Rochester Center for
Economic Research

Two Index Theorems for Bandit Problems

Jeffrey S. Banks   Rangarajan K. Sundaram

Working Paper No. 278
May 1991

University of
Rochester

Two Index Theorems for Bandit Problems

Jeffrey S. Banks and Rangarajan K. Sundaram

Rochester Center for Economic Research
Working Paper No. 278

May 1991

Two Index Theorems for Bandit Problems

Jeffrey S. Banks

and

Rangarajan K. Sundaram

University of Rochester

and

The Rochester Center for Economic Research
Working Paper No. 278

May 1991

Abstract: We prove that optimal index strategies continue to exist in independent–armed Bandit problems under geometric discounting, when either (a) there is a cost for switching between arms, or (b) the probability that an arm may "die" while in use is non–zero.

*The first author gratefully acknowledges financial support from the Sloan Foundation and the National Science Foundation.

1. Introduction and Summary

This paper studies the class of Bandit problems (cf. Berry and Fristedt, 1985) in which there are an arbitrary (finite) number of independent arms, and discounting is geometric over an infinite horizon. The standard version of this framework is detailed in Section 2. It is well–known that optimal strategies always exist in this model; indeed, the optimal strategy may be chosen to be an ***index strategy***. That is, it is possible to attach an index to each arm, with the index on an arm depending solely on the characteristics of that arm, which has the property that the arm with the highest index is an optimal choice at each point in time[1]. The construction of this index, which is known as the ***Dynamic Allocation Index (DAI)***, or the ***Gittins Index***, is described in section 2.3.

We consider two generalizations of this framework. In section 3, we consider the case where the decision–maker is required to pay a fixed cost for switching between arms;[2] this cost is allowed to be arm–specific. We define a modified version of the DAI for this problem in section 3.3. This modified index coincides with the DAI if all switching costs are zero. In section 3.4, we demonstrate that the strategy defined by this index is, indeed, optimal in the class of all strategies for the problem with switching costs.

In section 4, we consider the possibility that arms may "fail" or "die" with non–zero probability when in use, i.e., that they might switch to an absorbing state in which they produce a constant reward. This probability is allowed to be arm–specific. In section 4.2, we prove that an optimal index strategy exists in this case also. Indeed, as we explain there, the existence of an optimal index strategy in this case follows almost immediately from Whittle's (1982) proof of the Gittins–Jones Theorem once the

[1] This result was first proved by Gittins and Jones (1974). For details and extensions of this result, see Berry and Fristedt (1985), Whittle (1982), or Gittins (1989).

[2] Kolonko and Benzing (1983) have studied the special case of this setup where there are two arms, with one arm generating payoffs according to a known distribution, and the other according to a Bernoulli distribution with unknown mean. As Berry and Fristedt (1985, p.239) point out, the Bandit problem in this special case is just one of determining an optimal stopping rule.

appropriate spaces have been defined.

We note that an index theorem continues to hold if these models are combined, namely, when there is a cost of switching arms, and there is a non–zero probability of the failure of an arm.

These results are of special interest in a labor–market setting where the Bandit framework has found wide applicability. In one interpretation, the decision–maker in a Bandit problem is viewed as a worker searching over jobs, and the arms of the Bandit represent the set of available jobs. An alternative version models the decision–maker as an employer searching over potential employees (the arms). In either case, the payoffs from an arm indicate the fitness of the resulting "match"[3]. Allowing for a cost of changing jobs (or employees), and admitting the possibility of a unforeseen termination of the relationship (say, the bankruptcy of the firm, or the resignation of the worker) seem necessary extensions of the basic model in this context.

2. Bandit Problems: A Description

This section is divided into three parts. Subsection 2.1 sets up the notation and gathers preliminary definitions. Subsection 2.2 outlines the standard framework of Bandit problems. In subsection 2.3, the Theorem of Gittins and Jones (1974) is reviewed. Since both the Bandit framework and the Gittins–Jones Theorem are quite well–understood, our descriptions are necessarily terse. For greater detail than we provide, we refer the reader to Berry and Fristedt (1985), Gittins (1989), or Whittle (1982).

2.1. Notation and Definitions

$\mathfrak{D}$ will denote the space of probability measures on the real line $\mathbb{R}$, and $\mathfrak{D}^n$ the set of

[3] See Mortensen (1985) for more on the use of Bandits in the study of labor–markets. Our companion paper (Banks and Sundaram, 1991) also contains a discussion of various applications of the Bandit framework in economics and other fields.

ordered n–tuples of members of $\mathfrak{D}$, where n is any positive integer. A generic element of $\mathfrak{D}$ [resp. $\mathfrak{D}^n$] is denoted by p [resp. $P = (p_1,...,p_n)$]. $\mathfrak{D}$ is given the topology of convergence in distribution, and $\mathfrak{D}^n$ the corresponding product topology. It is well–known that under this topology $\mathfrak{D}$ inherits many of the topological features of $\mathbb{R}$: $\mathfrak{D}$ is separable, metrizable, complete, and its topology is determined by convergent sequences and their limits. The Borel field on $\mathfrak{D}$, and the product σ–field (of the Borel field on $\mathfrak{D}$) on $\mathfrak{D}^n$, are the only σ–fields that are used throughout.

The set of probability distributions on $\mathfrak{D}^n$ is denoted by $\mathfrak{D}(\mathfrak{D}^n)$, with generic element G. $\mathfrak{D}(\mathfrak{D}^n)$ is also given the topology of convergence in distribution. The Borel σ–field of $\mathfrak{D}(\mathfrak{D}^n)$ is again the only σ–field considered throughout.

2.2. Bandit Problems

A Bandit problem is a sequential scheduling problem in which a single decision–maker (hereafter, referred to as the ***principal***) faces K parallel projects, called the "arms" of the Bandit, indexed k = 1,...,K. At each instant t of discrete time, only one arm may be activated by the principal. The activation of an arm results in instant payoffs to the principal from that arm. However, the "true" payoff distribution associated with some or all of the arms may be a priori unknown to the principal, who begins, instead, with a prior specifying the initial belief regarding these distributions. Observations accumulated over the course of play are used to update the initial belief over these distributions. The principal's objective is to maximize the expected discounted payoff over the horizon of the model, by choosing a rule for the engagement of the projects in each period.

More formally, a ***Bandit problem*** is described by (i) an arbitrary finite number of arms K, indexed by k; (ii) a ***discount sequence*** $D = (\delta_0, \delta_1, ...)$, with the interpretation that the principal discounts period–t rewards by δ_t; and (iii) a distribution $G \in \mathfrak{D}(\mathfrak{D}^K)$, representing the principal's ***initial belief*** or the ***prior***.

We place the following conditions on this framework that are maintained throughout the paper:

(A1) The discount sequence is ***geometric***: $\delta_t = \delta^t$, for some $\delta \in [0,1)$.

(A2) The arms are ***independent***: the initial belief $G = G_1 x ... x G_K$; a generic element in the support of G_k will be denoted p_k.

(A3) Rewards are uniformly bounded[4]: there is a real number A such that for each k, it is the case that with G_k–probability one

$$\left[\int_{\mathbb{R}} |r| dp_k(r)\right] \leq A. \tag{2.1}$$

Let $\mathfrak{X}$ denote the subset of all distributions in $\mathfrak{D}$ that satisfy (A3); $\mathfrak{X}^K$ the set of ordered K–tuples of members of $\mathfrak{X}$; and $\mathfrak{D}(\mathfrak{X}^K)$ the set of all distributions on $\mathfrak{X}^K$ satisfying (A3). Since the Bandit problem is fully specified by $\delta \in [0,1)$ and $G \in \mathfrak{D}(\mathfrak{X}^K)$, we will refer to the Bandit as the (G,δ) Bandit. Formal measure theoretic details are omitted in the sequel; for these we direct the reader to Berry and Fristedt (1985, Ch.2).

Conditional Distributions:

For any $G \in \mathfrak{D}(\mathfrak{X}^K)$, let G(k,r) denote a version of the conditional probability distribution that arises when the reward r is witnessed on arm k; and let $G_k(r)$ denote a version of the marginal G_k of G when the reward r is witnessed from the arm k. Of course, since $G = G_1 x ... x G_K$, so $G(k,r) = G_1 x ... x G_k(r) x ... x G_K$. We note the important point that for each k, G(k,r) can be chosen to depend measurably on (r,G) (see, e.g., Berry and Fristedt, 1985, Lemma 2.2.1).

[4] This uniform boundedness assumption is stronger than required, but makes for considerable expositional simplification. In particular, it suffices to assume (see Berry and Fristedt, 1985, Ch.2) that each component p_k of $(p_1,...,p_K)$ has finite first absolute moment with G–probability one; and that this moment has finite G–expectation.

Histories and Strategies:

A ***t–history*** for the (G,δ) Bandit is a specification of the arms used in each period $\tau \in \{0,1,...,t-1\}$, and the consequent rewards witnessed. Let $H_0 = \phi$, and for $t \geq 1$, let H_t be the set of all t–histories for the Bandit. A ***strategy*** σ for the principal may be defined in one of two ways.

In the usual definition, σ specifies an action for the principal for each possible history. That is, σ is a sequence of measurable maps $\{\sigma_t\}$ such that $\sigma_0 \in \{1,...,K\}$ and for $t \geq 1$, $\sigma_t: H_t \rightarrow \{1,...,K\}$. An alternative definition of σ is based on the observation that each 1–history (k,r) from an arbitrary initial prior $G \in \mathfrak{D}(\mathfrak{X}^K)$ results in the new Bandit $(G(k,r),\delta)$. Hence, σ could be thought of as specifying an arm k to be played for each $F \in \mathfrak{D}(\mathfrak{X}^K)$, with two additional conditions appended. First, σ must depend measurably on $F \in \mathfrak{D}(\mathfrak{X}^K)$. Second, since any particular observation may have prior probability zero, versions of conditional distributions to be used must be chosen in advance. That these definitions are then equivalent is intuitively apparent; a formal proof may be found in Berry and Fristedt (1985).

Rewards and Optimal Strategies:

Each strategy σ defines in the obvious manner an expected period–t reward $r_t(\sigma;G)$ from the prior G. Hence, each strategy σ also defines a total expected reward $W(\sigma;G)$ from the prior G defined as[5]

$$W(\sigma;G) = \Sigma_{t=0}^{\infty}\, \delta^t r_t(\sigma;G). \qquad (2.2)$$

The objective is to find a strategy σ^* such that $W(\sigma^*;G) \geq W(\sigma;G)$ for all other strategies σ. When such an strategy exists it will be termed an ***optimal strategy***, and $W(\sigma^*;G)$ will

[5] W obviously depends on δ as well as G. However, since all our results are valid for all $\delta \in [0,1)$, we avoid complicating notation and suppress dependence on δ throughout.

be termed the ***value*** of the (G,δ) Bandit, and will be denoted simply by $V(G)$.

Existence of Optimal Strategies:

It is well–known[6] that in this class of Bandit problems, optimal strategies exist from any prior $G \in \mathfrak{D}(\mathfrak{X}^K)$. Moreover, the value function $V(.)$ satisfies

$$V(G) = \vee_{k=1}^{K} L_k V(G), \tag{2.3}$$

where

$$L_k V(G) = \int_{\mathfrak{X}} \int_{\mathbb{R}} \left[r + \delta V(G(k,r)) \right] p_k(dr) G_k(dp_k). \tag{2.4}$$

Indeed, V may be explicitly recovered using the following dynamic programming argument[7]. Let $\mathfrak{D}(\mathfrak{X}^K)$ denote the ***state space*** of the problem, and $\{1,...,K\}$ the ***action space***. ***Transitions*** from current states and actions into distributions over future states are implicitly defined through the (measurable) map that for each k, takes (r,G) into the conditional probability distribution G(k,r). Now, let $\mathfrak{C}$ be the space of all bounded measurable functions w from $\mathfrak{D}(\mathfrak{X}^K)$ to $\mathbb{R}$. Define the mapping $T:\mathfrak{C} \to \mathfrak{C}$ by

$$Tw(G) = \vee_{k=1}^{K} L_k w(G)$$

where, once again,

$$L_k w(G) = \int_{\mathfrak{X}} \int_{\mathbb{R}} \left[r + \delta w(G(k,r)) \right] p_k(dr) G_k(dp_k).$$

Endow $\mathfrak{C}$ with the sup–norm metric topology. Then, $\mathfrak{C}$ is evidently a complete metric space. That T is a contraction is easy to see. Consequently, T has a unique fixed point w^*

[6] See, for instance, Berry and Fristedt (1985).

[7] This argument implicitly invokes the equivalence between the given Bandit problem (which involves unknown parameters), and a Markovian dynamic programming problem (which does not). For a proof of this equivalence, see, for example, Rieder (1975), or Schael (1979).

such that $Tw^* = w^*$. By (2.3), we also have $TV = V$, so by the uniqueness of the fixed–point $V = w^*$.

The optimal strategy may also be obtained via this argument. Let H denote the correspondence of maximizers in (2.4), and let h be any measurable selection from H. Then, the strategy, which at any prior G picks the action h(G) is an optimal strategy.

2.3. The Dynamic Allocation Index

In this subsection, we describe a fundamental result due to Gittins and Jones (1974), which provides a sharper characterization of the optimal strategy in the class of Bandit problems identified above, than that provided by the dynamic programming argument. Fix a Bandit (G,δ) meeting the assumptions listed above. Pick an arm $k \in \{1,\dots,K\}$, and consider the optimal stopping problem in which the principal's sole choice in each period is to play the arm k for one more period, or to accept a "terminal reward" of $m \in \mathbb{R}$. Similar dynamic programming arguments to that used above show that the value to the principal of this stopping problem, denoted $V(.,m)$, can be obtained as the unique fixed point of a suitably defined contraction mapping; and that V satisfies

$$V(G_k;m) = \max\left[m, \int_{\mathfrak{D}}\int_{\mathbb{R}}\{r + \delta V[G_k(r);m]\}p_k(dr)G_k(dp_k)\right] \qquad (2.5)$$

where $G_k(r)$ is a version of the conditional probability distribution that arises from the prior G_k when the observation r is witnessed on arm k. The ***dynamic allocation index*** (DAI) on arm when the prior on the arm is G_k, denoted $\mu_k(G_k)$, is then defined as

$$\mu_k(G_k) \;=\; \inf\{m \mid V(G_k;m) = m\} \qquad (2.5)$$

Since rewards are uniformly bounded, it is evident that for m sufficiently large (say, $m > A$), we have $V(G_k;m) = m$; while for m sufficiently small (say, $m < -A$), we have $V(G_k;m) > m$. The DAI is, therefore, well–defined. The following result shows the importance of the DAI in Bandit problems:

Theorem 0 *(Gittins and Jones, 1974): The optimal initial selections in the* (G,δ) *Bandit are those arms* k *for which*

$$\mu_k(G_k) = \vee_{l=1}^{K} \mu_l(G_l). \qquad (2.6).$$

Remark: The importance of independent arms for this result is obvious. What is less apparent is the result in Berry and Fristedt (1985, Ch.6) that geometric discounting is, within broad limits, ***necessary*** for the validity of this result.

Whittle (1982) contains a proof of Theorem 0. In fact, Whittle defines the Bandit problem directly in terms of the Markovian dynamic programming problem, without any reference to the notion of prior beliefs[8]. The idea of his proof, which we adopt in sections 3 and 4, is to show that the total payoff from the DAI strategy satisfies equation (2.4). An appeal to the uniqueness of the fixed–point defining (2.4) then completes the argument.

3. Bandit Problems with Switching Costs

In this section we consider a generalization of the Bandit framework of section 2, by admitting the possibility of non–zero switching costs. Evidently, there are two ways of modelling switching costs. One option is to consider the situation where there is a fixed cost $c_k \geq 0$ of switching ***away*** from arm k, regardless of which arm is switched to. The other option is to consider a fixed cost $c_k \geq 0$ of switching ***to*** arm k, regardless of from which arm the switch occurs. We consider the former situation in this section. However, we note that when the switching costs are not arm–specific (i.e., we have $c_k = c$ for all k), then the two problems are formally equivalent. All the other assumptions of Section 2 are

[8] We note also that Berry and Fristedt (1985) provide a proof of Theorem 0 using Whittle's arguments, but without an explicit appeal to dynamic programming considerations.

maintained.

The definitions of histories and strategies remain unchanged from the previous section. The total expected payoff from a strategy σ requires the obvious modification that each time the strategy requires a change in the arm in use, the appropriate fixed cost has to be deducted. In section 3.1, we prove the existence of an optimal strategy for this problem. Subsection 3.2 consists of an example that explains why if an index theorem is to hold in this class of problems, the index cannot depend only on the characteristics of that arm. Subsection 3.3 then defines a modified version of the DAI for this problem. Finally, in subsection 3.4, we prove that the strategy defined by this modified index is an optimal strategy.

3.1 Existence of an Optimal Strategy

Once again, we exploit the equivalence between the given Bandit problem and a suitably defined Markovian dynamic programming problem. When switching costs are non–zero, the state of the Bandit problem at any time cannot adequately be described by the prior belief G. Rather, it is also important to know the arm that was used in the previous period. Therefore, letting Δ denote the set of all arms $\{1,\ldots,K\}$, we define $\Omega = \mathfrak{D}(\mathfrak{X}^K) \times \Delta$, to be the ***augmented state space***. Let $\mathfrak{C}^*$ denote the set of all bounded, measurable functions from Ω to $\mathbb{R}$. Endow $\mathfrak{C}^*$ with the sup–norm metric topology. We first introduce some shorthand notation that simplifies exposition in the sequel. For each $H \in \mathfrak{C}^*$, and $(G,j) \in \Omega$, define

$$r_k(G_k) = \int_{\mathfrak{X}} \int_{\mathbb{R}} r p_k(dr) G_k(dp_k), \tag{3.1}$$

and

$$E[H(G')\,|\,G,j] = \int_{\mathfrak{X}} \int_{\mathbb{R}} H[G(j,r),j] p_j(dr) G_j(dp_j). \tag{3.2}$$

Now, define the mapping L from $\mathfrak{C}^*$ to itself by

$$LH(G,k) \;=\; \bigvee_{j=1}^{K} L_jH(G,k), \tag{3.3}$$

where, for $j \neq k$, we have

$$L_jH(G,k) \;=\; r_j(G_j) + \delta E[H(G')\,|\,G,j] - c_k \tag{3.4}$$

and

$$L_kH(G,k) \;=\; r_k(G_k) + \delta E[H(G')\,|\,G,k]. \tag{3.5}$$

It is easy to see that L is a contraction, and hence has a unique fixed–point, denoted say, V. Evidently, V satisfies

$$V(G,k) \;=\; \bigvee_{j=1}^{K} L_jV(G,k), \tag{3.6}$$

where L_jV is defined as in (3.4)–(3.5) with V replacing H.

Routine dynamic programming arguments now show that V(G,k) is the value of the Bandit problem with the given switching costs, when the prior belief is given by G, and the arm the decision maker used in the previous period is k; and that any measurable selection ψ from the correspondence of maximizers in (3.4) defines an optimal strategy for this problem[9].

3.2. An Example

Let I(a) represent the distribution which places point mass at $a \in \mathbb{R}$. Consider a two–armed Bandit in which the prior on arm 1 is $p_1I(1) + (1-p_1)\,I(0)$, and that on arm 2 is $p_2I(1) + (1-p_2)I(0)$, where $p_1,p_2 \in (0,1)$, and $p_1 > p_2$. Let the cost of switching away from arm i be c_i, where $c_2 = c_1 = c$, say. Finally, suppose that $c > p_2/(1-\delta)$. We will show that there cannot exist an index strategy which is also an optimal strategy if the

[9] In the very first period of the problem, when no arms have yet been tried, the value of begining with the prior G is $V^*(G) = \max_j L_jV(G,j)$, and any j that attains this maximum is an optimal initial selection.

index on arm i is to depend only on its own characteristics, i.e., on p_i and c.

It is easy to see that the uniquely optimal strategy in this problem is simply to play arm 1 forever. Now consider any indices $\mu_i(p_i,c)$ for the arms. Since the arms are identical up to the prior, we must have $\mu_1 = \mu_2 = \mu$, say. Moreover, if μ is to have any chance of being optimal, it must be monotone in p (i.e., we must have $\mu(p,c) > \mu(q,c)$ if $p > q$), since an arm with prior p is always then more attractive than an arm with prior q. Therefore, we must also have $\mu(0,c) < \mu(p_2,c)$.

But now it is easy to see that the strategy defined by μ cannot be optimal. For although the strategy recommends arm 1 in period 1, it recommends a switch to arm 2 in period 2 if the first period reward is 0. This continuation is suboptimal, since $c > p_2/(1-\delta)$, by assumption. □

3.3. A Modified Index

The reason an index on arm k cannot depend solely on G_k and c_k, is that these arguments, taken collectively over k, still do not sufficiently capture the "state" of the Bandit problem at a point in time. Namely, at any given time calculation of the optimal continuation strategy requires knowledge of the priors (G_k), the switching costs (c_k), ***and the arm j that was in use in the previous period.***

In this subsection we define a modified version of the DAI that does rely on this additional piece of information. We note that our modified index will coincide with the DAI of Gittins and Jones (1974) if all switching costs are zero.

So consider an arm k, and let G_k be the prior belief on k. Let $\Omega_k = \mathfrak{X} \times \Delta$. Fix a terminal reward $m \in \mathbb{R}$. We introduce more shorthand notation in the spirit of (3.1)–(3.2). For any bounded measurable function H_k defined on Ω_k, let

$$EH_k[G_k' \mid G_k] = \int_{\mathfrak{X}} \int_{\mathbb{R}} H_k[G_k(r),k;m]p_k(dr)G_k(dp_k). \qquad (3.8)$$

Once again, using contraction mapping arguments it can be shown that there is a unique bounded measurable function $V_k(.;m)$ on Ω_k that satisfies

$$V_k(G_k,j;m) \;=\; \max\{m - c_j,\; L_kV_k(G_k,j,;m)\} \tag{3.9}$$

where, for $k \neq j$, we have

$$L_kV_k(G_k,j;m) = r_k(G_k) + \delta EV_k[G_k' \,|\, G_k] - c_j \tag{3.10}$$

and

$$L_kV_k(G_k,k;m) = r_k(G_k) + \delta EV_k[G_k' \,|\, G_k]. \tag{3.11}$$

Now define for each $(G_k,j) \in \Omega_k$,

$$\mu_k(G_k,j) = \inf\{m \,|\; V_k(G_k,j;m) \;=\; m - c_j\}. \tag{3.12}$$

For m large (say, $m \geq A + \max c_j$), we clearly have $V_k(.;m) = m$, while for m sufficiently small (say, $m < -A$), we also have $V_k(.;m) > m$. Therefore, $\mu_k(.)$ is well–defined. We will refer to $\mu_k(G_k,j)$ as the ***modified index on arm k*** when the state of the Bandit problem is given by (G,j). We note that for any k and j, $\mu_k(G_k,j)$ also depends on c_k and c_j. In the interests of notational brevity, we suppress this dependence.

3.4. Optimality of the Modified Index Strategy

We will prove the following Theorem that establishes the optimality of following the prescriptions of the modified index.

Theorem 1 *The optimal choices in the Bandit problem with switching costs at the state* (G,j) *are those arms* k *for which*

$$\mu_k(G_k,j) \;=\; \sup_{i \in \Delta} \mu_i(G_i,j). \tag{3.13}$$

The proof of Theorem 1 follows an analogous procedure to that used by Whittle (1982) in proving Theorem 0. For expositional ease, we organize the proof in a series of steps. First, we define a modified version of the original Bandit problem, which is parametrized by $M \in \mathbb{R}$. For "low" values of M, the value function $\phi(.;M)$ of the modified problem is independent of M, and coincides with the value function V of the original problem. Next, for each $M \in \mathbb{R}$, a function $U(.;M)$ is defined on Ω using the functions V_k of section 3.2. Then, it is shown that we have $U(.;M)$ is identically equal to $\phi(.;M)$. Hence, U coincides with V for "low" values of M. Finally, it is shown using properties of the function U, that an arm k is a maximum of the RHS of (3.6) at a state (G,j) if, and only if, (3.13) is met.

Step 1:

We begin with a modification of the original Bandit problem. Specifically, we add a (K+1)st arm to the Bandit, where this additional arm pays a constant reward of $M \in \mathbb{R}$. The cost of switching away from this arm is zero[10]. The usual arguments show that the value function $\phi(.;M)$ of this modified problem satisfies

$$\phi(G,k;M) = \max\{M - c_k, \max_{j \in \Delta} L_j\phi(G,k;M)\} \tag{3.14}$$

where, as usual,

$$L_j\phi(G,k;M) = r_j(G_j) + \delta E[\phi(G';M) \mid G,j] - c_k \tag{3.15}$$

and

$$L_k\phi(G,k;M) = r_k(G_k) + \delta E[\phi(G';M) \mid G,k] \tag{3.16}$$

Observe the important point that the uniform boundedness of rewards implies the existence of constants B and C such that for all $(G,k) \in \Omega$, we have (i) $\phi(G,k;M) = M$ for all $M \geq B$, and (ii) $\phi(G,k;M) = V(G,k)$ for all $M \leq C$, where V is the value function defined in section 3.1. Note also that if $M \geq B$, we must have $M \geq \mu_i(G_i,k)$ for any G_i and k.

Step 2:

Easy arguments, akin to those used in Whittle (1982, Theorem 14.2.1), or Berry and

[10] Note that the switching cost on this arm is really irrelevant, for if it is ever optimal to play this arm at any point, this must remain an optimal choice forever, since no new information is obtained by playing this arm.

Fristedt (1985, Lemma 6.1.2) show that for each k, the function $V_k(G_k,j;M)$ is non–decreasing and convex as a function of M for each fixed $(G_k,j) \in \Omega_k$. Consequently, the derivative with respect to M exists almost everywhere. For notational ease, let $v_k(G_k,j,;M) := \partial V_k(G_k,j;M)/\partial M$. For completeness, we shall assume that at points of ambiguity, v_k is given the value of the right derivative.

Step 3:

Define the function

$$U(G,k;M) \;=\; B \;-\; \int_M^B [\, \Pi_{j=1}^K v_j(G_j,k;m)]dm \;-\; c_k. \tag{3.14}$$

Fix any $j \in \Delta$. Define $D_j(G,k;M) := \Pi_{i \neq j} [v_i(G_i,k;M)]$. Note that for any $i \in \Delta$, $M \geq \mu_i(G_i,k)$ implies $v_i(G_i,k;M) \equiv 1$, by definition of $\mu_i(.)$. Therefore, for $M \geq \max_{i \neq j} \mu_i(G_i,k)$, we have $D_j(G,k;M) = 1$, and $d_m D_j(G,k;M) = 0$. Integrating (3.14) by parts, and noting that $V_j(.;B) = B - c_k$, we obtain

$$U(G,k;M) = V_j(G_j,k;M).D_j(G,k;M) + \int_M^\infty V_j(G_j,k;m)d_m D_j(G,k;m),$$

From the convexity of $V_i(G_i,k;.)$ in M, it follows now that

$$U(G,k;M) \geq M - c_k, \text{ with equality if } M \geq \max_i \mu_i(G_i,k).$$

Step 4:

We now show that the function U of step 3 is the same as the function ϕ of step 1; and that i maximizes the RHS of (3.14) whenever (3.13) holds. To this end, we hold (G,k) fixed, and suppress dependence of all functions on these arguments. Thus, $V_j(G_j,k;M)$ is written simply as $V_j(M)$, $\mu_j(G_j,k)$ as μ_j, $E[\phi(G';M)\,|\,G,k]$ as $E[\phi(M)]$, etc. Now, let

$$f_j(M) \;=\; V_j(M) - L_j V_j(M). \tag{3.17}$$

Note that $f_j \geq 0$, and $f_j(M) = 0$ if $M \leq \mu_j$. First consider the case when $j \neq k$. We have

$$L_jU(M) \;=\; r_j - c_k + \delta E[V_j(M)D_j(M)] + \delta E[\textstyle\int_M^\infty V_j(m)d_mD_j(m)].$$

Using the fact that $D_j(.)$ does not depend on G_j, we obtain

$$\begin{aligned} U(m) - L_jU(m) \quad &= c_k - r_j + \delta D_j(M).[V_j(M) - E[V_j(M)]] \\ &\qquad + \delta \textstyle\int_M^\infty [V_j(m) - E[V_j(m)]]d_mD_j(m). \end{aligned}$$

Since, $D_j(m) + \int_M^\infty d_mD_j(m) = 1$, this gives us

$$\begin{aligned} U(M) - L_jU(M) \quad &= D_j(M)f_j(M) + \textstyle\int_M^\infty f_j(m)d_mD_j(m) \\ &\geq 0. \end{aligned} \tag{3.19}$$

Now note that if $m \leq \mu_j$, then $f_j(m) = 0$; while, if $m \geq \max_{i \neq j} \mu_i$, then we also have $d_mD_j(m) = 0$. Putting these together, we obtain

$$U(M) - L_jU(M) = 0, \qquad \text{if } M_j \geq \max\{M,\ \max_{i\neq j} \mu_i\}. \tag{3.20}$$

An identical argument shows that for $j = k$, we again have $U(m) - L_jU(m) \geq 0$, with inequality if $\mu_j \geq m$, and $\mu_j \geq \mu_i$, for all $i \neq j$.

We now return to full notation. Recall that we have $U(G,k;M) \geq M - c_k$, with equality if $M \geq \max_{i\in\Delta} \mu_i(G_i,k)$ (see step 3). Together with (3.19)–(3.20), this implies

$$U(G,k;M) \;=\; \max\{M - c_k,\ \ \max_{j\in\Delta} L_jU(G,k;M)\} \tag{3.21}$$

where, letting $\mu^*(G,k,M) \;:=\; \max_{i\in\Delta} \{M, \max_{i\in\Delta} \mu_i(G_i,k)\}$,

$$\begin{aligned} U(G,k;M) \quad &= M - c_k, &&\text{if } M = \mu^*(G,k,M) &(3.22)\\ &= L_jU(G,k;M), &&\text{if } \mu_j(G_j,k) = \mu^*(G,k,M) &(3.23) \end{aligned}$$

Equation (3.21) implies that U must coincide with ϕ, for the contraction mapping defining ϕ has a unique solution. But ϕ itself coincides with V when $M \leq C$ (see step 1). Together with (3.22) – (3.23), this completes the proof of Theorem 1. □

4. Bandit Problems with Stochastic Termination

We now turn to the second of our generalizations of the model of section 2. We assume in this section that at the end of each period in which arm $k \in \{1,\dots,K\}$ is in use, there is a probability $\lambda_k \in [0,1]$ that it will "fail" or "die", i.e., transit into an absorbing state in which it produces a reward of a_k forever[11]. We assume that failure is observable by the principal, i.e., the principal recognizes immediately the fact that failure has occurred. Formally, let I(a) represent the distribution with point mass at a, for any $a \in \mathbb{R}$. Suppose $G \in \mathfrak{D}(\mathfrak{X}^K)$ is the prior at the beginning of a period in which arm k is the chosen arm, and suppose that the reward r is witnessed from arm k. Then, with probability $(1-\lambda_k)$ the new posterior on arm k will be $G_k(r)$; and with probability λ_k the arm will "fail" at the end of the period and the new posterior will be $I(a_k)$[12].

In section 4.1, we show the existence of an optimal index strategy in this class of problems. Indeed, this result is an immediate consequence of Whittle's (1982) proof of Theorem 0, once we have defined the state space of the problem appropriately. Section 4.2 then shows the importance of assuming that an arm fails only when it is in use. A counterexample there establishes that if failure can occur even when arms are not in play, then the use of index strategies may be suboptimal.

[11] As will become apparent, it suffices that in the absorbing state the arm produce rewards from any distribution whose mean is a_k.

[12] With transparent modifications of the arguments, it is easily seen that an index theorem is valid even if the arm may fail while in use, i.e., upon selection it generates a reward from the support of G_k with probability $(1-\lambda_k)$, and generates a reward a_k with probability λ_k.

4.1. Existence of an Optimal Index Strategy

We define the relevant probability spaces first in order to set up the above problem as a dynamic programming problem. Let $s_k \in \{0,1\}$ be an index that indicates if arm k is still active or not, where $s_k = 0$ implies the arm has not yet failed. In this notation, the state of arm k at any time can be represented by $(G_k,s_k) \in \mathcal{X} \times \{0,1\}$. The reward from arm k, denoted $R_k(G_k,s_k)$ is given by

$$\begin{aligned} R_k(G_k,s_k) &= r(G_k), \qquad \text{if } s_k = 0, \\ &= a_k, \qquad \text{if } s_k = 1. \end{aligned} \tag{4.1}$$

The state of arm k changes, of course, only when arm k is employed. In this case, there are two possibilities. If $s_k = 1$, then with probability one, the state remains at (G_k,s_k). While, if $s_k = 0$, and the reward r is witnessed from the arm that period, then with probability $(1-\lambda_k)$ the state will move to $(G_k(r),0)$, and with probability λ_k to $(G_k(r),1)$. Since the distribution of rewards r is calculable from knowledge of G_k, this gives rise in the obvious manner to the transition mechanism for (G_k,s_k). Denote this transition by $Q_k(\cdot\,|\,G_k,s_k)$. Finally, for notational convenience, let α_k denote a generic state (G_k,s_k) of arm k.

The dynamic programming problem is now easily defined. Let $\alpha := (\alpha_k)_{k\in\Delta}$ denote a generic state of the problem, and $\Delta = \{1,\ldots,K\}$ the action space. Since the state of arm k changes only when it is employed, the probabilities Q_k define the transition mechanism completely for this problem. Letting $(\alpha_{-k},\hat{\alpha}_k)$ denote the vector $\alpha = (\alpha_1,\ldots,\alpha_K)$ with α_k replaced by $\hat{\alpha}_k$, the usual methods now show that the value function V of this problem satisfies

$$V(\alpha) = \max_{j\in\Delta} L_jV(\alpha) \tag{4.2}$$

where

$$L_jV(\alpha) = R_j(\alpha_j) + \delta \int V(\alpha_{-k},\hat{\alpha}_k)Q_k(d\hat{\alpha}_k\,|\,\alpha_k). \tag{4.3}$$

We now define the indices $\mu_k(.)$ in the familiar way. Let $V_k(\alpha;M)$ denote the value of the stopping problem in which the initial state of arm k is α_k, and the principal is required in each period to choose between playing arm k one more period or accepting the terminal reward of M. Routine arguments show that the problem is well–defined. Now let

$$\mu_k(\alpha_k) \;=\; \inf\{M|\; V_k(\alpha_k;M) = M\}.$$

We have

Theorem 2 *The RHS of* (4.2) *is maximized by those* k *for which*

$$\mu_k(\alpha_k) \;=\; \max_{j\in\Delta} \mu_j(\alpha_j).$$

Proof: This is an immediate consequence of Whittle's (1982) proof of Theorem 0, which requires only that the problem is Markovian, and that the state of arm k change only when arm k is employed. □

4.2. An Example

When arms may "fail" even if not in use, the use of index strategies need not be optimal. Consider the following example in which there are two arms. Let $\delta = \frac{1}{2}$; $\lambda_1 = 1$, $\lambda_2 = 0$; and $a_1 = 0$. (a_2 is irrelevant since $\lambda_2 = 0$.) Suppose the initial belief on arm 1 places point mass at $\frac{1}{2}$; and that the belief on arm 2 is $pI(1) + (1-p)I(0)$. Direct calculation reveals that the index on arm 1 is $\frac{1}{2}$, while that on arm 2 is $4p/(1+p)$. For $p > \frac{1}{7}$, the index on arm 2 is strictly larger than $\frac{1}{2}$, so that an index strategy would indicate playing arm 2 in the first period. Since arm 1 fails at the end of period 1, the value of following the index strategy for $p > \frac{1}{7}$ is just $p/(1-\delta) = 2p$.

Now consider the strategy which begins with arm 1, and switches to arm 2 in period 2. The value of this strategy is evidently $[\frac{1}{2} + \delta 2p] = \frac{1}{2} + p$. Since $\frac{1}{2} + p > 2p$ whenever $p < \frac{1}{2}$, the index strategy is strictly suboptimal for $p \in (\frac{1}{7},\frac{1}{2})$. □

References

Banks, J.S. and R.K. Sundaram (1991) Denumerable–Armed Bandits, Working Paper, University of Rochester.

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

Gittins, J. (1989) ***Multi–Armed Bandit Allocation Indices,*** Wiley Intersciences, New York.

Gittins, J. and D. Jones (1974) A Dynamic Allocation Index for the Sequential Design of Experiments, in ***Progress in Statistics*** (J. Gani, et al, eds.), North Holland.

Kolonko and Benzing (1983) Sequential Design of Bernoulli Experiments Including Switching Costs, unpublished manuscript.

Mortensen, D. (1985) Job Search and Labor Market Analysis, in ***Handbook of Labor Economics,*** Vol.II (O. Ashenfelter and R. Layard, Eds.), North Holland, New York.

Whittle, P. (1982) ***Optimization over Time: Dynamic Programming and Stochastic Control*** Vol.I, Wiley, New York.