

Chapter 7. BANDIT PROBLEMS.

Bandit problems are problems in the area of sequential selection of experiments, and they are related to stopping rule problems through the theorem of Gittins and Jones (1974). In the first section, we present a description of bandit problems and give some historical background. In the next section, we treat the one-armed bandit problems by the method of the previous chapter. In the final section, we discuss the Theorem of Gittins and Jones which shows that the k -armed bandit problems may be solved by solving k one-armed problems. An excellent reference to bandit problems is the book of Berry and Fristedt, (1985).

§7.1 Introduction to the Problem. Consider a sequential decision problem in which at each stage there are k possible actions or choices of experiment. Choice of action j results in an observation being taken from the j th experiment, and you receive the numerical value of this observation as a reward. The observations you make may give you information useful in future choices of actions. Your goal is to maximize the present value of the infinite stream of rewards you receive, discounted in some way.

The name “bandit” comes from modeling these problems as a k -armed bandit, which is a slot machine with k arms, each yielding an unknown, possibly different distribution of payoffs. You do not know which arm gives you the greatest average return, but by playing the various arms of the slot machine you can gain information on which arm is best. However, the observations you use to gain information are also your rewards. You must strike a balance between gaining rewards and gaining information. For example, it is not good to always pull the arm that has performed best in the past, because it may have been that you were just unlucky with the best arm. If you have many trials to go and it only takes a few trials to clarify the matter, you can stand to improve your average gain greatly with only a small investment. Typically in these problems, there is a period of gaining information, followed by a period of narrowing down the arms, followed by a period of “profit taking”, playing the arm you feel to be the best.

A more important modeling of these problems comes from clinical trials in which there are k treatments for a given disease. Patients arrive sequentially at the clinic and must be treated immediately by one of the treatments. It is assumed that response from treatment is immediate so that the effectiveness of the treatment that the present patient receives is known when the next patient must be treated. It is not known precisely which one of the

treatments is best, but you must decide which treatment to give each patient, keeping in mind that your goal is to cure as many patients as possible. This may require you to give a patient a treatment which is not the one that looks best at the present time in order to gain information that may be of use to future patients.

To describe bandit problems more precisely, let the k reward distributions be denoted by $F_1(x|\theta_1), \dots, F_k(x|\theta_k)$ where $\theta_1, \dots, \theta_k$ are parameters whose exact values are not known precisely, but whose joint prior distribution is known and denoted by $G(\theta_1, \dots, \theta_k)$. Initially, an action a_1 is chosen from the set $\{1, \dots, k\}$ and then an observation, Z_1 , the reward for the first stage, is taken from the distribution F_{a_1} . Based on this information, an action a_2 is then taken from the same action space, and an observation, Z_2 , is taken from F_{a_2} and so on. It is assumed that given a_n and the parameters $\theta_1, \dots, \theta_k$, Z_n is chosen from F_{a_n} independent of the past. A decision rule for this problem is a sequence $A = (a_1, a_2, \dots)$ of functions adapted to the observations; that is, a_n may depend on past actions and observations,

$$a_n(a_1, Z_1, a_2, Z_2, \dots, a_{n-1}, Z_{n-1}).$$

It is hoped that no confusion results from using one symbol, a_n , to denote both the function of past observations and the action taken at stage n .

There is a discount sequence, denoted by $B = (\beta_1, \beta_2, \dots)$, such that the j th observation is discounted by β_j where $0 \leq \beta_j \leq 1$ for $j = 1, 2, \dots$. The total discounted return is then $\sum_1^\infty \beta_j Z_j$. The problem is to choose a decision rule A to maximize the expected reward, $E \sum_1^\infty \beta_j Z_j$. This problem is called the k -armed bandit problem. The one-armed bandit problem, mentioned in Exercise 1.4, is defined as the 2-armed bandit problem in which one of the arms always returns the same known amount, that is, the distribution F associated with one of the arms is degenerate at a known constant.

To obtain a finite value for the expected reward, we assume

- (1) each distribution, F_j for $j = 1, \dots, k$, has finite first moment, and
- (2) $\sum_1^\infty \beta_j < \infty$.

Two important special cases of the discount sequence are

- (1) the *n-horizon uniform* for which $\beta_1 = \dots = \beta_n = 1$ and $\beta_{n+1} = \beta_{n+2} = \dots = 0$, and
- (2) the *geometric* in which $B = (1, \beta, \beta^2, \beta^3, \dots)$, that is, $\beta_j = \beta^{j-1}$ for $j = 1, 2, \dots$

In the former, the payoff is simply $\sum_1^n Z_j$, the sum of the first n observations. The problem becomes one with a finite horizon which can in principle be solved by backward induction. In the latter, there is a time invariance in which the future after n stages looks like it did at the start except for the change from the prior distribution to the posterior distribution. We will treat mainly problems with geometric discount and independent arms, that is, prior distributions G for which $\theta_1, \dots, \theta_k$ are independent, so that an observation on one arm will not influence your knowledge of the distribution of any other arm.

First, we give a little historical background to add perspective to what follows. Early work on these problems centered mainly on the finite horizon problem with Bernoulli trials. These problems were introduced in the framework of clinical trials by Thompson (1933) for two treatments with outcomes forming Bernoulli trials with success probabilities having independent uniform prior distributions on $(0,1)$. Robbins (1952) reintroduced the problem from a non-Bayes viewpoint, and suggested searching for the minimax decision rule. In the Bernoulli case, he proposed the play-the-winner/switch-from-a-loser strategy, and discussed the asymptotic behavior of rules.

The first paper to search for Bayes decision rules for this problem is the paper of Bradt, Johnson and Karlin (1956). One of their important results, mentioned in Exercise 1.4, is that for the one-armed bandit with finite horizon and Bernoulli trials, if the known arm is optimal at any stage, then it is optimal to use that arm at all subsequent stages. Another important result is that for the 2-armed bandit with finite horizon and Bernoulli trials with success probabilities p_1 and p_2 , if the prior distribution gives all its weight to points (p_1, p_2) on the line $p_1 + p_2 = 1$, then the 1-stage look-ahead rule is optimal; that is, it is optimal to choose the arm with the higher present probability of success. In addition, they conjecture that if the prior distribution gives all its weight to two points (a, b) and (b, a) symmetrically placed about the line $p_1 = p_2$, then again the 1-stage look-ahead rule is optimal. This conjecture was proved to be true by Feldman (1962).

These are about the only cases in which the optimal rule is easy to evaluate. In particular, in the important practical case of independent arms, the difficulty of computation of the optimal rule hindered real progress. One important result for the 2-armed Bernoulli bandit with independent arms and finite horizon is the “stay-on-a winner” principle, proved in Berry (1972). This principle states that if an arm is optimal at some stage and if it proves to be successful at that stage, then it is optimal at the following stage also. This was proved for the one-armed Bernoulli bandit with finite horizon by Bradt, Johnson and Karlin (1956). This was the state of affairs when the theorem of Gittins and Jones showed that for the k -armed bandit with independent arms and geometric discount, the problem can be solved by solving k one-armed bandit problems.

We treat a problem of a somewhat more general structure than that indicated above by letting the returns for each arm be an arbitrary sequence of random variables, (not necessarily exchangeable). Thus for each arm, say arm j , there is assumed to be a sequence of returns, X_{j1}, X_{j2}, \dots with an arbitrary joint distribution, subject to the condition that $\sup_n EX_{jn}^+ < \infty$ for all $j = 1, \dots, k$. It is assumed that the arms are independent, that is, that the sets $\{X_{11}, X_{12}, \dots\}, \dots, \{X_{k1}, X_{k2}, \dots\}$ are independent sets of random variables. When an arm, say j , is pulled, the first random variable of the sequence X_{j1} is received as the reward. The next time j is pulled, the reward X_{j2} is received, etc. That the theorem of Gittins and Jones applies to this more general probabilistic structure has been proved by Varaiya, Walrand and Buyukkoc (1985). See also Mandelbaum (1986).

§7.2 The one-armed bandit. As a preliminary to the solution of the k -armed bandit with geometric discount and independent arms, it is important to understand the one-armed bandit. The one-armed bandit is really a bandit problem with two arms, but

one of the arms has a known i.i.d. distribution of returns, and so plays only a minor role. We first show how the one-armed bandit can be related to a stopping rule problem. We assume that arm 1 has an associated sequence of random variables, X_1, X_2, \dots with known joint distribution satisfying $\sup_n EX_n^+ < \infty$. For the other arm, arm 2, the returns are assumed to be i.i.d. from a known distribution with expectation λ . We take the discount sequence to be geometric, $B = (1, \beta, \beta^2, \dots)$ where $0 < \beta < 1$, and seek to find a decision rule $A = (a_1, a_2, \dots)$ to maximize

$$(1) \quad V(A) = E\left(\sum_1^{\infty} \beta^{j-1} Z_j | A\right).$$

First we argue that we may assume without loss of generality that the returns from arm 2 are degenerate at λ . Note that in the expectation above, any Z_j from arm 2 may be replaced by λ . However, the rule A may allow the actual values of these Z_j to influence the choice of the future a_j . But the statistician may produce his own private sequence of i.i.d. random variables from the distribution of arm 2 and use them in A in place of the actual values he sees. This produces a randomized decision rule which may be denoted by A^* . Use of A^* in the problem where the random variables are given to be degenerate at λ produces the same expected payoff as A does in the original problem.

The advantage of making this observation is that we may now assume that the decision rule A does not depend on Z_j when $a_j = 2$, since Z_j is known to be λ . Thus we assume that arm 2 gives a constant return of λ each time it is pulled.

We now show that if at any stage it is optimal to pull arm 2, then it is optimal to keep pulling arm 2 thereafter. This implies that if there exists an optimal rule for this problem, there exists an optimal rule with the property that every pull of arm 2 is followed by another pull of arm 2. Thus one need only decide on the time to switch from arm 1 to arm 2. This relates this problem to a stopping rule problem in which the stopping time is identified with the time of switching from arm 1 to arm 2. Without loss of generality, we may assume that arm 2 is optimal at the initial stage, and state the theorem as follows.

Theorem 1. *If it is initially optimal to use arm 2 in the sense that $\sup_A V(A) = V^* = \sup\{V(A) : A \text{ such that } a_1 = 2\}$, then it is optimal to use arm 2 always and $V^* = \lambda/(1 - \beta)$.*

Proof. For a given $\epsilon > 0$, find a decision rule A such that $a_1 = 2$ and $V(A) \geq V^* - \epsilon$. Then,

$$\begin{aligned} V(A) &= \lambda + \beta E\left(\sum_2^{\infty} \beta^{j-2} Z_j | A\right) \\ &= \lambda + \beta E\left(\sum_1^{\infty} \beta^{j-1} Z_{j+1} | A\right) \\ &= \lambda + \beta E\left(\sum_1^{\infty} \beta^{j-1} Z'_j | A^1\right) \\ &\leq \lambda + \beta V^*, \end{aligned}$$

where $A^1 = (a_2, a_3, \dots)$ is the rule A shifted by 1, and $Z'_j = Z_{j+1}$. Thus, we have $V^* - \epsilon \leq \lambda + \beta V^*$, or equivalently, $V^* \leq (\lambda + \epsilon)/(1 - \beta)$. Since $\epsilon > 0$ is arbitrary, this implies $V^* \leq \lambda/(1 - \beta)$, but this value is achievable by using arm 2 at each stage. ■

This theorem is also valid for the n -uniform discount sequence. In fact, Berry and Fristedt (1985) show that if the sequence $\{X_1, X_2, \dots\}$ is exchangeable, then an optimal strategy for the problem has the property that every pull of arm 2 is followed by a pull of arm 1 if, and essentially only if, the discount sequence is what they call regular. The discount sequence B is said to be *regular* if it has increasing failure rate, that is, if $\beta_n / \sum_n^\infty \beta_j$ is non-decreasing on its domain of definition. That the above theorem is not true for such discount sequences is easily seen: If $B = \{.1, 1, 0, 0, 0, \dots\}$, then B is regular, yet if X_1 is degenerate at 10, X_2 is degenerate at 0, and $\lambda = 0$, then clearly the only optimal strategy is to follow an initial pull of arm 2 with a pull of arm 1. Exactly what property of B is required for the above theorem seems to be unknown.

As a corollary, we see that there exists an optimal rule for this problem. It is either the rule that uses arm 2 at all stages, or the rule corresponding to the stopping rule $N \geq 1$ that is optimal for the stopping rule problem with payoff,

$$(2) \quad Y_n = \sum_1^n \beta^{j-1} X_j + \lambda \sum_{n+1}^\infty \beta^{j-1}.$$

In fact, we can say more.

Theorem 2. Let $\Lambda(\beta)$ denote the optimal rate of return for using arm 1 at discount β ,

$$(3) \quad \Lambda(\beta) = \sup_{N \geq 1} \frac{\mathbb{E}(\sum_1^N \beta^{j-1} X_j)}{\mathbb{E}(\sum_1^N \beta^{j-1})}.$$

Then arm 2 is optimal initially if, and only if, $\lambda \geq \Lambda(\beta)$.

Proof. By Theorem 1, we may restrict attention to decision rules A specified by a stopping time N which represents the last time that arm 1 is used. The payoff using N is $\mathbb{E}(\sum_1^N \beta^{j-1} X_j + \lambda \sum_{N+1}^\infty \beta^{j-1})$, which for $N = 0$ is $\lambda/(1 - \beta)$. Therefore, arm 2 is optimal initially if, and only if, for all stopping rules $N \geq 1$,

$$\mathbb{E}\left(\sum_1^N \beta^{j-1} X_j + \lambda \sum_{N+1}^\infty \beta^{j-1}\right) \leq \lambda/(1 - \beta)$$

or, equivalently,

$$\mathbb{E}\left(\sum_1^N \beta^{j-1} X_j\right) \leq \lambda \mathbb{E}\left(\sum_1^N \beta^{j-1}\right)$$

or, equivalently,

$$\mathbb{E}\left(\sum_1^N \beta^{j-1} X_j\right) / \mathbb{E}\left(\sum_1^N \beta^{j-1}\right) \leq \lambda.$$

This is equivalent to $\Lambda(\beta) \leq \lambda$. ■

The value $\Lambda(\beta)$ depends only on β and on the distribution of the returns from arm 1, X_1, X_2, \dots . It is called the Gittins index for arm 1 and it represents the indifference point: that value of λ for arm 2 in the one-armed bandit at which you would be indifferent between starting off on arm 1 and choosing arm 2 all the time.

§7.3 The Gittins Index Theorem. We return to the k -armed bandit with geometric discount and independent arms having returns denoted by

arm 1: $X(1, 1), X(1, 2), \dots$

arm 2: $X(2, 1), X(2, 2), \dots$

...

arm k : $X(k, 1), X(k, 2), \dots$

where it is assumed that the variables are independent between rows and that the first absolute moments exist and are uniformly bounded, $\sup_{k \geq 1, t \geq 1} \mathbb{E}|X(k, t)| < \infty$. The discount is β , where $0 \leq \beta < 1$, and we seek a decision rule $A = (a_1, a_2, \dots)$ to maximize the total discounted return,

$$(4) \quad V(A) = \mathbb{E}\left(\sum_{t=1}^{\infty} \beta^{t-1} Z_t | A\right).$$

For each arm we may compute a Gittins index,

$$(5) \quad \Lambda_j = \sup_{N \geq 1} \mathbb{E} \sum_{t=1}^N \beta^{t-1} X(j, t) / \mathbb{E} \sum_{t=1}^N \beta^{t-1} \quad \text{for } j = 1, \dots, k$$

where we suppress β in the notation for Λ since β is held constant throughout this section.

The celebrated theorem of Gittins and Jones (1974) states that for the k -armed bandit with geometric discount and independent arms, it is optimal at each stage to select the arm with the highest index. We give a proof of this theorem due to Varaiya, Walrand and Buyukkoc (1983) and adopt their didactic strategy of presenting the proof first in the special case in which there are just two arms ($k = 2$) and where all the random variables are degenerate. We denote the returns from arm 1 by $X(1), X(2), \dots$ and from arm 2 by $Y(1), Y(2), \dots$. Thus, it is assumed that $X(1), X(2), \dots$ and $Y(1), Y(2), \dots$ are bounded sequences of real numbers. Let the Gittins indices be denoted by

$$\Lambda_X = \sup_{j \geq 1} \sum_1^j \beta^{t-1} X(t) / \sum_1^j \beta^{t-1}$$

$$\Lambda_Y = \sup_{j \geq 1} \sum_1^j \beta^{t-1} Y(t) / \sum_1^j \beta^{t-1}.$$

Since $X(t)$ is assumed bounded, the series $\sum_1^n \beta^{t-1} X(t)$ converges, so that there exists a value of j , possibly ∞ , at which the supremum in the definition of Λ_X is taken on. In the following lemma, we suppose that s is this value of j so that $1 \leq s \leq \infty$.

Lemma 1. *Suppose the sequence $X(1), X(2), \dots$ is non-random and bounded. If $\Lambda_X = \sum_1^s \beta^{t-1} X(t) / \sum_1^s \beta^{t-1}$, then for all $j \leq s$,*

$$(6) \quad \sum_{t=j}^s \beta^{t-1} X(t) \geq \Lambda_X \sum_{t=j}^s \beta^{t-1},$$

and for s finite and all $j > s$,

$$(7) \quad \sum_{t=s+1}^j \beta^{t-1} X(t) \leq \Lambda_X \sum_{t=s+1}^j \beta^{t-1}.$$

Proof. Since we have both

$$\begin{aligned} \sum_1^s \beta^{t-1} X(t) &= \Lambda_X \sum_1^s \beta^{t-1}, \quad \text{and} \\ \sum_1^{j-1} \beta^{t-1} X(t) &\leq \Lambda_X \sum_1^{j-1} \beta^{t-1} \quad \text{for all } j, \end{aligned}$$

subtracting the latter from the former gives (6) when j is less than or equal to s , and gives (7) when $j > s$. ■

Inequality (6) implies that after $j - 1 < s$ stages have elapsed, the new Gittins index (which is at least as great as $\sum_j^s \beta^{t-1} X(t) / \sum_j^s \beta^{t-1}$) is at least as great as the original index, Λ_X . Similarly, (7) shows that after exactly s stages have elapsed, the new Gittins index can be no greater than the original one. In fact, the following proof shows that if $\Lambda_X \geq \Lambda_Y$ and $\Lambda_X = \sum_1^s \beta^{t-1} X(t) / \sum_1^s \beta^{t-1}$, then it is optimal to start with at least s pulls of arm 1.

Theorem 3. *If the sequences $X(1), X(2), \dots$ and $Y(1), Y(2), \dots$ are non-random and bounded, if $\Lambda_X \geq \Lambda_Y$ then it is optimal initially to use arm 1.*

Proof. Assume $\Lambda_X \geq \Lambda_Y$ and let A be any rule. We will show that there is a rule A' that begins with arm 1 and gives at least as great a value as A . Then it is clear that the supremum of $V(A)$ over all A is the same as the supremum of $V(A)$ over all A that begin with arm 1.

Find s such that $\Lambda_X = \sum_1^s \beta^{t-1} X(t) / \sum_1^s \beta^{t-1}$ where $1 \leq s \leq \infty$. Let $k(t)$ denote the number of times that rule A calls for the use of arm 2 just before arm 1 is used for the t th time, $t = 1, 2, \dots$. It may be that arm 2 is not used between pulls $t - 1$ and t of arm 1 so that $k(t) = k(t - 1)$ and it is possible that A does not use arm 1 t times, in which case $k(t)$ is defined as $+\infty$. We have $0 \leq k(1) \leq k(2) \leq \dots \leq \infty$. We define a new decision rule A' that starts out with s X 's followed by $k(s)$ Y 's and then, if $k(s) < \infty$,

following this with the same sequence of X 's and Y 's as in A . Subject to s and the first s $k(t)$'s being finite, the sequence of observations occurs in the following order for A :

$$Y(1), \dots, Y(k(1)), X(1), Y(k(1) + 1), \dots, Y(k(2)), X(2), \dots, Y(k(s)), X(s), Z(T + 1), \dots$$

where T is the time that the s th X occurs using decision rule A and $Z(T + 1), \dots$ represents the rest of this sequence beyond T . The sequence of observations occurs in the following order using A' :

$$X(1), X(2), \dots, X(s), Y(1), \dots, Y(k(1)), \dots, Y(k(2)), \dots, Y(k(s)), Z(T + 1), \dots$$

We are to show that $V(A') - V(A) \geq 0$. Write this difference as $V(A') - V(A) = \Delta(X) - \Delta(Y)$, where $\Delta(X)$ represents the improvement in the value caused by shifting the X 's forward, and $\Delta(Y)$ represents the loss due to shifting the Y 's backward. Thus,

$$\begin{aligned} \Delta(X) &= \sum_1^s \beta^{t-1} X(t) - \sum_1^s \beta^{t+k(t)-1} X(t) \\ &= \sum_1^s \beta^{t-1} (1 - \beta^{k(t)}) X(t) \\ &= \sum_{t=1}^s \beta^{t-1} X(t) \sum_{j=1}^t (\beta^{k(j-1)} - \beta^{k(j)}) \\ &= \sum_{j=1}^s (\beta^{k(j-1)} - \beta^{k(j)}) \sum_{t=j}^s \beta^{t-1} X(t) \\ &\geq \Lambda_X \sum_{j=1}^s (\beta^{k(j-1)} - \beta^{k(j)}) \sum_{t=j}^s \beta^{t-1} \\ &= \Lambda_X \sum_1^s \beta^{t-1} (1 - \beta^{k(t)}), \end{aligned}$$

where $k(0)$ represents 0, and the inequality follows from (6). It is important to note that this computation is valid even if some of the $k(t) = \infty$, that is, even if A did not contain s X 's. It is also valid if $s = \infty$. Similarly, we have

$$\begin{aligned} \Delta(Y) &= \sum_{j=1}^s \beta^{j-1} \sum_{t=k(j-1)+1}^{k(j)} \beta^{t-1} Y(t) - \beta^s \sum_{t=1}^{k(s)} \beta^{t-1} Y(t) \\ &= \sum_{j=1}^s (\beta^{j-1} - \beta^s) \sum_{t=k(j-1)+1}^{k(j)} \beta^{t-1} Y(t) \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^s \sum_{m=j}^s (\beta^{m-1} - \beta^m) \sum_{t=k(j-1)+1}^{k(j)} \beta^{t-1} Y(t) \\
&= \sum_{m=1}^s (\beta^{m-1} - \beta^m) \sum_{j=1}^m \sum_{t=k(j-1)+1}^{k(j)} \beta^{t-1} Y(t) \\
&= (1 - \beta) \sum_{m=1}^s \beta^{m-1} \sum_{t=1}^{k(m)} \beta^{t-1} Y(t) \\
&\leq \Lambda_Y (1 - \beta) \sum_{m=1}^s \beta^{m-1} \sum_{t=1}^{k(j)} \beta^{t-1} \\
&= \Lambda_Y \sum_{m=1}^s \beta^{m-1} (1 - \beta^{k(m)}).
\end{aligned}$$

The inequality follows from the definition of Λ_Y . This computation is also valid if $s = \infty$ or if some of the $k(t) = \infty$. Now, using the assumption that $\Lambda_X \geq \Lambda_Y$, we find that $\Delta(X) - \Delta(Y) \geq 0$, as was to be shown. ■

The proof of the general theorem is similar, but at the crucial step involving the inequalities, we will not be able to interchange summation and expectation because the limits of summation are random. To circumvent this difficulty, the following two lemmas will be used. We deal with possibly randomized stopping rules by allowing the increasing sequence of σ -fields,

$$(8) \quad \mathcal{F}(1) \subset \mathcal{F}(2) \subset \dots \subset \mathcal{F}(\infty)$$

to be such that $\mathcal{F}(t)$ is the σ -field generated by $X(1), \dots, X(t)$ and any number of other random variables independent of $X(t+1), X(t+2), \dots$. To say now that a random variable Z is $\mathcal{F}(t)$ -measurable means essentially that Z and $\{X(t+1), X(t+2), \dots\}$ are conditionally independent given $X(1), \dots, X(t)$. In particular, we have

$$(9) \quad \begin{aligned} \mathbb{E}X(t+1)Z &= \mathbb{E}(\mathbb{E}\{X(t+1)Z | X(1), \dots, X(t+1)\}) \\ &= \mathbb{E}(X(t+1)\mathbb{E}\{Z | X(1), \dots, X(t)\}) \end{aligned}$$

for any $\mathcal{F}(t)$ -measurable Z . This observation is useful in the proof of the following lemma.

Lemma 2. *Let $X(t)$ be a sequence of random variables such that $\sup_t \mathbb{E}|X(t)| < \infty$, let $0 < \beta < 1$ and let Λ_X denote the Gittins index. Then, for every stopping rule N , and every sequence of random variables $\alpha(t)$, $t = 1, 2, \dots$, such that $\alpha(t)$ is $\mathcal{F}(t-1)$ -measurable and $1 \geq \alpha(1) \geq \alpha(2) \geq \dots \geq 0$ a.s., we have*

$$(10) \quad \mathbb{E} \sum_{t=1}^N \alpha(t) \beta^{t-1} X(t) \leq \Lambda_X \mathbb{E} \sum_{t=1}^N \alpha(t) \beta^{t-1}.$$

Proof. Let $W(t) = \beta^{t-1}(X(t) - \Lambda_X)$. Then, the definition of Λ_X implies that for every stopping rule $N \geq 1$,

$$\mathbb{E} \sum_{t=1}^N W(t) \leq 0.$$

For any stopping rule N , $I(N \geq t)$ is $\mathcal{F}(t-1)$ -measurable. Hence, from (9),

$$\begin{aligned} \mathbb{E} \sum_{t=1}^N W(t) &= \mathbb{E} \sum_{n=1}^{\infty} I(N = n) \sum_{t=1}^n W(t) \\ &= \mathbb{E} \sum_{t=1}^{\infty} W(t) \sum_{n=t}^{\infty} I(N = n) \\ &= \mathbb{E} \sum_{t=1}^{\infty} W(t) \gamma(t) \leq 0, \end{aligned}$$

where $\gamma(t) = \mathbb{P}(N \geq t | X(1), \dots, X(t-1))$. Any sequence, $1 \geq \gamma(1) \geq \gamma(2) \geq \dots \geq 0$ a.s. with $\gamma(t)$ $\mathcal{F}(t-1)$ -measurable, determines a stopping rule N such that $\mathbb{P}(N \geq t | \mathcal{F}(t-1)) = \gamma(t)$. Hence, the hypothesis that $\mathbb{E} \sum_{t=1}^N W(t) \leq 0$ for all stopping rules N is thus equivalent to the hypothesis that $\mathbb{E} \sum_{t=1}^{\infty} W(t) \gamma(t) \leq 0$ for all sequences $\gamma(t)$ such that $\gamma(t)$ is $\mathcal{F}(t-1)$ -measurable and $1 \geq \gamma(1) \geq \gamma(2) \geq \dots \geq 0$ a.s. Now, since

$$\begin{aligned} \mathbb{E} \sum_{t=1}^N \alpha(t) W(t) &= \mathbb{E} \sum_{n=1}^{\infty} I(N = n) \sum_{t=1}^n \alpha(t) W(t) \\ &= \mathbb{E} \sum_{t=1}^{\infty} \alpha(t) W(t) \sum_{n=t}^{\infty} I(N = n) \\ &= \mathbb{E} \sum_{t=1}^{\infty} W(t) \gamma(t) \leq 0, \end{aligned}$$

where $\gamma(t) = \mathbb{E}\{\alpha(t)I(N \geq t) | X(1), \dots, X(t-1)\}$, the conclusion follows. ■

The next lemma provides the required generalization of (6) of Lemma 1. Since $\mathbb{E}|X_n|$ is assumed to be bounded, conditions A1 and A2 are satisfied and there exists an optimal stopping rule for every λ . In particular, there exists a rule N^* that attains the Gittins index.

Lemma 3. *Let $X(t)$ be a sequence of random variables such that $\sup_t \mathbb{E}|X(t)| < \infty$, let $0 < \beta < 1$, and let N^* denote a stopping rule that attains the Gittins index, Λ_X . Then, for every sequence of random variables $\xi(t)$, $t = 1, 2, \dots$ such that $\xi(t)$ is $\mathcal{F}(t-1)$ -measurable and $0 \leq \xi(1) \leq \xi(2) \leq \dots \leq 1$ a.s., we have*

$$(11) \quad \mathbb{E} \sum_{t=1}^{N^*} \xi(t) \beta^{t-1} X(t) \geq \Lambda_X \mathbb{E} \sum_{t=1}^{N^*} \xi(t) \beta^{t-1}.$$

Proof. Since the Gittins index is attained at N^* ,

$$\mathbb{E} \sum_{t=1}^{N^*} \beta^{t-1} X(t) = \Lambda_X \mathbb{E} \sum_{t=1}^{N^*} \beta^{t-1}.$$

From Lemma 2 with $\alpha(t) = 1 - \xi(t)$, we have

$$\mathbb{E} \sum_{t=1}^{N^*} (1 - \xi(t)) \beta^{t-1} X(t) \leq \Lambda_X \mathbb{E} \sum_{t=1}^{N^*} (1 - \xi(t)) \beta^{t-1}.$$

Subtracting the latter from the former gives the result. ■

We now turn to the general problem with k independent arms, and denote the sequence of returns from arm j by $X(j, 1), X(j, 2), \dots$ for $j = 1, \dots, k$. It is assumed that the sets $\{X(1, t)\}_{t=1}^{\infty}, \dots, \{X(k, t)\}_{t=1}^{\infty}$ are independent, and that $\sup_{j,t} \mathbb{E}|X(j, t)| < \infty$. We shall be dealing with random variables $k(j, t)$ that depend upon the sequence $X(j, 1), X(j, 2), \dots$ only through the values of $X(j, 1), \dots, X(j, t-1)$, though possibly on some of the $X(m, n)$ for $m \neq j$. Such random variables are measurable with respect to the σ -field generated by $X(j, 1), \dots, X(j, t-1)$, and $X(m, n)$ for $m \neq j$ and $n = 1, 2, \dots$. We denote this σ -field by $\mathcal{F}(j, t-1)$.

Any decision rule that at each stage chooses an arm that has the highest Gittins index is called a Gittins index rule.

Theorem 4. *For a k -armed bandit problem with independent arms and geometric discount, any Gittins index rule is optimal.*

Proof. Suppose that $\Lambda_1 = \max \Lambda_j$. Let A be an arbitrary decision rule. We prove the theorem by showing that there is a rule A' that begins with arm 1 and gives at least as great a value as A .

Let $k(j, t)$ denote the (random) time that A uses arm j for the t th time, $j = 1, \dots, k$, $t = 1, 2, \dots$, with the understanding that $k(j, t) = \infty$ if arm j is used less than t times. The value of A may then be written

$$\begin{aligned} V(A) &= \mathbb{E} \left\{ \sum_{t=1}^{\infty} \beta^{t-1} Z(t) \mid A \right\} \\ &= \mathbb{E} \sum_{j=1}^k \sum_{t=1}^{\infty} \beta^{k(j,t)-1} X(j, t). \end{aligned}$$

Let N^* denote the stopping rule that achieves the supremum in

$$\Lambda_1 = \sup_{N \geq 1} \mathbb{E} \left(\sum_1^N \beta^{t-1} X(1, t) \right) / \mathbb{E} \left(\sum_1^N \beta^{t-1} \right),$$

and let T denote the (random) time that A uses arm 1 for the N^* th time, $T = k(1, N^*)$. Define the decision rule A' as follows:

- (a) use arm 1 at times $1, 2, \dots, N^*$, and then if $N^* < \infty$,
- (b) use the arms $j \neq 1$ at times $N^* + 1, \dots, T$ in the same order as given by A , and then if $T < \infty$,
- (c) continue according to A from time T on.

Let $k'(j, t)$ denote the time when A' uses arm j for the t th time, so that $k'(1, t) = t$ for $t = 1, \dots, N^*$. Finally, let $m(j)$ denote the number of times that arm j is used by time T , so that $m(1) = N^*$. Then,

$$\begin{aligned}
 V(A') - V(A) &= \mathbb{E} \sum_{j=1}^k \sum_{t=1}^{m(j)} (\beta^{k'(j,t)-1} - \beta^{k(j,t)-1}) X(j, t) \\
 &= \mathbb{E} \sum_{t=1}^{N^*} (\beta^{t-1} - \beta^{k(1,t)-1}) X(1, t) \\
 &\quad - \mathbb{E} \sum_{j=2}^k \sum_{t=1}^{m(j)} \beta^{t-1} (\beta^{k(j,t)-t} - \beta^{k'(j,t)-t}) X(j, t) \\
 &= \mathbb{E} \sum_{t=1}^{N^*} \xi(t) \beta^{t-1} X(1, t) - \sum_{j=2}^k \mathbb{E} \sum_{t=1}^{m(j)} \alpha(j, t) \beta^{t-1} X(j, t)
 \end{aligned} \tag{12}$$

where

$$\begin{aligned}
 \xi(t) &= 1 - \beta^{k(1,t)-t}, \quad \text{and} \\
 \alpha(j, t) &= \beta^{k(j,t)-t} - \beta^{k'(j,t)-t}.
 \end{aligned}$$

Since $k(1, t) - t$ represents the number of times that an arm other than arm 1 has been pulled by the time the t th pull of arm 1 occurs, we have that $k(1, t) - t$ is $\mathcal{F}(1, t - 1)$ -measurable and nondecreasing in t a.s. so that $\xi(t)$ is $\mathcal{F}(1, t - 1)$ -measurable and $0 \leq \xi(1) \leq \xi(2) \leq \dots \leq 1$ a.s. Thus from Lemma 3, we have

$$\mathbb{E} \sum_{t=1}^{N^*} \xi(t) \beta^{t-1} X(1, t) \geq \Lambda_1 \mathbb{E} \sum_{t=1}^{N^*} \xi(t) \beta^{t-1}. \tag{13}$$

For $j > 1$, $\alpha(j, t) = \beta^{k(j,t)-t} (1 - \beta^{k'(j,t)-k(j,t)})$. Since $k(j, t) - t$ is $\mathcal{F}(j, t - 1)$ -measurable and nondecreasing and since $k'(j, t) - k(j, t)$ is equal to N^* minus the number of times arm 1 is pulled before the t th pull of arm j , and hence is $\mathcal{F}(j, t - 1)$ -measurable and nonincreasing, we find that $\alpha(j, t)$ is $\mathcal{F}(j, t - 1)$ -measurable and $1 \geq \alpha(j, 1) \geq \alpha(j, 2) \geq \dots \geq 0$. Hence from Lemma 2, we have for $j = 2, \dots, k$,

$$\mathbb{E} \sum_{t=1}^{m(j)} \alpha(j, t) \beta^{t-1} X(j, t) \leq \Lambda_j \mathbb{E} \sum_{t=1}^{m(j)} \alpha(j, t) \beta^{t-1}. \tag{14}$$

Combining (13) and (14) into (12) and recalling that $\Lambda_j \leq \Lambda_1$ for all $j > 1$, we find

$$V(A') - V(A) \geq \Lambda_1 \mathbb{E} \sum_{j=1}^k \sum_{t=1}^{m(j)} (\beta^{k'(j,t)-1} - \beta^{k(j,t)-1}).$$

This last expectation is zero since it is just $V(A') - V(A)$ with all payoffs put equal to 1.

■