of any finite arm problem. The last result in this section shows that the truncated Gittins index k-rule achieves the value of the countable arm problem in the limit. By the truncated rule, we mean the rule that selects arms $A_t \cap \{1, \ldots, N\}$ in period t. Denote this rule by $\hat{A}$ and the value corresponding to this rule by $V^{\hat{A}}(k, N)$.

Lemma 2.

1. $\lim_{N \to \infty} V^{\hat{A}}(k, N) = V(k)$,
2. $\lim_{N \to \infty} V(k, N) = V(k)$.

Proof. (1) Let $p^A(k, N) = \Pr\{N + 1 \in A_t \text{ for some } t\}$. Then with probability $1 - p^A(k, N)$, the realizations of A and $\hat{A}$ coincide. In the complementary case, the payoff difference is bounded since all the stage game payoffs are bounded and discounted. Hence we are done if we can show that $\lim_{N \to \infty} p^A(k, N) = 0$. But this follows readily from the observation that in a sampling bandit process, $\Pr\{m^i(\pi_t^i) > m_0, \forall t\} > 0$, as proven in Corollary 5.2 in Banks and Sundaram (1992) and the fact that the arms are statistically independent.

(2) Since the optimal value $V(k, N) \geq V^{\hat{A}}(k, N)$, the convergence of $V^{\hat{A}}(k, N)$ implies directly the convergence and asymptotic optimality of $V(k, N)$. □

4. Switching costs

In this section, we show that the result of Banks and Sundaram (1994) on the non-existence of an optimal index policy with switching costs disappears in the stationary bandit setting. The basic intuition is that the availability of untried arms makes it unnecessary to ever switch back to arms which were abandoned earlier. This observation combined with the result by Weitzman (1979) stating that an optimal index policy exists if switching costs are paid only on the first trial with a given arm shows that the Gittins index rule remains optimal in the stationary bandit setting. Theorem 1 then allows us to conclude that the Gittins index k-rule is optimal in the simultaneous allocation problem of k arms under switching costs.

Let c denote the switching cost to an arm and d denote the switching cost from an arm.

Theorem 2. The appropriately defined Gittins index k-rule is optimal in a stationary bandit problem with switching costs.

Proof. It is immediately verified that the index can be appropriately modified by deducting $(c + d)$ from the payoff of any untried arm in the initial period of its use. By Theorem 1, we know that an arm should be abandoned forever once its index falls below the (modified) index of the untried arms. After abandoning an arm, switching costs of $(c + d)$ would have to be deducted if the arm were to be reused in any future period. A fortiori, the strategy of Theorem 1 remains optimal and it is best to never fall back to an earlier abandoned arm whose Gittins index is strictly below the (modified) one of the untried arms. Since untried arms are available in all periods, the optimal strategy is to continue with an arm if and only if its index exceeds the (modified) one of the untried arms. □

5. Conclusion

In many economic problems such as the allocation of workers to firms, a stationary bandit situation is the most natural idealized description of the market. This note shows that while the Gittins index rule in a finite multi-armed bandit problem is not robust to many economically meaningful perturbations, the problem with infinitely many arms is much better behaved. A generalization of the Gittins index rule remains optimal in the allocation problem of k simultaneous experiments as well as in a problem with switching costs. We also show that the generalized Gittins index rule performs well in bandit problems with a large but finite number of arms. In particular, the payoff from the generalized Gittins index rule converges to that of the optimal policy as the number of arms tends to infinity.

On a more technical level, this note clarifies the reason for the failure of the Gittins index policy in finite arm bandits with multiple experiments. In the classical multi-armed bandit, a crucial requirement is that the employment of one arm leaves all other arms unaffected. With multiple experiments, the best available alternative in the next period to an individual arm depends on the choice of other experiments in the current period. But the Gittins index is calculated under the assumption that the payoffs from all but one arm are not affected by the choices in the current period. As a result, the Gittins index rule is not optimal with multiple experiments and finitely many arms. An infinite supply of ex ante identical arms guarantees that the relevant alternatives when an arm is to be abandoned do not depend on the past choices of arms and hence the optimality of the Gittins index rule is restored.

References

Banks, J.S., Sundaram, R.K., 1992. Denumerable-armed bandits. Econometrica 60, 1071–1096.
Banks, J.S., Sundaram, R.K., 1994. Switching costs and the Gittins index. Econometrica 62, 687–694.
Gittins, J., 1989. Allocation Indices for Multi-Armed Bandits. Wiley, London.
Varaiya, P.P., Walrand, J.C., Buyukkoc, C., 1985. Extensions of the multiarmed bandit problem: the discounted case. IEEE Transactions on Automatic Control AC-30, 426–439.
Weitzman, M., 1979. Optimal search for the best alternative. Econometrica 47, 641–654.
Whittle, P., 1982. Optimization Over Time, vol. 1. Wiley, Chichester.