Rochester Center for
Economic Research

Denumerable-armed Bandits

Banks, Jeffrey S. and Rangarajan K. Sundaram

Working Paper No. 277
May 1991

University of
Rochester

Denumerable-Armed Bandits

Jeffrey S. Banks and Rangarajan K. Sundaram

Rochester Center for Economic Research
Working Paper No. 277

May 1991

Denumerable–Armed Bandits

Jeffrey S. Banks

and

Rangarajan K. Sundaram

University of Rochester

and

The Rochester Center for Economic Research
Working Paper No. 277

May 1991

This supersedes RCER Working Paper # 235.

It is a pleasure to acknowledge the assistance of an anonymous referee whose numerous comments and suggestions have materially altered the body of this paper. We also benefited from the comments of a second anonymous referee, Prajit Dutta, David Levine, Bill Zame, and, especially, Mahmoud El–Gamal and Nicholas Kiefer. We would like to thank seminar audiences at Buffalo, CalTech, Carnegie–Mellon, George Mason, Harvard, Michigan, Johns Hopkins, Rochester, Washington–St. Louis, Western Ontario, and Yale, and participants at the III Buffalo–Cornell–Rochester conference in Ithaca, and the Mid West Mathematical Economics conference in Urbana–Champaign, for their input. The first author acknowledges financial support provided by the Sloan Foundation and the NSF.

1. Introduction and Summary

This paper studies the class of denumerable– (i.e., finite– or countably infinite–) armed Bandit problems with the characteristic that each arm available to the decision–maker generates rewards according to one of a finite number of distributions, called the ***types*** of the arm.[1] The arms are assumed ***independent:*** trying one arm is uninformative about the types of the other arms. It is also assumed that discounting by the decision–maker is geometric over an infinite horizon. No other restrictions are employed for obtaining the general results of the paper. In particular, the forms of the reward distributions and the number of types are allowed to be arbitrary and to vary across the arms, and the discount factor of the decision–maker is allowed to take on any value in [0,1).

It is well known that when the number of arms is finite, optimal strategies may be obtained through solving a family of stopping problems that associates with each arm an index, known as the ***Dynamic Allocation Index*** (DAI), or the ***Gittins Index***, where this index depends only on the current belief on that arm's type.[2] In section 3 (Lemma 3.1), we characterize the stopping problem defining the DAI. The resulting properties of this problem enable us to show (Lemma 3.2) that for each arm, the DAI is a continuous, quasi–convex function of the prior on the arm, which possesses in addition, a certain monotonicity property. These results, while of independent interest, prove valuable in the sequel.

Section 4 addresses the existence issue in the context of infinite–armed Bandits. In Theorem 4.1, we derive sufficient conditions for the existence of optimal strategies, indeed, optimal DAI strategies, in such a context. Two examples show that these conditions are not necessary (Example 4.1), but that they are minimally sufficient (Example 4.2).

[1] A number of our results do not depend on this finiteness restriction, as we explain at the end of this section.

[2] See, e.g, Berry and Fristedt (1985), Gittins (1989), or Whittle (1982).

Sections 5 and 6 are concerned with the characteristics of optimal strategies in denumerable–armed Bandit problems. In section 5, we analyze the stochastic process governing the the continuous play of an arm under the optimal strategy. Our main result in this section is that at each point in time, the arm selected by the optimal strategy will remain an optimal selection ***forever*** with strictly positive probability. More specifically, we show in Theorem 5.1 that for each arm that becomes optimal at some point, there must exist a type in the support of that arm with the following property: if that type were in fact the true type of the arm (i.e. the type generating the observed rewards), then the arm would survive forever with non–zero probability under the optimal strategy. Surprisingly, the fact that some type in the support of an arm may lead to survival forever with positive probability has no implications, in general, for the survival prospects of the arm under the "better" types in its support, that is, under types which generate a higher expected payoff. Example 5.1 illustrates this point. Each arm here may be one of the same three types. Under the optimal strategy, the best type fails in finite time with probability one, while the second–best type survives forever with probability one.

Section 6 turns to an examination of the trade–off between ***current*** reward maximization and the acquisition of information to better ***future*** decision making ability, a trade–off the literature on Bayesian learning has frequently emphasized as a hallmark of dynamic decision making under uncertainty.[3] In particular, our focus here is on how well optimal strategies measure up against ***myopic strategies,*** i.e., strategies that select the arm to be played in each period solely on the basis of the current (one–period) expected reward from the arms, thereby "ignoring" the possible information content of current actions. We show (Theorem 6.1) that in a restricted, but still surprisingly large, class of Bandit problems –specifically, all those in which each arm is one of the same two types– myopic

[3] See, e.g., Berry and Fristedt (1985, Ch.1), or Easley and Kiefer (1988). Indeed, Whittle, commenting on this trade–off, states that the Bandit framework "..embodies, in essential form, a conflict evident in all human action." (Whittle, 1982, p.210)

strategies are fully optimal, regardless of the reward distributions, the prior beliefs, ***or the discount factor***. That this result cannot, however, be extended is demonstrated by Example 6.1, which considers a situation in which each arm is one of the same three possible types and proves that myopic strategies are no longer optimal.

We also derive some implications of these results for the behavior of optimal plans. A simple consequence of Theorem 5.1 is that the probability that an arm will survive at least (t+1)–periods, ***conditional*** on its surviving at least t periods and being one of the aforementioned types which "survive forever", must converge to unity as t goes to infinity (Corollary 5.1). But even here, behavior need not be regular: the optimality of myopic strategies in the two–type case enables showing that this convergence could very well be non–monotone in t (Example 6.2). Example 6.2 also illustrates another interesting point: for a family of Bernoulli reward distributions – a frequently studied case in the Bandit literature – the stochastic process of the continued use of an arm follows a ***random walk*** on the non–negative integers, with an absorbing barrier at zero that signifies replacement of the arm.

A class of Bandit problems that has enjoyed some success in economics (notably in the theory of job–search and matching; see below) is a special case of our framework that we label ***Stationary Bandits***. A stationary Bandit is an infinite–armed Bandit in which all arms are *a priori* identical, namely, the set of possible types is the same across all arms, as is the prior belief concerning an arm's type. All our general results apply *in toto* to stationary Bandits of course; but the additional structure here also enables a sharper characterization. In particular, optimal strategies ***always*** exist in stationary Bandits, and never involve the ***recall*** of a previously selected and discarded arm (Corollary 4.1);[4] and the expected number of arms employed in an optimal strategy in a stationary Bandit is

[4] Many papers in labor economics ***impose*** the restriction that recall is not permissible, and while Jovanovic (1979), in his pioneering paper on job–matching, mentions that the "no–recall" result holds in his framework, his paper does not contain a proof.

finite, so that with probability 1 only a finite number of arms are ever used (Corollary 5.2).

The framework of stationary Bandits, as mentioned above, has been a popular framework for the analysis of decision making in labor markets. The parametrized "matching" models of Jovanovic (1979), Wilde (1979), and Viscusi (1979) all have as their scenario a worker who periodically receives information concerning her current job's true but unobservable characteristics.[5] In Jovanovic (1979), for instance, the productivity of the worker is job–specific, the worker's compensation in each period is her expected productivity, and the worker uses output observations to infer her true productivity with the current firm and thereby predict future wages from remaining with the current job. Moreover, all untried firms are *ex–ante* identical. The resulting optimization problem can evidently be viewed as a stationary Bandit, in which the jobs are the arms of the Bandit, and the worker's true productivity on a particular job is the arm's true type.

The motivation for the current project was itself an alternative interpretation of the stationary Bandit framework: as a median–voter model of repeated elections. Consider a single voter faced with a set of candidates from whom she elects one to be her political representative for the current period. The chosen representative (stochastically) generates per–period rewards for the voter as a function of some unobservable, candidate–specific parameters. The voter, through her ability to elect and observe candidates while in office, attempts to identify "good" candidates. Treating the candidates as the arms, and the candidate–specific parameters as the arms' true types, this forms a special case of the framework we study in this paper. It is worth noting that most models of repeated elections (for instance, Barro (1973), Ferejohn (1986), or Austen–Smith and Banks (1989)), study the voter's decision problem from a "moral hazard" perspective: the voter is attempting to control the actions of their current representative through their choice of

[5] cf. Mortensen (1985) for an in–depth survey of these and other search models in labor economics.

re–election rule.[6] In contrast, with the interpretation above, the Bandit framework becomes a model of adverse selection.

Several other economic (and non–economic) problems are also amenable to being modeled in the Bandit framework. For instance, an alternative labor market version of the stationary Bandit model is obtained by treating the arms of the Bandit as workers who differ in their productivity; and the decision–maker as a firm searching over these workers. As other examples, we mention general search problems involving non–durable experience goods, and models of dating and marriage.

Two important remarks are in order at this point. First, at the risk of repetition, we wish to emphasize that ***none*** of our general results require any special structure on the reward distributions or restrictions on the value of the discount factor. Second, the assumption of a ***finite*** number of types in the support of each arm is exploited in order to establish the quasi–convexity of the DAI in the prior. We are unclear on the extent to which this may be generalized. However, a number of our results – for instance, the existence of optimal plans in infinite–armed bandits under suitable conditions, or the continuity of the DAI – do not in any way depend on this finiteness restriction, and may easily, if at considerable notational cost, be generalized to arbitrary sets of types.

Finally, we briefly indicate the related theoretical literature. Three excellent summaries of results for finite–armed Bandit problems are the monographs by Berry and Fristedt (1985), who provide an exhaustive analysis of Bandits under general discount sequences; Gittins (1989), who discusses index theorems for Bandits;[7] and Pressman and

[6] Rogoff (1990) considers a repeated elections incomplete information game, where candidate "types" denote competency and where candidates while in office take actions which both effect the voter's utility and potentially signal their competence. However, a candidate's type is uncorrelated across electoral cycles, so there is no issue of the voter "learning" a candidate's type over time.

[7] In our companion paper (Banks and Sundaram, 1991) we supplement Gittins' results by demonstrating the optimality of (suitably defined) index strategies in Bandit problems when there are costs for switching between arms, and when arms may "die off" with positive probability when in use.

Sonin (1990), who focus on Bandits with ***dependent*** arms. There is also an extensive literature on optimal Bayesian learning in economic environments, e.g., Rothschild (1974), Easley and Kiefer (1988), McLennan (1988), and Feldman (1989). A question of primary interest in the latter has been whether optimally–acting individuals will, in the limit, learn the "truth", i.e., the parameter values actually driving the the model. Two interpretations of the learning question could be provided in the framework we have adopted; but learning cannot occur with certainty in either case. First, one could view the unknown parameter as the vector describing the true type of each of the arms. The main result of Section 5 shows that with positive probability the ***very first*** arm employed will be used forever; hence with positive probability the decision–maker only learns the true type of a single arm, so that it cannot be the case that learning occurs in this sense with probability one. A second interpretation would be to consider only whether the decision–maker would be able to identify an arm of the "best" type in the limit. But Example 5.1 which shows that, even in a stationary Bandit, the best type may last only finitely long with probability one, demonstrates that "learning" in this weaker sense need not occur either.

2. The Framework

The family of Bandit problems we study has the following structure. There are N independent arms, where $N \geq 2$ is either a positive integer or ∞. The set of all arms is denoted by $\mathfrak{N}$, with generic element i. Arm i may be one of a finite number K(i) of ***types.*** If the true type of arm i is $k \in \{1,...,K(i)\}$, then it generates rewards according to the density $f_k^i(.)$.[8] Let R_k^i denote the corresponding expected reward, i.e., $R_k^i = \int r f_k^i(r)dr$. We assume, without loss of generality, that these rewards are ordered for each i in the sense

[8] The use of continuous reward distributions is unnecessary for our results. With transparent modifications, all proofs continue to be valid if the reward distributions are instead discrete, i.e., have finite or countable support.

that $R^i_1 \geq R^i_2 \geq \ldots \geq R^i_{K(i)}$, with at least one inequality strict. We also assume that

$$R^* := \sup_{i,k} |R^i_k| < \infty. \tag{2.1}$$

We make no assumptions regarding common support of, or stochastic dominance in the reward distributions arising from, the densities (f^i_k).

In each period of an infinite horizon, a decision–maker (hereafter referred to as the ***principal***) must decide on the choice of arm to be employed that period. However, the true type of some or all of the arms (and, hence, the true reward distribution associated with those arms) may be *a priori* unknown to the principal. The principal begins with a vector of ***prior beliefs*** $P = (p(i))_{i \in \mathfrak{N}}$, where[9] $p(i) \in \Delta^{K(i)-1}$ represents the the principal's belief regarding the type distribution of arm i, *viz.*, the k–th coordinate of p(i) is the principal's prior probability that the true type of arm i is k.

The beliefs are updated using observed rewards as follows. Let $P^t = (p^t(i))_{i \in \mathfrak{N}}$ represent the principal's beliefs at the beginning of any period t, and suppose arm i is chosen that period and the reward r is witnessed. Then, by independence, the reward r reveals no information about the true types of arm $j \neq i$, so that we have $p^{t+1}(j) = p^t(j)$ for all $j \neq i$. For arm i the updated belief $p^{t+1}(i)$ is given by the ***Bayes map*** $\beta_i(p^t(i);r) = (\beta_{ik}(p^t(i);r)_{k=1,\ldots K(i)}$, where,

$$\beta_{ik}(p^t(i),r) = p^t_k(i).f^i_k(r) / \left[\Sigma^{K(i)}_{m=1} p^t_m(i).f^i_m(r)\right]. \tag{2.2}$$

A ***t–history*** for the Bandit is a description of the arm used in each period up to t and the corresponding rewards witnessed. Let H_t be the set of all possible t–histories. A ***strategy*** σ for the principal is a specification of the arm to be played in any period as a function of the initial belief and the history up to that period. Formally, σ is a sequence of

[9] For any finite integer n, Δ^{n-1} will denote the positive unit simplex in $\mathbb{R}^n$: $\Delta^{n-1} = \{x \in \mathbb{R}^n | \; x_i \geq 0, \text{ and } \Sigma_i x_i = 1\}$.

measurable maps $\{\sigma_t\}_{t=0}^{\infty}$ where $\sigma_0 \in \mathfrak{N}$, and for $t \geq 1$, $\sigma_t : H_t \to \mathfrak{N}$. Let Σ denote the set of all strategies.

The principal discounts future rewards geometrically, using the discount factor $\delta \in [0,1)$. Given the initial prior P, each strategy σ defines in the obvious (if notationally complex) way an expected t–th period reward $r_t(\sigma;P)$ for the principal. Hence, each strategy σ also defines a total expected reward $W(\sigma;P)$ as

$$W(\sigma;P) \;=\; \Sigma_{t=0}^{\infty}\, \delta^t r_t(\sigma;P). \tag{2.3}$$

The principal's objective is to find a strategy σ^* such that $W(\sigma^*;P) \geq W(\sigma;P)\ \forall\ \sigma \in \Sigma$. When such a strategy exists, it will be called an ***optimal strategy.***

Of special interest is a class of Bandit problems that we label ***Stationary Bandits.*** A stationary Bandit is an infinite–armed Bandit in which all arms are *a priori* identical, that is, for all $i \in \mathfrak{N}$, we have (i) $K(i) = K$, (ii) $f_k^i = f_k$, $k = 1,\ldots,K$,and (iii) $p(i) = \pi \in \Delta^{K-1}$. As we noted in the Introduction, stationary Bandits have been widely utilized in economics, notably in the labor–market literature. Evidently, stationary Bandits form a special case of the family of Bandit problems described above; consequently, all of our results retain their validity in this setting as well. But the additional structure provided by the assumption of *a priori* identical arms often enables a considerable strengthening of the results that we prove to hold in general. These are described at the end of each section.

3. The Dynamic Allocation Index

Gittins and Jones (1974) proved that for ***finite***–armed Bandit problems of the type detailed above, an optimal strategy can be obtained through solving a family of stopping problems, thereby associating with each arm an index which depends solely on the current prior belief on that arm. This index, the ***Dynamic Allocation Index*** or DAI (frequently also referred to as the ***Gittins Index***), plays a prominent role in our analysis of the framework outlined in Section 2. We describe in this section the construction of the DAI for a generic arm i, and derive some basic resulting properties. The proofs of all results in this section may be found in Appendix I.

For simplicity, we suppress the dependence of the various parameters on i. Suppose arm i is one of K types. Let the corresponding reward densities be denoted $(f_1,...,f_K)$, with associated expected rewards $R_1,...,R_K$. Let $p \in \Delta^{K-1}$ denote the prior belief on arm i; $R(p) = \Sigma_k p_k R_k$ the expected one period reward from playing arm i; and $f(p)(.) = \Sigma_k p_k f_k(.)$ the expected density of rewards.

Consider the optimal stopping problem in which the principal's options in each period are either to play the sole available arm i for another period, or to "stop" the process and receive a terminal reward of m. Standard arguments (e.g., Whittle (1982), Ross (1983)) establish for each m, the existence of a continuous function $V(.;m):\Delta^{K-1} \to \mathbb{R}$, such that V(p;m) is the value to the principal of this stopping problem when the prior on arm i is p and the terminal reward is m. Indeed, V(.;m) may be obtained as the unique fixed–point of the contraction mapping[10] $T:C(\Delta^{K-1}) \to C(\Delta^{K-1})$, where $C(\Delta^{K-1})$ is the space of all real–valued continuous functions on Δ^{K-1} endowed with the sup–norm topology, and, for $v \in C(\Delta^{K-1})$, Tv is defined by

$$Tv(p) = \max \{m, R(p) + \delta \int v[\beta(p;r)]f(p)(r)dr\}. \qquad (3.1)$$

Hence, V(.;m) satisfies at each p:

$$V(p;m) = \max.\{m, R(p) + \delta \int V[\beta(p;r);m]f(p)(r)dr\}. \qquad (3.2)$$

[10] The equivalence of the original stopping problem which involves unknown paramenters, and the dynamic programming problem for which the contraction is defined, is an intuitive result, but, as a referee pointed out to us, a highly non–trivial one. For a proof of this equivalence, see Rhenius (1974), Rieder (1975), or Schael (1979).

The following lemma collects some additional properties of this optimization problem that are important in characterizing the DAI:

Lemma 3.1 *i)* $V(.;m)$ *is convex in* p *for each* $m \in \mathbb{R}$.

ii) $V(p;.)$ *is convex and non–decreasing in* m *for each* p.

iii) $V(.;.)$ *is jointly continuous in* p *and* m.

The ***Dynamic Allocation Index of arm i*** when the prior on arm i is p, denoted M(p), is then defined as:

$$M(p) \;=\; \inf.\{m \in \mathbb{R} \mid V(p;m) = m\}. \tag{3.3}$$

Observe that if $m \geq R_1/[1-\delta]$, then we must also have $V(p;m) = m$, while evidently for $m < R_K/[1-\delta]$, $V(p;m) > m$. It follows that M(.) takes values in the compact set $[R_K/(1-\delta), R_1/(1-\delta)]$ and is, consequently, well–defined.

For $k = 1,\dots,K$, define e_k to be that element of Δ^{K-1} with 1 in the k–th place and zeros elsewhere. Recall that a real valued function h defined on a convex domain is said to be ***quasi–convex*** if for all $c \in \mathbb{R}$, the set $\{x \mid h(x) \leq c\}$ is convex. The following lemma gathers three properties of the DAI – continuity, quasi–convexity, and strict monotonicity along any ray through the "worst" prior – that play an important role in the sequel.

Lemma 3.2 *i)* $M(.)$ *is a continuous, quasi–convex function of* p.

ii) Let $p \in \Delta^{K-1}$, $p \neq e_K$. *Then,* $M(\lambda p + (1-\lambda)e_K)$ *is a strictly increasing function of* λ *for* $\lambda \in [0,1]$.

4. Existence of an Optimal Strategy

We begin with a statement of the celebrated Theorem of Gittins and Jones (1974) that establishes the existence of an optimal strategy when $\mathfrak{N}$ has a ***finite*** number of elements. Then, we show (Theorem 4.1) that this result extends in a straightforward manner to yield a general existence theorem for denumerable armed Bandits. Two simple examples show that the conditions under which Theorem 4.1 is established are not necessary, but are minimally sufficient.

So, let $M_i(.)$ represent the DAI function for arm i. We now have:[11]

Theorem 4.0 *(Gittins and Jones, 1974): Suppose* $\mathfrak{N}$ *consists of a finite number of elements* $\{1,...N\}$. *Then, the uniquely optimal class of strategies are those which at each time t pick any of the arms i for which*

$$M_i(p^t(i)) \;=\; \max\{M_j(p^t(j)) \mid j \in \mathfrak{N}\}, \tag{4.0}$$

where $P^t = (p^t(1),...,p^t(N))$ *is the vector of priors at time t.*

Two obvious problems arise if this result is to be extended to an infinite number of arms. Namely, (a) the supremum of the DAIs at the initial prior may not be attained, and (b) even if there is a well defined maximum at the initial beliefs, there may exist histories after which an "optimal" continuation does not exist. It turns out, however, that these are also the ***only*** problems that arise, and if they are ruled out an identical result to Theorem 4.0 may be shown to hold for infinite–armed bandits as well.

Some new definitions would help in stating the precise result. For each $j \in \mathfrak{N}$, let $\Sigma(j)$ denote the subset of strategies of Σ that begin with arm j. Let $V^*(P)$ be defined by $V^*(P) = \sup_{\sigma\in\Sigma} W(\sigma;P)$. Note that V^* is well defined for any P, since rewards are uniformly bounded (equation 2.1) and there is strict discounting. Call an arm i an ***optimal initial selection*** at P if it is true that $V^*(P) = \sup_{\sigma\in\Sigma(i)} W(\sigma;P)$. The proof of the

[11] It is worth noting that, within broad limits, the assumption of geometric discounting is also ***necessary*** for Theorem 4.0; see Berry and Fristedt (1985, Ch. 6).

following result may be found in Appendix II.

Theorem 4.1 *i) Arm* i *is an optimal initial selection at* P *if:*

$$M_i(p(i)) = \sup \{M_j(p(j)) \mid j \in \mathfrak{N}\} \tag{4.1}$$

ii) An optimal strategy exists from P *whenever* (4.1) *holds for infinitely many* i.

While part (ii) of Theorem 4.1 presents only sufficient conditions for the existence of an optimal strategy, it is easy to see that these conditions are not necessary, but are (almost) minimally sufficient.[12] Consider the following examples:

Example 4.1: Suppose arm 1 pays a reward of 1 with certainty, while arm n for $n \geq 2$ pays $(1 - 1/n)$ with certainty. Then, there is only a single arm that attains the maximum (arm 1), but it is evident that the uniquely optimal strategy is to pick arm 1 forever.

Example 4.2: Arms n for $n \geq 2$ are as in the previous example, while arm 1 either generates a reward of 2 with certainty or 0 with certainty. Let the prior probability of the first situation be p. It is evident that for p sufficiently close to 1, arm 1 is an optimal initial selection, but it is also clear that after the history in which the first period reward is 0, there is no optimal continuation strategy.

An easy consequence of Theorem 4.1 is the existence of an optimal "no recall" strategy in stationary Bandits:

[12] The precise condition which ***is*** minimally sufficient is the inelegant one that there should exist an m such that $M_i(p(i)) = m$ for infinitely many i, and $m < M_j(p(j))$ for only finitely many j.

Corollary 4.1: *Optimal strategies always exist in stationary Bandit problems. Moreover, the optimal strategy may be chosen to be one in which any arm that has been tried and discarded is never recalled.*

Proof: Since all arms are a priori identical, so they have the same DAI, denoted $M(\pi)$ (recall π is the prior on all arms), and existence follows from Theorem 4.1. The no–recall property is also immediate, since there are, after any history, an infinite number of arms with DAI $M(\pi)$, while an arm is discarded under the optimal strategy when, and only when, its DAI falls below $M(\pi)$. □

5. The Stochastic Process of Survival

We now turn to an examination of the stochastic process governing the repeated use of an arm. Specifically, we are interested in the distribution of the number of periods a generic arm will continue to remain optimal, once it has been chosen. The analysis below does not distinguish between finite– and infinite–armed bandits, since nothing depends on this distinction.

So let i be an arm that is an optimal choice at some vector of beliefs $P = (p(1),p(2),...)$, i.e., which is such that $M_i(p(i)) = \sup_{j\in\mathfrak{N}} M_j(p(j))$.[13] Let $m^* = \sup_{j\neq i} M_j(p(j))$. Under the optimal strategy, the arm i will be retained as long as the prior $p^t(i)$ on it satisfies $M_i(p^t(i)) \geq m^*$. Our aim in this section is to characterize the distribution of time for which this inequality will continue to hold.

For notational ease, we suppress the index i in what follows, and denote the initial prior p(i) on arm i by π. Let arm i be one of K possible types with reward densities $f_1,...,f_K$. We assume, without loss of generality, that at the initial prior we have $\pi_k > 0$ for $k = 1,...,K$, so that no type is redundant.

Recall that e_k denotes that element of Δ^{K-1} that has zeros in all but the k–th

[13] Such an arm will always exist, of course, if $\mathfrak{N}$ has only a finite number of elements.

place. There are two cases possible: $M(e_K) \geq m^*$, and $M(e_K) < m^*$. In the first case, we clearly also have $M(p) \geq m^*$ for all $p \in \Delta^{K-1}$, so that the arm will never be replaced regardless of the rewards it generates. The survival process is, therefore, trivial. In the sequel, we assume, consequently, that the second case holds, namely that $M(e_K) < m^*$. Note that we must have $M(e_1) > m^*$, for otherwise $M(\pi) \geq m^*$ is not possible.

Let $\Delta_R = \{p \in \Delta^{K-1} | M(p) < m^*\}$, and $\Delta_A = \{p \in \Delta^{K-1} | M(p) \geq m^*\}$. The following lemma gathers some properties of these sets, where these are immediate consequences of the continuity and quasi–convexity of M(.) [see lemma 3.2(i)].

Lemma 5.1: *Δ_R is a convex, open subset,while Δ_A is a closed subset, of Δ^{K-1}.*

We introduce some additional notation now, as well as a relatively informal description of the probability measures required to examine the survival process. A formal description may be found in Appendix III to this paper, where the main result of this section (Theorem 5.1) is proved.

Let $\text{supp}.f_k = \{r | f_k(r) > 0\}$ denote the support of f_k, $k = 1,...,K$, and let $\mathfrak{R} = \cup_{k=1}^{K} \text{supp}.f_k$. Define $\mathfrak{R}^t$ to be the t–fold Cartesian product of $\mathfrak{R}$, with generic element $r^t = (r_1,...,r_t)$. For each t, and for each $k \in \{1,...,K\}$, define the density F_k^t on $\mathfrak{R}^t$ by

$$F_k^t(r_1,...,r_t) = \Pi_{\tau=1}^{t} f_k(r_\tau). \tag{5.1}$$

Say that arm i ***survives at least t periods*** under the observed rewards $(r_1,...,r_t) \in \mathfrak{R}^t$ if the resulting sequence of posteriors $\{p^\tau\}_{\tau=1}^{t}$, calculated from the initial belief p(i) using these rewards, satisfies $M(p^\tau) \geq m^*$ for each $\tau = 1,...,t$. Let $\mathfrak{S}^t \subset \mathfrak{R}^t$ denote the set of all possible t–sequences of rewards under which an arm will survive at least t periods. This set is, of course, independent of the arm's true type.

Now, for k =1,...,K, and each positive integer t, let

$$Q_k(t) \; = \; \int_{\mathfrak{S}^t} F_k^t(r^t) dr^t. \tag{5.2}$$

$Q_k(t)$ is simply the probability that arm i will survive at least t periods, given that its true type is k. Let $U_k = \lim_t Q_k(t)$ be the probability that arm i will survive ***forever*** given that its true type is k. Note that U_k is well defined since $Q_k(t)$ is non–increasing in t. Finally, say that arm i survives forever with non–zero probability if $U_k > 0$ for some k = 1,...,K.

Our main result in this section is precisely that arm i must survive forever with non–zero probability.[14] Since both our choice of the initial prior P on the arms of the Bandit, and the choice of i from the set of initially optimal arms at P, were arbitrary, this result establishes that any arm which becomes optimal at some point will, with positive probability remain optimal forever. We emphasize the independence of this result from the choice of discount factor $\delta \in [0,1)$, and the form of the distributions (f_k).

Theorem 5.1: *There is* $k^* \in \{1,...,K\}$ *such that* $U_{k*} > 0$.

We sketch the arguments involved in proving Theorem 5.1 here. Consider the sequence of posterior beliefs $\{p^t\}$ on arm i that arise as observations on i are accumulated. [Under the optimal strategy, no observations on i are, of course, witnessed from i beyond the first t such that $M(p^t) < m^*$, i.e., that $p^t \in \Delta_R$.] Routine arguments establish that this sequence of posteriors must follow a Martingale process with respect to the probability measure P_π generated on the space of sample paths by the prior belief π. Intuitively, the Martingale property of beliefs results from the observation that the principal cannot expect his beliefs about the arm to change from one period to another in any predictable manner,

[14] Our original result was for the case K = 2, and employed a direct proof that, in fact, U_1 was non–zero. The outline of the proof for the case of general K, which we present in Appendix III, was provided by an anonymous referee.

so that today's expectation of tomorrow's belief must be today's belief itself.

Now, note that Δ_R is a convex set, and $\pi \in \text{int.}\Delta^{K-1}$ is a point not in this set, so there exists a linear functional l separating the two. Moreover, l divides Δ^{K-1} into two convex sets Δ_1 and Δ_2 such that $\Delta_1 \subset \Delta_A$, and $\Delta_R \subset \Delta_2$; and there exists a constant c such that $l(p) \geq c$ for all $p \in \Delta_1$, and $l(p) < c$ for all $p \in \Delta_2$. Consider a stronger rejection rule than that specified under M, namely, the one under which an arm is rejected in favor of an untried arm at the first t for which $l(p^t) < c$, i.e., for which $l(p^t) \in \Delta_2 \supset \Delta_R$.

Since l is linear, $l(p^t)$ is itself a Martingale. A fundamental result in the theory of Martingales (see Proposition A.1, Appendix III) states that with non–zero P_π –probability, $l(p^t)$ will stay above c forever, so that, in particular, $l(p^t)$ will stay in Δ_A forever with non–zero P_π–probability. Letting P_k denote the probability measure induced on the space of sample paths by the type–parameter k (i.e., by the belief e_k), it now follows as a simple consequence that with non–zero P_k–probability for some k, $M(p^t)$ will remain in Δ_A forever. The last statement is precisely that $U_k > 0$ for some k.

Now, let Z be the subset of $\{1,\ldots,K\}$ defined by $Z = \{k \mid M(e_k) \geq M(\pi)\}$. It appears a reasonable conjecture that all arms of type $k \in Z$ will survive forever with non–zero probability. Surprisingly, even a weaker version of this conjecture turns out to be false. Namely, the fact that an arm of type k* will survive forever with non–zero probability has no implications, in general, for the "better" types $k \in \{1,\ldots,k^*-1\}$.[15] In the example below, the sketch of which was provided us by a referee, ***a type 2 arm survives forever with probability 1,*** but a type 1 arm is rejected in ***finite time with probability 1.***

[15] Note, however, that arms of type $k \notin Z$ ***must*** fail in finite time with probability 1. This follows since the consistency of Bayes updating implies their true type will be revealed with probability 1 if they are played forever, so that $M(p^t)$ falls below $M(\pi)$ in finite time with probability 1.

Example 5.1: Consider a stationary Bandit in which each arm is one of the same three possible types. The initial belief is $P = \{\pi,\pi,...\}$, where $\pi \in \Delta^2$ will be specified shortly. The reward space is discrete and equals $\{0,1,2\}$. The reward probabilities associated with the types are described in the matrix below (ϵ is any number satisfying $0 < \epsilon < \frac{1}{4}$):

	Pr{r=0}	Pr{r=1}	Pr{r=2}
Type 1	ϵ	$\frac{1}{2}$	$\frac{1}{2} - \epsilon$
Type 2	0	1	0
Type 3	$\frac{1}{2} - \epsilon$	$\frac{1}{2}$	ϵ

Note that $R_1 = \frac{3}{2} - 2\epsilon \; > \; 1 = R_2 \; > \; \frac{1}{2} + 2\epsilon = R_3$. For any $p = (p_1,p_2,p_3) \in \Delta^2$, $p_3 > 0$, the Bayes updating rule for this problem has the important feature that

$$\beta_1(p,1)/\beta_3(p,1) \;=\; p_1/p_3, \tag{5.3}$$

where $\beta_k(p,r)$ refers to the k–th coordinate of $\beta(p,r)$. Fix any $\delta \in [0,1)$, and let M(.) represent the DAI function for this problem, where M(.) is, of course, the same for all arms. Recall that, by Corollary 4.1, any arm whose history has resulted in a prior p satisfying $M(p) < M(\pi)$ will never be recalled again.

We now specify the distribution π. Let the prior probability of a type 2 arm be any $\pi_2 \in (0,1)$. For $\pi^*(\pi_1) = (\pi_1,\pi_2,1-\pi_1-\pi_2)$, it is immediate from the Bayes updating formula that, as $\pi_1 \to 0$, we have $\beta_1(\pi^*(\pi_1),2) \to 0$ and $\beta_3(\pi^*(\pi_1),2) \to 1$. By the continuity of M(.) [lemma 3.2], we have:

$$\begin{aligned}\lim_{\pi_1 \to 0} M(\pi_1,\pi_2,1-\pi_1-\pi_2) \quad &= \; M(0,\pi_2,1-\pi_2) \\ &\geq \; \pi_2 R_2/(1-\delta) + (1-\pi_2)R_3/(1-\delta) \\ &> \; R_3/(1-\delta)\end{aligned}$$

$$= M(e_3) = \lim_{\pi_1 \to 0} M[\beta(\pi^*(\pi_1),2)] \qquad (5.4)$$

It easily follows that for $\pi_1 > 0$, but sufficiently small, we have $M(\pi^*(\pi_1)) > M[\beta(\pi^*(\pi_1),2)]$. Pick any such π_1, and let π be given by $(\pi_1,\pi_2,1-\pi_1-\pi_2)$. We show that, under π, the optimal strategy implies that a type two arm lasts forever with probability 1, while a type 1 arm is rejected in finite time with probability 1.

First, note that if an arm produces a reward of 2 in the very first period of its use (which only arms of types 1 and 3 will), then the arm will be replaced immediately by an untried arm, since, by construction, $M(\pi) > M[\beta(\pi,2)]$.

Second, if the arm produces a reward of 0 in the very first period of its use, it will again be replaced immediately since:

i) M(.) is strictly increasing on the ray joining e_3 and e_1 by lemma 3.2, and

ii) $\beta_1(\pi,0) < \beta_1(\pi,2)$, while $\beta_2(\pi,0) = \beta_2(\pi,2) = 0$.

Combining these, we have that $M[\beta(\pi,0)] < M[\beta(\pi,2)]$.

A type 1 arm will therefore survive the first period if, and only if, it produces a reward of 1. If this happens, however, the relative likelihoods of types 1 and 3 [i.e., the ratio $\beta_1(\pi,1)/\beta_3(\pi,1)$] remains unchanged. Direct calculation now reveals that, as a consequence, a reward of 2 in the second period again leads to period 3 beliefs of $\beta(\pi,2)$, while a reward of 0 leads to period 3 beliefs of $\beta(\pi,0)$. But this means that an arm will survive into the third period if, and only if, the reward in each of the first two periods is 1.

The argument evidently iterates. Survival occurs up to period t if, and only if, the reward in each of the first (t–1) periods is 1. Since the probability of a type 1 arm producing rewards of 1 forever is 0, such an arm must fail in finite time with probability 1.[16] On the other hand, a type two arm produces rewards of 1 forever with probability 1,

[16] Indeed, in this example the expected length of continuous use for a type 1 arm (as also for a type 3 arm) is just 4 periods.

and, hence, survives forever with probability 1.[17] □

Now, define $C_k(t+1;t)$ by $C_k(t+1;t) = Q_k(t+1)/Q_k(t)$, if $Q_k(t) > 0$, and $C_k(t+1;t) = 0$, otherwise. Then, $C_k(t+1;t)$ is the ***conditional*** probability that an arm of type k which has survived t periods will survive (t+1) periods. We have the following Corollary to Theorem 5.1:

Corollary 5.1: *If* $U_k > 0$, *then* $C_k(t+1;t) \to 1$ *as* $t \to \infty$.

Proof: Let $P_k(t) = Q_k(t) - Q_k(t+1)$ be the probability of lasting exactly t periods. We have,

$$\begin{aligned} C_k(t+1;t) &= Q_k(t+1)/Q_k(t) \\ &= [U_k + \Sigma^{\infty}_{\tau=t+1} P_k(\tau)]/[U_k + \Sigma^{\infty}_{\tau=t} P_k(\tau)] \\ &= 1 - [P_k(t)/(U_k + \Sigma^{\infty}_{\tau=t} P_k(\tau))]. \end{aligned} \tag{5.5}$$

Since $P(t)$ is summable, so $\lim_t P(t) = \lim_t \Sigma^{\infty}_{\tau=t} P_k(t) = 0$. Since $U_k > 0$, by hypothesis, the warranted result is proved. □

Somewhat curiously though, this convergence of $C_k(t+1;t)$ to unity need not be monotone in t. We provide an example of this in the next section (see Example 6.2 below).

Finally, to close this section, we note the following strong implication of Theorem 5.1 for stationary Bandit problems:

[17] A referee conjectured that survival of some type should imply survival of all better types if the reward distributions have common support. We have unable to prove this conjecture, or to come up with a counterexample.

Corollary 5.2: *In a stationary Bandit problem, the expected number of arms used in an optimal strategy is finite. In particular, with probability 1 the optimal strategy requires the use of only a finite number of arms.*

Proof: Since all arms are *a priori* identical, the probability that any arm will last forever from the time it is initially chosen is independent of the arm's identity. Let α denote this probability; by Theorem 5.1, $\alpha > 0$. Since an arm is never recalled under the DAI strategy if it has been tried and discarded, the probability that exactly n arms are used is clearly $\alpha(1-\alpha)^{n-1}$. Therefore, the expected number of arms that will be used is $\Sigma_{n=1}^{\infty}[n\alpha(1-\alpha)^{n-1}] = 1/\alpha < \infty$. □

6. Myopia and Optimality

In this section, we examine the trade–off in the optimal strategies of Bandit problems between ***current*** reward maximization and the acquisition of information that could enhance ***future*** decision–making. Specifically, we focus on the relationship between optimality and myopia.

A ***myopic strategy*** σ^m in a Bandit problem, is the strategy that at each time t, given the beliefs P^t at t, picks any of the arms i for which

$$R(p^t(i)) \;=\; \max\,\{R(p^t(j)) \mid j \in \mathfrak{N}\}. \qquad (6.1)$$

It is well–known that when $\delta > 0$, so that the future is not irrelevant, myopic strategies are in general suboptimal, precisely on account of their ignoring the information content of current actions.[18] In contrast, our main result in this section is that for a surprisingly large class of Bandit problems, myopic strategies are, indeed, optimal, ***regardless of the value of δ.*** As with our earlier results, this one also does not depend on the choice of any particular structure on the reward distributions.

[18] Berry and Fristedt (1985) contains several examples.

Consider the class of Bandit problems in which all arms have support on the same class of distributions. Namely, assume that for all $i \in \mathfrak{N}$, we have $K(i) = K$, and $f_k^i = f_k$ for all $k = 1,\ldots,K$. We now have the following generalization of the result in Banks and Sundaram (1990):[19]

Theorem 6.1: *σ^m is an optimal strategy when* $K = 2$.

Proof: For notational ease, we denote the prior on a generic arm by $p \in [0,1]$, rather than $(p_1,p_2) \in \Delta^1$, with the interpretation that p is the probability that the arm is a type 1 arm. Let M(.) denote the DAI function.[20] We show that the recommendations of σ^m always coincide with that of the DAI strategy, establishing the former's optimality. Specifically, we claim that

$$R(p(i)) = \sup_{j\in\mathfrak{N}} R(p(j)) \iff M((p(i)) = \sup_{j\in\mathfrak{N}} M(p(j)). \quad (6.2)$$

Indeed, note that R(.) is a strictly increasing function of p, since $R_1 > R_2$; and M(.) is a strictly increasing function of p by lemma 3.2(ii). Equation (6.2) follows. □

Remark: While σ^m is always well–defined for finite–armed Bandits, it, evidently, need not be well–defined in infinite–armed Bandits. Consequently, Theorem 6.1 may be read as stating that an optimal strategy exists in the latter case if, and only if, the myopic strategy is well–defined,[21] in which case the two coincide.

[19] In Banks and Sundaram (1990) we prove this myopia result for finite–armed Bandits, using similar characteristics of the the DAI; by Theorem 4.1 above, therefore, Theorem 6.1 follows. We include the (short) proof here merely for completeness.

[20] M(.) is, of course, independent of i since all arms have the same distributions in their support.

[21] That is, there is at least one arm that attains the myopic maximum after any history.

When K $\geq$ 3, myopic strategies need no longer be optimal. Consider the following example:

Example 6.1: We consider discrete reward distributions with outcome space $\{0,1\}$. There are three possible types indexed by k. Let $(1-q_k)$ and q_k be the probabilities of rewards of 0 and 1 respectively, from a type k arm. Note that $R_k = q_k$, k = 1,2,3. To complete the specification of the model, let $\pi = (1/3,1/3,1/3)$, and let $(q_1,q_2,q_3) = (3/4,1/2,1/4)$.

We will show the existence of a history that occurs with positive probability, after which a myopic continuation is strictly suboptimal. Let r^t denote the period t observed reward. Consider the (2n–1)–history in which $r^1 = ... = r^n = 1$, and $r^{n+1} = ... = r^{2n-1} = 0$. A simple application of the Bayes updating rule shows that the myopic strategy strictly favors retaining the arm that generates these rewards, at each point of this history, regardless of the value of n. For sufficiently large n, the resulting belief about the arm is of the form $p = (3\epsilon,1-4\epsilon,\epsilon)$ for some "small" $\epsilon > 0$; and $\epsilon \to 0$ as $n \to \infty$. Since $R(p) = (1+\epsilon)/2 > R(\pi)$, the myopic strategy strictly favors retaining the arm for period–2n also.

It is intuitively clear – and a formal proof is not very difficult – that $M(\pi) > R(\pi)/(1-\delta)$. It follows that $M(0,1,0) < M(\pi)$, since, evidently, $M(0,1,0) = R(0,1,0)/(1-\delta)$, while $R(0,1,0) = R(\pi)$. By the continuity of M(.) [see lemma 3.2(i)], therefore, it is also the case that $M(3\epsilon,1-4\epsilon,\epsilon) < M(\pi)$, for sufficiently small ϵ. But this implies the strict suboptimality of continuing with the arm with prior $(3\epsilon,1-4\epsilon,\epsilon)$ although, as noted above, the myopic strategy strictly favors it. □

One implication of this example is worth noting. Authors[22] have sometimes commented that stationary Bandits bear a strong resemblance to two–armed Bandits with one "known" arm,[23] using the argument that (by the no–recall property; see Corollary 4.1)

[22] For example, Mortensen (1985, p.878).

[23] That is, with one arm generating rewards according to a known, fixed distribution.

the optimal strategy in a stationary Bandit involves, at any time, only a choice between the prior p on the arm in use in the previous period, and the known, fixed distribution π, which represents the common prior on all untried arms. In point of fact, this resemblance is purely superficial, and has no deeper significance. For, it is well–known that in two–armed Bandit problems with a known arm, it is optimal to use the unknown arm whenever the expected reward from the unknown arm is at least as large as that from the known arm.[24] Translating, this should imply that it is optimal to retain the arm with prior p whenever $R(p) \geq R(\pi)$, which, as we have just seen above, is not always valid.

The optimality of myopic strategies when each arm is one of the same two possible distributions, enables a simple demonstration of our earlier claim (section 5) that while $C_k(t+1;t)$ may converge to unity, it need not do so monotonically in t. Consider the following example:

Example 6.2: The reward distributions are Bernoulli with q_k [resp. $(1-q_k)$] being the probability of a reward of 1 [resp. of 0] from a type k arm, k =1,2. Let $q_1 = 1 - q_2$. Applying Bayes' rule, an arm survives for at least t periods if, and only if, the reward sequence $r^t = (r_1^t,\dots,r_t^t)$ satisfies

$$s(r^t;\tau) \geq f(r^t;\tau), \quad \tau = 1,\dots,t, \tag{6.3}$$

where $s(r^t;\tau)$ [resp. $f(r^t;\tau)$] denotes the number of 1's [resp. 0's] in the first τ observations of r^t. It is now immediate that if t is odd, then $C_k(t+1;t)$ must be unity for both k, since to survive an odd number of periods, the number 1's generated must be at least one more than the number of 0's generated. On the other hand, for t even, $C_k(t+1;t)$ must be strictly less than unity for either k, since the history in which an equal number of 1's and

[24] Intuitively, this is obvious, for going with the unknown arm involves no current expected loss, and a possible information (hence, future rewards) gain.

0's have been generated through period t occurs with positive probability under either k.

The survival rule for this example has a second interesting implication, namely, that the stochastic process of the continued use of an arm of type k can be viewed as a ***Random Walk*** on the integers beginning at 1, with the probability of a "right" move (+1) equal to q_k, the probability of a "left" move (–1) equal to $(1-q_k)$, and with an absorbing barrier at zero. Standard results in the theory of Random Walks (see, e.g., Feller, 1968) tell us the following about these processes. Since $q_2 < 1/2$, so with probability 1 the random walk under q_2 will get absorbed at zero in finite time. The expected time to absorption (i.e., the expected length of time a type 2 arm will remain in continuous play) is $[1/(1-2q_2)] + 1$. On the other hand, since $q_1 > 1/2$, so the random walk with parameter q_1 will, with probability $[2q_1-1]/q_1 > 0$, never get absorbed. For instance, if $q_1 = 3/4$, and $q_2 = 1/4$, then a type 1 arm will last forever with probability 2/3, while a type 2 arm enjoys an expected length of continuous play of only 3 periods. Finally, if $\pi_1 = \pi_2 = 1/2$, then the expected number of arms used is 3. □

Appendix I

I.1. Proof of Lemma 3.1:

To prove part (i), we adopt the techniques of McLennan (1988). Let m be given. Define the mapping T on the space $C(\Delta^{K-1})$ as in section 3. For notational ease, define, for $w \in C(\Delta^{K-1})$, (i) $Gw(p) = \int w(\beta(p;r))f(p)(r)dr$, and (ii) $Hw(p) = R(p) + \delta Gw(p)$. We proceed in two steps.

Step 1: We show that if w is convex, then Tw is also convex. Let $p, p' \in \Delta^{K-1}$, and let $p^* = (1-\lambda)p + \lambda p'$ for some $\lambda \in (0,1)$. Define, for each r in the support of the densities (f_k), $e(r) \in (0,1)$ by $e(r).f(p^*)(r) = \lambda f(p')(r)$ [or, equivalently, $(1-e(r)).f(p^*)(r) = (1-\lambda)f(p)(r)$]. Note that $(1-e(r))\beta(p,r) + e(r)\beta(p'r) = \beta(p^*,r)$.

Suppose, now, that w is convex. Then,

$$\begin{aligned} Gw(p^*) &= \int w(\beta(p^*;r)f(p^*)(r)dr \\ &= \int w[(1-e(r))\beta(p,r) + e(r)\beta(p',r)]f(p^*)(r)dr \\ &\leq \int [(1-e(r))w(\beta(p,r)) + e(r)w(\beta(p'r))]f(p^*)(r)dr \\ &= \int (1-\lambda)w(\beta(p,r))f(p)(r)dr + \int \lambda w(\beta(p',r))f(p')(r)dr \\ &= (1-\lambda)Gw(p) + \lambda Gw(p'), \end{aligned} \tag{I.1}$$

where the inequality obtains by Jensen's Inequality for convex functions.

Now observe that since R(.) is linear, so Hw is also convex, whenever w is. As the maximum of convex functions in this case, Tw evidently inherits the convexity of w. This completes step 1.

Step 2. Let $\mathfrak{W}$ be the set of all convex w such that $w \leq Tw$. Evidently, $\mathfrak{W}$ is bounded above and non–empty. Let w* be defined by

$$w^* = \sup\{w \mid w \in \mathfrak{W}\}. \tag{I.2}$$

As the supremum of convex functions, w is convex.

Now, observe that T is a monotone operator: $v \leq w$ implies $Tv \leq Tw$. Therefore, for all $w \in \mathfrak{W}$, we have $Tw \in \mathfrak{W}$ as well, by definition of $\mathfrak{W}$. But $Tw^* \in \mathfrak{W}$ implies $w^* \leq$

Tw^*, while, by definition of w^*, $w^* \geq Tw^*$. Therefore, $Tw^* = w^*$, or w^* is a fixed–point of the mapping T. But T is a contraction and has a unique fixed–point. Therefore, it must be that $w^*(.) = V(.;m)$, proving lemma 3.1(ii).

Part (ii) of lemma 3.1, (the convexity and monotonicity of V in m for each fixed p) is established in Berry and Fristedt (1985, lemma 6.1.2), who also prove that V is continuous in m for each fixed p.

Finally, we turn to lemma 3.1(iii). Recall that V is continuous in m for each p, while the construction of V(.;m) as a fixed point of the contraction T, establishes continuity in p for each m. We show that this separate continuity of V, combined with its monotonicity in m, implies joint continuity in p and m.

So let $(p_n,m_n) \to (p,m)$. Define $h_n(.) = V(p_n,.)$, and $h(.) = V(p,.)$. We need to show that $h_n(m_n) \to h(m)$ as $n \to \infty$.

First, note that for each n, h_n is a continuous, non–decreasing function, as is h. Therefore, by Helley's Selection Theorem (Billingsley, 1978, p.290) there is a right–continuous, non–decreasing function h^* such that $h_n(\overline{m}) \to h^*(\overline{m})$ at each continuity point $\overline{m}$ of h^*. Note also that for each $\overline{m}$, $h_n(\overline{m}) \to h(\overline{m})$ by the separate continuity of $V(.;\overline{m})$.

First, we claim that $h^* = h$. To see this, note that since h^* is a monotone function, its values everywhere are completely determined by the dense set of its continuity points. But at any such $\overline{m}$, h^* and h must agree, by definition of these functions, establishing the claim.

Next, we claim that if h^* is continuous from the right [resp. left] at any $\overline{m}$, then for all $\overline{m}_n \to \overline{m}$, we have $\limsup_n h_n(\overline{m}_n) \leq h^*(\overline{m})$ [resp. $\liminf h_n(\overline{m}_n) \geq h^*(\overline{m})$]. This will establish that for all sequences $m_n \to m$, $h_n(m_n) \to h^*(m)$, since by the earlier claim, $h^* = h$, and h is, of course, continuous everywhere.

To see the claim, suppose first that h^* is right–continuous at $\overline{m}$. Pick a sequence

m_k such that $m_k > \overline{m}$ for each k and $m_k \to \overline{m}$, and such that for each k, m_k is a continuity point of h*. Since the continuity points of h* are dense this is possible. Fix any k. Since $\overline{m}_n \to \overline{m}$, so $\overline{m}_n < m_k$ for all k sufficiently large. Since each h_n is non–decreasing, so $h_n(\overline{m}_n) \leq h_n(m_k)$. Since m_k is a continuity point of h*, so $h_n(m_k) \to h^*(m_k)$ as $n \to \infty$. Combining the last two statements, $\limsup_n h_n(\overline{m}_n) \leq h^*(m_k)$. Since this holds for each m_k, and h* is continuous from the right by hypothesis, taking limits as $k \to \infty$ establishes one part of the claim. The other part is established by an analogous argument exploiting the left–continuity of h*. This completes the proof of joint–continuity. □

I.2. Proof of Lemma 3.2:

We begin with two claims:

Claim 1: $M(p) \geq R(p)/(1-\delta)$ for all $p \in \Delta^{K-1}$.

Proof: This is straightforward, and follows easily from the observation that $W(p;m) \geq R(p)/(1-\delta)$ for all p. Note that if $p = e_k$ for any k, then $M(p) = R(p)/(1-\delta)$.□

Next, for $p \in \Delta^{K-1}$, $m \in \mathbb{R}$, let

$$HV(p;m) = R(p) + \delta\int V[\beta(p;r);m]f(p)(r)dr. \qquad \text{(I.3)}$$

Claim 2: $HV(p;m) \lesseqqgtr m$ as $m \gtreqqless M(p)$.

Proof: If $p = e_k$ for some k, this is obvious, so suppose $p \neq e_k$ for any k. Let $m_k = M(e_k) = R_k/(1-\delta)$. Since $V(p;m) > m$ for any $m \leq m_K$, we must have $HV(p;m) > m$ for any $m \leq m_K$. Similarly, since $V(p;m) = m$ for any $m \geq m_1$, we must also have $HV(p;m) \leq m$ for any such m, and, indeed, it is not too difficult to see that we must have strict inequality here. HV(p;.) evidently inherits the properties of continuity and convexity in m from V(p;.). By continuity, it follows that there exists a value of m, say $m^* \in (m_K,m_1)$, such that $HV(p;m^*) = m^*$. Pick any such m*, and consider any m' such that $m' = \lambda m^* + (1-\lambda)m_1$, for $\lambda \in (0,1)$. Note that $m' > m^*$. The convexity of HV(p;.) now implies

$$\begin{aligned} HV(p;m') &\leq \lambda HV(p;m^*) + (1-\lambda)HV(p;m_1) \\ &< \lambda m^* + (1-\lambda)m_1 = m'. \end{aligned} \tag{I.4}$$

But this inequality shows that m^* must be unique; that is, there exists only one value of m^* satisfying $HV(p;m^*) = m^*$. Therefore, for $m < m^*$, we must have $HV(p;m) > m$, while for $m > m^*$, we must have $HV(p;m) < m$. [Otherwise, the Intermediate Value Theorem furnishes a contradiction.] Finally, since $V(p;m^*) = HV(p;m^*) = m^*$, and for $m < m^*$, we have $V(p;m) = HV(p;m) > m$, so it is the case that $m^* = M(p)$, proving claim 2.□

Returning to the proof of the lemma, let $p_n \to p$, and $m_n = M(p_n)$. Since M(.) takes values in a compact set, we may, without loss of generality, assume that $m_n \to m$. By lemma 3.1(i), $V(p_n;m_n) \to V(p;m)$. The joint continuity of V in its arguments evidently implies that HV is also continuous jointly in p and m. Therefore, $HV(p_n;m_n) \to HV(p;m)$. Since $m_n = V(p_n;m_n) = HV(p_n;m_n)$ for all n (the last equality obtaining by claim 2), so $m = V(p;m) = HV(p;m)$. By claim 2, this implies $m = M(p)$, establishing continuity of M(.).

To see quasi–convexity of M in p, let $p,p' \in \Delta^{K-1}$, and let $p_\mu = \mu p + (1-\mu)p'$. Assume, without loss of generality, that $M(p) \geq M(p')$. Then, we are required to show that $M(p_\mu) \leq M(p)$. Since $M(p) \geq M(p')$, we have $V(p';M(p)) = M(p)$ by claim 2, while, of course, $V(p;M(p)) = M(p)$. Since V is convex in p for each m by lemma 3.1(ii), we have

$$V(p_\mu;M(p)) \leq \mu V(p;M(p)) + (1-\mu)V(p';M(p)) = M(p). \tag{I.5}$$

Since it is true that $V(p_\mu;m) \geq m$ for any m, the foregoing implies $V(p_\mu;M(p)) = M(p)$, so by definition of M(.), $M(p_\mu) \leq M(p)$, proving quasi–convexity.

Finally, let $p \neq e_K$, and let $p(\lambda) = \lambda p + (1-\lambda)e_K$ for $\lambda \in (0,1)$. We show that $M(p) > M(p(\lambda)) > M(e_K)$. Observe that the quasi–convexity of M(.) on the "ray" $\{p(\lambda) | p(\lambda) = \lambda p + (1-\lambda)e_K$ for $\lambda \in (0,1)\}$, already implies that M(.) is non–decreasing on the ray, since e_K is a minimum for M(.) on this ray (indeed, on Δ^{K-1}). Combining these statements, (iii) easily follows.

Evidently, $M(e_K) = R_K/(1-\delta) < M(p)$. In proving lemma 1, we showed that the convexity of V(.;m) in p also implies the convexity of HV(.;m) in p. Therefore, we have

$$\begin{aligned} HV(p(\lambda);M(p)) &\leq \lambda HV(p;M(p)) + (1-\lambda)HV(e_K;M(p)) \\ &< \lambda M(p) + (1-\lambda)M(p) \\ &= M(p), \end{aligned} \tag{I.6}$$

since $HV(p;M(p)) = M(p)$ by definition, and $HV(e_K;M(p)) < M(p)$ by claim 2. But this implies, in turn, that $M(p(\lambda)) < M(p)$. Evidently, $M(p(\lambda)) \geq R(p(\lambda))/(1-\delta) > R_K/(1-\delta) = M(e_K)$, so $M(p) > M(p(\lambda)) > M(e_K)$. □

Appendix II: Proof of Theorem 4.1

For ease of exposition, we suppress dependence on the vector of priors P throughout. For each integer n, let Σ_n denote the subset of Σ that consists of strategies that use only one of the first n arms after any possible history. Let $V_n = \sup \{W(\sigma) | \sigma \in \Sigma_n\}$. We show as a first step that $V^* = \lim_n V_n$.

Since Σ_n can obviously be associated with the n–armed bandit problem in which only arms {1,...,n} are available (and the initial prior is the appropriate restriction of P to this set), Theorem 4.0 ensures the existence of $\sigma_n^* \in \Sigma_n$ such that $V_n = W(\sigma_n^*) \geq W(\sigma)\ \forall\ \sigma \in \Sigma_n$. Moreover, σ_n^* must be a DAI strategy as described in Theorem 4.0. It is evident that we must have $V_n \leq V_{n+1} \leq V^*$ for all n, since any strategy feasible in Σ_n is also feasible in Σ_{n+1} and Σ. Therefore, $\lim_n V_n$ is well–defined.

Let $\epsilon > 0$ be given. Pick $\sigma \in \Sigma$ such that $W(\sigma) \geq V^* - \epsilon/2$. By definition of V^*, such a σ may be seen to exist. Also pick $t(\epsilon)$ to be any positive integer that satisfies

$$\delta^{t(\epsilon)}R^*/[1-\delta] \leq \epsilon/4. \tag{II.1}$$

By (2.1) and the fact that $\delta < 1$, such a $t(\epsilon)$ also exists. Let $N(\epsilon)$ be the (finite) set of all possible arms σ may ever use in the first $t(\epsilon)$ periods, and pick n sufficiently large so that $N(\epsilon) \subset \{1,...,n\}$.

Pick any $m \geq n$. Consider the strategy $\sigma_m \in \Sigma_m$ that imitates σ for the first $t(\epsilon)$ periods and then proceeds arbitrarily. By definition of V_m we have

$$V_m \geq W(\sigma_m). \tag{II.2}$$

But by definition of R^*, and choice of $t(\epsilon)$, it is also true that

$$|W(\sigma) - W(\sigma_m)| \leq 2\delta^{t(\epsilon)}R^*/[1-\delta] \leq \epsilon/2. \tag{II.3}$$

So, certainly

$$|V_m - W(\sigma)| \leq \epsilon/2. \tag{II.4}$$

Therefore, we now have

$$\begin{aligned} |V^* - V_m| &\leq |V^* - W(\sigma)| + |W(\sigma) - V_m| \\ &\leq \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned} \tag{II.5}$$

Since $\epsilon > 0$ was arbitrary, we have shown that $V^* = \lim_n V_n$.

Recall that i attains the sup in (4.1). By Theorem 4.0, for any $n \geq i$, there exists a strategy $\sigma_n^* \in \Sigma_n$ that begins with arm i and satisfies $W(\sigma_n^*) = V_n$, i.e., $\sigma_n^* \in \Sigma(i)$ also. Therefore,

$$\sup_{\sigma\in\Sigma(i)} W(\sigma) \geq \lim_n V_n = V^*, \tag{II.6}$$

proving part (i) of the Theorem.

To see part (ii) of the Theorem, note that the presence of an infinite number of arms that attain the supremum in (4.1) at the initial belief implies that after any possible history, there is at least one arm that now attains the supremum. Let $m^* = \sup\{M_i(p(i)) | i \in \mathfrak{N}\}$. Let I^* be the infinite set of i for which $M_i(p(i)) = m^*$, and let $\{i_1, i_2, ...\}$ be any enumeration of the elements of I^*. Consider the strategy $\sigma(\infty)$ which begins with i_1 and switches from i_n to i_{n+1} at the first time (if ever) that the DAI of i_n falls below m^*. We show that $W(\sigma(\infty)) = V^*$, completing the proof of the Theorem.

For any N, let $\mu(N)$ denote the number of elements in the intersection of $\{i_1,i_2,...\}$ with $\{1,...,N\}$. For all N sufficiently large, $\mu(N) > 0$; and, since I^* has an infinite number of elements, $\mu(N) \to \infty$ as $N \to \infty$. Let $\sigma_N(\infty)$ be that strategy in Σ_N that follows $\sigma(\infty)$ as long as feasible (i.e., as long as the recommendations of $\sigma(\infty)$ are within $\{1,...,N\}$), and proceeds arbitrarily within $\{1,...,N\}$ otherwise. Evidently

$$|W(\sigma(\infty)) - W(\sigma_N(\infty))| \leq 2\delta^{\mu(N)}R^*/(1-\delta). \tag{II.7}$$

Since there is a DAI strategy that is optimal in Σ_N and coincides with $\sigma_N(\infty)$ for at least $\mu(N)$ periods it is also the case that

$$|V_N - W(\sigma_N(\infty))| \leq 2\delta^{\mu(N)}R^*/(1-\delta). \tag{II.8}$$

Combining these inequalities,

$$|W(\sigma(\infty)) - V_N| \leq 4\delta^{\mu(N)}R^*/(1-\delta) \to 0 \text{ as } N \to \infty. \qquad \text{(II.9)}$$

Equation (II.9) establishes the result since $V_N \to V^*$ as $N \to \infty$. □

Appendix III: Proof of Theorem 5.1.

This Appendix develops formally the ideas sketched in the text. The proof is in several steps.

Step 0: A Preliminary Result

The following Proposition is an immediate consequence of Proposition IV–3–12 of Neveu (1975).

Proposition A.1: *Let* (X_t) *be a uniformly bounded martingale on a probability space* $(\Omega,\mathfrak{F},P)$ *relative to the sub–sigma fields* $(\mathfrak{F}_t)$, *and let* X^* *denote the almost sure limit*[25] *of the martingale* (X_t). *Let* τ *be a stopping time for the martingale. Define the random variable* X_τ *by* $X_\tau(\omega) = X_{\tau(\omega)}(\omega)$, *if* $\tau(\omega)$ *is finite, and* $X_\tau(\omega) = X^*(\omega)$ *otherwise. Then,* $E[X_\tau] = E[X_1]$.

Step 1:

Recall that p(i) is denoted by π. By lemma 5.1, the set $\Delta_R = \{p \mid M(p) < m^*\}$ is a convex and relatively open subset of Δ^{K-1} [henceforth, just Δ]. Moreover, π, which is an interior point of Δ, is a point not in this set. Hence, the application of a standard separation argument implies the existence of a linear functional l on $\mathbb{R}^K$ and a constant $c \in \mathbb{R}$, such that the hyperplane$\{x \mid l(x) = c\}$ divides Δ into two convex subsets Δ_1 and Δ_2 with $\Delta_R \subset \Delta_1$, $\Delta_2 \subset \Delta_A$, and $l(p) < c$ for all $p \in \Delta_1$, $l(p) \geq c$ for all $p \in \Delta_2$. By these containment relations we have, of course, that $M(p) < M(\pi) \Rightarrow l(p) < l(\pi)$.

Step 2:

Recall that $\mathfrak{R}$ is the union of supp.f_k over k. Define (i) $\mathfrak{R}^t = X_{\tau=1}^{t}\mathfrak{R}$, (ii) $\mathfrak{R}^{-t} = X_{\tau=t+1}^{\infty} \mathfrak{R}$, and (iii) $\mathfrak{R}^* = X_{\tau=1}^{\infty} \mathfrak{R}$. Let $\mathfrak{F}(\mathfrak{R}^t)$ represent the Borel sigma field of $\mathfrak{R}^t$. Define the family $\{\mathfrak{F}^t\}$ of increasing sigma fields on $\mathfrak{R}^*$ by $\mathfrak{F}^t = \{A \subset \mathfrak{R}^* \mid A = C \times \mathfrak{R}^{-t};$

[25] This limit, of course, exists by the Martingale Convergence Theorem.

$C \in \mathfrak{F}(\mathfrak{R}^t)\}$. Let $\mathfrak{F}^* = V_{t=1}^{\infty} \mathfrak{F}^t$.

Next, let $Z = \{1,...,K\}$, and let $\mathfrak{F}(Z)$ denote the power set of Z. Finally, define $\Omega = Z \times \mathfrak{R}^*$, and endow Ω with the sigma–field $\mathfrak{F}(\Omega) = \sigma(\mathfrak{F}(Z) \times \mathfrak{F}^*)$.

For $k \in \{1,...,K\}$ and $A \in \mathfrak{F}^t$, let $P_k(A)$ be the probability under k of observing $(r_1,...,r_t) \in C$, where $A = C \times \mathfrak{R}^{-t}$. P_k is clearly calculable from knowledge of the density $f_k(.)$.

The measurable space of sample paths $\{\Omega, \mathfrak{F}(\Omega)\}$ may now be endowed with with the probability measure P_π which is the extension of $P(D \times A) = \Sigma_{k \in D} \pi_k P_k(A)$, for $D \in \mathfrak{F}(Z)$, $A \in \mathfrak{F}^*$. All almost–sure statements on sample paths $\omega = \{k,r_1,r_2,...\}$ are with respect to P_π.

Step 3:

Letting ϕ denote the empty set, let $\mathfrak{G}^t = \sigma(\mathfrak{F}^t \times \{\phi,Z\})$ for each t, and let $\mathfrak{G}^* = V_{t=1}^{\infty} \mathfrak{G}^t$. Then, the probability p_k^t the principal places on the parameter $k \in Z$ at time t can be written as $p_k^t = E[I_{\{k\} \times \mathfrak{R}^*} \mid \mathfrak{G}^t]$, where I denotes the indicator random variable. Since $I \leq 1$ a.s., it follows by Billingsley (1978, example 35.5, p.410) that p_k^t is a martingale with respect to the sigma–filtration $\mathfrak{G}^t$. An appeal to the Martingale Convergence Theorem (Billingsley, 1978, p.416) now reveals the existence of a random variable p_k^{∞} such that p_k^t converges to p_k^{∞} a.s.

Since Z is a finite set, and the preceding statements hold for each k, it now follows that there is a set F of sample paths with $P_\pi(F) = 1$, such that for each k, p_k^t converges to p_k^{∞} on F.

Finally, since linear functions of martingales are themselves martingales, we have that $\ell(p_1^t,...,p_K^t) := \ell(p^t)$ is also a (uniformly bounded) martingale, which converges a.s. to a limit random variable. Simple arguments show that this limit must be $\ell(p_1^{\infty},...,p_K^{\infty}) := \ell(p^{\infty})$.

Step 4:

Recall the definition of c in step 1. Define the (possibly extended–) integer–valued random variable τ by

$$\tau = \min \{t \mid \ell(p^t) < c\}$$

if this is well–defined, and set $\tau = \infty$, otherwise. It is easy to see that τ is a stopping time, i.e., $\{\tau = t\} \in \hat{\mathfrak{F}}^t$ for all t. Let the random variable $\ell(p_\tau)$ be defined by

$$\begin{aligned} \ell(p_\tau)(\omega) &= \ell(p_{\tau(\omega)}(\omega)), && \text{if } \tau(\omega) \text{ is finite} \\ &= \ell(p^\infty(\omega)), && \text{otherwise.} \end{aligned}$$

Since $\ell(p^t)$ is uniformly bounded a.s., the conditions of Proposition A.1 (step 0) are met. Therefore, $E[\ell(p_\tau)] = E[\ell(p^1)]$, and, of course, $E[\ell(p^1)] = c$, since $\ell(\pi) = c$

But this implies the existence of a set G with $P_\pi(G) > 0$ such that $\tau = \infty$ on G. For, the contrary would imply that τ is finite almost surely, which in turn implies $E[\ell(p_\tau)] < c$, a contradiction.

Step 5:

Finally, observe that by the definition of P_π (see step 2), there must exist a $k^* \in Z$, and $A \subset \mathfrak{R}^*$ such that $P_{k^*}(A) > 0$, for, otherwise, $P_\pi(G) > 0$ is not possible. But this just says that, conditional on its "true" type being k*, an arm will last forever with positive probability, if the rejection rule followed is that specified in step 1, namely, if the arm is replaced by an untried arm at the first t at which its prior p^t satisfies $\ell(p^t) < \ell(\pi) = c$. By construction, however, $M(p) < M(\pi) \Rightarrow \ell(p) < c$, and it now easily follows that under the rejection criterion specified by M(.) also, an arm of type k* will last forever with positive probability; or in the notation of Section 5.2 that $U_{k^*} > 0$. □

References

Austen–Smith, D. and J.S. Banks (1989) Electoral Accountability and Incumbency, in ***Models of Strategic Choice in Politics*** (P. Ordeshook, Ed.), University of Michigan Press, Ann Arbor.

Banks, J.S. and R.K. Sundaram (1990) A Class of Bandit Problems Yielding Myopic Optimal Strategies, ***Journal of Applied Probability***, forthcoming.

Banks, J.S. and R.K. Sundaram (1991) Two Index Theorems for Bandit Problems, Working Paper, University of Rochester.

Barro, R. (1973) The Control of Politicians: An Economic Model, ***Public Choice*** 14, 19–42.

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

Billingsley, P. (1979) ***Probability and Measure,*** Wiley, New York.

Easley, D. and N.M. Kiefer (1988) Controlling a Stochastic Process with Unknown Parameters, ***Econometrica*** 56, 1045–1064.

Feldman, M. (1989) On the Generic Non–convergence of Bayesian Actions and Beliefs, BEBR Working Paper, University of Illinois, Urbana–Champaign.

Feller, W. (1968) ***An Introduction to Probability Theory and its Applications,*** Vol. 1, Wiley, New York.

Ferejohn, J. (1986) Incumbent Performance and Electoral Control, ***Public Choice*** 50, 5–25.

Gittins, J. (1989) ***Allocation Indices for Multi–Armed Bandits,*** Wiley, London.

Gittins, J. and D. Jones (1974) A Dynamic Allocation Index for the Sequential Allocation of Experiments, in ***Progress in Statistics*** (J. Gani, et al, Eds.), North Holland, Amsterdam, pp.241–266.

Jovanovic, B. (1979) Job–Search and the Theory of Turnover, ***Journal of Political Economy*** 87, 972–990.

McLennan, A. (1988) Learning in a Repeated Statistical Decision Framework, Working Paper, University of Minnesota.

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.

Neveu, J. (1975) ***Discrete Parameter Martingales,*** North Holland, Amsterdam.

Pressman, and Y. Sonin (1990), ***Sequential Control with Partial Information***, Academic Press, New York.

Rieder, U. (1975) Bayesian Dynamic Programming, ***Advances in Applied Probability*** 7, 330–348.

Rhenius, (1974) Incomplete Information in Markovian Decision Models, ***Annals of Statistics***.

Rogoff, K. (1990) Equilibrium Political Budget Cycles, ***American Economic Review*** 80, 21–37.

Rothschild, M. (1974) A Two–Armed Bandit Theory of Market–Pricing, ***Journal of Economic Theory*** 9, 185–202.

Ross, S. (1983) ***Introduction to Dynamic Programming,*** Academic Press, New York.

Schael, M. (1979) On Dynamic Programming and Statistical Decision Theory, ***Annals of Statistics*** 7(2), 432–445.

Viscusi, W. (1979) Job–Hazards and Worker Quit Rates: An Analysis of Adaptive Worker Behavior, ***International Economic Review*** 20, 29–58.

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

Wilde, L. (1979) An Information–Theoretic Approach to Job Quits, in ***Studies in the Economics of Search*** (S. Lippman J. McCall, Eds.), North Holland, New York.