Incumbents, Challengers, and Bandits: Bayesian Learning in a Dynamic Choice Model

Banks, Jeffrey S. and Rangarajan K. Sundaram

Working Paper No. 235
July 1990

Incumbents, Challengers and Bandits: Bayesian Learning in a Dynamic Choice Model

Jeffrey Banks and Rangarajan Sundaram

Rochester Center for Economic Research
Working Paper No. 235

INCUMBENTS, CHALLENGERS, AND BANDITS: BAYESIAN LEARNING IN A DYNAMIC CHOICE MODEL*

Jeffrey S. Banks and Rangarajan K. Sundaram

Department of Economics
University of Rochester
Rochester, NY 14627

Working Paper No. 235

July 1990

*We thank Nicholas Kiefer for valuable comments on an earlier draft of this paper, and the National Science Foundation and Sloan Foundation for financial support of the first author.

1. Introduction and summary

In this paper we provide a general framework for the analysis of a class of dynamic choice models, and characterize the resulting optimal plans. We consider the problem faced by an individual, labelled the ***principal***, who must, in each period of a multi–period horizon, select an ***agent***, from a countable infinity of available candidates, for the performance of some task.[1] Rewards accrue to the principal as a consequence of his choice, and the principal's objective is to maximize the discounted sum of rewards over the horizon of the model; but the factor determining the distribution of rewards in any given period – the chosen agent's "type" – is unknown to the principal. Agents' types are independent and identically distributed, where the parameters determining this distribution may or may not be known to the principal.[2] Moreover, since information on an agent's type can be accumulated only through observations of rewards generated when the agent is employed, all untried agents appear ***ex ante*** identical to the principal. Finally, since our purpose in this paper is to examine the effect of uncertainty and information acquisition on the principal's optimal actions, we suppress the role of the agents by assuming there is only a single "action" available to the employed agent; thus we ignore the problem of the principal having to provide incentives for the agents.

---

[1]Strictly speaking, all we require is that there be more agents than the length of the principal's horizon, so that at least one untried agent is available in each period; however the horizon is assumed infinite for most of the paper.

[2]Note that if the principal is unaware of the "true" distribution of agent types in the population, a simultaneous learning issue emerges: rewards now provide information on not only the incumbent's true type, but also on the distribution of types in the population.

An immediate application of this model, and indeed the scenario which provides the motivation for this paper, is as a description of repeated elections, where the agents are potential representatives of some constituency, and the principal is the (median) voter. Most previous research on repeated elections, notably Barro (1973) Ferejohn (1986), and Austen–Smith and Banks (1989), study repeated elections from a "moral hazard" perspective: the voters are attempting to control the actions of their current representative through their choice of re–election rule.[3] In the current model, on the other hand, the problem faced by the voters is one of "adverse selection", in that the voters' problem is simply to learn which of the potential representatives is "best" at performing the given task of, eg., generating government–controlled benefits for the constituency. This learning aspect of the current model dovetails with the notion of "retrospective voting" due to Fiorina (1981), where voters attempt to infer which candidate will provide higher benefits in the next term as a function of the realized outcome from the previous term. The current model places retrospective voting in an optimal learning framework, where now the voters can use ***all*** previous realized outcomes to project which candidate will provide the highest (discounted) sum of benefits over the voters' horizon. Economic applications of our model include the decision problem faced by an employer where the current employee, and all potential employees, are of unknown quality; and as a variant of the standard job search model where the worker is uncertain about the reward distribution associated with each job (cf. Mortensen (1986) for a survey). In the latter application, for example, the worker receives information about the current job at regular intervals, and at each point in time has the choice of either remaining

---

[3]Other models of repeated elections include Alesina (1988) and Alesina and Spear (1988) on credible policy pronouncements, Ledyard (1988) on the transmission of information between candidates regarding voter preferences, and McKelvey and Riezman (1990) on the observance of seniority in legislatures.

with the current employer, selecting a new, untried employer, or returning to a previous employer.

Our model is closely linked to the extensive work on optimal Bayesian learning in economic environments, eg. Rothschild (1974), Easley and Kiefer (1988), Feldman (1989), McLennan (1988).[4] However, in contrast to our paper, where we aim to characterize the effect of uncertainty and learning on optimal actions, the optimal learning literature has focused, almost exclusively, on examining whether individuals in these models learn the "truth", ie. whether the stochastic process of beliefs converges to point–mass at the true value of the parameter governing rewards. In our model this unknown parameter corresponds to the infinite sequence describing the true type of each agent in the population; it is clear that, except in degenerate situations, asymptotic learning in the above sense is an impossibility in our model. On the other hand, from the point of view of our applications, the effect of the learning process on optimal behavior is an important issue, and it is this we attempt to address.

Section 2 below provides a formal description of the basic model, in which the distribution of agent–types in the population is presumed known to the principal. We show that one can without loss of generality restrict attention to "no recall" strategies by the principal, where such strategies select either the agent employed in the previous period (labelled the ***incumbent***), or else an untried agent (labelled a ***challenger***), but never a previously employed and discarded agent. This follows from our first result, which states that if there exists an optimal strategy in the space of "no recall" strategies, then such a strategy is optimal in the unrestricted space of strategies as well. This observation enables us to reformulate the optimization

---

[4]Our model is also related, to a lesser extent, to the literature on micro–rational expectations equilibrium; cf. Blume, Bray, and Easley (1982) for a survey.

exercise facing the principal as a dynamic programming problem, and in Section 3 an appeal to standard arguments now reveals the existence of an optimal "no recall" strategy for the principal. Moreover, this policy is ***stationary***, ie. such that the optimal decision on whether to retain the incumbent is a function only of the principal's belief on the incumbent's type.

In Sections 4 and 5 we turn to a characterization of the optimal policy. Our basic model can be viewed as an infinite–armed bandit problem, by associating each agent with an arm of the bandit;[5] indeed, such an analogy is used in proving the "no recall" property described above. Nonetheless, we show in Theorem 3 that a sharp characterization of the optimal stationary policy can be obtained in a manner very similar to the construction of the "Gittins index" for ***finite***–armed bandit problems (Gittins and Jones (1974)). Specifically, we show that by solving a family of two–armed bandit problems, where the principal has to decide at each point between accepting a fixed reward, thereby terminating the model, and retaining the incumbent for one more period, at the end of which the same option is again available, an "indifference reward level" can be associated with each possible prior belief on an agent's type. These indifference levels have the property that they completely determine the optimal choices at any point in time in the basic model: it is optimal to retain the incumbent in favor of a challenger if and only if the indifference reward level associated with the former is greater than that associated with the latter. Since two–armed bandit problems with one arm generating a known fixed payoff form the simplest class of bandit problems, this result also shows how the seemingly complex decision problem faced by the principal can be transformed into a family of relatively well–understood and elementary ones.[6]

---

[5]cf. Berry and Fristedt (1985) for a comprehensive study of finite–armed bandit problems.

[6]Actually, an even stronger analogy with two–armed bandit problems suggests itself.

A natural question of interest in the basic model concerns the extent to which the principal's optimal policy incorporates the consideration of ***future*** rewards into the ***current*** decision whether to retain the incumbent. Such trade–offs between current rewards and information acquisition that might better future decision–making ability are indeed an inherent part of all dynamic learning models (cf. Easley and Kiefer (1988)), and in Section 5 we examine this issue in the context of our model. As a benchmark against which to compare the optimal policy, we consider the set of ***myopic*** rules, ie. rules that completely ignore future rewards, and recommend retaining the incumbent if and only if the current expected reward associated with the incumbent is larger than that from the challenger. Our results show that a surprising divergence occurs when going from the case of exactly two possible agent types to three or more possible types. In the former case, Theorem 4 demonstrates that – regardless of the specification of any of the remaining parameters – optimal rules are ***always*** myopic. Thus, optimal considerations of information acquisition and enhancement of future rewards are achieved in this case by merely identifying the agent generating the largest one–period expected reward. A subsequent example then shows that optimal rules need not be myopic, and vice versa, when there are three or more agent types: in particular, it may be optimal to go with the incumbent (resp. challenger) even if myopic considerations favor the challenger (resp.

In the basic model, the "no recall" property implies that in each period the principal is choosing between the incumbent and a randomly chosen challenger. Since the type distribution of the latter is fixed and unchanging over time, this choice essentially amounts to that between the incumbent's prior and this known, fixed (over time) prior. Superficially, at least, this resembles a two–armed bandit problem with one arm generating payoffs according to a known fixed distribution. The results in Section 5 show that this analogy is ***purely*** superficial, and has no deeper significance; see footnote 8 below.

incumbent).[7] Therefore, depending on the number of possible agent types, myopic considerations may mean nothing or everything in the basic model.[8]

In Section 6 we turn to a generalization of the basic model, where we assume that not only the "true" type of any agent, but also the ***distribution*** of agent types in the population, is unknown to the principal. This complicates matters considerably: observed rewards now generate information not only on the actual type of the incumbent generating that reward, but also (since all agents, including the current incumbent, are drawn randomly from the population) on the true distribution of agent types. This simultaneous learning implies, of course, that the principal's prior beliefs on an agent's type and on the distribution of types in the population are not independent; and that the principal's beliefs regarding the type of any untried agent change over time as rewards are accumulated. To render the analysis tractable, we assume recall is no longer an option to the principal; thus in each period the only challengers are untried agents (as we explain below, in contrast to

---

[7]Indeed, the example also suggests that imposing any special structure on the distribution of rewards from each agent type cannot resolve this conclusion. In particular, the distributions in our example are actually ordered according to (first–order) stochastic dominance, surely the most natural condition under which one might hope for myopic optimal rules.

[8]These results highlight the differences between the basic model and two–armed bandit problems (with one known arm), since it is well–known that in the latter class of problems, myopic considerations always exert a one–way influence on optimal decisions. Namely, in such problems it is optimal to play the unknown arm whenever the current expected reward from it is larger, or not much smaller, than from the known arm; however the converse fails to hold. The reason for this is as follows: clearly, once the principal begins playing the known arm, he will continue to do so, since no new information is revealed to alter his decision. Therefore if the unknown arm gives a higher expected reward, the principal is better off today from playing this arm relative to playing the known arm, at that time. On the other hand, playing the unknown arm today generates information valuable to tomorrow's decision; thus the principal may prefer to play the unknown arm even when its expected reward is lower, since such a loss may be offset by a gain in future payoffs.

the basic model this assumption is now far from innocuous). It is quite straightforward to show that the results from Section 3 on the existence of an optimal stationary policy for the principal carry over ***in toto*** to this more general framework, with the one difference that the optimal policy bases its recommendation on the joint distribution of the principal's belief about the incumbent's type and the type–distribution of challengers. However, little else carries over from the basic model; in particular, an example shows that even in the simplest case where there are only two possible agent types and the true distribution of types in the population can take on only two values, optimal policies and myopic rules need not have anything in common. As in the three–type example in Section 5, it may be optimal to replace the incumbent even if the current expected reward from him is higher than that from the challenger (based on the principal's current perception of the type distribution in the population), and to retain him even if it is lower. Further, the example also reveals that the no–recall assumption may, unlike in the basic model, be binding: the principal may wish to recall an agent who was discarded earlier in favor of a challenger, since observations accumulated from that time might have changed the principal's perception of the population's type–distribution, thus lowering the "security level" associated with replacing the incumbent with a challenger below its earlier level.

Finally, in Section 7 we suggest a number of possible extensions and generalizations of our framework.

2. The basic model

2.1 Notation, Definitions, and Assumptions

In the interests of providing a single framework, applicable equally to the

various political and economic scenarios listed in the previous section, we couch the description of our model in the language of principl–agent theory, and maintain this terminology throughout the paper. An individual, whom we refer to as the ***principal***, has a "task" to be performed in each time period $t = 0, 1, 2 \ldots, T$; unless otherwise specified, we set $T = \infty$. In each period t, the principal must select a single ***agent*** for the performance of this task; we let $N = \{1, \ldots, i, \ldots\}$ denote the (infinite) set of available agents. In period t, the selected agent yields a ***reward*** $r^t \in \mathbb{R}$ to the principal. The distribution governing the realization of this reward is solely a function of the chosen agent's ***type***. Each agent may be one of a finite number of possible types $\{\omega_1, \ldots, \omega_K\} \equiv \Omega$, where $K \geq 2$. We denote by $f_k(\cdot) \equiv f(\cdot \mid \omega_k)$ the probability density of rewards accruing to the principal in any period if the chosen agent that period is of type ω_k. (Note that the likelihood of receiving particular levels of rewards is independent of the time period and the identity of the employed agent.) We assume that $f_j(\cdot)$ and $f_k(\cdot)$ have common support $[\underline{r}, \bar{r}]$, $j, k = 1, \ldots, K$. We also assume, without loss of generality, that Ω is ordered according to ***expected*** reward for each type, i.e., that $R_1 > R_2 > \ldots > R_K$, where

$$R_k = \int r \, f_k(r)dr, \; k = 1, \ldots K. \tag{1}$$

It is probably useful to provide at this point, a translation of the terms defined above into the appropriate ones for each application, since we do not touch on any specific application again in this paper. In the repeated–elections interpretation of this model, the principal is the (median) voter, the agents are potential representatives of the voter's constituency, and the rewards represent government–controlled benefits for the constituency. Thus, ***type*** in this context could refer to the agent's ability to steer these benefits in his constituents' favor. Similarly, in the job–search application, the terms principal, agents, and rewards refer

respectively to the searcher, jobs, and utility indices arising from various jobs.[9] Lastly, the agents could also be interpreted as various potential employees available to an employer, the principal, with the obvious interpretation that higher "types" represent better employer–employee matches.

Agents' types are independent and identically distributed according to $\pi \in P(\Omega)$, where for any set Δ, $P(\Delta)$ denotes the set of probability distributions over Δ. Thus $\pi = (\pi_1,...,\pi_K)$, where π_k is the ex ante probability an agent is type ω_k; we assume $\pi_k > 0 \ \forall \ k$. For now we also assume π is known with certainty by the principal; however we drop this assumption in Section 6.

Let $H^t = \{h^t = \{r^\tau\}_{\tau=1}^t\}$ denote the set of histories of length t of realizations, and define $H^0 = \phi$. A ***strategy*** for the principal is a sequence of functions $\sigma = (\sigma^1,...,\sigma^t,...)$ where for all t, σ^t: $H^{t-1} \longrightarrow N$. Thus $\sigma^t(h^{t-1}) \in N$ is the agent selected by the principal in period t upon observing the t–1 realizations of rewards described in h^{t-1}; let Σ denote the set of strategies. In selecting a strategy the principal wishes to maximize his total (discounted) sum of expected rewards over the T–period horizon, so for all $\sigma \in \Sigma$, let $W(\sigma) = E_\sigma[\sum_t \delta^{t-1} r^t]$ denote the ***worth*** of strategy σ, where $\delta \in (0,1)$ is the principal's discount factor.

Finally, we describe how the principal's belief about the agents' types evolve over time. At the beginning of period t, let $p^t(i) = (p_1^t(i),...,p_K^t(i))$ denote the principal's belief about agent i's type, and let $p^t = (p^t(1),...,p^t(i),...)$ denote his belief about all agents. Thus $p^0 = (\pi,...,\pi,...)$, and since agent types are independent, if agent i is not employed in period t then $p^{t+1}(i) = p^t(i)$. On the other hand, if agent i is employed in period t, then upon observing a reward of r^t

---

[9]This interpretation is, in a sense, more general than the simple job–search model (cf. Mortensen, 1986), fo rit admits the possibility that wages may not be the only distinguishing factor among jobs, and, moreover, permits recall.

the principal updates his beliefs about i in a Bayesian manner:

$$p^{t+1}(i)(r^t;p^t) = \beta(p(i),r) \equiv \left(\frac{p_k(i)f_k(r)}{\sum_j p_j(i)f_j(r)} \right)_{k=1,\ldots,K}, \qquad (2)$$

so that $\beta_k(p(i),r)$ is the principal's posterior probability agent i is type ω_k given the realized reward r and prior belief p(i). More specifically, we can write $p^{t+1}(.)$ as $p^{t+1}(h^t;\sigma)$, since the belief at the beginning of period t+1 is a function of the history of realizations, as well as the principal's selections up to period t (embodied in the strategy σ).

2.2 The "no recall" property

In general solving for an optimal strategy for the principal, ie. a strategy maximizing W(.), could be quite complicated. However, in this subsection we show that the principal can without loss of generality restrict attention to a simple class of strategies, namely those that never recall an agent who has been previously employed and then discarded. For all histories $h^t = \{r^\tau\}_{\tau=1}^t$, define $h_s^t = \{r^\tau\}_{\tau=1}^s$, $0 \leq s \leq t$; thus h_s^t is simply the first s realizations in the history h^t. We say that the strategy σ ***employs no recall at period*** t if $\forall\ h^{t-1} \in H^{t-1}$, $\sigma^t(h^{t-1}) = i$ and $\sigma^s(h_s^{t-1}) = i$, for some s, implies $\sigma^{s'}(h_{s'}^{t-1}) = i$ for all $s \leq s' \leq t-1$. Thus, σ employs no recall at t if $\sigma^t(h^{t-1})$ is either the previously employed agent, who has never been discarded, or else is an untried agent. We say σ is a ***no recall strategy*** if σ employs no recall at all t = 1,2,...; let Σ_{nr} denote the set of no recall strategies, and define σ_{nr} as a strategy such that $W(\sigma) = \sup_{\Sigma_{nr}} W(\sigma')$. For now we assume such a strategy exists; in the next section we prove its existence. We now show that σ_{nr} is optimal in the space of unrestricted strategies.

Suppose that only $n < \infty$ agents are available to the principal; since agents are identical ex ante we can think of this as restricting the principal's strategy space to $\Sigma_n = \{\sigma : \forall\ t,\ h^{t-1} \in H^{t-1},\ \sigma^t(h^{t-1}) \leq n\}$. Then the principal's decision problem is equivalent to an n–armed bandit problem (cf. Berry and Fristedt (1985)), with an "arm" associated with each available agent. Gittins and Jones (1974) show that an optimal strategy for such a problem can be characterized by the solution to a family of optimal stopping problems, one associated with each arm. In the current context, this would work as follows: as before let p(i) denote the principal's current belief about agent i's type, and suppose the principal's decision is to either employ agent i for the current period, or stop the process and accept a payment of $m \in \mathbb{R}$; further, if the principal employs the agent today, then he faces the same decision tomorrow (albeit with a potentially different belief). Now solve for the value m(p(i)), agent i's current "Gittins index", such that the principal is indifferent between employing agent i today and stopping the process, and do so for each of the n agents; note that m(.) is not indexed by i since agents are identical up to their type, nor is indexed by time since with geometric discounting and an infinite horizon the stopping problem is time–invariant. Then Gittins and Jones (1974) show that given the current belief about the n agents $p^t = (p^t(1),...,p^t(n))$, the optimal decision is to select the agent with the highest value of $m(p^t(i))$. Thus, letting $p^t(h^{t-1};\sigma)$ denote the Bayesian updated belief about the agents after the history h^{t-1}, we have that for all t, and all t–1 histories h^{t-1}, $\sigma^{*t}(n)(h^{t-1}) \in \underset{i}{\text{argmax}}\ m(p^t(i)(h^{t-1};\sigma^*))$. Note that although this definition looks circular, ie. σ^* appears on both sides, it is actually recursive in t, since $p^t(.;\sigma)$ requires only knowledge of the principal's decisions up to period t–1.

Now since all agents in the current model are identical ex ante, an implication of this result is that $\sigma^*(n)$ employs no recall through period n, or more precisely

that the principal never recalls a previously discarded agent until all available agents have been employed for at least one period. The reason is that if agent i has been employed only once, and has been discarded for an untried agent, then at the time i was dropped $m(p(i)) < m(\pi)$; further, since no new information about i has accrued, i's Gittins index is still $m(p(i))$. Thus, as long as there exists an untried agent, the principal will never recall agent i. Theorem 0 below can be seen as an immediate extension of this result to the case where there are an ***infinite*** number of agents: there exists an optimal strategy without recall, since there ***always*** exists an untried agent. However, the Theorem of Gittins and Jones is for ***finite***–armed bandits only, and we are unaware of any generalizations to the infinite–armed case. While we conjecture that such an extension should be possible (with some additional minor conditions[10]), we show that, in effect, their result does hold in the context of the current model.

Theorem 0. $W(\sigma_{nr}) = \sup_{\Sigma} W(\sigma)$.

Proof. Suppose not; let $\hat{\sigma}$ be such that $W(\hat{\sigma}) - W(\sigma_{nr}) = \epsilon > 0$, and define $t(\epsilon)$ as the unique value of t solving

$$\epsilon = \frac{\delta^t R_1}{1-\delta} .$$

Define the strategy σ' as follows: $\forall\ t \leq t(\epsilon)$, $\sigma^{t,} = \hat{\sigma}^t$, while for all $t > t(\epsilon)$ $\sigma^{t,}$ selects only those agents previously employed plus a set of untried agents, where the

---

[10]In a general infinite–armed bandit problem, one diffculty would be that an arm with maximum Gittins index need not exist, implying one additional condition may be that such a maximum does exist. Notice that this is not a problem in the current model: although there exist an infinite number of arms/agents, the fact that all are identical ex ante imples there is only a finite number of distinct values for the Gittins index at any point in time, since all untried agents have the ***same*** index.

sum of agents across these two groups is equal to $t(\epsilon)$. Thus σ' can be thought of as a strategy associated with a $t(\epsilon)$–armed bandit problem.

Claim 1. $|W(\hat{\sigma}) - W(\sigma')| < \epsilon$.

This follows since, i) $\hat{\sigma}$ and σ' agree on $t \leq t(\epsilon)$, ii) the maximum expected payoff from period $t(\epsilon)$ onwards, evaluated at $t = 0$, is

$$\delta^{t(\epsilon)}R_1 + \delta^{t(\epsilon)+1}R_1 + \delta^{t(\epsilon)+2}R_1 + \ldots$$

$$= \frac{\delta^{t(\epsilon)}R_1}{1-\delta} = \epsilon, \text{ and}$$

iii) the minimum expected payoff from $t(\epsilon)$ onwards is zero.

Recall $\sigma^*(n)$ is the optimal strategy when only n agents are available, ie. an n–armed bandit problem.

Claim 2. $W(\sigma^*(n))$ is strictly increasing in n.

Suppose n+1 agents are available, and let $\sigma(n;n+1)$ denote the Gittins index strategy which ignores one agent; clearly $W(\sigma(n;n+1)) = W(\sigma^*(n))$. By the Gittins and Jones Theorem, $\sigma^*(n+1)$ is the uniquely optimal strategy when n+1 agents are available, implying $W(\sigma^*(n+1)) > W(\sigma(n;n+1)) = W(\sigma^*(n))$.

Claim 3. $W(\sigma^*(t(\epsilon))) \geq W(\sigma')$.

This follows since both $\sigma^*(t(\epsilon))$ and σ' employ $t(\epsilon)$ agents, and by Gittins and Jones $\sigma^*(t(\epsilon))$ is the optimal strategy among those employing $t(\epsilon)$ agents.

Claim 4. $\forall\ n < \infty$, $W(\sigma_{nr}) \geq W(\sigma^*(n))$.

If not, then there exists t' such that $W(\sigma^*(t')) - W(\sigma_{nr}) = \lambda > 0$, implying (by Claim 3) $\forall\ t > t'$ $W(\sigma^*(t)) - W(\sigma_{nr}) > \lambda$. But since $\sigma^*(n)$ and σ_{nr} agree on the first n periods, $|W(\sigma^*(n)) - W(\sigma_{nr})|$ is bounded above by $\delta^n R_1/(1-\delta)$, which is less than λ for n sufficiently large.

Thus we have $W(\sigma_{nr}) \geq W(\sigma^*(t(\epsilon)))$ (by Claim 4)

$\geq W(\sigma')$ (by Claim 3)

and $|W(\hat{\sigma}) - W(\sigma')| < \epsilon$ (by Claim 1). Thus either $W(\sigma') \geq W(\hat{\sigma})$, in which case $W(\sigma_{nr}) \geq W(\hat{\sigma})$, or else $W(\hat{\sigma}) \geq W(\sigma')$, in which case $W(\hat{\sigma}) - W(\sigma_{nr}) < \epsilon$.

QED

Theorem 0 implies that the ***only*** previously employed agent the principal might select in period t is the agent employed in period t–1, whom we now refer to as the current ***incumbent***. Thus, we can without loss of generality recast the principal's problem as a binary choice problem at each point in time of either retaining the current incumbent, or else replacing him with a selection from the (infinite) set of untried and ex ante identical agents, whom we refer to as ***challengers***. Further, the only relevant information accruing from past realizations is the principal's belief about the current incumbent's type; thus, redefine $p = (p_1,\ldots,p_K) \in P(\Omega)$ as the belief about the current incumbent only, while $\pi \in P(\Omega)$ is the belief about current and future challengers. For all $p \in P(\Omega)$, let $f^p(.)$ be the expected density over

rewards given belief p: $f^p(r) = \sum_k p_k f_k(r)$; and let R(p) denote the expected reward given p: $R(p) = \sum_k p_k R_k$. Then the decision to replace the current incumbent with a challenger will affect the principal's immediate (expected) rewards to the principal, ie. either R(p) or $R(\pi)$, the distribution of rewards, either $f^p(.)$ or $f^\pi(.)$, and (through the realized reward and Bayes' Rule) the beliefs of the principal going into the next period, at which time the principal faces an identical decision problem.

3. Existence of an optimal policy

In this section we prove the existence of an optimal "no recall" strategy. As noted above, when restricting attention to no recall strategies the principal's problem is one of binary choice, ie. select either the incumbent or a challenger, and the only relevant variable in this decision is the principal's belief about the current incumbent's type. We now follow Easley and Kiefer (1988) and formulate this binary choice problem as a dynamic programming problem, whose components are:

i) the state space $S = P(\Omega)$;

ii) the action space $A = \{0,1\}$, where 0 denotes replacing the incumbent and 1 denotes retaining the incumbent;

iii) the (expected) reward function $R^*: S \times A \to \mathbb{R}$, with

$$R^*(p,1) = \int r f^p(r)dr = \sum_i p_i R_i = R(p)$$

$$R^*(p,0) = \int r f^\pi(r)dr = \sum_i \pi_i R_i = R(\pi) ;$$

iv) the transition probabilities $Q : S \times A \to P(S)$, where

$$Q(p,1)(X) = \text{prob } \{\beta(p,r) \in X \mid p\}$$

$$Q(p,0)(X) = \text{prob } \{\beta(\pi,r) \in X \mid \pi\}$$

for X a Borel subset of S.

The transition probabilities in (iv) are easily defined using the Bayes map β : $P(\Omega)\times[\underline{r},\bar{r}] \to P(\Omega)$ from eq. (2), and so we state the dynamic programming problem in terms of β rather than Q.

Our first result tailors the usual existence result of dynamic programming to the current environment.

Theorem 1.

i) The principal's optimization problem is well–defined. The value function $V : S \to \mathbb{R}$ associated with the problem is continuous on S, and for all $p \in S$ satisfies

$$V(p) = \max \{V(\pi),\ R(p) + \delta\int V(\beta(p,r))f^{p}(r)dr\}\ ; \tag{3}$$

ii) There exists a stationary optimal policy $\alpha^{*} : S \to A$.

Proof. Let $C(S,\mathbb{R})$ denote the space of continuous, bounded, real–valued functions on S, and for all $w \in S$ define the operator M on $C(S,\mathbb{R})$ by

$$\begin{aligned} Mw(p) = \max_{\alpha\in\{0,1\}} \{&\alpha R(p) + (1-\alpha)R(\pi) + \delta[\alpha\int w(\beta(p,r))f^{p}(r)dr \\ &+ (1-\alpha)\int w(\beta(\pi,r))f^{\pi}(r)dr]\}. \end{aligned} \tag{4}$$

It is easily seen that the Bayesian updating rule is continuous; hence Mw maps $C(S,\mathbb{R})$ into $C(S,\mathbb{R})$ by the standard arguments.

Since, (a) $v,w \in C(S,\mathbb{R})$ and $v \leq w$ implies $Mv \leq Mw$, and (b) $M(w + x) = Mw + \delta x$ for any constant x, it follows from Blackwell (1965) that M is a contraction given $\delta < 1$. Therefore M has a unique fixed point V:

$$V(p) = \max_{\alpha \in \{0,1\}} \{\alpha R(p) + (1-\alpha)R(\pi) + \alpha\delta \int V(\beta(p,r)) f^{p}(r)dr + (1-\alpha)\delta \int V(\beta(\pi,r)) f^{\pi}(r)dr\}. \quad (5)$$

It is immediate that we can translate this expression for V(.) into the more compact form given in the statement of the Theorem. Further, that V(.) is the value function of the problem also follows from standard dynamic programming arguments; cf. Maitra (1968).

Condition (ii) follows from the fact that the continuity of V(.) implies the correspondence of maximizers in (5) is upper–hemicontinuous, and hence admits a measurable selection. QED

In the next two sections we consider the characteristics of the optimal stationary policy α^{*}. It turns out that for the case K = 2, ie. two possible agent types, we get an extremely precise result, while for $K \geq 3$ the picture is a little murky. The proof of the former is facilitated by the following:

Theorem 2. The value function $V : S \to \mathbb{R}$ is convex.

Proof. The arguments follow McLennan (1988), and proceed in several steps. For all $w \in C(S,\mathbb{R})$ define

$$Hw(p) = \int w(\beta(p,r)) f^{p}(r)dr. \quad (6)$$

We first show that if $w \in C(S,\mathbb{R})$ is convex, so is Hw. Using this and the linearity of R(.) in p, we show that Hw convex implies Mw convex, where M is the contraction mapping from Theorem 1. Lastly, we show that this implies the existence of a convex $w^{*} \in C(S,\mathbb{R})$ such that $w^{*} = Mw^{*}$. Since V is the unique fixed point of M, this implies V is convex.

Let $w \in C(S,\mathbb{R})$ and $p^1,p^2 \in S$, and define $p = \gamma p^1 + (1-\gamma)p^2$ for some $\gamma \in (0,1)$.

Step 1. w convex implies Hw convex.

For each $r \in [\underline{r},\overline{r}]$ define $\epsilon(r) \in (0,1)$ by

$$\epsilon(r)f^p(r) = \gamma f^{p^2}(r) \tag{7}$$

(equivalently $(1 - \epsilon(r))f^p(r) = (1 - \gamma)f^{p^1}(r)$).

Observe that $\epsilon(r) \in (0,1)$ and

$$(1-\epsilon(r))\beta(p^1,r) + \epsilon(r)\beta(p^2,r) = \beta(p,r). \tag{8}$$

Suppose w is convex; then

$$Hw(p) = \int w(\beta(p,r)f^p(r)dr$$

$$= \int w((1-\epsilon(r))\beta(p^1,r) + \epsilon(r)\beta(p^2,r))f^p(r)dr \tag{9}$$

$$\leq \int [(1-\epsilon(r))w(\beta(p^1,r)) + \epsilon(r)w(\beta(p^2,r))]f^p(r)dr \tag{10}$$

(by Jensen's inequality)

$$= \int \frac{(1-\gamma)f^{p^1}(r)}{f^p(r)} w(\beta(p^1,r))f^p(r)dr$$

$$+ \int \frac{\gamma f^{p^2}(r)}{f^p(r)} w(\beta(p^2,r))f^p(r)dr \tag{11}$$

$$= (1-\gamma)H_w(p^1) + \gamma H_w(p^2), \text{ completing step 1.}$$

Step 2. Hw convex implies Mw convex.

This is a straightforward exercise using step 1 and the linearity of R(.) in p.

Step 3.

Let **W** be the set of all convex w such that $w \leq Mw$. Since R(.) is bounded and $\delta \in (0,1)$, **W** is non–empty (eg. take $w \equiv 0$) and bounded above. Define w^* by

$$w^*(p) = \sup_{w \in \mathbf{W}} w(p). \tag{12}$$

As the supremum of convex functions, w^* is convex as well.

But $w \leq Mw$ also implies $Mw \leq M(Mw)$ for any $w \in \mathbf{W}$. By step 2, w convex implies Mw convex, so for all $w \in \mathbf{W}$ we have $Mw \in \mathbf{W}$. However, by the definition of w^*, $Mw^* \in \mathbf{W}$ implies $w^* \geq Mw^*$, implying $w^* = Mw^*$. Since V is the unique fixed point of M, $V = w^*$, thus proving the convexity of V. QED

The intuition behind Theorem 2 is quite straightforward. Consider $p,p' \in P(\Omega)$, and let $p_\lambda = \lambda p + (1-\lambda)p'$, where $\lambda \in (0,1)$. Then p_λ can be viewed as the expected prior before it is known whether an event has occurred, where λ is the probability of the event's occurrence. On average, knowing whether or not the event has occurred cannot lower the value of the problem, which is precisely the convexity of V(.).

4. A bandit characterization of the optimal policy α^*

In Section 2 above we saw how the Theorem of Gittins and Jones (1974) generates a concise characterization of an optimal solution to a finite–armed bandit problem. In this Section we derive an analogous result for the current model; namely, the optimal policy α^* for the principal can be characterized by the solutions to a family of stopping problems. What makes this exercise tractable is the "no recall" property of an optimal strategy for the principal: rather than solving for the solutions to an ***infinite*** number of stopping problems (as might be the case in a

general infinite–armed bandit problem), we need only solve for ***two***, one for the current incumbent and one for a randomly chosen challenger.

Recall that in the stopping problem, the principal has the option of either allowing an agent to perform the task in the current period, or stopping the process and accepting a payment of $m \in \mathbb{R}$.[11] Further, if the agent is chosen, then in the next period the choice is between this same agent and the payment m, although the beliefs of the principal concerning the agent's type will undoubtedly differ. For any $m \in \mathbb{R}$, let $W(.;m)$ be the unique fixed point of the contraction mapping

$$Mw(p;m) = \max \{m,\ R(p) + \delta \int w(\beta(p,r);m) f^{p}(r)dr\}, \tag{13}$$

and for all $p \in P(\Omega)$ define $m(p) = \inf \{m : W(p;m) = m\}$. Thus, m(p) is the amount rendering the principal indifferent between selecting the agent, where the principal's current belief about the agent is p, and selecting the payment.

Theorem 3. $\alpha^{*}(p) = 1$ if and only if $m(p) \geq m(\pi)$.

Proof.

Step 1. $W(p;.) : \mathbb{R} \to \mathbb{R}$ is convex and increasing on $\mathbb{R}$.

$W(p;.)$ increasing in m is obvious. If g is any strategy that stays with m once it is selected (as an optimal policy must), it is trivial to see that the payoff under g is a linear increasing function of m. Since the optimal strategy is the payoff–supremum over all such g, and the supremum of non–decreasing linear functions is convex and non–decreasing, the result follows.

---

[11]Note that this problem is strategically equivalent to a two–armed bandit problem with one arm generating rewards according to a known fixed distribution, since in any optimal solution to the latter, if the "known" arm is uniquely optimal in any period, it remains uniquely optimal in all subsequent periods.

Step 2. $W(p;m) - W(p,m') \leq m - m'$ for all $m > m'$.

Let $g(.;m)$ be the optimal policy under m, and consider the policy $g(.;m)$ when m' is the outside option. Clearly,

$$W(p;m) \leq \hat{W}(p;m'|g(.;m)) + (m - m') \tag{14}$$

$$\leq W(p;m') + (m - m'), \tag{15}$$

where $\hat{W}(p;m'|g(.;m))$ is the payoff under $g(.;m)$ if the outside option is m'. The first inequality holds since the only difference between $W(p;m)$ and $\hat{W}(p;m|g(.;m))$ is as a consequence of taking the outside option. The second inequality follows immediately from the first.

Step 3. $[W(p;m) - m]$ is decreasing in m.

Follows from Step 2.

Step 4. Let $m = V(\pi)$; then $W(.;m) \equiv V(.)$.

Follows from the uniqueness of fixed points of the respective contraction mappings.

Step 5. $m(\pi) = V(\pi)$.

By Step 4, $W(\pi;V(\pi)) = V(\pi)$, so by definition of $m(\pi)$, $V(\pi) \geq m(\pi)$. Suppose $V(\pi) > m(\pi)$, and define

$$HW(p;m) = R(p) + \delta \int W(\beta(p,r);m) f^{p}(r)dr. \tag{16}$$

Since $W(.;V(\pi)) = V(.)$, so $HW(\pi;V(\pi)) = V(\pi)$ by the definition of $V(.)$.

We will show that for any m, if $m(\pi) < m$ then $HW(\pi;m) < m$. This will imply $HW(\pi;V(\pi)) < V(\pi)$, a contradiction. If $m > m(\pi)$, then

$$HW(\pi;m) = R(\pi) + \delta \int W(\beta(\pi,r);m) f^{\pi}(r)dr \tag{17}$$

$$= R(\pi) + \delta\int W(\beta(\pi,r);m(\pi))f^{\pi}(r)dr$$

$$+ \delta\{\int[W(\beta(\pi,r);m) - W(\beta(\pi,r);m(\pi))]f^{\pi}(r)dr\ \} \qquad (18)$$

$\leq m(\pi) + \delta(m - m(\pi))$ (by Step 2 and the definition of m(p))

$$= (1-\delta)m(\pi) + \delta m \qquad (19)$$

$< m$, since $\delta \in (0,1)$ and $m(\pi) < m$.

Step 6. $m(p) > m(\pi) <=> V(p) > V(\pi)$.

Suppose $m(p) > m(\pi)$ and $V(p) \leq V(\pi)$; then $V(p) = V(\pi)$. But $V(p) = W(p;m(\pi))$ (by Step 4), so $W(p;m(\pi)) = V(p) = V(\pi) = m(\pi)$. Therefore by the definition of $m(.)$, $m(p) \leq m(\pi)$, a contradiction.

Suppose now $m(p) \leq m(\pi)$ and $V(p) > V(\pi)$. Then $V(p) = W(p;m(\pi)) > V(\pi) = m(\pi)$, implying $W(p;m(\pi)) - m(\pi) > 0$. Since $W(p;m) - m$ is decreasing (by Step 2), it follows that $m(p) > m(\pi)$. QED

Thus Theorem 3 shows how α^*, the solution to an infinite–agent dynamic choice problem, can be characterized by the solution to a family of simple optimal stopping problems.

5. Can myopic decision rules be optimal?

A natural question concerns the extent to which the optimal policy α^* incorporates the consideration of future rewards into the current decision whether to retain the incumbent. We say a stationary decision rule $\alpha : S \to A$ is ***myopic*** if $\alpha(p) = 1$ if and only if $R(p) > R(\pi)$. Myopic decision rules are obviously a particularly simple class of rules in that they require ***no*** consideration of future

rewards, or, more precisely, that such a consideration is already imbedded in the maximization of current rewards. While it may be apparent that one can always find a set of parameters (eg. $\{f_k\}$,δ, or π, the most obvious case being $\delta = 0$) such that α^* is myopic, the next result shows that if there exists only two types of agents, then the optimal decision rule is ***always*** myopic.

Theorem 4. If $K = 2$, then $\alpha^*(.)$ is myopic; in particular, $\alpha^*(p) = 1$ if and only if $p_1 > \pi_1$.

Proof. To simplify notation let p be the probability the current incumbent's type is ω_1, and note that $\forall\ r \in [\underline{r},\bar{r}]$, $\beta(.,r)$ is strictly increasing.

Now by (3) we know that V(.) is bounded below by $V(\pi)$, and hence attains a global minimum at $p = \pi$. Also, $\beta(0,r) = 0$ and $\beta(1,r) = 1$ for all $r \in [\underline{r},\bar{r}]$; it follows then that since $R_1 > R_2$, V(.) attains a global minimum at $p = 0$ and a global maximum at $p = 1$. By the continuity and convexity of V(.) (Theorems 1 and 2), it is immediate that V(.) is constant on $[0,\bar{p}]$ and non–decreasing on $(\bar{p},1]$, where $\bar{p} \geq \pi$; further, $\forall\ p \in (\bar{p},1]$, $V(p) > V(\pi)$. Thus, α^* is of the form

$$\alpha^*(p) = \begin{cases} 0 & \text{if } p \leq \bar{p} \\ 1 & \text{if } p > \bar{p} \end{cases}. \tag{20}$$

We show that $\forall\ p \in (\pi,1]$, $V(p) > V(\pi)$, implying $\bar{p} = \pi$, and therefore proving the Theorem.

By (20) we can write $V(\pi)$ as

$$V(\pi) = R(\pi) + \delta\{\int_A V(\pi)f^{\pi}(r)dr + \int_B V(\beta(\pi,r))f^{\pi}(r)dr\}, \tag{21}$$

where $A = \{r : \beta(\pi,r) \leq \bar{p}\}$, $B = \{r : \beta(\pi,r) > \bar{p}\}$. Thus A and B partition $[\underline{r},\bar{r}]$,

and if $r \in A$ the current incumbent is replaced at the beginning of the next period, while if $r \in B$ the incumbent is retained. Further, since $\beta(\pi,r) \geq \pi$ if and only if $f_1(r) \geq f_2(r)$, and $\bar{p} \geq \pi$, it must be that $B \subseteq \{r : f_1(r) \geq f_2(r)\}$.

Next, since V(.) gives the payoff arising from the optimal policy, we know that for all values of p

$$V(p) \geq R(p) + \delta\{\int_A V(\pi)f^p(r)dr + \int_B V(\beta(p,r))f^p(r)dr\}, \tag{22}$$

where the sets A and B are the same as those in (21). The RHS of (22) is the payoff associated with retaining the current incumbent, replacing him in the next period if $r \in A$ and then proceeding according to α^*, and retaining him if $r \in B$ and then proceeding according to α^*. Thus

$$V(p) - V(\pi) \geq R(p) - R(\pi) + \delta\{V(\pi)\int_A [f^p(r) - f^\pi(r)]dr + \int_B V(\beta(p,r))f^p(r)dr - \int_B V(\beta(\pi,r))f^\pi(r)dr\}. \tag{23}$$

Let $p > \pi$; then $R(p) > R(\pi)$, so that $V(p) - V(\pi) > 0$ if the bracketed term in (23) is non–negative. To see this holds, note that

$$V(\pi)\int_A [f^p(r) - f^\pi(r)]dr + \int_B V(\beta(p,r))f^p(r)dr - \int_B V(\beta(\pi,r))f^\pi(r)dr$$

$$\geq V(\pi)\int_A [f^p(r) - f^\pi(r)]dr + \int_B V(\beta(\pi,r))[f^p(r) - f^\pi(r)]dr \tag{24}$$

(since $\beta(.,r)$ and V(.) non–decreasing imply $V(\beta(.,r))$ is non–decreasing)

$$\geq V(\pi)\int_A [f^p(r) - f^{\pi}(r)]dr + \int_B V(\pi)[f^p(r) - f^{\pi}(r)]dr \qquad (25)$$

(since $\forall\ p \in [0,1]\ V(p) \geq V(\pi)$, and $[f^p(.) - f^{\pi}(.)]$ is non–negative on the set B)

$$= V(\pi)\ [\int_{A\cup B} f^p(r)dr - \int_{A\cup B} f^{\pi}(r)dr] \qquad (26)$$

$$= 0. \qquad \text{QED}$$

Hence if there exist only two types of agents, the principal's optimal policy prescribes simply selecting the agent who provides the highest expected reward for the current period, thereby apparently ignoring consideration of future beliefs and future rewards. Further, the reward distributions, $f_1(.)$ and $f_2(.)$, are immaterial to the optimal policy beyond simply the determination of which type generates the highest expected 1–period reward.

Two additional features of the model restricted to two types are worth noting. The first is that the policy in Theorem 4 is optimal even if the model is one of finite, rather than infinite, time. The key is that if $p_1 > \pi_1$ then regardless of the realization of rewards the current incumbent will be preferred to the current challenger in the next time period. Thus, consider the optimal policy if the current incumbent is replaced, and alter this policy by keeping the current incumbent and subsequently following the former policy with respect to retaining the agent. Thus, if in the next period the current challenger would be retained, retain the current incumbent as well; on the other hand if the current challenger would be replaced, replace the current incumbent. Then the principal receives a higher expected payoff today from retaining the current incumbent, and will either receive a higher expected payoff (if both would be retained) or the same expected payoff tomorrow.

Continuing this argument, we see that, regardless of when the procedure ends, retaining the current incumbent if $p_1 > \pi_1$ is optimal.

The second feature of this model is that the optimal decision rule in the space of ***rewards***, as opposed to beliefs, has the following simple property: if r is such that $f_1(r) \geq f_2(r)$, then the incumbent is necessarily retained regardless of p. This follows by Bayesian updating: if $p_1 > \pi_1$ and r is such that $f_1(r) \geq f_2(r)$, then $\beta_1(p,r) > p_1 > \pi_1$, implying $R(\beta(p,r)) > R(\pi)$ and by the Theorem the incumbent is retained; if on the other hand $f_1(r) < f_2(r)$ then the optimal decision rule depends on the relative magnitudes of $f_1(r)$ and $f_2(r)$ as well as the previous belief p. Thus, for example, if $[\underline{r},\overline{r}] = [0,1]$, $f_1(r) = 3r^2$, and $f_2(r) = 2r$, then if $r \geq 2/3$ the incumbent is necessarily retained.

On the other hand it is easily seen how Theorem 4 and the ensuing conclusions do not extend to the case where $K \geq 3$. Consider the following 4–period example, where $\Omega = \{\omega_1,\omega_2,\omega_3\}$, and the reward distributions are Bernoulli: let $q_k = \text{prob}\{r = 1 : \omega = \omega_k\}$, $1 - q_k = \text{prob}\{r = 0 : \omega = \omega_k\}$, and note that $R_k = q_k$. Let the prior on challengers be $\pi = (1/3,1/3,1/3)$, and $q = (q_1,q_2,q_3) = (3/4,1/2,1/4)$. Suppose r^1, the realized reward in period 1, is 1, in which case the incumbent is retained, and $r^2 = 0$. Applying Bayes' rule we see that the principal's belief about the incumbent at the beginning of period 3 is $p = (.3,.4,.3)$, implying in particular $R(p) = R(\pi)$. Should the principal retain the incumbent? The answer is "no", according to the following argument: with either the incumbent or the challenger the probability of $r^3 = 0$, which would entail selecting the challenger at the beginning of period 4, is 1/2; further if $r^3 = 0$ the expected payoff is the same regardless of the $t = 3$ decision. Now if the current incumbent is retained and r^3 turns out to be 1, then Bayes' rule implies $\beta(p,1) = (9/20,8/20,3/20)$, giving an expected reward of $R(\beta(p,1)) = 23/40$; if on the other hand the challenger is selected and $r^3 = 1$, then $\beta(\pi,1) = (1/2,1/3,1/6)$, giving an expected reward of $R(\beta(\pi,1)) = 7/12 > 23/40$.

Hence the expected payoff in period 3 is the same regardless of whether the incumbent is retained, while the expected payoff in period 4, evaluated at the beginning of period 3, is strictly greater if the incumbent is replaced. Thus, even though the principal is indifferent with regard to myopic or current payoffs, he is not with regard to future payoffs.

The key to this example is that if $p_1 = p_3$, as is the case here, than R(p) is necessarily equal to $R(\pi)$, and the next period's reward (if $r^3 = 1$) will be higher from keeping the current incumbent only if $p_1 > \pi_1 = 1/3$. However given the symmetry of the reward distributions, and since the principal has observed the same number of "1's" and "0's", he continues to believe $p_1 = p_3$, but now has shifted weight (relative to the prior) onto p_2, thus guaranteeing that p_1 will be less than π_1. Further, we can modify the example by "perturbing" p so that R(p) is strictly greater than $R(\pi)$ and yet the incumbent is still replaced. To do so use the above analysis as that pertaining to the last two periods preceded by a string of x periods with r = 1 rewards followed by a string of x–1 periods with r = 0 rewards. Hence in all previous periods the expected 1–period reward from the incumbent, is greater than that of the challengers, so we can assume the originally chosen agent is still the current incumbent. Then for x large we get that $p_1 = p_3 + \epsilon$ for some small ϵ, implying $R(p) = R(\pi) + \rho$ for ρ small, but p_2 will be close to 1. Therefore p_1 will be less than π_1, implying from above that $R(\beta(\pi,1))$ will always be strictly greater than $R(\beta(p,1))$, and therefore that for x large enough (and consequently ρ small enough) replacing the incumbent even though $R(p) > R(\pi)$ will be optimal.

As noted in Section 1, this gives a prescription which never occurs in the two–armed bandit problem with one known arm. In the latter case, if the unknown arm gives a higher (expected) payoff today than the known arm, then the principal is better off with the unknown arm today, and cannot be any worse off tomorrow, since he can still select the known arm at that time regardless of the realized

reward. In the current model, however, the current challenger, while giving a fixed expected payoff and being always available, is itself an unknown arm, and as such may give a different payoff tomorrow depending on the realized reward. Indeed, this is precisely the moral of the above example: today's challenger might look better than tomorrow's challenger or today's incumbent from ***tomorrow's*** perspective, even if today's challenger looks worse than today's incumbent from ***today's*** perspective. Of course, as noted above this logic fails when there are only two types of agent, in which case if today's incumbent looks better than today's challenger, then the former will look better than the latter for all possible realizations of rewards.

6. Adding uncertainty about the distribution of types

Suppose we now modify the principal's problem by assuming that, in addition to not knowing any agent's true type, the principal is also uncertain about π, the likelihood a challenger is of a particular type. In this case, upon observing a reward from the current incumbent the principal will not only update his beliefs about the former's type, but also about π, and hence about future challengers. This follows since the current incumbent's type was also drawn according to π, although at present the principal may have more information about the incumbent's type than about future challengers.

We make two simplifying assumptions. The first is that we restrict attention to "no recall" strategies, so that the principal's optimization problem can again be formulated as a dynamic programming problem. As we show below by way of example, in contrast to the basic model this assumption is no longer without loss of generality: since the principal is now learning about his "outside options", ie. the value of π, he may wish to return to a previously discarded agent, who now may

look attractive relative to a randomly chosen challenger. Therefore Theorem 0 will not necessarily generalize to the current environment.

The second assumption is that the principal "knows" $\pi \in \Pi = \{\pi(1),...,\pi(J)\}$; of course, the case where $J = 1$ corresponds to the previous model. The principal's belief space is now $P(\Omega \times \Pi)$, where $\lambda \in P(\Omega \times \Pi)$ gives the joint distribution over the current incumbent's type as well as over the population distribution. Thus λ_{kj} denotes the probability associated with the joint event $\{\omega = \omega_k, \pi = \pi(j)\}$, $k = 1,...,K$, $j = 1,...,J$. The marginal of λ on Ω, which yields the principal's prior on the incumbent's type, will be denoted $p(\lambda)$, and the marginal of λ on Π, ie. the principal's prior over the "true" population distribution, will be denoted $\mu(\lambda)$:

$$\text{prob } \{\omega = \omega_k : \lambda\} = p_k(\lambda) = \sum_j \lambda_{kj} , \tag{27}$$

$$\text{prob } \{\pi = \pi(j) : \lambda\} = \mu_j(\lambda) = \sum_k \lambda_{kj} . \tag{28}$$

Given a prior $\lambda \in P(\Omega \times \Pi)$ and an observation $r \in [\underline{r},\bar{r}]$, the principal again updates his beliefs according to Bayes' rule:

$$\beta_{kj}(\lambda,r) = \frac{\lambda_{kj} f_k(r)}{\sum_m \sum_q \lambda_{mq} f_m(r)} , \quad k = 1,...,K, \ j = 1,...,J. \tag{29}$$

In particular,

$$p(\beta(\lambda,r)) = \left(\frac{f_k(r)\Sigma_j \lambda_{kj}}{\Sigma_m \Sigma_q f_m(r) \lambda_{mq}} \right)_{k=1,...,K}$$

$$= \left(\frac{f_k(r) p_k(\lambda)}{\sum_m f_m(r) p_m(\lambda)} \right)_{k=1,...,K} \tag{30}$$

gives the updated belief about the current incumbent, while

$$\mu(\beta(\lambda,r)) = \left(\frac{\Sigma_k f_k(r)\lambda_{kj}}{\sum_m \sum_q f_m(r)\lambda_{mq}} \right)_{j=1,\ldots,J} \tag{31}$$

gives the updated belief about the population distribution. Note that the updated belief about the current incumbent can be expressed as a simple function of the prior belief about the incumbent, as in the earlier model where π was known.

Given λ, we denote the joint distribution obtained when the principal replaces the incumbent with a challenger by $h(\lambda)$, ie. the new prior $h(\lambda)$ is the joint distribution of the new incumbent's type and the population distribution. The belief $h(\lambda)$ is defined as follows: for $\mu \in P(\Pi)$, denote by $\bar{p}(\mu)$ the expected distribution of types under μ of a new draw from the population, ie.

$$\bar{p}_k(\mu) = \text{prob}\{\omega = \omega_k : \mu\} = \sum_j \pi_k(j)\mu_j \,, \; k = 1,\ldots,K; \tag{32}$$

when $\mu = \mu(\lambda)$ for some λ, we denote $\bar{p}(\mu)$ by $\bar{p}(\lambda)$. Then $h(\lambda)$ is obtained from λ by simply replacing $p(\lambda)$ with $\bar{p}(\lambda)$:

$$h_{kj}(\lambda) = \text{prob}\{\omega = \omega_k, \; \pi = \pi(j)\} = \pi_k(j)\mu_j, \quad k = 1,\ldots,K, \; j = 1,\ldots,J, \tag{33}$$

and note that $h(h(\lambda)) = h(\lambda)$.

The principal is now faced with the following problem at the beginning of every period: whether to continue with the present incumbent, so that the updated prior next period after an observation of r is $\beta(\lambda,r)$, or to replace the incumbent, in which case the new starting prior is $h(\lambda)$ and the Bayesian updated belief will be $\beta(h(\lambda),r)$. As before, the problem may be transformed into a dynamic programming problem, with

i) <u>state space</u> $\tilde{S} = P(\Omega \times \Pi)$

ii) action space $\tilde{A} = \{0,1\}$

iii) (expected) reward function $\tilde{R}^* : \tilde{S}\times\tilde{A} \to \mathbb{R}$, where

$$\tilde{R}^*(\lambda,1) = \int r f^{p(\lambda)}(r)dr = R(p(\lambda))$$

$$\tilde{R}^*(\lambda,0) = \int r f^{\bar{p}(\lambda)}(r)dr = R(\bar{p}(\lambda))$$

iv) transition probabilities $\tilde{Q} : \tilde{S}\times\tilde{A} \to P(\tilde{S})$, where

$$\tilde{Q}(\lambda,1)(X) = \text{prob } \{\beta(\lambda,r) \in X \mid \lambda\}$$

$$\tilde{Q}(\lambda,0)(X) = \text{prob } \{\beta(h(\lambda),r) \in X \mid h(\lambda)\} ,$$

where X is a Borel subset of $\tilde{S}$ and $\tilde{Q}(.)$ is definable from the Bayesian updating rule.

Generalizing the arguments in Section 3, we get the analogous results:

Theorem 5.

i) The principal's optimization problem is well–defined. The value function $\tilde{V} : \tilde{S} \to \mathbb{R}$ associated with the problem is continuous on $\tilde{S}$, and for all $\lambda \in \tilde{S}$ satisfies

$$\tilde{V}(\lambda) = \max \{\tilde{V}(h(\lambda)),\ R(p(\lambda)) + \delta \int \tilde{V}(\beta(\lambda,r)) f^{p(\lambda)}(r)dr\}. \qquad (34)$$

ii) There exists a stationary optimal policy $\tilde{\alpha}^* : \tilde{S} \to \tilde{A}$.

iii) The value function $\tilde{V} : \tilde{S} \to \mathbb{R}$ is convex.

While a bandit characterization analogous to that in Section 4 eludes us for the model with unknown π, we can show how the results in Section 5 do not generalize. Specifically, consider the following two–type, 3–period example, where J = 2, ie. π can take on only one of two values, $\Pi = \{\pi(1),\pi(2)\}$, and where we let $p(\lambda) \in [0,1]$ denote the probability the current incumbent's type is ω_1 Then we can write

$p(\beta(\lambda,r))$, the belief about the current incumbent upon observing a reward of r, given prior λ, as

$$p(\beta(\lambda,r)) = \frac{f_1(r)p(\lambda)}{f_1(r)p(\lambda) + f_2(r)(1-p(\lambda))} . \tag{35}$$

Similarly, after some manipulations we can write $\bar{p}(\beta(\lambda,r))$, the belief about tomorrow's challenger given prior λ and upon observing reward r, as

$$\bar{p}(\beta(\lambda,r)) = \frac{\Sigma_k \Sigma_j f_k(r)\pi(j)\lambda_{kj}}{\Sigma_m \Sigma_q f_m(r)\lambda_{mq}} \tag{36}$$

$$= \frac{f_1(r)\bar{p}(\lambda) + [f_2(r) - f_1(r)]\ \Sigma_j \pi(j)\lambda_{2j}}{f_1(r)\,p(\lambda) + f_2(r)(1-p(\lambda))} ; \tag{37}$$

and finally,

$$\bar{p}(\beta(h(\lambda)),r) = \frac{f_1(r)\bar{p}(\lambda) + [f_2(r) - f_1(r)]\ \Sigma_j \pi(j)h_{2j}}{f_1(r)\bar{p}(\lambda) + f_2(r)(1-\bar{p}(\lambda))} \tag{38}$$

is the belief about tomorrow's challenger, given reward r, if the current incumbent is replaced by the current challenger.

The following result is useful in characterizing the principal's optimal policy in this example.

Lemma. Suppose $\lambda = h(\lambda)$, implying $p(\lambda) = \bar{p}(\lambda)$; then $p(\beta(\lambda,r)) \geq \bar{p}(\beta(\lambda,r))$ if and only if $f_1(r) \geq f_2(r)$.

Proof. Follows immediately from (35), (37), and the fact that $\Sigma_j \pi(j)\lambda_{2j} > 0$. QED

Note: this is as in the 2–type model in Section 5.

Let α^t denote the principal's decision at time t, and let $\hat{p}(r^1,\alpha^1)$ denote the principal's belief about the agent selected at t = 2 given his t = 1 decision, the realized reward r_1 (and subsequent t = 2 decision).[12] Then the expected payoff in t = 2, evaluated at the beginning of period 2, is

$$\int r_2[\hat{p}(r_1)f_1(r_2) + (1-\hat{p}(r_1))f_2(r_2)]dr_2$$

$$= \int r_2 f_2(r_2)dr_2 + \hat{p}(r_1)\int r_2[f_1(r_2) - f_2(r_2)]dr_2$$

$$= R_2 + \hat{p}(r_1)[R_1 - R_2]. \tag{39}$$

Evaluating this payoff at t = 1, we get

$$R_2 + [R_1 - R_2]\int \hat{p}(r_1)[\tilde{p}(\alpha^1)f_1(r_1) + (1-\tilde{p}(\alpha^1))f_2(r_1)]dr_1 \tag{40}$$

where $\tilde{p}(\alpha^1)$ is the belief about the t = 1 incumbent given the principal's decision at the beginning of the t = 1 period.

Now suppose $p(\lambda) = \bar{p}(\lambda)$, implying $R(p(\lambda)) = R(\bar{p}(\lambda))$, ie. the current period expected rewards are the same from either the incumbent or the challenger, and the principal is simply interested in selecting the t = 1 agent to maximize the expected reward in period t = 2. If the incumbent is retained at the beginning of period 1, ie. $\alpha^1 = 1$, then by the Lemma the decision rule at t = 2 will be to retain the incumbent if $f_1(r) \geq f_2(r)$ and replace with the challenger otherwise; further, since

---

[12]The two time periods are to be considered the last two periods of a finite horizon model. The distribution of "initial" beliefs used below could, at the risk of complicating notation, have been derived from identical underlying priors on all agents.

$p(\lambda) = \bar{p}(\lambda)$ this is also the decision rule if $\alpha^1 = 0$, ie. the current incumbent is replaced. Let $A = \{r : f_1(r) < f_2(r)\}$ and $B = \{r : f_1(r) \geq f_2(r)\}$; then, using the Lemma and eqs. (35), (37), and (38) to rewrite eq. (40), we get the following expression for expected $t = 2$ payoffs, evaluated at $t = 1$, if $\alpha^1 = 1$:

$$R_2 + [R_1 - R_2]\{\int_B f_1(r)p(\lambda)dr + \int_A (f_1(r)\bar{p}(\lambda) + [f_2(r)-f_1(r)]\Sigma_j \pi(j)\lambda_{2j})dr\}$$

$$= R_2 + [R_1 - R_2]\{\bar{p}(\lambda) + \Sigma_j \pi(j)\lambda_{2j}\int_A [f_2(r)-f_1(r)]dr\} \tag{41}$$

since $p(\lambda) = \bar{p}(\lambda)$. Similarly, if $\alpha^1 = 0$, we get

$$R_2 + [R_1 - R_2]\{\int_B f_1(r)\bar{p}(\lambda)dr + \int_A (f_1(r)\bar{p}(\lambda) + [f_2(r)-f_1(r)]\Sigma_j \pi(j)h_{2j})dr\}$$

$$= R_2 + [R_1 - R_2]\{\bar{p}(\lambda) + \Sigma_j \pi(j)h_{2j}\int_A [f_2(r)-f_1(r)]dr\}. \tag{42}$$

Canceling terms, we see that $\alpha^1 = 1$ is optimal if and only if

$$\Sigma_j \pi(j)\lambda_{2j} \geq \Sigma_j \pi(j)h_{2j} = \Sigma_j \pi(j)[1-\pi(j)]\mu_j$$
$$= \Sigma_j \pi(j)[1-\pi(j)][\lambda_{1j} + \lambda_{2j}], \tag{43}$$

or,

$$\Sigma_j \pi(j)[\ (1-\pi(j))\lambda_{1j} - \pi(j)\lambda_{2j}] \leq 0. \tag{44}$$

Now let $\pi(1) = .1$, $\pi(2) = .9$, and suppose λ is the following:

	$\pi(1)$	$\pi(2)$
ω_1	10/200	86/200
ω_2	95/200	9/200

Then $p(\lambda) = \bar{p}(\lambda) = 96/200$, and

$$\Sigma_j \pi(j)[\ (1-\pi(j))\lambda_{1j} - \pi(j)\lambda_{2j}] = 4/2000 > 0;$$

hence $\alpha^{*1} = 0$.

Alternatively, suppose λ is the following:

	$\pi(1)$	$\pi(2)$
ω_1	30/200	66/200
ω_2	75/200	29/200

Then again $p(\lambda) = \bar{p}(\lambda) = 96/200$, but now

$$\Sigma_j \pi(j)[\ (1-\pi(j))\lambda_{1k} - \pi(j)\lambda_{2j}] = -\ 14.85/200 < 0,$$

implying $\alpha^{*1} = 1$.

Then, by perturbing λ slightly, we can have $R(p(\lambda)) > R(\bar{p}(\lambda))$ but $\alpha^{*} = 0$, and $R(p(\lambda)) < R(\bar{p}(\lambda))$ but $\alpha^{*} = 1$, ie. the incumbent having a higher 1–period expected payoff is neither necessary nor sufficient for the principal to retain him. In particular, even with only two types of agents and two possible values of π, myopic decision rules may not be optimal. This example, along with the example at the end of Section 5, shows how Theorem 4 is sensitive to both the assumption of only 2 types and to the assumption that π is known with certainty by the principal.

Notice in addition that if in the perturbed example $R(p(\lambda)) > R(\overline{p}(\lambda))$ but $\alpha_1^* = 0$, and the reward is such that the $t = 1$ challenger is retained, then the principal would rather recall the previous incumbent, since for $p(\lambda)$ arbitrarily close to $\overline{p}(\lambda)$, $p(\beta(\lambda,r))$ will be greater than $p(\beta(h(\lambda)),r)$ for some values of r. Therefore adding the ability to recall previously employed agents will non–trivially effect the analysis of the principal's optimal policy when π is unknown; ie. Theorem 0 will not generalize to the current environment as well.

7. Extensions and generalizations

The framework we have offered in this paper admits generalization and further investigation in several interesting directions. Perhaps the most obvious one is to allow agents a choice of actions, eg. effort levels, with the usual stipulation that higher costs accrue to agents from actions yielding (probabilistically) higher rewards to the principal. This would, of course, convert the one–person optimization model in the current paper into an infinite–player stochastic game. Several questions regarding the characteristics of the resulting equilibria now arise. First, continuing to assume the principal is limited to an "up or out" choice and cannot offer differential contracts to separate agent types, what (if any) are the conditions under which a "natural" equilibrium separation occurs in the agents' action space, with higher types invariably selecting higher actions? Second, in the two–type situation, are there Markov–perfect equilibria in which an analog of Theorem 4 holds, implying the principal's equilibrium strategy is myopic? Third, if the principal is allowed to offer differential contracts, what kind of contracts arise in equilibrium? In particular, are "separating" contracts a robust possibility?

An alternative avenue of exploration concerns the generalized model of Section 6, and a more comprehensive characterization of the principal's optimal policy. It would be especially interesting to know if a generalized Gittins index–type characterization is available for this problem along the lines of that provided in Section 3 for the basic model. A second open question within this framework pertains to "learning" in the limit; namely, whether (or under what conditions) the marginal of the principal's belief on the distribution of agent–types converges to point–mass at the "true" distribution.

Finally, although we have provided a reasonably exhaustive characterization of the optimal policy in the basic model, we have left unanswered several questions concerning properties of the model arising from this policy. These include the expected time to dismissal for different agent–types, whether "higher" types last on average longer than "lower" types, etc.

References

Alesina, A. (1988) Credibility and policy convergence in a two–party system with rational voters, American Economic Review 78:796–806.

Alesina, A. and S. Spear (1988) An overlapping generations model of electoral competition, Journal of Public Economics 37:359–379.

Austen–Smith, D. and J. Banks (1989) Electoral accountability and incumbency, in Models of Strategic Choice in Politics, ed. by P. Ordeshook. Ann Arbor: University of Michigan Press.

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. London: Chapman and Hall.

Blackwell (1965) Discounted dynamic programming, Annals of Mathematical Statistics 36:226–235.

Blume, L., M. Bray, and D. Easley (1982) An introduction to the stability of rational expectations equilibrium, Journal of Economic Theory 26:313–317.

Easley, D. and N. Kiefer (1988) Controlling a stochastic process with unknown parameters, Econometrica 56:1045–1064.

Feldman, M. (1989) On the generic nonconvergence of Bayesian actions and beliefs, BEBR working paper, University of Illinois, Urbana–Champaign

Ferejohn, J. (1986) Incumbent performance and electoral control, Public Choice 50:5–25.

Fiorina, M. (1981) Retrospective voting in American national elections. New Haven: Yale University Press.

Ledyard, J. (1989) Information aggregation in two–candidate elections, in Models of Strategic Choice in Politics, ed. by P. Ordeshook. Ann Arbor: University of Michigan Press.

Maitra, A. (1968) Discounted dynamic programming in compact metric spaces, Sankhya Ser A 30:211–216.

McKelvey, R. and R. Reizman (1990) Seniority in legislatures, working paper, California Institute of Technology.

McLennan (1988) Learning in a repeated statistical decision framework, mimeo, University of Minnesota.

Mortensen, D. (1986) Job search and labor market analysis, in Handbook of Labor Economics, Vol. II, ed. by O. Ashenfelter and R. Layard. New York: North–Holland.

Rothschild, M. (1974) A two–armed bandit theory of market pricing, Journal of Economic Theory 9:185–202.