

Generalized Bandit Problems¹

Rangarajan K. Sundaram
Department of Finance
Stern School of Business
New York University
New York, NY 10012

May 27, 2003

¹The questions addressed in this paper grew out of my work with Jeff Banks on bandit problems and their applications (Banks and Sundaram (1992a, 1992b, 1994)) and owe much to many discussions I had with him on this subject. I also had the benefit of several discussions with Andy McLennan, especially regarding the material in Sections 4 and 6 of this paper.

1 Introduction

The framework of multi-armed bandit problems is among the most widely-used models for the study of optimal information acquisition and “learning” by economic agents. An important factor in this popularity is the considerable degree of analytical tractability possessed by even very general formulations of problems in this framework. In large part, this tractability derives from the remarkable Theorem of Gittins and Jones (1974), which asserts that in any independent-armed bandit problems with geometric discounting over an infinite horizon, it is possible to associate with each arm an index, known as the *Gittins index*, with the property that a strategy in the bandit problem is an optimal strategy if, and only if, it (almost) always involves playing an arm with the highest value of the Gittins index at that point. (Such a strategy is called a *Gittins index strategy*.) The feature that gives this result especial potency is that the Gittins index on an arm depends solely on characteristics of that arm (and on the rate of discounting), but not on any other feature of the problem under study.

The fact that indices of the arms can be computed separately and then put together to generate the optimal strategy makes the Gittins-Jones Theorem a powerful tool in the characterization of optimal plans. For instance, Banks and Sundaram (1992a) show that the Gittins index function possesses certain curvature and continuity properties. Exploiting these properties, they show that in a very large class of bandit problems, it is the case that with non-zero probability, the optimal strategy will involve the play of just a single arm of the bandit *forever*. Moreover, this need *not* be the “best” type of arm (i.e., one that would be optimal under complete information); indeed, it could even be the case that arms that are optimal under complete information will be discarded in finite time *with probability one*, under the optimal strategy. Similarly, Banks and Sundaram (1992b) demonstrate—once again by identifying and exploiting properties of the Gittins index function—the existence of a class of bandit problems in which optimal strategies are *myopic*, i.e., in which regardless of the rate of discounting, the optimal strategies remain the same as those that are optimal at a discount factor of zero. Since the latter are trivial to compute, characterizing optimal behavior is made simple.

Given the importance of the Gittins-Jones Theorem in characterizing optimal plans, a natural question that arises concerns the extent to which the assumptions of the theorem may be relaxed, without doing violence to its conclusions. A number of papers in the literature have addressed various aspects of this question. On the positive side, it has been shown that an index theorem continues to hold when new projects are continually appearing (Whittle, 1981); when project stages are variable and the bandit problem is a semi-Markovian decision problem (Gittins, 1979); for the scheduling of priorities in a queue (see Whittle, 1982, Ch. 14.12); and in problems that resemble the bandit framework such as the “Pandora’s Box” (or “treasure-hunt”) problem (Weitzman, 1979; Whittle, 1982, Ch.14.10). On the downside, the stationarity of the underlying decision problem is very important: Berry and Fristedt (1985) show that for the Gittins index to be able to identify optimal strategies in all bandit problems, it is not only sufficient that discounting be geometric, it is also *necessary*.

This paper summarizes and generalizes some recent work in this area, and presents some new

results on this issue. Three questions are studied, all motivated by economic considerations.

In a number of economic applications, notably in the field of job-search and matching, it is natural to posit the presence of an *infinite* number of available arms. In section 4, we examine the extension of the Gittins-Jones result to such a setting, from its original assumption of a finite number of available arms. When the number of arms is not finite, the Gittins index strategy may itself cease to be well-defined, i.e., there may exist a set of histories which occur with positive probability, such that after any of these histories, it is no longer possible to identify an arm that attains the supremum of the Gittins indices following that history. We begin, therefore, with the identification of a set of conditions that are both necessary and sufficient for the Gittins index strategy to be well-defined from a given initial state (Theorem 4.1). Then, we show (Theorem 4.2) that regardless of the cardinality of the set of arms, as long as the Gittins index strategy is well-defined from a given initial state, a strategy is optimal from that state if, and only if, it is a Gittins index strategy. To complete the analysis, we examine the converse question: if the Gittins index strategy is not well-defined, is it the case that an optimal strategy does not exist from that initial state? It appears a strong conjecture that the answer to this question is in the affirmative, but this paper provides only a partial answer (Theorem 4.3). Two conjectures, which are of independent interest, but will also help provide a more complete answer are stated.

The standard formulation of the bandit problem limits the decision-maker to only choosing one arm in each period. This precludes the possibility of “parallel search” by the decision-maker, in which several options may be learnt about simultaneously. In section 5, we examine a simple generalization of the bandit framework, which allows the decision-maker to play a fixed number $k \geq 2$ arms per period, and examine the validity of a conjecture that arises naturally: whether it is the case that optimal strategies involve the decision-maker playing the arms with the k highest values of the Gittins index in each period. Somewhat unfortunately, this conjecture turns out to be false. We describe a three-arm problem in which $k = 2$. At the given initial state in this example, arm 3 has the highest index, arm 2 the next highest, and arm 1 the lowest, but we show that beginning with arms 1 and 3 strictly dominates beginning with arms 2 and 3.

Finally, in section 6, we consider a generalization of the bandit framework that is of considerable interest for economic analysis: the introduction of a cost for switching between arms.¹ Indeed, it is difficult to imagine a relevant decision problem in which the decision-maker may move between alternatives in a costless manner. Since switching costs do not affect the model’s stationarity, it appears a strong conjecture that a suitably generalized version of the Gittins index will continue to identify all optimal strategies in this problem. Moreover, as Weitzman (1979) has shown, optimal index strategies do exist in the related problem of “Pandora’s Box” which involves switching costs. Unfortunately, this conjecture also turns out to be false: Banks and Sundaram (1994) show that in the presence of switching costs, it is no longer possible to define any index on the arms which will unfailingly identify optimal strategies. In section 6 of this paper, we provide an alternative proof of their theorem.

The rest of this paper is organized as follows. Section 2 describes a typical bandit problem and

¹Despite their obvious appeal, bandit problems with switching costs have not been studied very widely. The literature consists only of a few papers including Kolonko and Benzing (1983) and Agarwal, et al (1991).

identifies some useful classes of strategies in these problems; it also provides a brief description of some of the principal applications of the bandit framework to economic analysis. Section 3 describes the construction of the Gittins index, and reviews the Gittins-Jones Theorem on the optimality of Gittins index strategies. Then, as stated above, section 4 focusses on infinite-armed bandits; section 5 on bandits with multiple plays allowed in each period; and section 6 on bandit problems with switching costs. The Appendix contains proofs of results omitted in the main text.

2 Bandit Problems

2.1 Description of the Framework

An independent-armed bandit problem with geometric discounting (hereafter, simply *bandit problem*) is specified by the following objects:

1. A finite set $I = \{1, \dots, n\}$ of *arms* of the bandit, with generic element i .
2. A tuple $C_i = (X_i, r_i, q_i)$ for each arm i , where
 - (a) X_i , a Borel subset of some Polish (i.e., complete, separable, metric) space, describes the set of possible *states* of arm i ;
 - (b) $r_i : X_i \rightarrow \mathbb{R}$ is a bounded measurable function describing the instantaneous reward from arm i ; and
 - (c) q_i represents a family of transition probabilities on X_i , i.e., for each $x_i \in X_i$, $q_i(\cdot | x_i)$ is a probability measure on X_i , and for each fixed Borel subset G of X_i , $q_i(G | \cdot)$ is a measurable function from X_i to $[0, 1]$.
3. A *discount factor* $\delta \in [0, 1)$.

We will denote such a problem by $(I, (C_i)_{i \in I}, \delta)$, or, suppressing dependence on δ , sometimes simply by $(I, (C_i)_{i \in I})$. We also define $X = \times_{i \in I} X_i$, and denote by $x = (x_1, \dots, x_n) \in X$ a typical state of the bandit problem.

There is a single decision-maker in the problem, who must decide, in each period $t = 0, 1, 2, \dots$, of an infinite horizon, the arm $i \in I$ of the bandit to be activated that period. This decision is made with full knowledge of the history of play upto that point, including the state $x_t = (x_{1t}, \dots, x_{nt})$ at the beginning of period t . If the decision-maker chooses arm i in period t , two things happen; first, he receives an instantaneous reward of $r_i(x_{it})$; second, the state of arm i moves to its period- $(t + 1)$ value x_{it+1} which is realized according to the distribution $q_i(\cdot | x_{it})$. The state of all unused arms remains frozen, so we have $x_{jt+1} = x_{jt}$ for all $j \neq i$. The decision-maker discounts future rewards by $\delta \in [0, 1)$ and aims to maximize total discounted expected reward over the infinite horizon.

Formally, a t -history, denoted h_t , is a list of the states of all the arms in each period upto t , the arm chosen in each of those periods, the consequent reward witnessed, and the period- t state. A strategy σ for the decision-maker is a sequence $\{\sigma_t\}$, where for each t , σ_t specifies the action to be taken (i.e., the arm to be activated) in period- t as a measurable function of the history upto t . Let Σ denote the set of all strategies for the bandit problem.

Each strategy σ and each initial state $x \in X$ define, in the obvious way, a t -th period expected reward, denoted $r_t[\sigma, x]$, for the decision-maker for each t . The worth of the strategy σ from the initial state x , denoted $W(\sigma)(x)$, is defined as $\sum_{t=0}^{\infty} \delta^t r_t[\sigma, x]$. A strategy σ^* is *optimal* from an initial state x if its worth from x is maximal amongst all possible strategies, i.e., we have $W(\sigma^*)(x) \geq W(\sigma)(x)$ for all $\sigma \in \Sigma$. If a strategy is optimal from all $x \in X$, then we simply say it is an *optimal strategy*.

The value function $V: X \rightarrow \mathbb{R}$ of the problem is defined by $V(x) = \sup_{\sigma \in \Sigma} W(\sigma)(x)$. Note that, as the supremum of the set of attainable rewards, the value function is well-defined whether or not an optimal strategy exists since rewards are bounded and there is strict discounting (i.e., $\delta < 1$.) On the other hand, it is obvious that if an optimal strategy does exist, then V must be equal to the worth $W(\sigma^*)(\cdot)$ of any optimal strategy.

2.2 The “Classical” Framework

The framework we have described above treats each arm of the bandit as an abstract Markov process whose “state” remains frozen when the arm is not in use. In what we shall call the “classical framework” of bandit problems, these states of an arm correspond to the belief of a decision-maker regarding the “true” distribution of rewards from that arm. We provide a brief description of this framework here. For omitted details, we refer the reader to the excellent monograph of Berry and Fristedt (1985).

Let D denote the set of all probability distributions on the real line, and let Δ represent the set of all probability distributions on D . Both D and Δ are given the topology of convergence in distribution. In the classical framework, the space Δ represents the state space X_i for any arm i , and is interpreted as the space of possible beliefs regarding the true distribution of rewards from arm i . To wit, each arm i is thought of as generating rewards according to some probability $p \in D$. $F \in \Delta$ then denotes the decision-maker’s belief (or *prior*) regarding the likelihood of these different distributions.

Thus, in the classical framework, the set Δ^n , defined as the n -fold Cartesian product of Δ and endowed with the product topology, arises as the state space of the bandit problem. Let $F = (F_1, \dots, F_n)$ denote a typical element of Δ^n . The reward functions $r_i: X_i \rightarrow \mathbb{R}$ are defined in the obvious manner as

$$r_i(F_i) = \int_{\Delta} \int_{\mathfrak{R}} r p_i(dr) F_i(dp_i)$$

Finally, there remains the definition of the transition probabilities. For any $F_i \in \Delta$, let $F_i[r]$ represent a version of the conditional probability distribution (i.e., the *posterior* belief) that arises

when arm i is used and the reward r is witnessed. We assume that for each F_i and r , some version has been chosen and is held fixed. (Berry and Fristedt (1985) show that $F_i[r]$ can be chosen to depend measurably on i and r .) The posteriors $F_i[r]$ then implicitly define the transition probabilities q_i on the space $X_i(= \Delta)$: given any Borel subset \mathcal{B} of Δ , the probability $q_i(\mathcal{B}|F_i)$ of the posterior belief being in \mathcal{B} given the prior belief F_i , is simply the probability under F_i of observing a reward r such that $F_i[r] \in \mathcal{B}$.

2.3 Markovian Strategies and Index Strategies

We close this section with a brief discussion of two subclasses of strategies of especial interest in the sequel: stationary Markovian strategies (hereafter, simply Markovian strategies), and strategies defined through an “index.” Given any t -history h_t , let $x(h_t) = (x_1(h_t), \dots, x_n(h_t))$ denote the period- t vector of states under h_t .

A *stationary Markovian strategy* is a strategy σ in which at each t , σ_t depends on h_t only through $x(h_t)$, but not through t or any other details of the history upto t . Such a strategy can evidently be represented completely by a measurable function $g: X \rightarrow I$, with the interpretation that $g(x(h_t)) \in I$ is the arm to be in played if the t -history h_t has occurred—and indeed, every such function gives rise to a stationary Markovian strategy. Abusing notation, we will refer to such a strategy by simply the function g .

An *index* for the bandit problem is a function λ that associates with each arm i and each state $x_i \in X_i$, a real number $\lambda(x_i, C_i)$. Loosely speaking, the number $\lambda(x_i, C_i)$ may be interpreted as the “worth” of an arm i whose characteristics are given by C_i and whose current state is x_i , but we will not push this interpretation.

A strategy σ is said to be an *index strategy* with respect to the index λ if at each t and each t -history h_t upto t , we have $\sigma_t(h_t) \in \{i \in I \mid \lambda(x_i(h_t), C_i) \geq \lambda(x_j(h_t), C_j) \forall j \in I\}$, i.e., if σ always selects one of the arms whose index, calculated according to λ is maximal at that point.

Intuitively speaking, the notions of Markovian strategies and index strategies appeal to different aspects of the framework we have defined. The former owe their interest to the fact that the optimization problem facing the decision-maker in our model is itself a stationary Markovian one, i.e., the current state of the arms encapsulates all relevant information concerning current and future payoff possibilities. Thus, stationary Markovian strategies constitute a natural starting point for analysis in these problems.

The idea of using an “index” to define strategies, on the other hand, has its genesis in the independence assumption that when any arm i is in use, the states of all remaining arms are frozen. To wit, suppose that in some independent-armed bandit problem, it is optimal to discard the current incumbent arm i in favor of some arm j if the history h transpires when i is in use; and that it is optimal to replace i with a third arm k , if the history that is witnessed from i is instead h' . Since no information is being obtained on j (or k or any other arm) when arm i is in use, it seems very plausible—indeed, almost “obvious”—that there must also exist an optimal strategy which chooses arm j after the history h' also. Put differently, the converse seems very

improbable: that it is optimal to continue with j after the history h , but not after the history h' . The conjecture that an index exists which gives rise to an optimal strategy is a stronger version of this statement: it asserts that this continuation arm j can be chosen by calculating its “worth” in some manner, independent of the other available arms.

2.4 Applications of the Bandit Framework

The bandit framework has found a vast degree of applicability in economics and allied disciplines. We present a brief description of simple formulations of a few of these in this section, in part to motivate some of the generalizations considered later in the paper.

2.4.1 Market Pricing: Learning the Demand Curve

It is common in most of economic theory to assume that firms and managers act under perfect knowledge of market conditions when choosing price and output decisions. One of the first papers to move away from this perfect information formulation was Rothschild (1974) who introduced the bandit framework to economic theory.

Rothschild considers the problem of a seller who does not know the parameters of the “true” demand curve he faces. Specifically, there are a finite number of possible prices (p_1, \dots, p_n) the seller can charge. Associated with each price p_i is a “probability of purchase” π_i (the probability that a consumer faced with the price p_i will actually buy the good). If the vector (π_1, \dots, π_n) is known, then the seller maximizes expected profits simply by setting the price equal to that i at which $p_i\pi_i$ is maximized. When the π_i 's are unknown, however, the seller faces a bandit problem in the classical framework where the n prices act as the n arms. The reward from choosing arm i is either p_i (if a sale occurs) or 0 (if one does not).² Each time the seller chooses a price p_i , he learns something about the probability of purchase at that price.

Rothschild shows that when $n = 2$ (the case he studies throughout), there is a non-zero probability *under the optimal strategy* that the seller will choose one arm and stick to it *forever*. He further shows that this arm could be the *inferior* arm (i.e., the arm that would not be optimal under full information). Banks and Sundaram (1992a) generalize these findings substantially. They prove that both results remain true under very general conditions in independent-armed bandit problems in the classical framework, i.e., there is a non-zero probability that an arm that becomes optimal at some point will remain optimal forever, and as in Rothschild's case this need not be the “best” type of arm, that is the arm that would be optimal under full information. Indeed, a particularly strong version of this last result is true. Banks and Sundaram provide an example of a bandit with a countable infinity of arms where each arm is one of the same three possible types, and show that in any optimal strategy in this example, the best type of arm (the one with the highest expected reward) will be rejected in finite time with probability one, while

²The situation is, thus, analogous to one where the arms generate rewards according to Bernoulli distributions with unknown probabilities.

the arm that would be second-best under complete information survives forever with probability one.

2.4.2 Job Search and Matching

The framework of bandit problems has also proved a popular one for the analysis of decision making in labor markets, notably in the theory of job-search and matching (see Mortensen (1985) for references). A typical “matching” model has the following structure: there is a single decision-maker (the “worker”) who faces a infinite number of *a priori* identical potential employers. The productivity of the worker is match-specific: it depends on the firm she opts to work for. However, the productivity levels of some or all of the matches are *a priori* unknown. Instead, both firm and worker begin with (the same) belief regarding her productivity, which is updated as information comes in. In the meanwhile, the worker receives as compensation the expected value of her match as given by the current belief on this score.

It is easy to see that, treating the firms as arms of the bandit, and the productivity of a match as the “true” distribution of rewards from that arm, the optimization problem facing the worker is precisely a bandit problem in the classical framework, the only difference being the assumption of an infinite number of available arms (firms). However, it is shown in Banks and Sundaram (1992a) that the independent-armed bandit framework we have outlined above is easily generalized to accomodate a countable infinity of arms, and existence of optimal strategies may be guaranteed under very general conditions that includes the case where all arms are *a priori* identical.

In addition to existence of optimal strategies, Banks and Sundaram (1992a) also show that the assumption of an infinite number of *a priori* identical arms has the following implications. First, there is always at least one optimal “no recall” strategy which never re-uses a once-tried and discarded arm. Second, the expected number of arms used in any optimal strategy is finite, so that, with probability one, the worker uses only a finite number of arms. And finally, when the true distribution of rewards from any match is one of only two possible types, *myopic strategies* (strategies which select the arm to be played purely on the basis of *current* expected rewards from each arm) are fully optimal.

2.4.3 Technology Choice and Learning-By-Doing

Consider the problem faced by a firm which has available to it two possible modes of production. The first (the “old” technology) has associated with it a constant marginal cost of production, denoted say c . The second (the “new” technology) has an initial marginal cost that is higher than c , but this marginal cost declines over time as the technology is used owing to a learning-by-doing effect. For definiteness, assume that the marginal cost of production using the new technology is given by

$$g(d_0, d_1, d_2, \tau; \zeta) = d_0 + \zeta d_1 e^{-d_2 \tau} \tag{2.1}$$

where $d_0, d_1, d_2 > 0$; $d_0 < c < d_0 + d_1$; τ is the number of periods the second technology has been used; and ζ is a noise term with known distribution Z and mean 1. Note that if $\tau = 0$ (i.e., the new technology has not yet been used), then the (expected) marginal cost of production using this technology is $d_0 + d_1 > c$, while as $\tau \rightarrow \infty$, this marginal cost approaches $d_0 < c$. Moreover, d_2 parametrizes the *rate* of decline of this cost. The firm produces one unit of output per period, and wishes to minimize the sum of discounted expected costs over an infinite horizon.

For notational ease, let θ represent the parameter vector (d_0, d_1, d_2) . When the vector θ is known with certainty, the resulting optimization problem facing the firm is a simple one: since costs are declining under the new technology and are constant under the old, if ever it becomes optimal to switch to the new technology remaining with this choice forever must also be optimal. Thus, the firm only needs to calculate the discounted expected cost of using the old technology forever, which is $c/(1 - \delta)$, and compare it with the cost of switching to the new technology forever, which is $\sum_{t=0}^{\infty} \int g(\theta, t, \zeta) dZ(\zeta)$. The smaller of the two then is the firm's optimal choice. The choice problem becomes non-trivial, however, if some or all of these parameters are not known with certainty. Moreover, the problem cannot be solved by putting it into the classical bandit framework, since the firm's belief (denoted, say F) regarding the vector θ is *not* sufficient to describe the state of the problem at any point. Rather, it is also important to know the number of periods τ the new technology has been used so far. Indeed, even the *current* marginal cost of production $\int \int g(\theta, \tau, \zeta) dF(\theta) dZ(\zeta)$ cannot be calculated without this piece of information.

Nonetheless, the problem can be cast as a bandit problem in the framework of subsection 2.1. To do so, let arm 1 be defined by $X_1 = \{c\}$, $r_1(c) = -c$, and $q_1(\{c\}|c) = 1$. This denotes the fact that arm 1 is in a constant state c , from which the rewards are $-c$. As regards arm 2, let \mathcal{F} denote the space of all possible beliefs regarding the parameter vector θ , and let $X_2 = \mathcal{F} \times \{t, \infty, \epsilon, \dots\}$ be the state space of arm 2. The expected reward $r_2(F, \tau)$ from using arm 2 at the state (F, τ) can be written as

$$r_2(F, \tau) = - \int \int g(\theta, \tau, \zeta) dF(\theta) dZ(\zeta).$$

Finally, suppose $F\{w\}$ represents the posterior belief regarding the value of θ given the prior F and the observation $w = -g(\theta, \tau; \zeta)$. Then, given the current state (F, τ) , if the reward w is witnessed from this arm, the new state of the arm is given by $(F\{w\}, \tau + 1)$. Since the distribution of w is completely specified by knowledge of F and τ , this implicitly defines the transition probabilities q_2 on arm 2.

The vectors (X_1, r_1, q_1) and (X_2, r_2, q_2) complete the specification of this problem as a bandit problem.

2.4.4 Agency Problems

The framework of bandit problems can also be thought of as representing a problem of pure agency, in which each arm of the bandit represents an agent and the different distributions in

the support of that arm represent the different types the agent can be. The problem facing the decision-maker is then to select an agent whose type is “best” (from the decision-maker’s point of view) subject to the implicit search costs involved.

One interpretation arises from treating the decision-maker as the median voter in some constituency, the arms as the various candidates offering themselves for election, and the payoffs from an arm as the government-controlled benefits that become available to the constituency. Thus, a candidate’s type is an index of that candidate’s ability to divert funds to his constituency. More generally, by expanding the state space to include variables apart from beliefs (as in the subsection above), the impact of features such as term-limits and incumbency effects could also be studied.

Alternatively, a different labor-market interpretation of the bandit problem from that presented above is obtained by treating the decision-maker as a firm or employer searching over workers (arms). Here, once again, the generalized bandit framework is particularly useful, since the state of an arm (i.e., the characteristics of the worker) could be used to incorporate a variety of detail apart from the worker’s expected productivity.

3 Optimal Strategies and the Gittins Index

The existence of optimal strategies in the bandit problem—in fact, the existence of Markovian strategies that are optimal—is easily ascertained using standard arguments from dynamic programming. These arguments are briefly outlined in subsection 4.1 below. Following this, we describe in subsection 4.2 the construction of the *Dynamic Allocation Index* (better known today as the *Gittins Index*) of Gittins and Jones (1974), and state their powerful theorem that a strategy in a bandit problem is an optimal strategy if, and only if, it is an index strategy with respect to the Gittins Index.

3.1 Markovian Optimal Strategies

The description of the bandit problem as a dynamic programming problem, and the obtaining of the value function and an optimal strategy using a contraction mapping argument, are wholly routine. Our exposition in this section is correspondingly brief.

As a first step, recall that since rewards are bounded, the value $V(x) = \sup_{\sigma \in \Sigma} W(\sigma)(x)$ is well defined from any x . A standard argument establishes that at any $x \in X$, V must satisfy the time-consistency principle embodied in the “Bellman Equation” for this problem:

$$V(x) = \bigvee_{i \in I} L_i V(x) \tag{3.1}$$

where,

$$L_i V(x) = r_i(x_i) + \delta \int V(x_{-i}, \hat{x}_i) q_i(d\hat{x}_i | x_i). \quad (3.2)$$

(Here, in obvious notation, (x_{-i}, \hat{x}_i) refers to the vector x , but with x_i replaced by \hat{x}_i .)

To recover the function V , we use the following procedure. Let $M(X)$ be the space of all bounded measurable functions on X , endowed with the sup-norm. Define the map $T: M(X) \rightarrow M(X)$ as follows: for $w \in M(X)$, let

$$Tw(x) = \bigvee_{i \in I} L_i w(x) \quad (3.3)$$

where

$$L_i w(x) = r_i(x_i) + \delta \int w(x_{-i}, \hat{x}_i) dq_i(\hat{x}_i | x_i). \quad (3.4)$$

Let $v, w \in M(X)$ and $a \in \mathbb{R}$. If $v \geq w$, then $L_i v \geq L_i w$ for all i , so certainly $Tv \geq Tw$. Moreover, it is evident that $T(v + a) = Tv + \delta a$. Since $\delta < 1$, it follows from Blackwell (1965) that T is a contraction on the complete metric space $M_i(X)$, and hence has a unique fixed point w^* . Since V also meets this equation, we must have $w^* = V$.

Finally, let g denote any measurable selection from the correspondence of maximizers of (3.1)-(3.2).³ It follows easily from the recursive nature of these equations that the total worth of the stationary Markovian strategy defined using g (and denoted simply by g) is, in fact, V . Thus, any measurable selection defined in this manner is a Markovian optimal strategy, establishing simultaneously the existence of optimal strategies, and the existence of Markovian optimal strategies.

3.2 The Gittins Index

The Theorem of Gittins and Jones (1974) is perhaps the single most powerful result in the literature on bandit problems. It asserts that in all independent-armed bandit problems under geometric discounting, the set of *all* optimal strategies may be recovered by solving a family of optimal stopping problems, that associate with each arm an index known as the *Dynamic Allocation Index*, or, more popularly, as the *Gittins Index*. The feature which gives this result especial power is that the index on an arm depends solely on the characteristics of that arm, and not on any other feature of the problem.

The Gittins index is constructed as follows. Let a bandit problem $(I, (C_i)_{i \in I}, \delta)$ be given, where, of course, $C_i = (X_i, r_i, q_i)$. Select any arm i . Consider the following optimal stopping

³The existence of such a selection can be shown by appealing to standard selection theorems.

problem involving arm i . In each period (conditional on the termination option not having been accepted yet) the decision-maker is given the choice between (a) terminating the problem and accepting a fixed terminal reward of $m \in \mathbb{R}$, or (b) continuing with playing arm i for one more period. Standard techniques from the theory of dynamic programming show that this problem is well defined from any initial state $x_i \in X_i$, and that the value of the problem $V(x_i, C_i, m)$ from the initial state x_i is a measurable function on X_i that satisfies at each x_i :

$$V(x_i, C_i, m) = \max\{m, r_i(x_i) + \delta \int V_i(\hat{x}_i, C_i, m) dq_i(\hat{x}_i|x_i)\}. \quad (3.5)$$

The *Gittins Index* $\mu(x_i, C_i)$, on an arm i , whose characteristics are given by C_i and which is currently in the state x_i , is then defined as

$$\mu(x_i, C_i) = \inf\{m \mid V(x_i, C_i, m) = m\}. \quad (3.6)$$

Since r_i is a bounded function by hypothesis, it is easy to see that for m sufficiently large we must have $V(x_i, C_i, m) = m$ for all $x_i \in X_i$; while, for $-m$ sufficiently large it must be the case that $V(x_i, C_i, m) > m$ at all x_i . It follows that the Gittins Index is well defined. The following result establishes the importance of this index:

Theorem 3.1 (Gittins and Jones (1974)) *The optimal selections in the bandit problem $(I, \delta, (C_i)_{i \in I})$ at the state (x_1, \dots, x_n) are those arms i which are Gittins index maximal at that state, i.e., which are such that*

$$\mu(x_i, C_i) = \bigvee_{j \in I} \mu(x_j, C_j). \quad (3.7)$$

Equivalently, a strategy σ for a bandit problem $(I, \delta, (C_i)_{i \in I})$ is an optimal strategy for that problem if, and only if, the set of histories on which its recommendations differ from the set of Gittins Index maximal arms following that history has probability zero.

4 Infinite-Armed Bandits and the Gittins Index

In many economic applications (such as the literature on matching models mentioned above), it is natural to allow for an *infinite* number of available arms, to leave open the possibility that there is at least one untried arm available each period. It would, therefore, be interesting to know if the Gittins-Jones Theorem can be extended to cover this case also.

Increased applicability is only one reason why one might wish to allow for an infinite number of arms in the bandit problem. Perhaps a deeper one is the intuitive feeling that, as an expression of the independence between arms, the “correct” statement of the Gittins-Jones Theorem should

not depend on the cardinality of the set of available arms. More precisely, consider the statement given in section 2.2 that in an independent-armed bandit problem, if it is optimal to switch from arm i to arm j after the history h (on arm i), and to switch to arm k after the history h' , it “should” be optimal to switch to j after h' also. The Gittins-Jones Theorem shows that this statement is indeed correct when the number of arms is finite; but it is unclear why finiteness of the set of available arms should matter for the truth of this statement.⁴

We examine the optimality of Gittins index strategies in bandits with a (possibly) infinite set of arms in this section. As before, we denote a bandit problem by $(I, (C_i)_{i \in I})$, but, for the purposes of this section alone, we will no longer assume that I is a finite set. Rather, I will be allowed to be of arbitrary cardinality. Notation is otherwise largely unchanged. The index i will continue to denote a generic member of I , and $C_i = (X_i, r_i, q_i)$ will denote the characteristics of a generic arm i . In addition, the following assumption will be maintained throughout this section to ensure that total discounted rewards under an arbitrary strategy are finite:

Assumption 1: There is $A \in \mathbb{R}$ such that $\sup_{x_i \in X_i, i \in I} |r_i(x_i)| \leq A$.

For expositional reasons, we break up the discussion in this section into two parts. In subsection 4.1, we present a simple example to show that when the cardinality of the set of available arms is allowed to be infinite, the Gittins index strategy need not be well-defined: there could exist a set of histories of positive probability under any of which there is no longer an arm attaining the supremum of the indices. We then identify a set of conditions that are necessary *and* sufficient for the index strategy to be well defined from a given initial state. Subsection 4.2 then focusses on the optimality of Gittins index strategies when there are an infinite number of arms.

4.1 The Gittins Index Strategy

Any strategy in the bandit problem must prescribe the continuation arm to be picked after any history. Thus, if we are to examine the optimality of the Gittins index strategy, we must first ensure that it is really a “strategy,” i.e., that after any history, there is at least one arm that attains the supremum of the set of indices. It is a simple matter to construct seemingly well-behaved bandit problems with a *compact* set of arms in which this is not the case. Consider the following:

Example 4.1 Let $I = [0, 1]$. For each $i \in I$, let $X_i = [0, 1]$, and let $r_i(x_i) = x_i$ for all $x_i \in X_i$. Finally, define the transition probabilities by $q_i(\{1\}|x_i) = 1 - q_i(\{0\}|x_i) = x_i/i$. Since the characteristics of all arms are the same, we denote the index on arm

⁴In Banks and Sundaram (1992, Theorem 4.1) it is shown that the Gittins index strategies continue to identify all the optimal selections at every point, when the number of available arms is allowed to be *countably* infinite. While sufficient for most applications, this result still begs the question of whether the intuition given here has a validity that is independent of the cardinality of the set of available arms.

i at the state x_i by simply $\mu(x_i)$. Suppose now that the initial state is given by $x_i = i/2$ for all i . A straightforward calculation shows that

$$\mu\left(\frac{i}{2}\right) = \left(\frac{i}{(1-\delta)(2-2\delta+\delta i)} \right)$$

so that arm 1 has the highest index at the initial state. However, after the 1-history in which the state of arm 1 moves to 0 (a 1-history which occurs with probability $1/2$), the index on arm 1 drops to zero; it is a trivial matter to see that there is no longer an arm that attains the supremum of the indices at this new state.

In the rest of this subsection, we focus on identifying a set of conditions which guarantee that the Gittins index strategy will be well-defined from a given fixed initial state $\bar{x} = (\bar{x}_i)_{i \in I}$. The Gittins index of arm i at the fixed initial state \bar{x}_i is denoted $\bar{\mu}_i$.

Consider the following strategy in the optimal stopping problem used to define the Gittins index on arm i , when m is the terminal reward:

Select arm i initially. In each subsequent period, stay with arm i if the index μ_i on arm i at the beginning of the period satisfies $\mu_i \geq m$. At the first point where $\mu_i < m$, accept the terminal reward m .

Denote this strategy by $\sigma_i[m]$. For $t = 1, 2, \dots$, let $H_i(t, m)$ denote the set of t -histories in this stopping problem (from the fixed initial state \bar{x}_i) under which $\sigma_i[m]$ accepts the terminal reward in period t . Also, denote by $p_i(t, m)$ the probability of the set $H_i(t, m)$, and by $R_i(t, m)$ the total expected discounted reward from arm i over this period.

Theorem 4.1 *The Gittins index strategy is well defined from the initial state $\bar{x} = (\bar{x}_i)_{i \in I}$ if, and only if, either*

- (a) *there are infinitely many arms $i \in I$ such that $\bar{\mu}_i \geq m^*$, or*
- (b) *there is $i \in I$ such that $\bar{\mu}_i \geq m^*$ and $p_i(t, m^*) = 0$ for all t ,*

where $m^* = \sup\{m \mid \bar{\mu}_i \geq m^* \text{ for infinitely many } i \in I\}$.

Remark Under the convention that the supremum of the empty set is $-\infty$, this result applies when I is finite also.

Proof We begin by proving that under (a) or (b), the Gittins index strategy is well defined. First suppose (a) holds. Define the sequences m_t , I_t and J_t inductively as follows. Let $m_1 = \sup_{i \in I} \bar{\mu}_i$, and set $I_1 = \{i \in I \mid \bar{\mu}_i = m_1\}$ and $J_1 = I - I_1$. For $t \geq 1$, let $m_{t+1} = \sup\{\bar{\mu}_i \mid i \in J_t\}$, $I_{t+1} = \{i \in J_t \mid \bar{\mu}_i = m_{t+1}\}$, and $J_{t+1} = J_t - I_{t+1}$.

It is obvious from the definition of m^* and the hypothesis that (a) holds that I_1 is non-empty. If I_1 contains an infinite number of elements, the Gittins index strategy is evidently well defined.

Suppose I_1 contains only a finite number of elements. Then, we claim, I_2 must be non-empty, and, in fact, that I_{t+1} must be non-empty whenever $\bigcup_{\tau=1}^t I_\tau$ contains only a finite number of elements. For, suppose the contrary. Clearly, we must then have $m^* < m_{t+1}$, for if $m^* \geq m_{t+1}$ and $\nexists i$ such that $\bar{\mu}_i = m_{t+1}$, (a) would be violated. On the other hand, since there is no i such that $\bar{\mu}_i = m_{t+1}$, it also follows that for any $\epsilon > 0$, there are infinitely many i such that $\bar{\mu}_i \in (m_{t+1} - \epsilon, m_{t+1}]$. For ϵ sufficiently small, m^* is not in this interval, implying that there exists $\hat{m} > m^*$ such that $\bar{\mu}_i \geq \hat{m}$ for infinitely many i . This violates the definition of m^* , and establishes the claim. It trivially follows from this claim that the Gittins index strategy is well defined from the initial state x .

On the other hand, suppose (b) holds; assume without loss that (a) does not. The set $I^* = \{j \in I \mid \bar{\mu}_j \geq m^*\}$ contains at least i , and since (a) is violated contains at most a finite number of elements. Moreover, after any history, the index on arm i will always be at least as large as m^* with probability one, so in following the Gittins index strategy, there will never be a call to play an arm not in I^* . It follows that, since I^* is finite, the Gittins index strategy is well defined.

Now suppose both (a) and (b) are false. If the set I^* of the previous paragraph is empty, then it follows from the definition of m that there is also no j such that $\bar{\mu}_j = \sup_{i \in I} \bar{\mu}_i$. So suppose that I^* is non-empty. Since (a) is violated, I^* can contain at most a finite number of elements. By (b) being false, for each arm j in this set, there is a positive probability that when j is played, the index on j falls strictly below m^* in finite time. Thus, with non-zero probability, the indices on all these arms will fall below m^* in finite time. From the definition of m^* , it follows that there is no longer any i attaining the supremum of the indices. \square

4.2 Optimality of the Gittins Index Strategy

The first result of this subsection shows that whenever the Gittins index strategy is well-defined, a strategy in the bandit problem is an optimal strategy if, and only if, it always recommends picking an arm that is Gittins index maximal (except, possibly, after a set of histories of collective probability zero). As usual, $W(\sigma)(x)$ will denote the total discounted reward to the decision-maker under the strategy σ from the initial state x . Observe that since Assumption 1 holds, the value function $V(x) = \sup_{\sigma} W(\sigma)(x)$ is well-defined from every initial state x , even if an optimal strategy does not exist from that state. Given an initial state x , and the history h_t from x , let $x(h_t) = (x_i(h_t))_{i \in I}$ represent the resulting new vector of states.

Theorem 4.2 *Suppose that the Gittins index strategy is well defined from the initial state $x \in X$, i.e., that after any t -history h_t from x , there exists $i \in I$ such that $\mu_i(x_i(h_t)) = \bigvee_{j \in I} \mu_j(x_j(h_t))$. Then, a strategy σ satisfies $W(\sigma)(x) = V(x)$ if, and only if,*

$$\sigma_t(h_t) \in \left\{ i \in I \mid \mu_i(x_i(h_t)) = \bigvee_{j \in I} \mu_j(x_j(h_t)) \right\}$$

except possibly after a set of histories of probability zero.

Remark The proof of this theorem actually follows from a close repetition of arguments in the original Gittins-Jones proof. It appears a plausible conjecture that the reason this has gone unnoticed is that much of the recent literature (including Berry and Fristedt, 1985) uses the shorter and more elegant proof of the Gittins-Jones Theorem due to Whittle (see Whittle, 1982). However, unlike the Gittins-Jones proof, Whittle's arguments rely in an essential way on the finiteness of the set of available arms.

Proof See Appendix A. \square

Theorem 4.2 establishes that provided the Gittins index strategy is well-defined, only strategies coinciding with it almost surely can be optimal. However, this result leaves unanswered the question of what happens if the Gittins index strategy is *not* well defined. In particular, is it the case that if the Gittins index strategy is not well defined, optimal strategies no longer exist? The following theorem establishes that in a large class of problems, this is indeed the case:

Theorem 4.3 *Suppose that for each $i \in I$ and each $x_i \in X_i$ it is the case that $q(\cdot|x_i)$ has countable support, i.e., there is a countable subset Y_i of X_i (possibly depending on x_i) such that $q(Y_i|x_i) = 1$. Then, a strategy is optimal in the bandit problem if, and only if, it is a Gittins index strategy. In particular, optimal strategies fail to exist whenever the Gittins index strategy is not well defined.*

Proof Under the stated hypothesis, any strategy in the bandit problem will involve only the use of (at most) a countable number of arms, since with probability one, there are only a countable number of distinct t -histories for any t . Given any countable subset J of I , a strategy σ_J is optimal in the bandit problem $(J, (C_j)_{j \in J})$ if, and only if, it is a Gittins index strategy. Theorem 4.3 follows. \square

On the other hand, the assumption of countable support seems restrictive, and it appears a strong conjecture that Theorem 4.3 is valid even without this restriction. I have not been able to prove this conjecture or to provide a counterexample. However, the following related—but simpler—conjectures, which pertain to ϵ -optimal strategies rather than fully optimal ones, would help provide an answer if true. (Recall that a strategy σ^* is ϵ -optimal from the initial state x if it is the case that $W(\sigma^*)(x) \geq W(\sigma)(x) - \epsilon$ for all $\sigma \in \Sigma$.)

Conjecture 4.4 *For all $\epsilon > 0$, there is $\eta > 0$ such that if a strategy σ always picks an arm whose index is within η of the supremum of the indices, then σ is ϵ -optimal.*

Conjecture 4.5 *For all $\epsilon > 0$, there is $\eta > 0$ such that if σ is ϵ -optimal, then it always picks an arm whose index is within η of the supremum of the indices at that point.*

5 Multiple Plays

We return now to the framework of section 2 where the set of arms I is finite, $I = \{1, \dots, n\}$. As discussed in the Introduction, one somewhat restrictive feature of this framework is that only one arm may be activated by the decision-maker at any point in time. In this section we consider a situation where the decision-maker is allowed to play k arms each period, where $1 \leq k < n$ is a given, fixed number, and examine whether it is the case (as intuition would suggest) that the optimal strategy consists of playing the arms with the k highest Gittins indices in each period.

Thus, the *action space* for the decision-maker in this set-up consists of the set of all possible combinations of k of the n arms. Let \mathcal{C} denote this set. A typical action will be denoted $C = (c_1, \dots, c_k)$ where $c_i \in I$ for each i (and, of course, $c_i \neq c_j$ if $i \neq j$). Given an action $C \in \mathcal{C}$, the *reward* $R(x, C)$ from taking the action C at the state $x = (x_1, \dots, x_n) \in X = \times_{i \in I} X_i$ is given by:

$$R(x, C) = \sum_{i \in C} r_i(x_i)$$

When the action C is taken at a state x , the state of all untried arms (i.e., arms $j \notin C$) remain frozen, while the state of arm $i \in C$ moves to a new state according to the distribution $q_i(\cdot | x_i)$. More formally, let $q_i(\cdot | x_i, C)$ be the measure $q_i(\cdot | x_i)$ if $i \in C$, and let $q_i(\cdot | x_i, C)$ be the measure $\chi_i(x_i)$ that places point-mass on x_i if $i \notin C$. Then, the transition probability $Q(\cdot | x, C)$ from taking the action C at the state x in this modified bandit problem is simply given by the product measure

$$Q(\cdot | x, C) = q_1(\cdot | x_1, C) \times \dots \times q_n(\cdot | x_n, C).$$

The tuple $\{X, \mathcal{C}, R, Q\}$ then represents the decision-maker's optimization problem as a Markovian dynamic programming problem with state space X , action space \mathcal{C} , reward function R , and transition probability Q .

In subsection 6.1, we establish the existence of Markovian optimal strategies in this problem. Subsection 6.2 then turns to the central question of this section: is it the case that at all points it is optimal to select the k arms with the highest Gittins indices at that point?

5.1 Markovian Optimal Strategies

As in previous sections, the existence of Markovian optimal strategies follows from a straightforward application of dynamic programming results:

Theorem 5.1 *The dynamic programming problem $\{X, \mathcal{C}, R, Q\}$ is well defined. The value function $V: X \rightarrow \mathbb{R}$ is a measurable function that satisfies the Bellman Principle of Optimality at*

each $x \in X$:

$$V(x) = \bigvee_{C \in \mathcal{C}} L_C V(x) \quad (5.1)$$

where

$$\begin{aligned} L_C V(x) &= r(x, C) + \delta \int V(\hat{x}) dQ(\hat{x}|x, C) \\ &= \sum_{i \in C} r_i(x_i) + \delta \int V(\hat{x}_1, \dots, \hat{x}_n) dq_1(\hat{x}_1|x_1, C) \cdots dq_n(\hat{x}_n|x_n, C). \end{aligned}$$

Any Markovian strategy defined through a measurable selection from the correspondence of optimizers of equation (5.1) defines a optimal strategy.

5.2 Optimality of Gittins-Index Strategies

When $k = 1$, the Gittins-Jones theorem shows that any optimal strategy involves (almost) always picking an arm with the highest current value of the Gittins index. In this section we provide a counter-example which shows that it is not necessarily optimal to play the arms with the k highest values of the Gittins index when $k > 1$.

The example has 3 arms: $I = \{1, 2, 3\}$, and $\mathcal{C} = \{(1, 2), (1, 3), (2, 3)\}$. The state space X_i of each arm i is the unit interval $[0, 1]$. The reward functions $r_i: X_i \rightarrow \mathbb{R}$ are given by $r_i(x_i) = ix_i$, $x_i \in [0, 1]$, $i = 1, 2, 3$. Finally, the transition probabilities $q_i(\cdot|x_i)$ are given by $q_i(\{1\}|x_i) = 1 - q_i(\{0\}|x_i) = x_i$, $x_i \in [0, 1]$, $i = 1, 2, 3$.

Some simple calculation shows that the indices on the three arms are given by the following expressions:

$$\mu_1(x_1) = \frac{x_1}{(1 - \delta)(1 - \delta + x_1\delta)} \quad (5.2)$$

$$\mu_2(x_2) = \frac{2x_2}{(1 - \delta)(1 - \delta + x_2\delta)} \quad (5.3)$$

$$\mu_3(x_3) = \frac{3x_3}{(1 - \delta)(1 - \delta + x_3\delta)} \quad (5.4)$$

Let $(\bar{x}_1, \bar{x}_2, \bar{x}_3) \in [0, 1]^3$ denote the initial state of this bandit problem. We begin by identifying a set of conditions under which, conditional on optimal continuations in either case, it is strictly

preferable to choose arms 1 and 3 in the first period, rather than arms 2 and 3. Then we will show that it is possible to choose the initial state to satisfy these conditions and simultaneously also have $\bar{\mu}_1 < \bar{\mu}_2 < \bar{\mu}_3$, where $\bar{\mu}_i = \mu_i(\bar{x}_i)$ is the Gittins index of arm i at this initial state.

Consider first the case where arms 2 and 3 are used in the first period at the initial state. Then, there are four possible continuation states that could result at the beginning of the second period: $(\bar{x}_1, 1, 1)$, $(\bar{x}_1, 0, 1)$, $(\bar{x}_1, 1, 0)$, and $(\bar{x}_1, 0, 0)$, with respective probabilities $\bar{x}_2\bar{x}_3$, $(1-\bar{x}_2)x_3$, $x_2(1-\bar{x}_3)$, and $(1-\bar{x}_2)(1-\bar{x}_3)$. Since all arms yield a reward of zero when they are in the state 0, and strictly positive rewards when they are not, the optimal continuation strategy is obvious in each case: continue with arms 2 and 3 forever if the new state is $(\bar{x}_1, 1, 1)$; switch to arms 1 and 3 forever if the new state is $(\bar{x}_1, 0, 1)$; switch to arms 1 and 2 forever if the new state is $(\bar{x}_1, 1, 0)$; and play either 1 and 2, or 1 and 3, if the new state is $(\bar{x}_1, 0, 0)$. The expected continuation payoffs are:

$$\begin{aligned} V(\bar{x}_1, 1, 1) &= 5/(1-\delta) \\ V(\bar{x}_1, 0, 1) &= (3+\bar{x}_1)/(1-\delta) \\ V(\bar{x}_1, 1, 0) &= (2+\bar{x}_1)/(1-\delta) \\ V(\bar{x}_1, 0, 0) &= \bar{x}_1/(1-\delta) \end{aligned}$$

Thus, the total value $L_{(2,3)}V(\bar{x}_1, \bar{x}_2, \bar{x}_3)$ of starting with arms 2 and 3 is:

$$\begin{aligned} L_{(2,3)}V(\bar{x}_1, \bar{x}_2, \bar{x}_3) &= 3\bar{x}_3 + 2\bar{x}_2 + \delta[\bar{x}_2\bar{x}_3V(\bar{x}_1, 1, 1) + \bar{x}_3(1-\bar{x}_2)V(\bar{x}_1, 0, 1) \\ &\quad + \bar{x}_2(1-\bar{x}_3)V(\bar{x}_1, 1, 0) + (1-\bar{x}_2)(1-\bar{x}_3)V(\bar{x}_1, 0, 0)] \\ &= 3\bar{x}_3 + 2\bar{x}_2 + \delta(1-\delta)^{-1}[5\bar{x}_2\bar{x}_3 \\ &\quad + (3+\bar{x}_1)\bar{x}_3(1-\bar{x}_2) + (2+\bar{x}_1)\bar{x}_2(1-\bar{x}_3) + \bar{x}_1(1-\bar{x}_2)(1-\bar{x}_3)] \end{aligned}$$

Now suppose instead that arms 1 and 3 were used in the first period. The the four possible second period states are $(1, \bar{x}_2, 0)$, $(0, \bar{x}_2, 1)$, $(0, \bar{x}_2, 0)$, and $(1, \bar{x}_2, 1)$, which occur with the respective probabilities $\bar{x}_1(1-\bar{x}_3)$, $(1-\bar{x}_1)\bar{x}_3$, $(1-\bar{x}_1)(1-\bar{x}_3)$, and $\bar{x}_1\bar{x}_3$. The optimal continuation in three of these cases is obvious: if $(1, \bar{x}_2, 0)$ occurs, the optimal continuation is to play arms 1 and 2 forever; if $(0, \bar{x}_2, 1)$ occurs, the optimal action is to play 2 and 3 forever; while if $(0, \bar{x}_2, 0)$ occurs, any continuation is optimal that involves playing arm 2 forever. The case where the state at the beginning of the second period is $(1, \bar{x}_2, 1)$ is a little more complicated. There are two possibilities here.⁵ One option is to play arms 1 and 3 at this state; if this is an optimal choice, it must remain optimal forever, since all the states remain frozen, and the continuation reward from employing this choice forever is $4/(1-\delta)$. A second option is to play arms 2 and 3, and to switch to arms 1 and 3 forever if, and only if, the state on arm 2 moves to 0; the continuation

⁵There are more, but these are the only two relevant possibilities.

reward from this option is $(3 + 2\bar{x}_2 + \delta(1 - \bar{x}_2))/(1 - \delta)$. It is easily seen that the first option is strictly preferable as long as $\bar{x}_2 < (1 - \delta)/(2 - \delta)$. Assuming this to be the case (we will ensure it later), the four continuation values that result are

$$\begin{aligned} V(0, \bar{x}_2, 1) &= (3 + 2\bar{x}_2)/(1 - \delta) \\ V(1, \bar{x}_2, 0) &= (1 + 2\bar{x}_2)/(1 - \delta) \\ V(0, \bar{x}_2, 0) &= 2\bar{x}_2/(1 - \delta) \\ V(1, \bar{x}_2, 1) &= 4/(1 - \delta) \end{aligned}$$

So, the value of beginning with arms 1 and 3 is:

$$\begin{aligned} L_{(1,3)}V(\bar{x}_1, \bar{x}_2, \bar{x}_3) &= 3\bar{x}_3 + \bar{x}_1 + \delta(1 - \delta)^{-1}[4\bar{x}_1\bar{x}_3 + \bar{x}_3(1 - \bar{x}_1)(3 + 2\bar{x}_2) \\ &\quad + \bar{x}_1(1 - \bar{x}_3)(1 + 2\bar{x}_2) + 2\bar{x}_2(1 - \bar{x}_1)(1 - \bar{x}_3)]. \end{aligned}$$

Subtracting $L_{(1,3)}V$ from $L_{(2,3)}V$ and cancelling common terms results in the following:

$$L_{(2,3)}V(\bar{x}_1, \bar{x}_2, \bar{x}_3) - L_{(1,3)}V(\bar{x}_1, \bar{x}_2, \bar{x}_3) = 2\bar{x}_2 - \bar{x}_1 + \frac{\delta}{1 - \delta}\bar{x}_2\bar{x}_3 \quad (5.5)$$

This difference is negative provided:

$$\bar{x}_1 > \bar{x}_2 \left(2 + \frac{\delta}{1 - \delta}\bar{x}_3 \right). \quad (5.6)$$

Thus, as long as (5.6) holds (and $\bar{x}_2 < (1 - \delta)/(2 - \delta)$), it is the case that beginning with arms 2 and 3 is dominated by the situation where the decision-maker begins with arms 1 and 3.

Now consider the following parametrization. Let $\delta = 1/2$, $\bar{x}_1 = 7/12$, $\bar{x}_2 = 1/4$, and $\bar{x}_3 = 1/6$. At these values, using the expressions (5.2)-(5.4) for the indices on the three arms, it is readily calculated that $\bar{\mu}_1 = 28/19$, $\bar{\mu}_2 = 8/5$, and $\bar{\mu}_3 = 12/7$, so we have $\bar{\mu}_1 < \bar{\mu}_2 < \bar{\mu}_3$. In particular, the two arms with the highest Gittins indices are arms 2 and 3.

Note that at the given values of the parameters, we have $(1 - \delta)/(2 - \delta) = 1/3 > 1/4 = \bar{x}_2$. Thus, (5.5) applies and we have

$$L_{(2,3)}V(\bar{x}_1, \bar{x}_2, \bar{x}_3) - L_{(1,3)}V(\bar{x}_1, \bar{x}_2, \bar{x}_3) = \frac{1}{2} - \frac{7}{12} + \frac{1}{24} = -\frac{1}{24} < 0$$

so that beginning with arms 1 and 3 and continuing optimally is strictly preferable to beginning with arms 2 and 3 and continuing optimally. \square

6 Switching Costs in the Bandit Framework

Another variant of the basic bandit framework that is of considerable interest concerns the introduction of costs for switching between arms. Indeed, it is difficult to imagine a relevant economic decision problem in which the decision-maker may costlessly move between alternatives. Unfortunately, Banks and Sundaram (1993) show that it is not possible, in the presence of switching costs, to define an index on the arms such that the resulting strategy is invariably optimal. We provide in this section an alternative proof of their non-existence result.

Recall that in our current framework, the tuple $C_i = (X_i, r_i, q_i)$ completely describes arm i . We introduce switching costs by including in this tuple two real numbers c_i and d_i with the interpretation that c_i is the cost of switching *to* arm i (from any other arm), and d_i is the cost of switching away *from* arm i (to any other arm). Thus, the characteristics of arm i are given by the quintuple $(X_i, r_i, q_i, c_i, d_i)$. To distinguish this from the case of no switching costs, we will denote this quintuple by C_i^* . Note that in this scenario, the total cost of a switch from arm i to arm j is given by $d_i + c_j$.⁶

When switching costs are allowed to be non-zero, the attractiveness of an arm evidently changes depending on whether or not that arm is the one “currently in use” (i.e., whether or not it is the arm that was used in the previous period by the decision-maker). For it is obvious that in comparing two otherwise identical arms, of which one was used in the previous period, the one that was in use must be more attractive than the one that was idle. This motivates the following modification in our definition of an “index” on the arms:

Definition *An index in the presence of switching costs is any function λ which specifies for a generic arm i a value $\lambda(x_i, C_i^*, s_i)$, where C_i^* denotes the characteristics of arm i , x_i is the current state of arm i , and $s_i \in \{0, 1\}$ is a variable that denotes whether ($s_i = 1$) or not ($s_i = 0$) i is the arm currently in use.*

An index λ will be said to be optimal in the presence of switching costs if the strategy it induces is optimal in every bandit problem $\{I, (C_i^*)_{i \in I}, \delta\}$, in which switching costs are possibly non-zero.

6.1 Markovian Optimal Strategies

Let a bandit problem $\{I, (C_i^*)_{i \in I}, \delta\}$ be given. As in the previous sections, the existence of Markovian strategies that are optimal in the bandit problem can be ascertained using standard arguments from the theory of stationary dynamic programming. Indeed, the only modification in the arguments that is required from those used in section 4 lies in the definition of the state space to be used.

⁶It is apparent that the framework described here is the most general framework of switching costs in which the existence of an optimal index may reasonably be expected. Under more inclusive situations (such as, say, where the cost of switching from arm i to arm j is given by c_{ij}), it is clearly not possible to have an optimal index strategy if the index on arm i is to depend solely on the characteristics of arm i .

When switching costs are allowed to be non-zero, the state of the bandit problem cannot be adequately described by just the vector of current states $(x_i)_{i \in I}$ of the individual arms (i.e., optimal continuations cannot be calculated with just this knowledge); rather it is also important to know the arm that was in use in the previous period. Thus, in the representation of the bandit problem as a dynamic programming exercise, the state space is given by $\Delta = X \times I$, where as always $X = \times_{i \in I} X_i$. Routine arguments now show that

Theorem 6.1 *The value function $V: \Delta \rightarrow \mathbb{R}$ of the problem $\{I, (C_i^*)_{i \in I}\}$ satisfies the Bellman optimality equation at all $(x, j) \in \Delta$:*

$$V(x, j) = \bigvee_{i \in I} L_i V(x, j), \quad (6.1)$$

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

$$L_i V(x, j) = r_i(x_i) - c_i - d_j + \delta \int V((x_{-i}, \hat{x}_i), i) dq_i(\hat{x}_i | x_i), \quad (6.2)$$

while

$$L_j V(x, j) = r_j(x_j) + \delta \int V(x_{-j}, \hat{x}_j) dq_j(\hat{x}_j | x_j). \quad (6.3)$$

The Markovian strategy defined through any measurable selection from the correspondence of maximizers of (6.1) is an optimal strategy.

6.2 Optimal Index Strategies

The following result, the main result of this section, shows that optimal index strategies no longer exist when switching costs are allowed to be non-zero:

Theorem 6.2 *There is no index λ such that the strategy induced by λ is optimal in every bandit problem $\{I, (C_i^*)_{i \in I}\}$.*

The proof of this theorem occupies the rest of this section, and comes in two stages. In the first stage a simple argument is used to show that any bandit problem in which there are both costs of switching “to” and costs of switching “from” is equivalent to another bandit problem in which there are only costs of switching “to.” In the second stage, we consider bandit problems with only costs of switching “to,” and use a *reductio ad absurdum* method to show that an optimal index strategy cannot exist on this class of bandit problems.

Let a bandit $B = \{I, (C_i^*)_{i \in I}\}$ be given. Define the bandit $\tilde{B} = \{I, (\tilde{C}_i^*)_{i \in I}\}$ from B as follows. For each $i \in I$, let $\tilde{X}_i = X_i$; $\tilde{r}_i(x_i) = r_i(x_i) + (1 - \delta)d_i$ for all $x_i \in X_i$; $\tilde{q}_i = q_i$; $\tilde{c}_i = c_i + d_i$; and, finally, $\tilde{d}_i = 0$. Note that \tilde{B} only has costs of switching “to.”

The only difference between the bandits B and \tilde{B} is that in the bandit B , a cost of d_i is paid every time a switch away from arm i occurs, while in bandit \tilde{B} , d_i is paid “in advance” when the switch to i occurs, but an additional reward of $(1 - \delta)d_i$ is received in each period that arm i is in use. It follows that if arm i is used for t contiguous periods, the present value of the total switching cost paid in the bandit B is $c_i + \delta^t d_i$. In bandit \tilde{B} , this total cost is $(c_i + d_i)$; however, an additional reward of $(1 - \delta)d_i$ is received in each of these t periods, so that the *net* cost paid under \tilde{B} is $c_i + d_i(1 - \sum_{s=0}^{t-1} \delta^s(1 - \delta)) = c_i + \delta^t d_i$, which is exactly the same cost as in the bandit B . The equivalence of B and \tilde{B} is immediate now, completing the first stage of the proof.

We proceed to the second stage of our proof. Suppose that an optimal index did exist on the class of bandit problems with only costs of switching “to.” Denote this index by λ . We will show the presence of a contradiction.

Some new notation will help simplify exposition since all the arms we consider here involve similar characteristics. For $a, b \in \mathbb{R}$, $x \in [0, 1]$, and $c \geq 0$, let $\{x\chi(a) + (1 - x)\chi(b), c\}$ denote the arm with state space $[0, 1]$ and initial state x ; reward function $r(x) = xa + (1 - x)b$; transition probabilities $q(\{1\}|z) = 1 - q(\{0\}|z) = z$ for all $z \in [0, 1]$; and switching cost c .⁷ As a further simplification, let $\{\chi(a), c\}$ denote an arm with switching cost c that pays a reward of a in each period with certainty. In this notation, $\lambda(\{x\chi(a) + (1 - x)\chi(b), c\}; s)$ will denote the value of the hypothetical optimal index on the arm $\{x\chi(a) + (1 - x)\chi(b), c\}$ at the state $s \in \{0, 1\}$. (Recall that $s = 1$ signifies that the arm is currently in use.)

We proceed in a series of claims that establish properties that any optimal index λ must satisfy. Then we show that these properties are not mutually consistent. The intuition underlying this construction is quite straightforward: if there is any possibility of switching back to the arm currently in use after a switch away from it (say, the switch back depends on the realizations from the other arms), then the index on the incumbent arm must be *increasing* in the cost of switching to it, since a higher cost of switching back ought to make the decision-maker more reluctant to leave the arm. On the other hand, if switching back is a zero-probability event (say, because the worst observations on the other arms would still dominate the one currently in use), the arm’s index must be independent of the cost of switching back to it. Thus, the index must depend in a non-trivial way on the characteristics of the other arms, and this is not possible.

Claim 6.1 *For any $a \in \mathbb{R}$, we have $\lambda(\{\chi(a), 0\}; 1) = \lambda(\{\chi(a), 0\}; 0)$.*

Proof Consider a two-armed bandit in which both arms are specified by $\{\chi(a), 0\}$. Then, regardless of the arm currently in use, either arm is an optimal continuation, so the claim follows by the presumed optimality of λ . \square

⁷That is, $(x\chi(a) + (1 - x)\chi(b), c)$ is an arm which pays $(xa + (1 - x)b)$ in the first period of its use, and thereafter either pays a forever (which happens with probability x) or b forever (which happens with probability $(1 - x)$).

Claim 6.2 For any $a \in \mathbb{R}$ and $\epsilon > 0$, we have $\lambda(\{\chi(a + \epsilon), 0\}; 0) > \lambda(\{\chi(a), 0\}; 0)$.

Proof Pick any $b < a$, and consider a three-armed bandit in which arm 1 is $\{\chi(b), 0\}$, arm 2 is $\{\chi(a), 0\}$, and arm 3 is $\{\chi(a + \epsilon), 0\}$. Suppose arm 1 is the one currently in use. It is immediate that selecting arm 3 is the uniquely optimal continuation, from which the claim follows. \square

Claim 6.3 For any $c \geq 0$, $\epsilon > 0$, and $a \in \mathbb{R}$, $\lambda(\{\chi(a + \epsilon), 0\}; 0) > \lambda(\{\chi(a), c\}; 1)$.

Proof Consider a two-armed bandit in which arm 1 is given by $\{\chi(a), c\}$ and arm 2 is given by $\{\chi(a + \epsilon), 0\}$. Suppose arm 1 is the one currently in use. The uniquely optimal continuation strategy is obviously to play arm 2 forever, and the claim obtains. \square

Claim 6.4 For any $c > 0$, it is the case that for almost every $a \in \mathbb{R}$, we have $\lambda(\{\chi(a), 0\}; 1) \geq \lambda(\{\chi(a), c\}; 1)$.

Proof Pick any $c > 0$. By claim 6.2, $\lambda(\{\chi(a), 0\}; 0)$ is a strictly increasing function of a , and so is continuous almost everywhere. Pick any continuity point \hat{a} . From claim 6.3, we have $\lambda(\{\chi(\hat{a} + \epsilon), 0\}; 0) > \lambda(\{\chi(\hat{a}), c\}; 1)$ for every $\epsilon > 0$, so taking limits as $\epsilon \rightarrow 0$, we obtain $\lambda(\{\chi(\hat{a}), 0\}; 0) \geq \lambda(\{\chi(\hat{a}), c\}; 1)$. By claim 6.1, $\lambda(\{\chi(\hat{a}), 0\}; 0) = \lambda(\{\chi(\hat{a}), 0\}; 1)$, and the result follows. \square

The claims leading upto claim 6.4 relied on a series of comparisons in which coming back to the incumbent arm (if a switch away from it occurred) was a zero-probability event. The next two claims now use another series of comparisons, in which there is a non-zero chance of coming back to the incumbent arm (depending on the observations from the other arms). This leads to a contradiction to claim 6.4, by showing that $\lambda(\{\chi(a), c\}; 1)$ must be strictly increasing in c for $c \geq 0$.

Claim 6.5 Suppose $a > b$. Then, $\lambda(\{x\chi(a) + (1 - x)\chi(b), c\}; 0)$ must be increasing in x .

Proof Let $x_1, x_2 \in [0, 1]$ with $x_1 > x_2$. Pick $\alpha \in \mathbb{R}$ to satisfy

$$[ax_1 + b(1 - x_1) - (1 - \delta)c] > [ax_2 + b(1 - x_2) - (1 - \delta)c].$$

For $c^* \in \mathbb{R}$, consider a two-armed bandit in which arm 1 is given by $\{\chi(\alpha), c^*\}$ and the second arm is given by $\{x_i\chi(a) + (1 - x_i)\chi(b), c\}$. Suppose also that arm 1 is the one currently in use. When c^* is sufficiently large, it is uniquely optimal to switch to the second arm and stay there forever if $i = 1$; and to stay with arm 1 forever if $i = 2$. Since λ is optimal by hypothesis, we must have:

$$\lambda(\{x_1\chi(a) + (1 - x_1)\chi(b), c\}; 0) > \lambda(\{\chi(\alpha), c^*\}; 1) > \lambda(\{x_2\chi(a) + (1 - x_2)\chi(b), c\}; 0),$$

so we indeed have $\lambda(\{x_1\chi(a) + (1 - x_1)\chi(b), c\}; 0) > \lambda(\{x_2\chi(a) + (1 - x_2)\chi(b), c\}; 0)$, as required. \square

Claim 6.6 For and $a \in \mathbb{R}$, $\lambda(\{\chi(a), c\}; 1)$ must be increasing in c .

Proof Let $c_1 > c_2$. Pick $\alpha, \beta \in \mathbb{R}$ to satisfy

$$\begin{aligned}\alpha &> a \\ \beta &< (a - (1 - \delta)c_1).\end{aligned}$$

Since $c_1 > c_2$, we also have $\beta < (a - (1 - \delta)c_1) < (a - (1 - \delta)c_2)$. Define $x_i \in [0, 1]$ by:

$$x_i = \left(\frac{(1 - \delta)(a - \beta + \delta c_i)}{\alpha(1 - \delta)\beta - \delta a + \delta(1 - \delta)c_i} \right)$$

Note that $1 > x_1 > x_2 > 0$. Consider the two-armed bandit in which arm 1 is given by $\{\chi(a), c_i\}$ and arm 2 by $\{x_i\chi(\alpha) + (1 - x_i)\chi(\beta), 0\}$. Suppose also that arm 1 is the one currently in use. A simple calculation shows that the choice of x_i implies either arm is an optimal continuation in this problem. It follows that we must have

$$\lambda(\{\chi(a), c_i\}; 1) = \lambda(\{x_i\chi(\alpha) + (1 - x_i)\chi(\beta), 0\}; 0).$$

This establishes the result by claim 6.5, since $x_1 > x_2$. \square

Claims 6.4 and 6.6 are in evident contradiction, completing the proof of the theorem. \square

A Proof of Theorem 4.2

The proof takes the form of several lemmata leading to the key Lemma A.4 which asserts that if arm i attains the supremum of the indices at the initial state and j does not, then beginning with i and continuing with the index strategy is strictly superior to beginning with j and continuing with the index strategy. We employ the notation of Theorem 4.3 throughout. Fix an initial state x at which the conditions of the Theorem are satisfied, and denote by $\bar{\mu}_i$ the initial values of the Gittins indices $\mu_i(x_i)$.

Recall the definition of the strategy $\sigma_i[m]$ in the stopping problem defining the Gittins index on arm i , and of the expressions $H_i(t, m)$, $p_i(t, m)$, and $R_i(t, m)$. Let $H_i(\infty, m)$ be the set of histories under which $\sigma_i[m]$ never accepts the terminal reward (i.e., on which the index stays above m forever), and let $p_i(\infty, m)$ be the probability of $H_i(\infty, m)$. Finally, let $R_i(\infty, m)$ be the total expected discounted reward from arm i conditional on $H_i(\infty, m)$.

Lemma A.1 *For any $i \in I$ and $m \in \mathbb{R}$, it is the case that*

$$\sum_{\tau=1}^{\infty} p_i(\tau, m) R_i(\tau, m) + p_i(\infty, m) R_i(\infty, m) > m \left(1 - \sum_{\tau=1}^{\infty} \delta^\tau p_i(\tau, m) \right) \quad (1.4)$$

if $m < \bar{\mu}_i$; that equality holds in (4.1) if $m = \bar{\mu}_i$; and that the reverse strict inequality holds if $m > \bar{\mu}_i$.

Proof The total discounted reward associated with the strategy $\sigma_i[m]$, denoted $A_i[m]$, is evidently given by the following:

$$A_i[m] = \sum_{\tau=1}^{\infty} p_i(\tau, m) (R_i(\tau, m) + \delta^\tau m) + p_i(\infty, m) R_i(\infty, m) \quad (1.5)$$

Since it is uniquely optimal to select arm i initially when $m < \bar{\mu}_i$, we have $A_i[m] > m$ when $m < \bar{\mu}_i$; rearranging this inequality yields precisely the first inequality in the statement of the lemma. The other two are obtained similarly: since either arm is an optimal initial choice when $m = \bar{\mu}_i$, we have $A_i[m] = m$ in this case; and finally, since accepting the terminal reward right away is the unique optimal action when $m > \bar{\mu}_i$, we have $A_i[m] < m$ when $m > \bar{\mu}_i$. \square

Some further notation is unfortunately required for the next two lemmata. For any $j \in I$, denote by $\sigma(j)$ the strategy in the bandit problem that begins by selecting arm j initially, and after the first period proceeds by choosing the arm with the highest Gittins Index in each period. Additionally, let $\sigma(ij)$ be the strategy that begins with arm i , stays with arm i until the first point where the index on arm i drops below its original value of $\bar{\mu}_i$, then switches to arm j and stays with arm j until the index on arm j falls below the original index $\bar{\mu}_i$ of arm i , and then proceeds by choosing the arm with the highest Gittins index in each period. Recall that $W(\sigma)(x)$ denotes the worth of the strategy from the initial state x . Since x is fixed, we suppress dependence on it.

Lemma A.2 $W(\sigma(ij)) = W(\sigma(ji))$ if $\bar{\mu}_i = \bar{\mu}_j$.

Proof A little algebra shows that

$$\begin{aligned}
W(\sigma(ij)) &= p_i(\infty, \bar{\mu}_i)R_i(\infty, \bar{\mu}_i) + \sum_{t=1}^{\infty} p_i(t, \bar{\mu}_i)R_i(t, \bar{\mu}_i) \\
&\quad + \sum_{t=1}^{\infty} \delta^t p_i(t, \bar{\mu}_i) \left(p_j(\infty, \bar{\mu}_i)R_j(\infty, \bar{\mu}_i) + \sum_{\tau=1}^{\infty} p_j(\tau, \bar{\mu}_i)R_j(\tau, \bar{\mu}_i) \right) \\
&\quad + \sum_{t=1}^{\infty} \sum_{\tau=1}^{\infty} p_i(t, \bar{\mu}_i)p_j(\tau, \bar{\mu}_i)E_{ij}^*(t, \tau, \bar{\mu}_i) \\
&= A_i[\bar{\mu}_i] + \sum_{t=1}^{\infty} \delta^t p_i(t, \bar{\mu}_i)A_j[\bar{\mu}_i] + \sum_{t=1}^{\infty} \sum_{\tau=1}^{\infty} p_i(t, \bar{\mu}_i)p_j(\tau, \bar{\mu}_i)E_{ij}^*(t, \tau, \bar{\mu}_i)
\end{aligned}$$

where $E_{ij}^*(t, \tau, \bar{\mu}_i)$ is the continuation value of the strategy $\sigma(ij)$ conditional on the arm i having fallen below $\bar{\mu}_i$ for the first time in t periods, and arm j having fallen below $\bar{\mu}_i$ for the first time in τ periods. Similarly, we also have

$$W(\sigma(ji)) = A_j[\bar{\mu}_i] + \sum_{\tau=1}^{\infty} \delta^\tau p_j(\tau, \bar{\mu}_i)A_i[\bar{\mu}_i] + \sum_{\tau=1}^{\infty} \sum_{t=1}^{\infty} p_j(\tau, \bar{\mu}_i)p_i(t, \bar{\mu}_i)E_{ji}^*(\tau, t, \bar{\mu}_i)$$

where E_{ji}^* is defined in the same way as E_{ij}^* with the obvious changes. It is easy to see that, by the independence of the arms, and the assumption that $\bar{\mu}_i = \bar{\mu}_j$, we must have $E_{ij}^* = E_{ji}^*$. In turn, this gives us the following:

$$W(\sigma(ij)) - W(\sigma(ji)) = A_i[\bar{\mu}_i] \left(1 - \sum_{\tau=1}^{\infty} \delta^\tau p_j(\tau, \bar{\mu}_i) \right) - A_j[\bar{\mu}_i] \left(1 - \sum_{t=1}^{\infty} \delta^t p_i(t, \bar{\mu}_i) \right)$$

Substituting for $A_i[m]$ and $A_j[m]$ and appealing to Lemma A.1 completes the proof. \square

Lemma A.2 establishes in particular that if two arms have the same index, the order in which they are used does not matter in evaluating the index strategy. We will now show that if arm i attains the supremum of the indices while arm j does not, the strategy $\sigma(ij)$ is strictly preferable to the strategy $\sigma(j)$.

Lemma A.3 $W(\sigma(ij)) > W(\sigma(j))$ whenever $\bar{\mu}_i = \sup_{k \in I} \bar{\mu}_k > \bar{\mu}_j$.

Proof Under the stated hypothesis that $\bar{\mu}_i = \sup_{k \in I} \bar{\mu}_k$, $W(\sigma(j))$ can be thought of as the strategy that begins with arm j , switches to i at the first time $t = 1, 2, \dots$, at which the index on j drops below $\bar{\mu}_i$, then stays with i until the index on i also drops below $\bar{\mu}_i$, and proceeds by picking at each subsequent stage the arm with the highest Gittins index at that point. Thus, after some algebra, we obtain,

$$W(\sigma(j)) = A_j[\bar{\mu}_i] + \sum_{\tau=1}^{\infty} \delta^\tau p_j(\tau, \bar{\mu}_i) A_i[\bar{\mu}_i] + E_{ji}^*$$

where E_{ji}^* is defined exactly as previously. So,

$$W(\sigma(ij)) - W(\sigma(j)) = A_i[\bar{\mu}_i] \left(1 - \sum_{\tau=1}^{\infty} \delta^\tau p_j(\tau, \bar{\mu}_i) \right) - A_j[\bar{\mu}_i] \left(1 - \sum_{t=1}^{\infty} \delta^t p_i(t, \bar{\mu}_i) \right)$$

By Lemma A.1, $A_i[\bar{\mu}_i] = \bar{\mu}_i(1 - \sum_{t=1}^{\infty} \delta^t p_i(t, \bar{\mu}_i))$, while since $\bar{\mu}_i > \bar{\mu}_j$ by hypothesis, another appeal to Lemma A.1 gives $A_j[\bar{\mu}_i] < \bar{\mu}_i(1 - \sum_{\tau=1}^{\infty} \delta^\tau p_j(\tau, \bar{\mu}_i))$. Substituting these in the expression above, we obtain $W(\sigma(ij)) - W(\sigma(j)) > 0$, as required. \square

In words, Lemma A.3 establishes that if arm i is the maximum of the indices at the initial state while arm j is not, then there is a strategy beginning with i that does strictly better than the strategy that begins with j and proceeds by picking in each subsequent period, one of the arms with the highest Gittins index at that point. It is now relatively easy to show that:

Lemma A.4 *If $\bar{\mu}_i = \sup_{k \in I} \bar{\mu}_k$, and $\bar{\mu}_i > \bar{\mu}_j$, then $W(\sigma(i)) > W(\sigma(j))$.*

Proof Under the stated hypotheses, the previous lemma shows that $\sigma(ij)$ strictly improves on $\sigma(j)$. Of course, $\sigma(ij)$ initially selects arm i , which has the highest of the Gittins indices at the initial state. We first claim that if, at any point, $\sigma(ij)$ recommends continuing with an arm that does *not* have the highest index at that point, then there is an alternative continuation strategy which recommends continuing with one of the arms with the highest index at that point, and which does strictly better in the continuation than $\sigma(ij)$.

To see this, note that under $\sigma(ij)$ the decision-maker always picks the arm with the largest Gittins index at each point in time, at least until the first time a state is reached where the strategy recommends that arm j be selected. For notational ease, call this new state $\bar{x} = (\bar{x}_i)_{i \in I}$. (Of course, \bar{x} differs from the initial state x only in the value of the i -th coordinate x_i .) At this point carrying on with $\sigma(ij)$ amounts to using $\sigma(j)$ from the new initial state \bar{x} . If, however, arm j is not the arm with the maximal index at \bar{x} (say there is $k \in I$ such that $\mu_k = \sup_{l \in I} \mu_l > \mu_j$, where all the indices are evaluated at \bar{x}), then Lemma A.3 shows that it is strictly better to use the strategy $\sigma(kj)$ in the continuation, where, of course, $\sigma(kj)$ is defined with respect to the new state \bar{x} . This proves the claim.

Iterating on the claim shows that the Gittins index strategy of picking the arm with the highest Gittins index at each point is strictly superior to picking j initially and continuing with the Gittins index strategy, if j is not the arm with the highest Gittins index. \square

To complete the proof of Theorem 4.2, let σ be any strategy in the bandit problem. Consider the strategy σ^T which imitates σ upto period T , and then proceeds with the Gittins index strategy. It is easy to see that σ^T is well defined: after any T -history, at most T of the arms have had their initial states altered. Since the Gittins index strategy was well defined at the initial state x , it continues to be well defined at this resulting new state which differs from x in at most T coordinates. And since this last statement is true for any T -history, σ^T itself is well defined.

For any $\epsilon > 0$, the total discounted expected rewards under σ and σ^T can be made to differ by less than ϵ by choosing T sufficiently large. (This follows since rewards are uniformly bounded.) Now consider any $(T - 1)$ -history in which in period $(T - 1)$, σ (and, therefore σ^T) picks an arm j that does not have the highest index at that point. Since σ^T continues from period T on with the Gittins index strategy, Lemma A.2 shows that the total reward conditional on this history can be strictly improved over σ^T , by instead picking an arm with the highest index at this point, i.e., by continuing from period $(T - 1)$ with the Gittins index strategy. It easily follows that $\sigma^{(T-1)}$ strictly improves over σ^T , unless σ^T almost surely picks only arms with the highest index at all points from period $(T - 1)$ onwards, i.e., unless effectively σ^T and $\sigma^{(T-1)}$ are the same. In turn, this implies $W(\sigma) - W(\sigma^{(T-1)}) < \epsilon$, also.

Iterating back to zero, we obtain $W(\sigma) - W(\sigma^0) < \epsilon$, where, of course, σ^0 is the Gittins index strategy. Since ϵ is arbitrary, we finally see that the Gittins index strategy does no worse than the strategy σ . Finally, since σ is also arbitrary, this shows that the Gittins index strategy is, indeed, an optimal strategy.

To see the other part of Theorem 4.2—that every optimal strategy is also a Gittins index strategy—consider any optimal strategy σ , and suppose it is not a Gittins index strategy. Let $t \in \{0, 1, 2, \dots\}$ denote any date such that with positive probability σ fails to use a Gittins index maximal arm for the first time at time t . (Since σ is not a Gittins Index strategy, such a date t must exist.) Consider the strategy $\hat{\sigma}$ that imitates σ upto, and including, period t , and then proceeds according to the Gittins index strategy. Since the Gittins index strategy is also an optimal strategy, and σ is optimal by hypothesis, it follows that $\hat{\sigma}$ must also be an optimal strategy. But by Lemma A.4, the Gittins index strategy strictly improves upon $\hat{\sigma}$ in the continuation reward from period t onwards, since σ uses arms that are not Gittins index maximal with positive probability in period t . This implies $\hat{\sigma}$, and therefore σ , cannot be optimal strategies, a contradiction, completing the proof. \square

References

- [1] Agrawal, R.; M.V. Hegde, and D. Teneketzis (1988) Asymptotically Efficient Adaptive Allocation Rules for the Multiarmed Bandit Problem with Switching Costs, *IEEE Transactions on Optimal Control* 33(10), 899-906.
- [2] Berry, D. and B. Fristedt (1985) *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, London.
- [3] Banks, J.S. and R.K. Sundaram (1992a) Denumerable-Armed Bandits, *Econometrica* 60(5), 1071-1096.
- [4] Banks, J.S. and R.K. Sundaram (1992b) A Class of Bandit Problems Yielding Myopic Optimal Strategies, *Journal of Applied Probability*, 625-632.
- [5] Banks, J.S. and R.K. Sundaram (1994) Switching Costs and the Gittins Index, *Econometrica* 62(3), 687-694.
- [6] Basu, A; A. Bose, and J.K. Ghosh (1990) An Expository Review of Sequential Design and Allocation Rules, Technical Report 90-08, Department of Statistics, Purdue University.
- [7] Feldman, D (1962) Contributions to the "Two-Armed Bandit" Problem, *Annals of Mathematical Statistics* 33, 847-856.
- [8] Feldman, M. and M. Spagat (1993), Optimal Learning with Costly Adjustment, Working Paper, Department of Economics, Brown University.
- [9] Gittins, J. (1979) Bandit Processes and Dynamic Allocation Indices, *Journal of the Royal Statistical Society, Series B* 41, 148-164.
- [10] Gittins, J. (1989) *Allocation Indices for Multi-Armed Bandits*, Wiley, London.
- [11] Gittins, J. and D. Jones (1974) A Dynamic Allocation Index for the Sequential Allocation of Experiments, in (J. Gani, et al, Eds.) *Progress in Statistics*, North Holland, Amsterdam.
- [12] Kolonko, M. and H. Benzing (1983) The Sequential Design of Bernoulli Experiments including Switching Costs, unpublished.
- [13] Lai, T.L. and H. Robbins (1985) Asymptotically Efficient Adaptive Allocation Rules, *Advances in Applied Mathematics* 6, 4-22.
- [14] Mortensen, D. (1985) Job Search and Labor Market Analysis, in (O. Ashenfelter and J. Layard, Eds.) *Handbook of Labor Economics* Vol. II, North Holland, New York.
- [15] Pressman, E.L. and I.M. Sonin (1990) *Sequential Control with Partial Information*, Academic Press, New York.

- [16] Rieder, U. (1975) Bayesian Dynamic Programming, *Advances in Applied Probability* 7, 330-348.
- [17] Rothschild, M. (1974) A Two-Armed Bandit Theory of Market Pricing, *Journal of Economic Theory* 9, 185-202.
- [18] Schäl, M. (1979) Dynamic Programming and Statistical Decision Theory, *Annals of Statistics* 7(2), 432-445.
- [19] Vicusi, W. (1979) Job Hazards and Worker Quit Rates: An Analysis of Adaptive Worker Behavior, *International Economic Review* 20, 29-58.
- [20] Weizman, M.L. (1979) Optimal Search for the Best Alternative, *Econometrica* 47, 641-654.
- [21] Whittle, P. (1981) Arm-Acquiring Bandits, *Annals of Probability* 9(2), 284-292.
- [22] Whittle, P. (1982) *Optimization Over Time: Dynamic Programming and Stochastic Control*, Vol. I, Wiley, New York.