Econometrica, Vol. 60, No. 5 (September, 1992), 1071–1096

DENUMERABLE-ARMED BANDITS

BY JEFFREY S. BANKS AND RANGARAJAN K. SUNDARAM[1]

This paper studies the class of denumerable-armed (i.e. finite- or countably infinite-armed) bandit problems with independent arms and geometric discounting over an infinite horizon, in which each arm generates rewards according to one of a finite number of distributions, or "types." The number of types in the support of an arm, as also the types themselves, are allowed to vary across the arms. We derive certain continuity and curvature properties of the dynamic allocation (or Gittins) index of Gittins and Jones (1974), and provide necessary and sufficient conditions under which the Gittins-Jones result identifying all optimal strategies for finite-armed bandits may be extended to infinite-armed bandits. We then establish our central result: at each point in time, the arm selected by an optimal strategy will, with strictly positive probability, remain an optimal selection forever. More specifically, for every such arm, there exists (at least) one type of that arm such that, when conditioned on that type being the arm's "true" type, the arm will survive forever and continuously with nonzero probability. When the reward distributions of an arm satisfy the monotone likelihood ratio property (MLRP), the survival prospects of an arm improve when conditioned on types generating higher expected rewards; however, we show how this need not be the case in the absence of MLRP. Implications of these results are derived for the theories of job search and matching, as well as other applications of the bandit paradigm.

KEYWORDS: Bandits, Gittins index, survival, monotone likelihood ratio property, stationary bandits, job search.

1. INTRODUCTION AND SUMMARY

THIS PAPER STUDIES THE CLASS of denumerable-armed (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 densities, called the *types* of the arm.[2] The expected reward from each type of each arm is assumed to be defined and finite. The arms are assumed independent: trying one arm is uninformative about the types of the other arms. Finally, it is assumed that discounting by the decision-maker is geometric over an infinite horizon. No other restrictions are employed. In particular, the forms of the reward densities and the number of types are allowed to be arbitrary and to

[1] This paper has benefited immeasurably from the detailed comments and suggestions offered by Andy McLennan as a (then anonymous) referee. We are also very grateful to Martin Hellwig and two other anonymous referees for their careful reading of earlier drafts and their suggestions; as also to seminar audiences at Buffalo, Caltech, Carnegie-Mellon, Columbia, George Mason, Harvard, Hoover Institution, McMaster, Michigan, Johns Hopkins, Rochester, Toronto, UCLA, UC-Riverside, Washington-St. Louis, Western Ontario, and Yale, for their input. In particular, we would like to thank Prajit Dutta, Mahmoud El-Gamal, Nicholas Kiefer, David Levine, and Bill Zame. The first author gratefully acknowledges financial support provided by the Sloan Foundation and the NSF. The final draft of this paper was completed when the second author was visiting the California Institute of Technology, and he would like to thank them for their hospitality.

[2] With transparent modifications, all of our results remain valid if, instead of reward densities, we had discrete reward distributions (i.e., those with finite or countable support).

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.[3] 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(i), we establish that strategies using the DAI once again uniquely identify the class of all optimal strategies. This enables us to provide (Theorem 4.1(ii)) necessary and sufficient conditions under which optimal strategies exist.

Section 5 turns to the main focus of this paper, the question of "survival." We examine here the stochastic process governing 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, Theorem 5.1 shows 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 and continuously with nonzero probability as an optimal selection.

A natural conjecture, in view of Theorem 5.1, is that if some type in the support of an arm survives forever with positive probability, all "better" types (i.e., those generating a higher expected reward) should also survive forever with nonzero probability. Somewhat surprisingly, this conjecture is false. 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.

We then turn to an examination of conditions under which the "negative" conclusion of Example 5.1 may be avoided. In Theorem 5.2, we show that if the reward distributions on an arm satisfy the *monotone likelihood ratio property* (MLRP), then the probability of surviving at least t periods (for any positive integer t) is higher for better types in the support of that arm. In particular, the "best" type has the highest probability, and the "worst" type the lowest. As a corollary, it follows that if some type survives forever with positive probability, then all better types also survive forever with at least the same probability.

[3] See, e.g., Gittins and Jones (1974), Berry and Fristedt (1985), or Whittle (1982). Weitzman (1979) proves an analogous result for the "Pandora's box" (or "treasure hunt") problem, where the reward distributions are degenerate (implying the true type of an arm is known with certainty after one play of the arm).

Two 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 densities or restrictions on the value of the discount factor (other than of course the MLRP assumption in Theorem 5.2). 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, we note that 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 depend on this finiteness restriction.

A special case of our framework 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, 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, (i) optimal strategies *always* exist in stationary bandits; (ii) optimal strategies exist which never recall a previously selected and discarded arm; and (iii) the expected number of arms employed in an optimal strategy in a stationary bandit is *finite*, so that with probability 1 only a finite number of arms are ever used.

The framework of stationary bandits has been a popular framework for the analysis of decision making in labor markets. The parameterized "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.[4] 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.[5]

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,

[4] Cf. Mortensen (1985) for an in-depth survey of these and other search models in labor economics.

[5] It is worth noting that many papers in labor economics *impose* the restriction that recall of a tried and discarded arm 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.

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 in contrast to most models of repeated elections (e.g., Barro (1973), Ferejohn (1986), or Austen-Smith and Banks (1989)) which study the voter's decision problem from a "moral hazard" perspective, the stationary bandit framework is a model of adverse selection.

Several other economic (and noneconomic) 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 nondurable experience goods, and models of dating and marriage.

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; and Pressman and Sonin (1990), who focus primarily on bandits with *dependent* arms. In addition, Basu, et al. (1990) provide a comprehensive survey of recent papers in this field. 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 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 parameters 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.

The paper is organized as follows: Section 2 sets up the framework, and gathers notation and definitions. Section 3 introduces and characterizes the DAI. Section 4 is concerned with existence of optimal plans, while Section 5 focuses on the question of "survival." Section 6 concludes with a description of some open questions and unresolved conjectures. Proofs of results that are omitted in the main body of the paper may be found in the Appendices.

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 $k \in \{1, \ldots, K(i)\}$, then it generates rewards according to the density $f_k^i(\cdot)$. 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 that $R_1^i \geq R_2^i \geq \cdots \geq R_{K(i)}^i$. We also assume that

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

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

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[6] $p(i) \in \Delta^{K(i)-1}$ represents the principal's belief regarding the type distribution of arm i, viz., the kth 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* β_i: $\Delta^{K(i)-1} \times \mathbb{R} \to \Delta^{K(i)-1}$. This map is defined by $\beta_i(p^t(i); r) = (\beta_{ik}(p^t(i); r))_{k=1,\ldots K(i)}$, where

$$\beta_{ik}(p^t(i), r) = p_k^t(i) \cdot f_k^i(r) \Bigg/ \left[\sum_{m=1}^{K(i)} p_m^t(i) \cdot f_m^i(r) \right] \tag{2.2}$$

if the quantity on the right-hand side is well defined (i.e., the denominator is nonzero), and is arbitrary otherwise.

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

[6] For any finite integer n, Δ^{n-1} will denote the positive unit simplex in $\mathbb{R}^n$:

$$\Delta^{n-1} = \left\{ x \in \mathbb{R}^n \mid x_i \geq 0, \text{ and } \sum_i x_i = 1 \right\}.$$

period. Formally, σ is a sequence of measurable maps $\{\sigma_t\}_{t=0}^{\infty}$ where $\sigma_0 \in \mathfrak{R}$, and for $t \geq 1$, σ_t: $H_t \to \mathfrak{R}$. 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 tth period reward $r_t(\sigma; P)$ for the principal. Hence, each strategy σ also defines a total expected reward $W(\sigma; P)$ as

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

The principal's objective is to find a strategy σ^* such that $W(\sigma^*; P) \geq W(\sigma; P)$ for all $\sigma \in \Sigma$. 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 $i \in \mathfrak{R}$, we have (i) $K(i) = K$, (ii) $f_k^i = f_k$, $k = 1, \ldots, K$, and (iii) $p(i) = \pi \in \Delta^{K-1}$. 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 (also frequently 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 $(f_1, \ldots, f_K)$, with associated expected rewards $R_1, \ldots, R_K$. Let $p \in \Delta^{K-1}$ denote the prior belief on arm i; $R(p) = \sum_k p_k R_k$ the expected one period reward from playing arm i; and $f(p)(\cdot) = \sum_k p_k f_k(\cdot)$ 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 $V(\cdot; m)$: $\Delta^{K-1} \to \mathbb{R}$, 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, $V(\cdot;m)$ may be obtained as the unique fixed-point of the contraction mapping[7] 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

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

Hence, $V(\cdot;m)$ satisfies at each p:

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

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

LEMMA 3.1: (i) *$V(\cdot;m)$ is convex in p for each m.*
(ii) *$V(p;\cdot)$ is convex and nondecreasing in m for each p.*
(iii) *$V(\cdot;\cdot)$ 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:

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

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

For $k = 1, \ldots, K$, define e_k to be that element of Δ^{K-1} with 1 in the kth 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 $\{x | h(x) \leqslant c\}$ 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) *$M(\cdot)$ 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

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

straightforward manner to yield a general existence theorem for denumerable armed bandits. A subsequent example then shows that if the conditions of Theorem 4.1 are violated, optimal strategies need not exist.

So, let $M_i(\cdot)$ represent the DAI function for arm i. The following result establishes the optimality of *index strategies*, i.e. strategies that at each point select any of the arms with the highest DAI at that point.[8]

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

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

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

REMARK: Since any optimal strategy continues to remain optimal if its recommendations are altered on a set of histories of collective probability zero, the uniqueness claim in Theorem 4.0 should be understood modulo this proviso. Namely, that a strategy is optimal if, and only if, the set of histories on which its recommendation differs from the DAI-maximal arm(s) has probability zero.

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 a DAI-maximal arm 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 expected rewards are uniformly bounded (equation 2.1) and there is geometric 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 following result may be found in Appendix II.

THEOREM 4.1: (i) *Arm i is an optimal initial selection at P if and only if*:

$$(4.1) \qquad M_i(p(i)) = \sup\{M_j(p(j)) | j \in \mathfrak{N}\}.$$

(ii) *The only optimal strategies are those which always select a DAI-maximal arm, except possibly after a set of histories of collective probability zero. In*

[8] It is worth noting that, within broad limits, the assumption of geometric discounting is also *necessary* for Theorem 4.0; see Berry and Fristedt (1986, Ch. 6).

particular, an optimal strategy exists from P if, and only if, either
(a) *there are infinitely many arms i such that* $M_i(p(i)) \geq M^*$, *or*
(b) *there is an arm i such that* $M_i(e_{K(i)}) \geq M^*$,
where $M^* = \sup\{m \mid m \leq M_j(p(j))$ *for infinitely many* $j\}$.

REMARK 1: It is readily seen that the conditions (a) and (b) are *sufficient* for the index strategy to be well-defined, i.e., for there to exist a DAI-maximal arm after any history. A little reflection shows that these conditions are *necessary* as well. For, suppose both conditions were violated. If there are no arms i such that $M_i(p(i)) \geq M^*$, then evidently we are done. So suppose there is a finite set of arms I such that $M_i(p(i)) \geq M^*$ if and only if $i \in I$. Then, it must be the case under an index strategy, that with positive probability the index on all arms $i \in I$ drop strictly below M^* in finite time. (This follows from the assumption that (b) is violated.) By definition of M^*, the continuation index strategy is no longer well defined since for any $\varepsilon > 0$, there are infinitely many arms whose indices are in $(M^* - \varepsilon, M^*)$, but none equal to M^*.

REMARK 2: It is easy to construct examples where the conditions of Theorem 4.1(ii) are not met, and no optimal strategies, therefore, exist. Consider the following:

Example 4.1: Suppose arm 1 either generates a reward of 2 with certainty or 0 with certainty, while arm n for $n \geq 2$ pays $(1 - 1/n)$ 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.

REMARK 3: The assumption that the number of arms is countable (as opposed to a set of arms of arbitrary cardinality) is used only at a single point in the proof of Theorem 4.1. It seems likely that this condition can be dropped and the result generalized, but we do not have a proof. For a more detailed description of the issues involved, see the remark following the proof of Theorem 4.1 in Appendix II.

An immediate consequence of Theorem 4.1 is the existence of an optimal "no recall" strategy for stationary bandit problems:

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, they have the same DAI, denoted $M(\pi)$ (recall π is the prior on all arms), and existence follows from Theorem 4.1. Since, after any history, there are an infinite number of arms with DAI $M(\pi)$, the following strategy is an optimal "no recall" strategy: begin with

arm 1, and move from arm i to arm $i+1$ at the first time when the prior $p(i)$ on arm i satisfies $M(p(i)) < M(\pi)$. *Q.E.D.*

Note, however, that even in a stationary Bandit problem, there may exist optimal strategies that are not "no recall" strategies. For instance, an optimal strategy may choose to drop an arm i, in favor of an untried arm j, whenever the index on i is less than or equal to $M(\pi)$, but to return to it if the index on j drops below $M(\pi)$, and the index on i at the time it was dropped was exactly equal to $M(\pi)$.

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.

Let i be an arm that is an optimal choice at some vector of beliefs $P = (p(1), p(2), \ldots)$, i.e., which is such that $M_i(p(i)) = \sup_{j \in \mathfrak{N}} M_j(p(j))$.[9] 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 $f_1, \ldots, f_K$. We assume, without loss of generality, that at the initial prior we have $\pi_k > 0$ for $k = 1, \ldots, K$, so that no type is redundant.

Recall that e_k denotes that element of Δ^{K-1} that has zeros in all but the kth 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 $M(\cdot)$ [see Lemma 3.2(i)].

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

We introduce some additional notation now, and a relatively informal description of the probability measures required to examine the survival process.

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

A formal description may be found in Appendix III to this paper, where the results of this section are proved.

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

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

Say that arm i *survives at least* t *periods* under the observed rewards $(r_1, \ldots, r_t) \in \mathfrak{R}^t$ if the resulting sequence of posteriors $\{p^\tau\}_{\tau=1}^t$, calculated from the initial belief $p(i)$ using these rewards, satisfies $M(p^\tau) \geqslant m^*$ for each $\tau = 1, \ldots, t$. 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 $k = 1, \ldots, K$, 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 nonincreasing in t. Finally, say that arm i survives forever with nonzero probability if $U_k > 0$ for some $k = 1, \ldots, K$.

Our main result in this section is precisely that arm i must survive forever with nonzero probability.[10] 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* $k^* \in \{1, \ldots, K\}$ *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. Routine arguments (as employed, e.g., by Easley and Kiefer (1988)) 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.

[10] Our original result was for the case $K = 2$, and employed a direct proof that, in fact, U_1 was nonzero. An outline of the proof for the case of general K was suggested to us by Andy McLennan, who also provided the sketch of Example 5.1.

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 nonzero 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 nonzero 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 nonzero 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 $Z = \{k | M(e_k) \geq M(\pi)\}$. It appears a reasonable conjecture that all arms of type $k \in Z$ will survive forever with nonzero 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 nonzero probability has no implications, in general, for the "better" types $k \in \{1, \ldots, k^* - 1\}$.[11] In the example below, *a type 2 arm survives forever with probability* 1, but a type 1 arm is rejected 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 $P = \{\pi, \pi, \ldots\}$, 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 $0 < \varepsilon < (1/4)$):

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

Note that $R_1 = \frac{3}{2} - 2\varepsilon > 1 = R_2 > \frac{1}{2} + 2\varepsilon = R_3$. Fix any $\delta \in [0, 1)$, and let $M(\cdot)$ represent the DAI function for this problem, where $M(\cdot)$ is, of course, the same for all arms. Recall that, by Corollary 4.1, the optimal strategy may be chosen to be a no-recall index strategy.

The reasoning underlying this example comes in two parts. First, 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

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

[11] 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.

i.e., the relative probabilities of types 1 and 3 are unaffected by a reward of 1. Combining (5.3) with the fact that a type 2 arm generates a reward of 1 with probability 1, it is readily seen that the posterior on an arm that has generated a reward sequence of 1's followed by a reward of 2, is simply $\beta(\pi, 2)$ regardless of the length of the sequence of 1's. Similarly, after any sequence of 1's, the first reward of 0 always leads to the posterior $\beta(\pi, 0)$.

Second, we show that, if in the specification of $\pi = (\pi_1, \pi_2, 1 - \pi_1 - \pi_2)$, π_1 is chosen to be sufficiently small, then we can satisfy $M(\beta(\pi, 2)) < M(\pi)$. The intuition behind the second step is that, given the relatively small probability of a type 1, a reward of 2 acts as a "signal" that the arm could be a type 3. This makes continuing to play the arm an unattractive option.

Combining these steps, it easily follows that an arm will be discarded at the first time it generates a reward of 2. To complete the example, we show the intuitive result that $M(\beta(\pi, 0)) < M(\beta(\pi, 2))$, so that an arm will also be discarded the very first time it generates a reward of 0.

We first show that π can be specified to satisfy $M(\beta(\pi, 2)) < M(\pi)$. Let the prior probability of a type 2 arm be any $\pi_2 \in (0, 1)$. Fixing π_2, define for each $\pi_1 \in [0, 1 - \pi_2]$, $\pi^*(\pi_1) = (\pi_1, \pi_2, 1 - \pi_1 - \pi_2)$. 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 $M(\cdot)$ [Lemma 3.2], we have:

$$
\begin{aligned}
(5.4) \qquad \lim_{\pi_1 \to 0} M(\pi_1, \pi_2, 1 - \pi_1 - \pi_2) &= M(0, \pi_2, 1 - \pi_2) \\
&\geqslant \pi_2 R_2/(1 - \delta) + (1 - \pi_2) R_3/(1 - \delta) \\
&> R_3/(1 - \delta) \\
&= M(e_3) = \lim_{\pi_1 \to 0} M\left[\beta(\pi^*(\pi_1), 2)\right].
\end{aligned}
$$

It easily follows that for $\pi_1 > 0$, but sufficiently small, we have $M(\pi^*(\pi_1)) > M[\beta(\pi^*(\pi_1), 2)]$. Picking any such π_1, this step is complete.

Next, note that (a) $M(\cdot)$ is strictly increasing on the ray joining e_3 and e_1 by Lemma 3.2, and (b) $\beta_1(\pi, 0) < \beta_1(\pi, 2)$, while $\beta_2(\pi, 0) = \beta_2(\pi, 2) = 0$. Combining these, we have $M[\beta(\pi, 0)] < M[\beta(\pi, 2)]$.

It now follows easily that 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.[12] On the other hand, a type two arm produces rewards of 1 forever with probability 1, and, hence, survives forever with probability 1.

We now state a sufficient condition under which $U_k > 0$ implies $U_l > 0$ for all $l \in \{1, \ldots, k - 1\}$. Some new definitions are required. We say the densities $(f_1, \ldots, f_K)$ possess the *monotone likelihood ratio property* (MLRP) if for all

[12] 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.

$k, l \in \{1, \ldots, K\}$ such that $k < l$, and for all $a, b \in \operatorname{supp} f_k \cup \operatorname{supp} f_l$ such that $a > b$,

$$f_k(a) \cdot f_l(b) - f_k(b) \cdot f_l(a) \geq 0.$$

In particular, the likelihood ratio $f_k(r)/f_l(r)$ is nondecreasing in r whenever $f_l(r) \neq 0$.[13] We note that two of the most frequently used distributions in decision problems, namely the Bernoulli and the normal (with known variance and unknown mean) satisfy MLRP, as do many others. Say that f_k *stochastically dominates* f_l if for all increasing functions $h: \mathbb{R} \to \mathbb{R}$,

$$\int h(r) f_k(r)\, dr \geq \int h(r) f_l(r)\, dr.$$

Ross (1983) proves that MLRP implies stochastic dominance; hence assuming $(f_1, \ldots, f_K)$ satisfy MLRP guarantees that for all types k, l of an arm such that $k < l$, we have that f_k stochastically dominates f_l.

THEOREM 5.2:[14] *Suppose the reward densities satisfy the MLRP. Then, for any positive integer t, it is the case that*

$$Q_1(t) \geq Q_2(t) \geq \cdots \geq Q_K(t).$$

The proof of Theorem 5.2 (found in Appendix III) essentially consists of two parts. First, we show that MLRP and the resulting stochastic dominance imply that at any prior a "cutoff rule" is optimal; namely, that for each $p \in \Delta^{K-1}$, there exists an $\alpha(p)$ in the closure of $\mathfrak{R}$ (possibly equal to $\inf \mathfrak{R}$ or $\sup \mathfrak{R}$) such that $M[\beta(p, r)] \geq m^*$ if, and only if, $r \geq \alpha(p)$. The second part shows that if a cutoff rule is optimal, then stochastic dominance implies that $Q_k(t)$ must be decreasing in k.

As an immediate consequence of Theorem 5.2, we obtain the following Corollary:[15]

COROLLARY 5.1: *Under the conditions of Theorem* 5.2, $U_1 \geq U_2 \geq \cdots \geq U_K$, *with* $U_1 > 0$.

Hence under MLRP the "best" type of an arm lasts forever with positive probability, and "better" types last forever at least as often as do "worse" types (where recall the "better than/worse than" ordering is determined by the types' expected one-period reward).

[13] Cf. Grossman and Hart (1983) and Milgrom (1981) for economic applications employing the MLRP assumption.

[14] We are grateful to Martin Hellwig for suggesting that we examine monotone likelihood ratios as a condition under which the conclusions of Theorem 5.2 obtain.

[15] In Banks and Sundaram (1991) we provide a further characterization of the survival process, including an examination of the conditional probabilities of a type k arm surviving at least $t+1$ periods given that it has survived t periods. We show that for any k such that $U_k > 0$, these conditional probabilities must converge to unity, but that this convergence could be highly non-monotone even assuming the densities satisfy MLRP.

We next turn our attention to an implication of Theorem 5.1 for stationary bandits:

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 $\sum_{n=1}^{\infty}[n\alpha(1-\alpha)^{n-1}] = 1/\alpha < \infty$. *Q.E.D.*

We conclude this section with an example which illustrates the predictions of Corollary 5.2. To solve for the principal's optimal strategy we employ a result from Banks and Sundaram (1990) establishing the optimality of "myopic" strategies (i.e., strategies that recommend the arm to be played in each period purely as a function of the immediate expected reward from each arm) when all arms of the bandit are one of the same two possible types. Formally, the *myopic strategy* $\sigma(m)$ is the strategy that at each time t, given the beliefs P^t at t, picks any of the arms i for which

$$R(p^t(i)) = \max\{R(p^t(j)) | j \in \Re\}.$$

THEOREM 5.3[16] (Banks and Sundaram (1990)): *Suppose* $K(i) = 2$ *for all* i, *and that* $f_k^i = f_k$ *for all* $i \in \Re$, $k = 1, 2$. *Then,* $\sigma(m)$ *is an optimal strategy whenever it is well-defined.*

Example 5.2: Consider a stationary bandit problem in which each arm is one of the same two possible types. The reward distributions from either type 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 $q_1 = 1 - q_2$. As usual, we assume that type 1 arms are "better", so we have $q_1 > 1/2 > q_2$. Applying Bayes' rule, and invoking Theorem 5.3, an arm survives for at least t periods if, and only if, the reward sequence $r^t = (r_1^t, \ldots, r_t^t)$ satisfies

$$s(r^t; \tau) \geq f(r^t; \tau) \qquad (\tau = 1, \ldots, t),$$

[16] Several points are worth noting about this result. First, Banks and Sundaram (1990) actually prove this result for *finite*-armed bandits; however, their proof shows that the recommendations of the myopic strategy and the DAI strategy always coincide under the given conditions, so that whenever the conditions of Theorem 4.1(ii) above are satisfied myopic strategies are well-defined and optimal in denumerable-armed bandits as well. Second, Theorem 5.3 is valid regardless of the value of $\delta \in [0, 1)$, in particular even for δ arbitrarily close to 1. Third, the assumption of only two possible types is crucial; in Banks and Sundaram (1991) we provide a three-type example (with, in fact, the reward densities satisfying MLRP) and show that myopic strategies are strictly suboptimal.

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.

The survival rule for this example has the interesting implication 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 $q_2 < 1/2$, 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 $q_1 > 1/2$, 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.

6. CONCLUSION

The results of this paper have shown that the optimal strategies in independent armed bandit problems with geometric discounting have strong properties. Many open questions remain.

Within the context of this paper, there is the issue of the generalization of our existence result from one retaining validity when the number of arms is countable, to one holding even when the number of arms has arbitrary cardinality. This question is not, perhaps, of technical interest alone, since, as a referee remarked, one feels that the "right" proof of the Gittins-Jones result ought not to depend on the cardinality of the set of arms. The remark at the end of Appendix II offers some intuition why, in the latter case, it is possible that attention may be restricted to a countable set of arms without loss of generality. If this is true, than Theorem 4.1 may be invoked showing that an optimal index strategy continues to exist under suitable conditions.

Secondly, there are many possible generalizations of the bandit framework which would make them more widely applicable in economic settings. Perhaps the most compelling is to introduce a cost of switching between arms. It is certainly difficult to imagine a relevant economic decision problem in which the decision-maker may costlessly switch between alternatives. It is therefore of interest to inquire as to whether a suitably defined index strategy would continue to be optimal; and the extent to which the results of this paper would continue to remain valid.

One very restrictive aspect of bandit problems, especially from the point of view of applications, is that only one arm may be used in a given period. This precludes situations in which the decision maker may learn about several options simultaneously. There are (at least) two possible ways to resolve this issue. The first is to allow the decision maker to sample up to $m \geqslant 1$ arms in

each period, where m is a priori fixed.[17] An alternative framework, which may be viewed as a model of research and development under budgetary constraints, is the following. The decision maker is modeled as distributing a single unit of available "effort" over the arms (research projects). The more effort that is put into an arm, the more reliable the reading from that arm regarding the true payoff prospects of that arm (say, the variance of observations is inversely proportional to the effort put into the arm). Secondly, if no effort is put into an arm, no reward is witnessed. When an additional constraint is appended that all the effort has to be put into a single arm, we obtain the framework of this paper.

Dept. of Economics, University of Rochester, Rochester, NY 14627, U.S.A.

Manuscript received August, 1990; final revision received March, 1992.

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 $e(r)\cdot f(p^*)(r) = \lambda f(p')(r)$ [or, equivalently, $(1-e(r))\cdot f(p^*)(r) = (1-\lambda)f(p)(r)$]. The significance of this choice of $e(r)$ is that $(1-e(r))\beta(p,r) + e(r)\beta(p'r) = \beta(p^*,r)$ for any r, since for each $k \in \{1,\dots,K\}$, we have

$$\begin{aligned}\beta_k(p^*,r) &= p_k^* f_k(r)/f(p^*)(r)\\ &= [(1-\lambda)p_k + \lambda p_k']f_k(r)/f(p^*)(r)\\ &= \{(1-\lambda)f(p)(r)/f(p^*)(r)\}\cdot\beta_k(p,r)\\ &\quad + \{\lambda f(p')(r)/f(p^*)(r)\}\cdot\beta_k(p',r)\\ &= (1-e(r))\beta_k(p,r) + e(r)\beta_k(p',r).\end{aligned}$$

Suppose, now, that w is convex. Then,

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

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

[17] Of course, this framework may be viewed as a bandit problem with $\binom{n}{m}$ dependent arms, but there does not appear to be much gain conceptually from doing so.

Now observe that since $R(\cdot)$ is linear, Hw is also convex whenever w is. As the maximum of convex functions, Tw then inherits the convexity of w. This completes *Step 1*.

Step 2: Let $\mathfrak{W}$ be the set of all convex w such that $w \leqslant Tw$. Evidently, $\mathfrak{W}$ is bounded above and nonempty. Let w^* be defined by

$$w^*(p) = \sup\{w(p) | w \in \mathfrak{W}\}. \tag{I.2}$$

As the pointwise supremum of convex functions, w^* is convex. We will first show that $w^* \in \mathfrak{W}$, so that $w^* \leqslant Tw^*$.

First, note that T is a monotone operator: $v \leqslant w$ implies $Tv \leqslant Tw$. Therefore, $w^*(p) = \sup\{w(p) | w \in \mathfrak{W}\} \leqslant \sup\{Tw(p) | w \in \mathfrak{W}\}$, by definition of $\mathfrak{W}$. In addition, by the monotonicity of T, we also have $\sup\{Tw(p) | w \in \mathfrak{W}\} \leqslant Tw^*(p)$. Combining these observations, we see that $w^* \leqslant Tw^*$, and so $w^* \in \mathfrak{W}$.

Now, for all $w \in \mathfrak{W}$, we have $w \leqslant Tw$, so by the monotonicity of T, we also have $Tw \leqslant T(Tw)$. This means that $Tw \in \mathfrak{W}$ whenever $w \in \mathfrak{W}$. In particular, $Tw^* \in \mathfrak{W}$. Therefore, by the definition of w^*, we must have $w^* \geqslant Tw^*$. Summing up, we have $Tw^* = w^*$, 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 the case that $w^*(\cdot) = V(\cdot; m)$, proving Lemma 3.1(i).

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 $V(\cdot; m)$ 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 $h_n(\cdot) = V(p_n, \cdot)$, and $h(\cdot) = V(p, \cdot)$. 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, nondecreasing function, as is h. Therefore, by Helly's Selection Theorem (Billingsley (1979, p. 290)) there is a subsequence of h_n (again denoted by h_n), and a right-continuous, nondecreasing function h^* such that $h_n(\bar{m}) \to h^*(\bar{m})$ at each continuity point $\bar{m}$ of h^*. Note also that for each $\bar{m}$, $h_n(\bar{m}) \to h(\bar{m})$ by the separate continuity of $V(\cdot; \bar{m})$.

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 $\bar{m}$, 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 $\bar{m}$, then for all $\bar{m}_n \to \bar{m}$, we have $\limsup_n h_n(\bar{m}_n) \leqslant h^*(\bar{m})$ [resp. $\liminf h_n(\bar{m}_n) \geqslant h^*(\bar{m})$]. 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 $\bar{m}$. Pick a sequence m_k such that $m_k > \bar{m}$ for each k and $m_k \to \bar{m}$, 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 $\bar{m}_n \to \bar{m}$, $\bar{m}_n < m_k$ for all k sufficiently large. Since each h_n is nondecreasing, $h_n(\bar{m}_n) \leqslant h_n(m_k)$. Since m_k is a continuity point of h^*, $h_n(m_k) \to h^*(m_k)$ as $n \to \infty$. Combining the last two statements, $\limsup_n h_n(\bar{m}_n) \leqslant h^*(m_k)$. 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. *Q.E.D.*

I.2. Proof of Lemma 3.2

We begin with two claims:

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

PROOF: For any m, and for all p, it is the case that $V(p; m) \geqslant R(p)/(1-\delta)$, since the right-hand side is merely the expected payoff from the strategy that never chooses the terminal reward after any history. Since $V(p; M(p)) = M(p)$, Claim 1 follows. Note that if $p = e_k$ for any k, then $M(p) = R(p)/(1-\delta)$. *Q.E.D.*

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

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

CLAIM 2: $HV(p;m) \lesseqgtr m$ as $m \lesseqgtr M(p)$.

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 $m \leqslant m_K$, we must have $HV(p;m) > m$ for any $m \leqslant m_K$. Similarly, since $V(p;m) = m$ for any $m \geqslant m_1$, we must also have $HV(p;m) \leqslant m$ for any such m, and, indeed, it is not too difficult to see that we must have strict inequality here. $HV(p;\cdot)$ evidently inherits the properties of continuity and convexity in m from $V(p;\cdot)$. 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 $HV(p;\cdot)$ now implies

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

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. *Q.E.D.*

Returning to the proof of the lemma, let $p_n \to p$, and $m_n = M(p_n)$. Since $M(\cdot)$ 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), $m = V(p;m) = HV(p;m)$. By Claim 2, this implies $m = M(p)$, establishing continuity of $M(\cdot)$.

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 $M(p) \geqslant M(p')$. Then, we are required to show that $M(p_\mu) \leqslant M(p)$. Since $M(p) \geqslant M(p')$, 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

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

Since it is true that $V(p_\mu;m) \geqslant m$ for any m, the foregoing implies $V(p_\mu;M(p)) = M(p)$, so by definition of $M(\cdot)$, $M(p_\mu) \leqslant M(p)$, proving quasi-convexity. This completes the proof of part (i) of Lemma 3.2.

Now, 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 $M(\cdot)$ on the "ray" $\{p(\lambda) | p(\lambda) = \lambda p + (1-\lambda)e_K$ for $\lambda \in (0,1)\}$, already implies that $M(\cdot)$ is nondecreasing on the ray, since e_K is a minimum for $M(\cdot)$ on this ray (indeed, on Δ^{K-1}). Combining these statements, part (ii) of Lemma 3.2 easily follows.

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

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

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, $M(p(\lambda)) \geqslant R(p(\lambda))/(1-\delta) > R_K/(1-\delta) = M(e_K)$, so $M(p) > M(p(\lambda)) > M(e_K)$. *Q.E.D.*

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 $\{1,\ldots,n\}$ 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 $\varepsilon > 0$ be given. Pick $\sigma \in \Sigma$ such that $W(\sigma) \geq V^* - \varepsilon/2$. By definition of V^*, such a σ may be seen to exist. Also pick $t(\varepsilon)$ to be any positive integer that satisfies

$$\text{(II.1)} \qquad \delta^{t(\varepsilon)} R^*/[1-\delta] \leq \varepsilon/8,$$

where R^* is defined by (2.1). Since R^* is finite and $\delta < 1$, such a $t(\varepsilon)$ exists. Let $\eta > 0$ be such that $\eta \cdot R^*/[1-\delta] \leq \varepsilon/8$. Finally, pick any n large enough to ensure that the probability that σ uses an arm not in $\{1,\ldots,n\}$ in the first $t(\varepsilon)$ periods is smaller than η. (Note that this step depends on the assumption that the number of arms is countable.)

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

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

Since the probability that σ_m will be able to imitate σ in each of the first $t(\varepsilon)$ periods is at least $1-\eta$, and the maximum penalty in any period from not being able to do so is $2R^*$ (which occurs when the arm suggested by σ gives a reward of R^*, while all available options to σ_m yield $-R^*$), it is also true that

$$\begin{aligned}\text{(II.3)} \qquad W(\sigma) - W(\sigma_m) &\leq 2\eta R^*/[1-\delta] + (1-\eta)2\delta^{t(\varepsilon)} R^*/[1-\delta] \\ &\leq 2\eta R^*/[1-\delta] + 2\delta^{t(\varepsilon)} R^*/[1-\delta] \\ &\leq \varepsilon/4 + \varepsilon/4 = \varepsilon/2.\end{aligned}$$

So, certainly, from the definitions of $t(\varepsilon)$ and η,

$$\text{(II.4)} \qquad W(\sigma) - V_m \leq \varepsilon/2.$$

Therefore, we now have

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

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

Now, suppose 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 $W(\sigma_n^*) = V_n$. Therefore,

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

proving the "if" part of Theorem 4.1(i).

To check the "only if" part, suppose there were j such that $M_j(p(j)) > M_i(p(i))$. Let $\sigma_i^n \in \Sigma(i) \cap \Sigma_n$ be that strategy that begins with arm i, and then proceeds optimally within $\{1,\ldots,n\}$, and let $L_i V_n$ represent the corresponding payoff $W(\sigma_i^n)$. It is immediate from Theorem 4.0 that $V_n - L_i V_n > 0$, for all $n \geq \max\{i, j\}$. In fact, it is straightforward to verify from Whittle's (1982, p. 216, eq. 6) explicit expression for $V_n - L_i V_n$ the intuitive result that $V_n - L_i V_n$ remains bounded away from zero as $n \to \infty$. It is also clear that $L_i V_n \to \sup_{\sigma \in \Sigma(i)} W(\sigma)$, establishing Theorem 4.1(i).

To see part (ii) of the Theorem, suppose first that condition (a) is satisfied. Let $I = \{i \in \mathfrak{N} \mid M_i(p(i)) \geq M^*\}$. Define the subsets $\{I_d\}$ of I iteratively as follows. $I_1 = \{i \in I \mid M_i(p(i)) \geq M_j(p(j))$ for all $j \in I\}$, and for integers $d \geq 2$, $I_d = \{i \in I \mid M_i(p(i)) \geq M_j(p(j))$ for all $j \in I - I_1 - \ldots - I_{d-1}\}$. From the definition of M^* and the maintained hypothesis that condition (a) is satisfied, it is easily seen that I_d is nonempty for each d. Therefore there is an enumeration $\{i(1), i(2), \ldots\}$ of the elements of I such that $M_{i(n)}[p(i(n))] \geq M_{i(n+1)}[p(i(n+1))]$ for all n. Let $\sigma(\infty)$ be the index strategy, i.e., the strategy which begins with $i(1)$ and chooses at any point, one of the arms with the highest index at that point. By the above arguments, $\sigma(\infty)$ is well-defined. We show that $W(\sigma(\infty)) = V^*$.

Pick any positive integer n, and choose an integer $\mu(n)$ sufficiently large so that $\{i(1), \ldots, i(n)\} \subset \{1, \ldots, \mu(n)\}$. Let $\sigma_{\mu(n)}(\infty)$ be that strategy in $\Sigma_{\mu(n)}$ that follows $\sigma(\infty)$ as long as feasible (i.e., as long as the recommendations of $\sigma(\infty)$ are within $\{1, \ldots, \mu(n)\}$), and proceeds arbitrarily within $\{1, \ldots, \mu(n)\}$ otherwise. Since $\sigma(\infty)$ and $\sigma_{\mu(n)}(\infty)$ coincide for at least n periods, we have

$$\text{(II.7)} \qquad \left|W(\sigma(\infty)) - W\left(\sigma_{\mu(n)}(\infty)\right)\right| \leqslant 2\delta^n R^*/(1-\delta).$$

Since there is a DAI strategy that is optimal in $\Sigma_{\mu(n)}$ and coincides with $\sigma_{\mu(n)}(\infty)$ for at least n periods, it is also the case that

$$\text{(II.8)} \qquad \left|V_{\mu(n)} - W\left(\sigma_{\mu(n)}(\infty)\right)\right| \leqslant 2\delta^n R^*/(1-\delta).$$

Combining these inequalities,

$$\text{(II.9)} \qquad \left|W(\sigma(\infty)) - V_{\mu(n)}\right| \leqslant 4\delta^n R^*/(1-\delta) \to 0 \quad \text{as} \quad n \to \infty.$$

Taking limits as $n \to \infty$, and noting that $V_{\mu(n)} \to V^*$ as $n \to \infty$ (since $\mu(n) \to \infty$ as $n \to \infty$), the desired result is established.

So now suppose condition (a) does not hold, but condition (b) does. Let $I^* = \{i \in \mathfrak{N} \mid M_i[p(i)] \geqslant M^*\}$. Since there is at least one arm i for which $M_i[e_{K(i)}] \geqslant M^*$, I^* is nonempty. Moreover, I^* is at most finite since condition (a) does not hold. By Theorem 4.0, there is an optimal index strategy when the set of arms is restricted to I^*. Denote this strategy by σ^*, and let V_{I^*} be the associated value. Pick any N such that $I^* \subset \{1, \ldots, N\}$. Invoking Theorem 4.0 once again, there is an optimal index strategy when the set of arms is restricted to $\{1, \ldots, N\}$. By definition of M^*, and since I^* contains at least one i such that $M_i[e_{K(i)}] \geqslant M^*$, this strategy will never use an arm outside I^*. Therefore, $V_N = V_{I^*}$. Since $V_N \to V^*$ as $N \to \infty$, we are done.

Finally, suppose both conditions (a) and (b) were violated. Then, it must be the case (see Remark 1 after the statement of Theorem 4.1) that there exists a set of histories that occur with positive probability, after which the index strategy is not well-defined, i.e., after any of these histories, there is no arm whose DAI attains the supremum of the indices. It follows from part (i) of the Theorem that no optimal continuation strategy now exists. Since the set of histories admitting no optimal continuation has positive probability, no optimal strategy can exist in the original problem either, completing the proof of the Theorem. *Q.E.D.*

REMARK: The assumption that there are only a countable number of arms was used to prove that $V^* = \lim_n V_n$, specifically to ensure that if n was chosen large enough, then the probability that σ would use an arm not in $\{1, \ldots, n\}$ in the first $t(\varepsilon)$ periods could be made less than η. Note that if an index strategy was optimal even if the set of arms was of arbitrary cardinality, it would immediately imply that attention could be restricted to a countable subset of the arms (those with the highest indices) without loss of generality. On the other hand, the following informal argument seems to suggest that even in bandit problems where the number of arms is of arbitrary cardinality, one may restrict oneself, without loss of generality, to a countable subset. Suppose for simplicity, that an optimal strategy does exist, and say it begins with arm i. Since the arms are independent, when arm i is in use no information is being accumulated about arms $j \neq i$. Therefore, if it is optimal to discard i in favor of j after some history h, and i in favor of k after some other history h', then it "must" be optimal to switch to j after the history h' also. Similarly, in considering a switch away from j (but not to i) it "should" suffice to have only one arm to which all switches occur. And so on. Of course, this argument is, in one sense, merely restating the intuition behind why an index strategy is likely to be optimal in this case also, but it appears to point to a generalization of Theorem 4.1. Note that the conditions of the Theorem would still remain sufficient conditions if these arguments are valid.

APPENDIX III

III.1: Proof of Theorem 5.1

We develop 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*[18] *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 $p \in \Delta_1$, $l(p) \geqslant c$ 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 $\mathfrak{F}^t = \{A \subset \mathfrak{R}^* \mid A = C \times \mathfrak{R}^{-t};\ C \in \mathfrak{F}(\mathfrak{R}^t)\}$. Let $\mathfrak{F}^* = V_{t=1}^\infty \mathfrak{F}^t$.

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

The measurable space of sample paths $\{\Omega, \mathfrak{F}(\Omega)\}$ may now be endowed 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 $\omega = \{k, r_1, r_2, \ldots\}$ 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 placed by the principal 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 $I \leqslant 1$ a.s., it follows by Billingsley (1979, 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 (1979, 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 $l(p_1^t, \ldots, p_K^t) \coloneqq l(p^t)$ is also a (uniformly bounded) martingale, which converges a.s. to a limit random variable. Simple arguments show that this limit must be $l(p_1^\infty, \ldots, p_K^\infty) \coloneqq l(p^\infty)$.

Step 4: Recall the definition of c in Step 1. Define the (possibly extended-) integer-valued random variable τ by

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

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

$$l(p_\tau)(\omega) = l(p_{\tau(\omega)}(\omega)), \qquad \text{if } \tau(\omega) \text{ is finite}$$

$$= l(p^\infty(\omega)), \qquad \text{otherwise.}$$

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

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 $E[l(p_\tau)] < c$, 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 $l(p^t) < l(\pi) = c$. By construction, however, $M(p) < M(\pi) \Rightarrow l(p) < c$, and it now easily follows that under the rejection criterion specified by $M(\cdot)$ also, an arm of type k^* will last forever with positive probability; or in the notation of Section 5.2 that $U_{k^*} > 0$. *Q.E.D.*

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

III.2: Proof of Theorem 5.2

We begin with a definition. For $p, p' \in \Delta^{K-1}$, say that *p prior-dominates p'* if, for all $l \in \{1, \ldots, K\}$, it is the case that $\sum_{k=1}^{l} p_k \geq \sum_{k=1}^{l} p'_k$.

CLAIM 1: *For $a, b \in \Re$, $\beta(p, a)$ prior-dominates $\beta(p, b)$ whenever $a > b$.*

PROOF: Since $\beta_k(p, r) = p_k f_k(r) / [\sum_{j=1}^{K} p_j f_j(r)]$, we have to show that

$$\sum_{k=1}^{l} p_k f_k(a) \Big/ \sum_{j=1}^{K} p_j f_j(a) \geq \sum_{k=1}^{l} p_k f_k(b) \Big/ \sum_{j=1}^{K} p_j f_j(b),$$

for all $l \in \{1, \ldots, K\}$. Cross-multiplying, and canceling common terms, this is the same as

$$\left[\sum_{k=1}^{l} p_k f_k(a)\right]\left[\sum_{j=l+1}^{K} p_j f_j(b)\right] \geq \left[\sum_{k=1}^{l} p_k f_k(b)\right]\left[\sum_{j=l+1}^{K} p_j f_j(a)\right]$$

or,

$$\sum_{k=1}^{l} \sum_{j=l+1}^{K} \left[p_k p_j f_k(a) f_j(b)\right] \geq \sum_{k=1}^{l} \sum_{j=l+1}^{K} \left[p_k p_j f_k(b) f_j(a)\right].$$

But, for each $k \in \{1, \ldots, l\}$ and $j \in \{l+1, \ldots, K\}$, $f_k(a) f_j(b) \geq f_k(b) f_j(a)$ by the MLRP. This establishes the claim. *Q.E.D.*

CLAIM 2: *$f(p)$ stochastically dominates $f(p')$ if p prior-dominates p'.*

PROOF: This is a straightforward consequence of the densities $(f_1, \ldots, f_K)$ being ordered according to stochastic dominance. *Q.E.D.*

LEMMA A.1: *For any $p \in \Delta^{K-1}$, and $a, b \in \Re$, $M[\beta(p, a)] \geq M[\beta(p, b)]$ if $a > b$.*

PROOF: Intuitively, this follows from the facts that (i) $a > b$ implies $\beta(p, a)$ prior-dominates $\beta(p, b)$, and, therefore, (ii) $f(\beta(p, a))$ stochastically dominates $f(\beta(p, b))$. For expositional continuity, a formal proof is postponed to the end of this section.

REMARK: Lemma A.1 is the critical step in the first part of the proof of Theorem 5.2. A slightly different proof of this lemma may be obtained by establishing that the conditions of Ross (1983, Ch. VII, Prop. 5.4) are satisfied. Since this Lemma is important to the overall proof (and is also, perhaps, of independent interest), we include the proof here so as to keep the exposition self-contained.

CLAIM 3: *For any $p \in \Delta^{K-1}$, there is $\alpha(p) \in Cl(\Re)$ such that $M[\beta(p, r)] \geq m^*$ iff $r \geq \alpha(p)$.*

PROOF: Immediate consequence of Lemma A.1. *Q.E.D.*

CLAIM 4: *For any $p \in \Delta^{K-1}$, $\alpha[\beta(p, a)] \leq \alpha[\beta(p, b)]$ if $a > b$.*

PROOF: Follows from Lemma A.1, and the definition of $\alpha(\cdot)$. *Q.E.D.*

Now, fix $p \in \Delta^{K-1}$. We introduce some new notation. First, for each t, and $r^t = (r_1, \ldots, r_t) \in \Re^t$, define $B^t(r^t)$ inductively as follows. Let $B^1(r^1) = \beta(p, r_1)$, and for $t \geq 1$, let $B^t(r^t) = \beta[B^{t-1}(r^{t-1}), r_t]$. Further, let the functions k^t be inductively defined as follows: $k^0 = \alpha(p)$, and for $t \geq 1$, $k^t(r^t) = \alpha[B^t(r^t)]$. Then, (i) $B^t(r^t)$ is simply the posterior belief on the arm when the initial belief was p, and the rewards $r^t = (r_1, \ldots, r_t)$ were observed, and, (ii) $k^t(r^t)$ is the cut-off level in

period $t+1$ for retaining the arm, given the initial prior p, and the rewards r^t through period t. Iterating on the arguments used to prove Lemma A.1, it is simple to show the following claim.

CLAIM 5: *If* $r^t(1), r^t(2) \in \mathfrak{R}^t$, *and* $r^t(1) \geq r^t(2)$, *then* $k^t(r^t(1)) \leq k^t(r^t(2))$.

Now fix t. We will show that $Q_k(t) \geq Q_l(t)$ whenever $k < l$. So fix such a k and l. Recall the definitions of F_k^t and $\mathfrak{S}^t$ from Section 5. Using the notation introduced above, we can write $\mathfrak{S}^t$ as $\{r^t \in \mathfrak{R}^t | r_1 \geq k^0$, and $r_\tau \geq k^{\tau-1}(r_1, \ldots, r_{\tau-1})$ for $\tau = 2, \ldots, t\}$. Note that

$$Q_k(t) = \int_{\alpha(p)}^{\infty} \int_{k^1(r^1)}^{\infty} \int_{k^2(r^2)}^{\infty} \cdots \int_{k^{t-1}(r^{t-1})}^{\infty} F_k^t(r^t)\, dr^t.$$

Some further notation will greatly simplify the arguments to follow. Define:

$$h^{t-1}(r^{t-1}) = \int_{k^{t-1}(r^{t-1})}^{\infty} f_k(r)\, dr$$

and, inductively, for $\tau = 2, \ldots, t-1$,

$$h^{t-\tau}(r^{t-\tau}) = \int_{k^{t-\tau}(r^{t-\tau})}^{\infty} \left[h^{t-\tau+1}(r^{t-\tau}, r) f_k(r)\right] dr.$$

Now define the functions $g^1, \ldots, g^{t-1}$ exactly as above, but with f_k replaced by f_l. In this notation, then,

$$Q_k(t) = \int_{\alpha(p)}^{\infty} h^1(r) f_k(r)\, dr,$$

$$Q_l(t) = \int_{\alpha(p)}^{\infty} g^1(r) f_l(r)\, dr.$$

Finally, note that by Claim 5, it is the case that for each τ, $h^{t-\tau}(r^{t-\tau-1}, r)$ and $g^{t-\tau}(r^{t-\tau-1}, r)$ are both nondecreasing in r. Summing up, we have

$$\begin{aligned} h^{t-1}(r^{t-1}) &= \int_{k^{t-1}(r^{t-1})}^{\infty} f_k(r)\, dr \\ &\geq \int_{k^{t-1}(r^{t-1})}^{\infty} f_l(r)\, dr, \\ &= g^{t-1}(r^{t-1}), \end{aligned}$$

where the inequality follows from stochastic dominance. So,

$$\begin{aligned} h^{t-2}(r^{t-2}) &= \int_{k^{t-2}(r^{t-2})}^{\infty} \left[h^{t-1}(r^{t-2}, r) f_k(r)\right] dr \\ &\geq \int_{k^{t-2}(r^{t-2})}^{\infty} \left[g^{t-1}(r^{t-2}, r) f_k(r)\right] dr \\ &\geq \int_{k^{t-2}(r^{t-2})}^{\infty} \left[g^{t-1}(r^{t-2}, r) f_l(r)\right] dr \\ &= g^{t-2}(r^{t-2}), \end{aligned}$$

where the first inequality obtains since $h^{t-1} \geq g^{t-1}$, and the second since $g^{t-1}(r^{t-2}, \cdot)$ is nondecreasing and f_k stochastically dominates f_l. The argument evidently iterates, and we obtain $h^1(\cdot) \geq g^1(\cdot)$, and finally, therefore, $Q_k(t) \geq Q_l(t)$. *Q.E.D.*

Finally, we turn to the proof of Lemma A.1.

Proof of Lemma A.1

We have to show that for any $p \in \Delta^{K-1}$, and any $a, b \in \Re$ such that $a > b$, we have $M(\beta(p,a)) \geq M(\beta(p,b))$. This is the same as showing that, for any terminal reward of $m \in \mathbb{R}$, we have $V[\beta(p,a);m] \geq V[\beta(p,b);m]$. So let m be given. We will use an induction proof constructed as follows. First, we will consider the stopping problem when the horizon is truncated to T periods, $T = 0, 1, 2, \ldots$. ($T = 0$ corresponds to the case when there is no play at all.) Equivalently, this may be considered as altering the discount sequence to $\{1, \delta, \delta^2, \ldots, \delta^{T-1}, 0, 0, \ldots\}$. Letting $V_T(\cdot;m)$ denote the value of this problem, we will show that for all T, for any $p \in \Delta^{K-1}$, and for all $a, b \in \Re$ with $a > b$, we must have $V_T[\beta(p,a);m] \geq V_T[\beta(p,b);m]$. An appeal to Theorem 2.5.1 of Berry and Fristedt (1985, p. 40) then shows that $V_T(\cdot;m) \to V(\cdot;m)$ as $T \to \infty$. Thus, $V[\beta(p,a);m] \geq V[\beta(p,b);m]$, whence the lemma follows.

First, note that $V_T[\beta(p,a);m] \geq V_T[\beta(p,b);m]$ is true for $T = 0$, for any $p \in \Delta^{K-1}$, and $a > b$, since both expressions are 0. Suppose now that it holds for $\tau = 0, 1, \ldots, T$, i.e., for all such τ, and for any $p \in \Delta^{K-1}$, and any $a, b \in \Re$ such that $a > b$, we have $V_\tau[\beta(p,a);m] \geq V_\tau[\beta(p,b);m]$. Pick any $p \in \Delta^{K-1}$, and any $a, b \in \Re$ with $a > b$. Let $\pi = \beta(p,a)$ and $\pi' = \beta(p,b)$. We will show that $V_{T+1}(\pi;m) \geq V_{T+1}(\pi';m)$. If the optimal choice in this stopping problem at the prior π' picks the terminal reward m, then the inequality evidently holds, so suppose the optimal strategy picks the arm with prior π'. The value of this strategy is

$$V_{T+1}(\pi';m) = R(\pi') + \delta \int V_T[\beta(\pi',r);m] f(\pi')(r)\, dr,$$

while it is also true that

$$V_{T+1}(\pi;m) \geq R(\pi) + \delta \int V_T[\beta(\pi,r);m] f(\pi)(r)\, dr.$$

Suppressing dependence on m, it is therefore true that

$$\begin{aligned} V_{T+1}(\pi) - V_{T+1}(\pi') &\geq \{R(\pi) - R(\pi')\} \\ &\quad + \delta \int \{V_T[\beta(\pi,r)] f(\pi)(r) - V_T[\beta(\pi',r)] f(\pi')(r)\}\, dr \\ &= \{R(\pi) - R(\pi')\} + \delta \int V_T[\beta(\pi,r)] \{f(\pi)(r) - f(\pi')(r)\}\, dr \\ &\quad + \delta \int \{V_T[\beta(\pi,r)] - V_T[\beta(\pi' r)]\} f(\pi')(r)\, dr. \end{aligned}$$

We will now argue that each of the three terms on the right-hand side is nonnegative, completing the induction step. The first term is obviously nonnegative since π prior-dominates π', and we have stochastic dominance in the reward distributions. Now, $V_T[\beta(\pi,r)]$ is increasing in r by the induction hypothesis. Moreover, since π prior-dominates π', $f(\pi)$ stochastically dominates $f(\pi')$ by Claim 2. So, the second term is also nonnegative. Pick any $r \in \Re$, and note that $\beta(\pi,r) = \beta(\beta(p,a),r) = \beta(\beta(p,r),a)$, while $\beta(\pi',r) = \beta(\beta(p,b),r) = \beta(\beta(p,r),b)$. Thus, letting $\rho = \beta(p,r)$, we have $V_T[\beta(\pi,r)] = V_T[\beta(\rho,a)]$, and $V_T[\beta(\pi' r)] = V_T[\beta(\rho,b)]$. Since $a > b$, the induction hypothesis implies that

$$V_T[\beta(\pi,r)] = V_T[\beta(\rho,a)] \geq V_T[\beta(p,b)] = V_T[\beta(\pi' r)].$$

This implies, of course, that the third term is also nonnegative, establishing $V_{T+1}(\pi;m) \geq V_{T+1}(\pi';m)$, which in turn proves the lemma. *Q.E.D.*

REFERENCES

AUSTEN-SMITH, D., AND J. S. BANKS (1989): "Electoral Accountability and Incumbency," in *Models of Strategic Choice in Politics* (P. Ordeshook, Ed.). Ann Arbor: University of Michigan Press, 121–148.

BANKS, J. S., AND R. K. SUNDARAM (1990): "A Class of Bandit Problems Yielding Myopic Optimal Strategies," *Journal of Applied Probability*, forthcoming.

——— (1991): "Denumerable-Armed Bandits," RCER Working Paper #277, University of Rochester.

BARRO, R. (1973): "The Control of Politicians: An Economic Model," *Public Choice*, 14, 19–42.
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.
BERRY, D., AND B. FRISTEDT (1985): *Bandit Problems: Sequential Allocation of Experiments*. London: Chapman and Hall.
BILLINGSLEY, P. (1979): *Probability and Measure*. New York: Wiley.
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. New York: Wiley.
FEREJOHN, J. (1986): "Incumbent Performance and Electoral Control," *Public Choice*, 50, 5–25.
GITTINS, J. (1989): *Allocation Indices for Multi-Armed Bandits*. London: Wiley.
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.). Amsterdam: North Holland, pp. 241–266.
GROSSMAN, S., AND O. HART (1983): "An Analysis of the Principal-Agent Problem," *Econometrica*, 51, 7–46.
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.
MILGROM, P. (1981): "Good News and Bad News: Representation Theorems and Applications," *Bell Journal of Economics*, 12, 380–391.
MORTENSEN, D. (1985): "Job-Search and Labor Market Analysis," in *Handbook of Labor Economics*, Vol. II (O. Ashenfelter and R. Layard, Eds.). New York: North Holland, 849–919.
NEVEU, J. (1975): *Discrete Parameter Martingales*. Amsterdam: North Holland.
PRESSMAN, E. L., AND I. M. SONIN (1990): *Sequential Control with Partial Information*. New York: Academic Press.
RHENIUS, D. (1974): "Incomplete Information in Markovian Decision Models," *Annals of Statistics*, 2, 1327–1334.
RIEDER, U. (1975): "Bayesian Dynamic Programming," *Advances in Applied Probability*, 7, 330–348.
ROTHSCHILD, M. (1974): "A Two-Armed Bandit Theory of Market-Pricing," *Journal of Economic Theory*, 9, 185–202.
ROSS, S. (1983): *Introduction to Stochastic Dynamic Programming*. New York: Academic Press.
SCHÄL, 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.
WEITZMAN, M. (1979): "Optimal Search for the Best Alternative," *Econometrica*, 47, 641–654.
WHITTLE, P. (1982): *Optimization Over Time: Dynamic Programming and Stochastic Control*, Vol. I. New York: Wiley.
WILDE, L. (1979): "An Information-Theoretic Approach to Job Quits," in *Studies in the Economics of Search* (S. Lippman and J. McCall, Eds.). New York: North Holland, 35–52.