On the Value of Randomizing and Limiting Memory in Repeated Decision-Making under Minimal Regret

Karl H. Schlag[1]

March 2003

[1] Economics Department, European University Institute, Via della Piazzuola 43, 50133 Florence, Italy, Tel: 0039-055-4685951, email: schlag@iue.it

Abstract

We search for behavioral rules that attain minimax regret under geometric discounting in the context of repeated decision making in a stationary environment where payoffs belong to a given bounded interval. Rules that attain minimax regret exist and are optimal for Bayesian decision making under the prior where learning can be argued to be most difficult. Minimax regret can be attained by randomizing using a linear function of the previous payoffs. For myopic individuals, minimax regret behavior requires only one round of memory, for intermediate discount factors two rounds of memory suffice to attain minimax regret.

JEL classification: D81, D83.

1 Introduction

Decision making is an elementary part of human behavior. It is the foundation of any model of strategic interaction. The theory of decision making thus influences directly or indirectly almost any economic prediction. Under what we call rational decision making today (von Neumann-Morgenstern 1944, Savage 1972), the decision maker first specifies a prior probability distribution over the set of states that may occur. Then he selects the action that maximizes expected utility and updates this initial prior after any new information arrives. We call a decision maker *Bayesian* if he behaves according to this procedure. From the start the Bayesian approach has been criticized. In particular it has been questioned whether individuals are able to form such priors and whether they have the ability and time to perform the necessary calculations. These objections are particularly relevant when stakes are low, time is scarce and priors are diffuse (cf Simon, 1982).

In the following we select behavior according to a worst case analysis based on regret which extends work of Berry and Fristedt (1985) who themselves build on Gibbons (1952) and Wald (1950). We measure performance of a rule according to the largest regret it induces among all priors and select for rules with smallest maximal regret. We show how this approach relates to a Bayesian decision maker is endowed with the prior under which learning is most difficult.

We use a distribution free approach which means that we do not invoke priors. This simplifies the task of the individual as he does not have to calculate a new rule each time he faces a similar decision problem. Unfortunately there are only few results on selecting rules for decision making using a distribution free method such as Börgers et al. (2001).

The setting of this paper is as in the classic multi-armed bandit problem where a decision maker repeatedly chooses (in consecutive rounds) one of a finite number of actions. Each choice yields a random payoff which is drawn according to an action dependent distribution which is stationary and independent of previous occurrences. We assume that payoffs belong to $[0,1]$ but our results generalize immediately to any given bounded interval $[\alpha,\omega] \subset \mathbb{R}$. The classic multi-armed bandit specification includes a prior over these payoff distributions.

A *behavioral rule* is a description of which action the individual chooses next given his previous observations. The behavioral rule has n round memory if current behavior does not depend on choices or payoffs obtained more than n rounds ago. A *(randomized) rule* is a probability measure over deterministic behavioral rules. A *multi-action decision problem* is the specification of a payoff distribution for each action. A *prior* is a probability measure over the set of multi-action decision problems.

We assume that the individual is risk neutral and that future payoffs are discounted with a given discount factor $\delta \in (0, 1)$. Again our results generalize immediately if instead we consider von Neuman-Morgenstern utility that is contained in a bounded interval. A *worst case prior* is a prior that maximizes over all priors the difference between the expected discounted payoff obtained if the underlying distributions of each arm are known and the expected discounted payoff achieved by the optimal rule. Here optimality refers to standard payoff maximization while updating the prior over time. The *regret* of a given rule under a given prior is defined as the difference between the expected discounted payoff obtained if the underlying distributions of each arm are known and the expected discounted payoff achieved by the given rule. Thus, the worst case prior maximizes the regret of the optimal rule over all priors.

In this paper we are interested in selecting a rule without making assumptions on the prior and choose to select according to *minimax regret* (Wald, 1950, Gibbons, 1952). The idea is to learn when there are incentives to learn. We evaluate the performance of a given rule by the maximal regret it yields over all priors and then select for the rule that yields the minimal value of the maximal regret.

A behavioral rule is *linear* if choice probabilities are linear observed payoffs. Typically linear rules will not emerge from Bayesian learning as linear rules typically involve randomizing between actions when all payoffs observed belong to $(0, 1)$. On the other hand, we provide evidence that Bayesian optimal rules typically do not involve randomizing.

Symmetric randomized rules are randomized rules whose behavior does not depend on the labels of the actions. Under a symmetric rule each action is played equally likely in the first round. Notice that randomized rules that are symmetric need not have only symmetric rules in their support.

We extend results obtained by Berry and Fristedt (1985) for Bernoulli two-armed bandits to our setting that includes also more than two actions and a continuum of payoffs. (i) There exists a randomized rule that only has linear symmetric rules in its support and that attains minimax regret. (ii) There exists a worst case prior that is a probability measure over Bernoulli decision problems only. (iii) Any rule that attains minimax regret is optimal for a Bayesian facing a worst case prior. The theorem also provides a method to characterize minimax regret and worst case priors in terms of a Nash equilibrium of a zero-sum game where the player minimizes regret and nature maximizes. Berry and Fristedt (1985) prove their result by first establishing the existence of a Nash equilibrium and then invoking standard minimax results for zero-sum games. However, as they also point out, regret is no longer continuous when payoffs are allowed to be in $[0, 1]$ instead of only in $\{0, 1\}$ and consequently existence of Nash equilibria can no longer be easily established. We avoid this problem with a very simple trick and show that it is enough to use minimax regret rules obtained for the Bernoulli case if one extends these linearly to rules on $[0, 1]$.

Given the above, minimax regret can be considered a way to select among the rules that are Bayesian optimal under some prior. Unfortunately, a large part of the literature on Bayesian decision-making assumes that the arms are independent which need not be true.

In the rest of the paper we consider two arms only. First we provide some useful techniques for finding minimax behavior. Assume that there is a rule with finite memory that attains minimax regret with a symmetric worst case prior that has only two two-armed decision problems in its support. Then we prove that the worst case prior equals Q_0 where Q_0 is the symmetric prior that puts equal weight on the two deterministic two-action decisions in which one arm yields payoff 1 and the other payoff 0. So we only have to check Q_0 as a worst case prior if we are interested in simple rules and simple priors. In the proof we use existing results by Kakigi (1983) on Bayesian optimal rules under two-point distributions. Next we show that there exists a critical value $\delta_1 \approx 0.61$ such that Q_0 is not a worst case prior if $\delta > \delta_1$. This result applies for any minimax regret behavior and is proven using Taylor expansions near Q_0.

Next we investigate minimax regret behavior when δ is small. It is intuitive that Q_0 is the worst case prior when δ is sufficiently small as it maximizes the minimum regret in the first round. In fact, the following symmetric linear single round memory rule is Bayesian optimal against Q_0. This rule specifies to repeat the previous action with probability equal to the payoff attained. This rule has been proposed by Robbins (1952) as a very simple rule to use when $\delta = 1$ and payoff are in $\{0, 1\}$ who also derived its maximal regret for this case (see also Tsetlin, 1961). We find that this single round memory rule attains minimax regret with Q_0 as worst case prior if $\delta \leq \sqrt{2}-1 \approx 0.41$ but does not attain minimax regret for higher values of δ. Our result for small δ stands in contrast to the fact that Bayesian optimal rules do not necessarily have finite memory when δ is small. It is simply that the worst case prior does not require more memory.

We also investigate two round memory rules and find that the linearization of a rule proposed by Robbins (1956) (again for $\delta = 1$ and payoffs in $\{0,1\}$) attains minimax regret if and only $\delta \leq \delta_1$ where δ_1 is the same value obtained above. Q_0 is the worst case prior when $\delta \leq \delta_1$ and thus together with our previous results that we find that Q_0 is a worst case prior if and only if $\delta \leq \delta_1$. The selected two round memory rule has the *stay with a winner property* in the sense that the same action is played again whenever the highest payoff is obtained. It chooses the same arm again in the next two rounds whenever payoff 1 is obtained and otherwise switches arms each time 0 is obtained.

Finally we investigate for which values of δ between $\sqrt{2}-1$ and δ_1 memory of the payoff two rounds ago is not necessary to achieve minimax regret. We find that there is a cutoff $\delta_0 \approx 0.54$ such that this is only possible if $\delta < \delta_0$. The selected symmetric linear rule has the stay with a winner property, specifies to switch arms after payoff 0 unless the same arm is chosen twice in which case arms are switched with probability approximately equal to 0.16.

We proceed as follows. Section two introduces decision problems and strategies. Section three contains the necessary definitions. In Section four we supply the main characterization result of minimax regret behavior and worst case priors. In the final Section five we analyze separately rules that attain minimax regret among those with single round memory, two round memory and two round action memory.

2 Decision Problems and Rules

Let ΔY denote the set of probability measures over the set Y. A *multi-action decision problem* (W,P) consists of a finite set of actions W (with $|W| \geq 2$) and a measurable payoff distribution $P_i \in \Delta\mathbb{R}$ for each action $i \in W$. Sometimes we will index parameters by the decision problem D they refer to, e.g. write $P_i(D)$ instead of P_i. In the following fix W and consider only payoff distributions belonging to $\Delta[0,1]$.[1] The set of all multi-action decision problems will be denoted by $\mathcal{D}$. A *multi-armed bandit* is described by a finite set of actions W and by a prior (or probability measure) $Q \in \Delta\mathcal{D}$ over the set of multi-action decision problems with action set W. We add the term "Bernoulli" if realized payoffs only belong to $\{0,1\}$ where the payoffs 0 and 1 are referred to as *failure* and *success* respectively.[2] The set of all Bernoulli multi-action decision problems will be denoted by $\mathcal{D}_0$.

Let $\pi_i(D) = \int x dP_i(x,D)$ denote the expected payoff of choosing action i when facing the multi-action decision problem D. Given $D \in \mathcal{D}$ and a permutation ι of W let $D_\iota \in \mathcal{D}$ be the multi-action decision problem defined by permuting the labels of the actions in D using ι, i.e. $P_i(D_\iota) = P_{\iota(i)}(D)$ for $i \in W$. For a given multi-armed bandit $Q \in \Delta\mathcal{D}$ let Q_ι be the distribution defined by interchanging the labels of the actions in Q. A prior Q is called *symmetric* if $Q = Q_\iota$ holds for all permutations ι of W. The set of symmetric priors over a subset $\mathcal{Z}$ of $\mathcal{D}$ will be denoted by $\Delta_p\mathcal{Z}$. The symmetric prior Q will be called a *symmetric two point prior* if there exist $0 \leq v < w \leq 1$ such that $Q(v,w)\left(\tilde{D}\right) = \frac{1}{2}$ for $\tilde{D} \in \mathcal{D}_0$ with $\pi_1\left(\tilde{D}\right) = v$ and $\pi_2\left(\tilde{D}\right) = w$. We also write Q_0 instead of $Q(0,1)$.

Consider an individual who repeatedly faces the same multi-armed bandit Q. In each of a sequence of rounds the individual is asked to choose an action from W. Before the first round nature selects the multi-armed decision problem $\left(W, \tilde{P}\right)$ the individual is facing according to the prior Q. Choice of action i in round t yields a payoff realized according to $\tilde{P}_i$ that is drawn independently of previous choices and payoff realizations of the individual.

A rule is the formal description of how this individual makes his choice as a function of his previous experience. A *behavioral rule* is a mapping $f : \emptyset \cup_{m=1}^{\infty} \times_{k=1}^{m} \{W \times [0,1]\} \rightarrow W$ where $f(\emptyset)_i$ is the probability of choosing action i in the first round and $f(a_1, x_1, .., a_m, x_m)_i$ is the probability of choosing action i in round $m+1$ after choosing action a_k and receiving payoff x_k in round k for $k = 1, .., m$. A *deterministic rule* is a behavioral rule in which the choice in each round is always deterministic.. The set of all deterministic rules will be denoted by $\mathcal{F}$. A *(randomized) rule* ϕ is an element of $\Delta\mathcal{F}$.

[1] Our results can be applied to payoff distributions over a given bounded interval $[\alpha, \omega]$ by first rescaling payoffs using the linear transformation $x \longmapsto \frac{x-\alpha}{\omega-\alpha}$.

[2] The machine learning literature (cf Naremdra and Thathachar, 1989) refers to this situation as the P-model. In the Q-model and in the S-model the support of the payoff distribution is finite and infinite respectively.

We say that the behavioral rule f has n *round memory* if the behavior of f in round $m+1$ does not depend on a_k or x_k for $k \leq m-n$. f has *finite memory* if there exists n such that f has n round memory. f has n *round action memory* if f has n round memory where the behavior of f in round $m+1$ does not depend on x_k for $k \leq m-1$.

Given a behavioral rule f and a permutation ι of W let $\iota(f)$ be the behavioral rule that is derived from f by permuting actions with ι, i.e., $\iota(f)(\emptyset)_i = f(\emptyset)_{\iota(i)}$ and $\iota(f)(a_1, x_1, .., a_m, x_m)_i = f(\iota(a_1), x_1, .., \iota(a_m), x_m)_{\iota(i)}$. Then f is called *symmetric* if $\iota(f) \equiv f$ for all permutations ι of W. The set of all symmetric rules are denoted by $\mathcal{F}_p$. A randomized rule ϕ is called *symmetric* if $\phi(T) = \phi(\{\iota(f) \text{ s.t. } f \in T\})$ holds for all permutations ι of W and for all sets of deterministic rules T. The set of symmetric randomized rules will be denoted by $\Delta_p\mathcal{F}$. Notice that $\Delta\mathcal{F}_p \subsetneqq \Delta_p\mathcal{F}$.

A behavioral rule f is called *linear* if $f(a_1, x_1, .., a_m, x_m)_i$ is linear in x_k for all $k = 1, .., m$ and all m which means that

$$f(a_1, x_1, .., a_m, x_m)_i = \sum_{j_1=0}^{1} .. \sum_{j_m=0}^{1} \left[\Pi_{k=1}^{m} (j_k x_k + (1-j_k)(1-x_k))\right] f(a_1, j_1, .., a_m, j_m)_i$$

holds for all m and for all $a_i \in W$ and $x_i \in [0,1]$, $i = 1, .., m$. The set of all linear deterministic rules will be denoted by $\mathcal{F}^L$.

The behavioral rule f will be attributed the *'stay with a winner'* property if $f(a_1, x_1, .., a_m, 1)_{a_m} = 1$ for all a_k, x_k, $k = 1, .., m-1$, all a_m and all m.

3 Selection

Assume from now on that the individual is risk neutral and discounts future payoffs with discount factor $\delta \in (0,1)$.[3] For a given rule ϕ and a given decision problem D let $p_i^{(n)} = p_i^{(n)}(\phi, D)$ be the probability of choosing action $i \in W$ in round n unconditional on previous choices. Then $\pi^\delta(\phi, D) := (1-\delta)\sum_{n=1}^{\infty}\sum_{i \in W} p_i^{(n)}(\phi, D)\, \pi_i(D)$ is the discounted value of future payoffs.

The *regret* (or opportunity loss) of a rule ϕ when facing the multi-action decision problem D is defined as $L_\phi(D) := \max_{i \in W}\{\pi_i(D)\} - \pi^\delta(\phi, D)$. Regret is a measure of the loss due to ignorance of the true state of affairs.

[3] Our analysis also applies to agents that are not risk neutral by replacing payoffs with von Neumann-Morgenstern utilities as long as utility is bounded.

If the individual faces a known multi-armed bandit $\tilde{Q}$ then he chooses a rule $\phi^* \in \arg\max \int \pi_\phi^\delta (D)\, d\tilde{Q}(D)$. We call ϕ^* with this property *Bayesian optimal under* $\tilde{Q}$. We will call $\bar{Q}$ a *worst case prior* if it maximizes the expected regret of this individual, i.e. if $\bar{Q} \in \max_{Q \in \Delta\mathcal{D}} \int L_{\phi^*(Q)}(D)\, dQ(D)$ where $\phi^*(Q)$ is a Bayesian optimal rule under Q. Simplifying we obtain that $\bar{Q}$ is a worst case prior if and only if $\bar{Q} \in \max_{Q \in \Delta\mathcal{D}} \min_{\phi \in \Delta\mathcal{F}} \int L_\phi(D)\, dQ(D)$.

If W is known but the prior $\tilde{Q}$ is unknown then according to Savage (1972) the individual specifies a subjective prior $\hat{Q}$ and chooses a Bayesian optimal rule under $\hat{Q}$. An alternative approach is to assume that the individual selects a rule that minimizes among all rules the maximum among all decision problems (with action set W) the expected regret. More specifically, we say that ϕ *attains minimax regret* if $\phi^* \in \arg\min_{\phi \in \Delta\mathcal{F}} \sup_{D \in \mathcal{D}} L_\phi(D)$.

4 General Results

We present our central theorem on the characterization of minimax regret behavior and worst case priors.

Proposition 1 *i) There exists a worst case prior in $\Delta_p\mathcal{D}_0$ and a rule in $\Delta\mathcal{F}_p^L$ that attains minimax regret. The value of minimax regret is strictly positive.*

ii) $\phi^* \in \Delta\mathcal{F}_p^L$ attains minimax regret and $Q^* \in \Delta_p\mathcal{D}_0$ is a worst case prior if and only if

$$\int L_{\phi^*}(D)\, dQ(D) \leq \int L_{\phi^*}(D)\, dQ^*(D) \leq \int L_\phi(D)\, dQ^*(D) \;\; \forall \phi \in \Delta\mathcal{F}_p^L \; \forall Q \in \Delta_p\mathcal{D}_0.$$

(iii) $\phi^* \in \Delta\mathcal{F}$ attains minimax regret and $Q^* \in \Delta\mathcal{D}$ is a worst case prior if and only if

$$\int L_{\phi^*}(D)\, dQ(D) \leq \int L_{\phi^*}(D)\, dQ^*(D) \leq \int L_\phi(D)\, dQ^*(D) \;\; \forall \phi \in \Delta\mathcal{F} \; \forall Q \in \Delta\mathcal{D}. \qquad (1)$$

In particular, this means that any rule that attains minimax regret is Bayesian optimal under any worst case prior. The above generalizes findings that Berry and Fristedt (1985) obtained already for Bernoulli two-armed bandits. The only additional insight for this special case is that we prove the existence of a randomized symmetric rule that attains minimax regret instead of simply existence of a symmetric randomized rule that possibly includes non symmetric rules in its support.

Proof. We first review the results of Berry and Fristedt (1985) obtained for Bernoulli two-armed bandits which are statement (i) and the 'if' statements of (ii) and (iii) while replacing $\Delta\mathcal{F}_p^L$ by $\Delta_p\mathcal{F}^L$. They introduce a topology on the set of strategies and then show that a Nash equilibrium (ϕ^*, Q^*) exists in the zero sum game where the individual chooses a rule to minimize regret and nature chooses a prior to maximize regret. If (ϕ^*, Q^*) is a such a Nash equilibrium (i.e. (1) holds when restricted to the case of $|W| = 2$ and $Q \in \Delta\mathcal{D}_0$) then

$$\int L_{\phi^*}(D)\, dQ^*(D) = \max_{Q\in\Delta\mathcal{D}_0} \int L_{\phi^*}(D)\, dQ(D) \geq \min_{\phi\in\Delta\mathcal{F}} \max_{Q\in\Delta\mathcal{D}_0} \int L_{\phi}(D)\, dQ(D)$$
$$\geq \max_{Q\in\Delta\mathcal{D}_0} \min_{\phi} \int L_{\phi}(D)\, dQ(D) \geq \min_{\phi\in\Delta\mathcal{F}} \int L_{\phi}(D)\, dQ^*(D) = \int L_{\phi^*}(D)\, dQ^*(D)$$

which proves the 'if' statement of (iii) for Bernoulli two-armed bandits. Berry and Fristedt (1985) also ensure the existence of a strictly positive lower bound on the value of minimax regret so this completes (i) for Bernoulli two-armed bandits. Quasi-concavity of $\max_{Q\in\Delta\mathcal{D}_0} \int L_{\phi}(D)\, dQ(D)$ as a function of ϕ shows that $\Delta_p\mathcal{F} \cap \arg\min_{\phi\in\Delta\mathcal{F}} \max_{Q\in\Delta\mathcal{D}_0} \int L_{\phi}(D)\, dQ(D) \neq \emptyset$. Similarly, quasi-convexity of $\min_{\phi\in\Delta\mathcal{F}} \int L_{\phi}(D)\, dQ(D)$ as a function of Q is used to show that $\Delta_p\mathcal{D}_0 \cap \arg\max_{Q\in\Delta\mathcal{D}_0} \min_{\phi\in\Delta\mathcal{F}} \int L_{\phi}(D)\, dQ(D) \neq \emptyset$. Finally, the 'if' statement of (ii) follows from the fact that $\Delta_p\mathcal{D}_0 \cap \arg\max_{Q\in\Delta\mathcal{D}_0} \int L_{\phi^*}(D)\, dQ(D) \neq \emptyset$ if $\phi^* \in \Delta_p\mathcal{F}$ and similarly, $\Delta_p\mathcal{F} \cap \arg\min_{\phi\in\Delta\mathcal{F}} \int L_{\phi}(D)\, dQ^*(D) \neq \emptyset$ if $Q^* \in \Delta_p\mathcal{D}_0$.

The above can be generalized to Bernoulli multi-armed bandits immediately. In the following we will show that it also holds when payoffs are not restricted to $\{0, 1\}$. Let $(\phi^*, Q^*) \in \Delta\mathcal{F}^L \times \Delta\mathcal{D}_0$ be a Nash equilibrium (that exists) of the zero-sum game when restricting attention to $\mathcal{D}_0$. Since ϕ randomizes over linear rules, $\max_{Q\in\Delta\mathcal{D}_0} \int L_{\phi^*}(D)\, dQ(D) = \max_{Q\in\Delta\mathcal{D}} \int L_{\phi^*}(D)\, dQ(D)$ and $Q^* \in \Delta\mathcal{D}_0$ implies that $\min_{\phi\in\Delta\mathcal{F}^L} \int L_{\phi}(D)\, dQ^*(D) = \min_{\phi\in\Delta\mathcal{F}} \int L_{\phi}(D)\, dQ^*(D)$ and hence (1) holds. Notice furthermore that the "if statement" of (iii) holds as stated by the same proof as when we considered only $\mathcal{D}_0$. Part (i) and the "if statement" of (ii) then also follow as above.

Consider now the 'only if' statements of (ii) and (iii). If ϕ^* attains minimax regret and Q^* is a worst case prior then

$$\int L_{\phi^*}(D)\, dQ^*(D) \leq \sup_{Q\in\Delta\mathcal{D}} \int L_{\phi^*}(D)\, dQ(D) = \min_{\phi\in\Delta\mathcal{F}} \sup_{Q\in\Delta\mathcal{D}} \int L_{\phi}(D)\, dQ(D)$$
$$\max_{Q\in\Delta\mathcal{D}} \inf_{\phi\in\Delta\mathcal{F}} \int L_{\phi}(D)\, dQ(D) = \inf_{\phi\in\Delta\mathcal{F}} \int L_{\phi}(D)\, dQ^*(D) \leq \int L_{\phi^*}(D)\, dQ^*(D)$$

so the claim follows as we know that $\min_{\phi\in\Delta\mathcal{F}}\sup_{Q\in\Delta\mathcal{D}_0}\int L_\phi(D)\,dQ(D)=\max_{Q\in\Delta\mathcal{D}_0}\inf_{\phi\in\Delta\mathcal{F}}$ $\int L_\phi(D)\,dQ(D)$ holds.

Finally we need to show that $\Delta_p\mathcal{F}^L$ can be replaced by $\Delta\mathcal{F}_p^L$ in the statements above. Assume that $\phi^*\in\Delta_p\mathcal{F}^L$ attains minimax regret. Let $\phi^+\in\Delta\mathcal{F}_p^L$ be such that $\phi^+(a,..)\equiv\phi^*(a,..)$. Then $\int L_{\phi^*}(D)\,dQ(D)=\int L_{\phi^+}(D)\,dQ(D)$ holds for all $Q\in\Delta_p\mathcal{D}_0$ which proves the statement. ■

The above undermines the usefulness of linearity for attaining minimax regret. Notice that linear rules typically involve randomizing after round one, for instance whenever only interior payoffs in $(0,1)$ are obtained in all previous rounds then behavior is either random or all payoffs are "ignored". More specifically, if f is a linear rule and $x_k\in(0,1)$ for all n then either $f(a_1,x_1,..,a_n,x_n)_c$ is independent of $x_1,..,x_n$ or $f(a_1,x_1,..,a_n,x_n,..,a_m,x_m)\notin W$. In contrast we now show that Bayesian optimal rules typically do not involve randomizing.

Proposition 2 *For almost all symmetric priors there is some payoff $z\in(0,1)$ that can occur in any round where no Bayesian optimal rule will randomize after receiving z.*

Proof. Consider a symmetric prior $Q\in\Delta_p\mathcal{D}$ such that there exists a payoff $z\in(0,1)$ that can occur for any D drawn under Q and that reveals that the current arm is best, i.e. $P(\pi_c(D)>\pi_d(D)\,|\,\text{arm } c \text{ yields } z, D \text{ unknown but drawn using prior } Q)=1$, $c\neq d$. Notice that the set of such priors lies dense in $\Delta_p\mathcal{D}$. Consider any $f\in\arg\min_{f\in\mathcal{F}}\int L_f(D)\,dQ(D)$ and any history $(a_1,x_1,..,a_{m-1},x_{m-1})$ that can arise under f for some D drawn under Q. Then $f(a_1,x_1,..,a_m,z)_{a_m}=1$. ■

4.0.1 Necessary Conditions in Two-Armed Bandits

Given Proposition 1 minimax regret can be considered a method to select among rules that are Bayesian optimal under some prior. Unfortunately we cannot utilize the bulk of the two-armed bandit that concerns only independent arms as we do not expect that worst case prior have this property. We start by investigating when a finite memory rule can attain minimax regret under a symmetric prior with a two point distribution.

Proposition 3 *Consider the case of two-armed bandits. If the linear symmetric n round memory rule ϕ^* attains minimax regret and $\arg\max_{D\in\mathcal{D}_0:\pi_a(D)>\pi_b(D)}L_{\phi^*}(D)$ is single valued then*

Q_0 is a worst case prior.

Proof. Let $\{D_0\} = \arg\max_{D \in \mathcal{D}_0 : \pi_a(D) > \pi_b(D)} L_{\phi^*}(D)$. Since ϕ^* is assumed to attain minimax regret, $\pi_b(D_0) < \pi_a(D_0)$. Hence, the symmetric two point prior $Q(\pi_b(D_0), \pi_a(D_0))$ is a worst case prior.

We now analyze optimal behavior under a symmetric two point prior $Q(v, w)$.

Kakigi (1983) shows that the following symmetric rule is optimal in such a symmetric two point prior. Choose action c in round n if the beliefs based on experience up to round n indicate that the probability that $\pi_c > \pi_d$ is greater than 0.5 where $\{c, d\} = \{a, b\}$.

Samaranayake (1992) shows that the two arms are negatively correlated after any history. As the support of each marginal distribution has two elements we can apply Proposition 5.2 in (Samaranayake, 1992) to show that the individual strictly prefers action c over d after c yielded a success and strictly prefers action d over c after c yielded a failure. This means that the rule of Kakigi (1983) is the unique optimal behavior whenever the updated prior does not equal 0.5.

It is clear that the optimal behavior does not have finite round memory when $0 < v < w < 1$.

Assume $v = 0$ and $w \in (0, 1)$. Then a Bayesian optimal behavior is given by the symmetric two round memory rule f^* that has the staying with a winner property and where $f^*(c, 0)_c = f^*(c, 0, d, 0)_d = 0$ and $f^*(c, *, c, 0)_c = 1$ for $c \neq d$. Let z be the future value after only failure obtained previously then $z = (1 - \delta)\frac{1}{2}w + \frac{1}{2}w\delta w + \left(1 - \frac{1}{2}w\right)\delta z$ so $z = w\frac{1-\delta+w\delta}{2-2\delta+w\delta}$ and hence $L_{f^*} = w - z = \frac{(1-\delta)w}{2-2\delta+w\delta}$ which for given δ obtains its unique maximum when $w = 1$. Thus, $Q(0, w)$ is never a worst case prior if $w < 1$.

Finally, assume $v > 0$ and $w = 1$. Here a Bayesian optimal behavior is given by the symmetric single round memory rule f^* that has the staying with a winner property and where $f^*(c, 0)_c = 0$. Let x be the future value of payoffs after only achieving successes in the previous rounds with the worse arm. Then $x = (1 - \delta)v + v\delta x + (1 - v)\delta$ so $x = \frac{\delta+v-2\delta v}{1-\delta v}$ and hence $L_{f^*} = 1 - \frac{1}{2} - \frac{1}{2}\frac{\delta+v-2\delta v}{1-\delta v} = \frac{1}{2}\frac{(1-\delta)(1-v)}{1-\delta v}$ which for given δ obtains its unique maximum when $v = 0$. Thus, $Q(v, 1)$ is never a worst case prior if $v > 0$. ■

Proposition 4 *Consider two-armed bandits. Then Q_0 is not a worst case prior for $\delta > -\frac{1}{2} + \frac{1}{2}\sqrt{5}$.*

Proof. Consider a symmetric rule f^* that attains minimax regret with Q_0 as worst case prior. Then $\frac{d}{d\pi_a}L(1,0) \geq 0 \geq \frac{d}{d\pi_b}L(1,0)$. In the following we will consider only regret in the two two-armed Bernoulli decisions with $(\pi_a, \pi_b) \in \{(1,v), (w,0)\}$ for $v, w \in (0,1)$. Since f^* is symmetric, $f^*(\emptyset) = \frac{1}{2}$. Since Q_0 is a worst case prior we have $f^*(c,x)_c = x$ for $c \in \{a,b\}$ and $x \in \{0,1\}$.

When facing $(\pi_a, \pi_b) = (1,v)$ then we will ignore events when arm b yields two successes. Similarly, when facing $(\pi_a, \pi_b) = (w,0)$ then we will ignore events when arm a yields two failures. Given these restrictions we can assume that f^* plays a best response whenever possible against these two particular decision problems. So once we observe two successes of the same arm then we are choosing the better arm and lock in on that action. In particular this means that when facing $(\pi_a, \pi_b) = (1,v)$, if we choose the better action in the first round then this action is chosen forever. Similarly, we know that arm a is the better arm after observing $(b,0)$ and $(a,1)$.

Let $x = f^*(c,1,c,0)_d$ and $y = f^*(d,0,c,0)_d$ for $c \neq d$. Consider payoffs when facing $(\pi_a, \pi_b) = (w,0)$. Then

$$\begin{aligned}
\pi^\delta &= (1-\delta)\frac{1}{2}w + (1-\delta)\,\delta\frac{1}{2}(1+w)\,w + \frac{1}{2}w\delta^2 w + \frac{1}{2}w^2\delta^2 w \\
&\quad + (1-\delta)\,\delta^2\left(\frac{1}{2}w(1-w)\,x + \frac{1}{2}(1-w)\,y + \frac{1}{2}(1-w)(1-y)\right)w \\
&\quad + \delta^3\left(\frac{1}{2}(1-x)\,w^2(1-w)\,w + \frac{1}{2}w(1-w)\,xw\right) \\
&\quad + \frac{1}{2}(1-w)\,\delta^3 w\left((1-y) + yw + (1-y)\,w + y\right) + O\left((1-w)^2\right)
\end{aligned}$$

where the expressions refer to the payoffs in round one and two, continuation payoff starting round three after the events $(b,0,a,1)$ and $(a,1,a,1)$, round three payoffs after $(a,1,a,0)$, $(a,0,b,0)$ and $(b,0,a,0)$ and continuation payoffs starting round four after $(a,1,a,0,a,1)$, $(a,1,a,0,b,0)$ and after $(a,0,b,0,b,0)$, $(a,0,b,0,a,1)$, $(b,0,a,0,a,1)$ and $(b,0,a,0,b,0)$. Here we assume that $f^*(a,0,b,0,b,0)_b = f^*(b,0,a,0,b,0)_b = 0$ following the Bayesian optimal behavior against $Q(0,w)$. Consequently,

$$L_{f^*} = \frac{1}{2}(1-\delta) - \frac{1}{2}(1-\delta)\left(1-\delta-\delta^2-x\delta^2\right)(1-w) + O\left((1-w)^2\right)$$

and hence $1-\delta-\delta^2-x\delta^2 \geq 0$ is necessary for Q_0 to be a worst case prior. However, $1-\delta-\delta^2 < 0$ for $\delta > -\frac{1}{2}+\frac{1}{2}\sqrt{5}$ so for values of δ with this property there is no value of x under which Q_0 is a worst case prior. ■

We combine the above to obtain the following.

Corollary 5 *Consider two armed bandits and assume $\delta > -\frac{1}{2}+\frac{1}{2}\sqrt{5}$. Then either there is no n round memory rule that attains minimax regret or the worst case prior* $\arg\max_{D\in\mathcal{D}_0:\pi_a(D)>\pi_b(D)} L_{\phi^*}(D)$ *is not single valued for any ϕ^* that attains minimax regret.*

Next we derive behavior in the first three rounds for a symmetric rule that attains minimax regret at the critical discount factor.

Proposition 6 *Assume that f^* is a symmetric rule that attains minimax regret when $\delta = -\frac{1}{2}+\frac{1}{2}\sqrt{5}$. Then $f^*(c,0)_c = 0$, $f^*(c,1,c,0)_c = 1$, $f^*(c,0,d,1)_d = 1$, $f^*(c,1,c,1,c,0)_c = 1$, $f^*(c,1,c,0,c,0)_c = 0$, $f^*(c,0,d,1,d,0)_d = 1$, $f^*(c,0,d,0,c,0)_c = 0$ if $f^*(c,0,d,0)_c > 0$, $f^*(c,0,d,0,d,0)_d = 0$ if $f^*(c,0,d,0)_d > 0$ and f^* does not switch after any success in the first three rounds. In particular, f^* does not have n round action memory for any n.*

In particular, rules suggested by Robbins (1956) or Isbell (1959) for $n > 2$ do not attain minimax regret when $\delta = -\frac{1}{2}+\frac{1}{2}\sqrt{5}$.

Proof. First we provide the analogous calculations as in the proof of Proposition 4 when facing $(\pi_a,\pi_b) = (1,v)$. We calculate π^δ where we do not explicitly calculate events where two successes of the worse arm occur. Then

$$\pi^\delta = \frac{1}{2}+\frac{1}{2}(1-\delta)v+\frac{1}{2}(1-v)\delta+\frac{1}{2}v(1-v)x\delta^2+\frac{1}{2}v(1-v)(1-x)(1-v)\delta^3+O\left(v^2\right)$$

where the expressions refer to the event $(a,1,a,1,...)$, the payoff in round one from choosing arm b and the events $(b,0,a,1,a,1,...)$, $(b,1,b,0,a,1,a,1,...)$ and $(b,1,b,0,b,0,a,1,a,1,...)$. Consequently

$$L_{f^*} = \frac{1}{2}(1-\delta)-\frac{1}{2}(1-\delta)\left(1-\delta-(1-x)\delta^2\right)v+O\left(v^2\right)$$

and hence $1-\delta-\delta^2+x\delta^2 \geq 0$ is necessary if Q_0 is a worst case prior.

Looking a bit more carefully at the above calculations as well as those in the proof of Proposition 4 it is easily verified that Q_0 is not a worst case prior if one of the conditions in the statement of the proposition do not hold. ■

5 Rules attaining minimax regret in two-armed bandits

In the following we will consider two arms only and search for rules that attain minimax regret when δ is small. Let $W = \{a, b\}$. Given our above result we search for Nash equilibria of the zero-sum game. If ϕ is symmetric then each action is played with probability 0.5 in the first round and we obtain $L_\phi(D) = (1-\delta)\,0.5\,|\pi_a - \pi_b| + (1-\delta)\,o(\delta)$. This gives us intuition that the decision problem that maximizes regret satisfies $\{\pi_a, \pi_b\} = \{0, 1\}$. We do not try to prove this directly but use this intuition to motivate our search for situations where the worst case prior puts equal weight on the two two-armed decision problems in which one arm yields payoff 1 and the other arm yields payoff 0. Let Q_0 denote this prior. Bayesian optimal rules with finite memory for facing Q_0 are easily computed. All we then check is whether Q_0 maximizes regret of such a Bayesian optimal rule.

5.1 Single round memory

Proposition 7 *The linear symmetric single round memory rule that has the stay with a winner property and that satisfies $f(a,0)_a = 0$ attains minimax regret if and only if $\delta \le \sqrt{2} - 1$. This rule yields*

$$\pi^\delta = \frac{1}{2}(\pi_a + \pi_b) + \frac{1}{2}\delta \frac{1}{1 + \delta(1 - \pi_a - \pi_b)}(\pi_a - \pi_b)^2 \quad .$$

No other single round memory rule attains minimax regret when $\delta = \sqrt{2} - 1 \approx 0.41$.

Notice that Bayesian optimal rules generally do not have finite memory even when δ is small. For instance, as argued in the proof of Proposition 3, any Bayesian optimal rule under the two point distribution $Q(v, w)$ with $0 < v < w < 1$ does not have finite round memory.

Proof. Let D_c be the two-action decision problem with $P_c(1) = P_d(0) = 1$ for $d \in W \setminus \{c\}$ and let Q_0 be the prior such that $Q_0(D_c) = 0.5$ for $c \in W$. Then it follows immediately that the single round memory rule described above is the unique symmetric linear Bayesian optimal rule under Q_0.

In the following we show how to derive the above expression for π^δ. Let z_c be the discounted future value of payoffs conditional on choosing action c. Then $z_a = (1-\delta)\pi_a + \delta\pi_a z_a + (1-\pi_a)\delta z_b$. Similar expression for z_b and solving yields the above.

Finally, given $\pi_a > \pi_b$ we obtain

$$\frac{d}{d\pi_a}L = \frac{1}{2}\frac{1+2\delta-4\delta\pi_a+\delta^2-4\delta^2\pi_a+2\delta^2\pi_a^2+4\delta^2\pi_a\pi_b-2\delta^2\pi_b^2}{\left(1+\delta-\delta\pi_a-\delta\pi_b\right)^2}$$

where the enumerator is decreasing in π_a. If $\pi_a = 1$ then the enumerator is also increasing in π_b. So evaluating the enumerator at $\pi_a = 1$ and $\pi_b = 0$ we obtain $1-2\delta-\delta^2$ which has the positive root $-1+\sqrt{2}$. Hence $\frac{d}{d\pi_a}L \geq 0$ holds for all π_a and π_b if $\delta \leq -1+\sqrt{2}$. On the other hand, if $\delta > -1+\sqrt{2}$ then $\frac{d}{d\pi_a}L < 0$ when $\pi_a = 1$ and $\pi_b = 0$.

Similarly for $\pi_a > \pi_b$ we obtain

$$\frac{d}{d\pi_b}L = -\frac{1}{2}\frac{\left(1+\delta-2\delta\pi_a\right)^2}{\left(1+\delta-\delta\pi_a-\delta\pi_b\right)^2}$$

which shows that Q_0 is a worst case prior as long as $\delta \leq -1+\sqrt{2}$.

Finally, it can be verified for the selected rule, denote this by ϕ^*, that $\arg\max_{D\in\mathcal{D}_0:\pi_a(D)>\pi_b(D)} L_{\phi^*}(D)$ is single valued for all $\delta \in (0,1)$. Thus, by Corollary 5 ϕ^* does not attain minimax regret for $\delta > -1+\sqrt{2}$. As ϕ^* is the unique symmetric linear single round memory rule that is Bayesian optimal against Q_0 the statement is proven. ■

5.2 Two round memory

We find that the linearization of the rule suggested by Robbins (1956) for use in Bernoulli two-armed decisions when $\delta = 1$ attains minimax regret when δ is not too large. When payoffs are in $\{0,1\}$ this rule prescribes to switch back and forth until the first success is obtained and then to only switch after two consecutive failures.

Proposition 8 *Consider the linear symmetric two round memory rule that has the stay with a winner property and that satisfies* $f(*,0,c,0)_c = 0$ *and* $f(c,1,c,0)_c = 1$ *which yields*

$$\pi^{\delta,1} = \frac{1}{2}\left(\pi_1+\pi_2\right)+\frac{1}{2}\delta\frac{\left(\pi_1-\pi_2\right)^2\left(1+\delta-\delta\left(\pi_1+\pi_2\right)\right)}{\delta^2\left(1-\pi_1\right)^2+\delta^2\left(1-\pi_2\right)^2+\left(1-\delta\right)\left(1+\delta\left(2-\pi_1-\pi_2\right)\right)}.$$

This rule attains minimax regret if and only if $\delta \leq -\frac{1}{2}+\frac{1}{2}\sqrt{5}$. *No other two round memory rule attains minimax regret when* $\delta = -\frac{1}{2}+\frac{1}{2}\sqrt{5} \approx 0.62$.

The rule selected above behaves in Bernoulli two-armed decisions like the one suggested by Robbins (1956) for use when $\delta = 1$ with the only alteration that the decision maker randomizes in the first round.

Proof. Consider a two round memory rule with the stay with a winner property that is Bayesian optimal against Q_0. Then $f(\emptyset)_a = 0.5$, $f(c,0)_c = 0$ and $f(c,1)_c = f(*,*,c,1)_c = 1$. Let $z = f(c,0,c,0)_d$, $x = f(c,1,c,0)_d$ and $y = f(d,0,c,0)_d$ for $c \neq d$.

Starting round two the rule has six states $a0a$, $a1a$, $a0b$, $b0b$, $b1b$ and $b0a$ where cxd denotes the state in which the present action is d and the previous action was c which yielded $x \in \{0,1\}$. Then the probability of being in these states in round 2 equals 0, $\frac{1}{2}\pi_a$, $\frac{1}{2}(1-\pi_a)$, 0, $\frac{1}{2}\pi_b$ and $\frac{1}{2}(1-\pi_b)$ respectively. Given the transition matrix M equal to

$$\begin{matrix}
(1-\pi_a)(1-z) & (1-\pi_a)(1-x) & 0 & 0 & 0 & (1-\pi_a)(1-y) \\
\pi_a & \pi_a & 0 & 0 & 0 & \pi_a \\
(1-\pi_a)z & (1-\pi_a)x & 0 & 0 & 0 & (1-\pi_a)y \\
0 & 0 & (1-\pi_b)(1-y) & (1-\pi_b)(1-z) & (1-\pi_b)(1-x) & 0 \\
0 & 0 & \pi_b & \pi_b & \pi_b & 0 \\
0 & 0 & (1-\pi_b)y & (1-\pi_b)z & (1-\pi_b)x & 0
\end{matrix}$$

we obtain for $\pi_a > \pi_b$ that

$$L(\pi_a,\pi_b) = \pi_a - \frac{1}{2}(1-\delta)(\pi_a+\pi_b) - (1-\delta)\delta \begin{pmatrix} \pi_a & \pi_a & \pi_b & \pi_b & \pi_b & \pi_a \end{pmatrix} (Id - \delta M)^{-1} \xi$$

where ξ is the vector of probabilities in round two.

It can be verified that

$$\begin{aligned}
\frac{d}{d\pi_a} L(1,0) &= \frac{1}{2}\frac{1-(3-z)\delta + (-x+2-2z)\delta^2 + \left(-zy-x-xz+xy+y^2+z\right)\delta^3 + (z-y)(y-x)\delta^4}{1-\delta+\delta z} \\
\frac{d}{d\pi_b} L(1,0) &= -\frac{1}{2}\frac{(1-\delta)\left(1-2\delta+\delta z+(x-z)\delta^2\right)}{1-\delta+\delta z}
\end{aligned}$$

We search for maximal δ such that $\frac{d}{d\pi_a}L(1,0) \geq 0 \geq \frac{d}{d\pi_b}L(1,0)$. From the second inequality we obtain $x \geq z - (1-2\delta+\delta z)/\delta^2$. Since $\frac{d}{d\pi_a}L(1,0)$ is decreasing in x we replace x by $z - (1-2\delta+\delta z)/\delta^2$ in the enumerator of $\frac{d}{d\pi_a}L(1,0)$ to obtain that

$$\left(2-4\delta+\left(-2\delta^3-3\delta^2+3\delta\right)z+\left(-\delta^4+\delta^2\right)z^2\right)+\delta\left(2\delta^3 z-\delta^2 z+2\delta^2+\delta-\delta z-1\right)y+\delta^3(1-\delta)y^2 \geq 0$$

It is easily verified that the left had side of the above inequality is convex in z and y and hence there are four cases $z, y \in \{0,1\}$ to check. It is then directly verified that the inequality is violated for $\delta < 0.6$ when $z = 0$ while it holds for $\delta \leq -\frac{1}{2} + \frac{1}{2}\sqrt{5}$ when $z = 1$ and either $y = 0$ or $y = 1$.

Consider the rule with $y = 1$, $z = 1$ and $x = 0$. This yields

$$L = \frac{1}{2} \frac{(\pi_a - \pi_b)\left(1 + \delta - 2\delta\pi_a - 2\delta^2\pi_a + 2\delta^2\pi_a^2\right)}{1 + \delta - \delta\pi_a - \delta\pi_b - \delta^2\pi_a - \delta^2\pi_b + \delta^2\pi_a^2 + \delta^2\pi_b^2}.$$

Assume $\delta \leq -\frac{1}{2} + \frac{1}{2}\sqrt{5}$. By first showing that $\frac{d}{d\pi_b}L \leq 0$ and then that $\frac{d}{d\pi_a}L \geq 0$ holds when $\pi_b = 0$ it can easily be verified that $(\pi_a, \pi_b) = (1, 0)$ is the unique maximizer of L conditional on $\pi_a > \pi_b$.

The alternative rule with $y = 0$, $z = 1$ and $x = 0$ we obtain for $\pi_2 = 0$

$$L = \frac{1}{2}\pi_1 \frac{\delta + \delta^3 + 1 + 2\delta^2\pi_1^2 - 2\delta\pi_1 - 3\delta^2\pi_1 - 3\delta^3\pi_1 + \delta^2 + 2\delta^3\pi_1^2}{1 + \delta - 2\delta^2\pi_1 - 2\delta^3\pi_1 + \delta^2\pi_1^2 - \delta\pi_1 + \delta^3\pi_1^2 + \delta^3 + \delta^2}$$

and $\frac{d^2}{(d\pi_1)^2}L = \delta(2\delta - 1)(1 + \delta)^2$ if $\pi_1 = 1$ so this rule does not attain minimax regret if $\delta > \frac{1}{2}$.

Finally, it can be verified for the selected rule, denote this by ϕ^*, that $\arg\max_{D \in \mathcal{D}_0 : \pi_a(D) > \pi_b(D)} L_{\phi^*}(D)$ is single valued for all $\delta \in (0, 1)$. Thus, by Corollary 5 ϕ^* does not attain minimax regret for $\delta > -\frac{1}{2} + \frac{1}{2}\sqrt{5}$. ■

Combining the above result with Proposition 4 we obtain:

Corollary 9 *Q_0 is a worst case prior if and only if $\delta \leq -\frac{1}{2} + \frac{1}{2}\sqrt{5}$.*

5.3 Two round action memory

Proposition 10 *There exists δ_0 with $\delta_0 \approx 0.544$ such that:*

(i) If $\delta \leq \delta_0$ then the linear symmetric two round action memory rule that has the stay with a winner property and that satisfies $f(c, \cdot, c, 0)_c = \frac{1-\delta_0}{\delta_0} \approx 0.84$ and $f(c, \cdot, d, 0)_d = 0$ for $c \neq d$ attains minimax regret.

(ii) If $\delta_0 < \delta \leq -\frac{1}{2} + \frac{1}{2}\sqrt{5}$ then there is no two round action memory rule that attains minimax regret.

Proof. Consider a two round action memory rule with the stay with a winner property that is Bayesian optimal against Q_0. Then $f(\emptyset)_a = 0.5$, $f(c, 0)_c = 0$ and $f(c, 1)_c = f(d, x, c, 1)_c = 1$. Let $\lambda = f(c, *, c, 0)_d$ and $\mu = f(c, *, d, 0)_c$ for $c \neq d$.

Starting round two the rule has four states *aa*, *ab*, *bb*, *ba* where *cd* denotes the state in which the present action is d and the previous action was c. Let v_n, w_n, y_n and z_n be the probabilities of being in these states in rounds $n \geq 2$. Then $v_2 = \frac{1}{2}\pi_a$, $w_2 = \frac{1}{2}(1-\pi_a)$, $y_2 = \frac{1}{2}\pi_b$ and $z_2 = \frac{1}{2}(1-\pi_b)$. Given the transition matrix M equal to

$$\begin{matrix} \pi_a + (1-\pi_a)(1-\lambda) & 0 & 0 & \pi_a + (1-\pi_a)(1-\mu) \\ (1-\pi_a)\lambda & 0 & 0 & (1-\pi_a)\mu \\ 0 & \pi_b + (1-\pi_b)(1-\mu) & \pi_b + (1-\pi_b)(1-\lambda) & 0 \\ 0 & (1-\pi_b)\mu & (1-\pi_b)\lambda & 0 \end{matrix}$$

we obtain $\begin{pmatrix} v_{n+1} & w_{n+1} & y_{n+1} & z_{n+1} \end{pmatrix}^T = M \begin{pmatrix} v_n & w_n & y_n & z_n \end{pmatrix}^T$ and hence

$$\begin{aligned} L &= \max\{\pi_a, \pi_b\} \\ &\quad -\frac{1}{2}(1-\delta)(\pi_a+\pi_b) - (1-\delta)\delta \begin{pmatrix} \pi_a & \pi_b & \pi_b & \pi_a \end{pmatrix} (Id - \delta M)^{-1} \begin{pmatrix} v_2 & w_2 & y_2 & z_2 \end{pmatrix}^T \end{aligned}$$

where $Id \in \mathbb{R}^{4,4}$ is the identity matrix.

The explicit expression for L is too elaborate to present here but it is easily verified for $\pi_a > \pi_b$ that

$$\begin{aligned} \frac{d}{d\pi_a} L|_{(\pi_a,\pi_b)=(1,0)} &= \frac{1}{2}\frac{\left(1-3\delta+\delta\lambda+2\delta^2-3\delta^2\lambda-\delta^3\lambda^2-\delta^4\lambda^2\right)+2\delta^4\lambda\mu+\left(\delta^3-\delta^4\right)\mu^2}{1-\delta+\delta\lambda} \\ \frac{d}{d\pi_b} L|_{(\pi_a,\pi_b)=(1,0)} &= -\frac{1}{2}\frac{(1-\delta)(1-2\delta+\delta\lambda)}{1-\delta+\delta\lambda} \end{aligned}$$

In the following we search values of λ and μ that maximize the largest value of δ such that $\frac{d}{d\pi_a} L|_{(\pi_a,\pi_b)=(1,0)} \geq 0$ and $\frac{d}{d\pi_b} L|_{(\pi_a,\pi_b)=(1,0)} \leq 0$ holds. Let λ_0, μ_0 and δ_0 be the solutions to this problem. It follows that $\mu_0 = 1$ which yields $\frac{d}{d\pi_a} L|_{(\pi_a,\pi_b)=(1,0)} = \frac{1}{2}\left(-\delta^3\lambda+\delta^3-\delta^2\lambda-2\delta+1\right)$. So we are looking for λ_0 and δ_0 such that $1-2\delta_0+\delta_0\lambda_0 = 0$ and $-\delta_0^3\lambda+\delta_0^3-\delta_0^2\lambda_0-2\delta_0+1=0$. Solving these two equations yields $\lambda_0 = \frac{2\delta_0-1}{\delta_0}$ and

$$\delta_0 = \sqrt[3]{\left(\frac{17}{27}+\frac{1}{9}\sqrt{33}\right)} - \frac{2}{9\sqrt[3]{\left(\frac{17}{27}+\frac{1}{9}\sqrt{33}\right)}} - \frac{1}{3} \approx 0.54369.$$

Thus, for $\delta > \delta_0$ either $\frac{d}{d\pi_a} L|_{(\pi_a,\pi_b)=(1,0)} < 0$ or $\frac{d}{d\pi_b} L|_{(\pi_a,\pi_b)=(1,0)} > 0$ which means for $\delta > \delta_0$ that either Q_0 is not a worst case prior. Combining this with Proposition 8 we have proven part (ii).

In the following we consider $\delta \leq \delta_0$, $\lambda = \lambda_0$, $\mu = 1$ and $\pi_a > \pi_b$ and will prove that L attains its maximum at $(\pi_a, \pi_b) = (1, 0)$.

First we will prove that $\frac{d}{d\pi_b}L \leq 0$. Let $\pi_a = 1 - w$. Then

$$L = \frac{1}{2}\frac{\begin{array}{c}(1-w-\pi_b)+(1-w-\pi_b)(\lambda_0 w+w-2+\pi_b+\lambda_0-\pi_b\lambda_0)\,\delta \\ +(\lambda_0-1)(-1+w+\pi_b)(-\lambda_0 w+\pi_b\lambda_0 w-2w+\pi_b w+1-\pi_b)\,\delta^2 \\ +w(\lambda_0-1)^2(-1+\pi_b)(-1+w+\pi_b)\,\delta^3\end{array}}{1-(1+\pi_b\lambda_0-\lambda_0-\lambda_0 w)\,\delta-w(\lambda_0-1)(\lambda_0+1)(-1+\pi_b)\,\delta^2-w(\lambda_0-1)^2(-1+\pi_b)\,\delta^3}$$

Now also assume $\pi_b = 0$. Then

$$\begin{aligned}\frac{d}{d\pi_b}L &= -\frac{1}{2}\left(1-(1-w-\lambda_0 w)\,\delta-w(1-\lambda_0)\,\delta^2\right) * \\ &\quad \frac{\left(1-(2-\lambda_0)\,\delta+\delta\left(1+\lambda_0(1-\delta)+\lambda_0^2\delta+(1-\lambda_0)^2\delta^2\right)w-(1-\lambda_0)\,\delta^2 w^2\right)}{\left(1+\lambda_0 w\delta-w(1-\lambda_0)\,\delta^2\right)\left(1-(1-\lambda_0-\lambda_0 w)\,\delta-w\left(1-\lambda_0^2\right)\delta^2+w(1-\lambda_0)^2\delta^3\right)}\end{aligned}$$

The second factor in the enumerator is the only one that can take negative values. Looking at this term we find that $\frac{d}{d\pi_b}L|_{(\pi_a,\pi_b)=(1,0)} \leq 0$ implies $\frac{d}{d\pi_b}L|_{\pi_b=0} \leq 0$ for all π_b. We also obtain

$$\frac{d}{d\pi_b}\frac{d}{d\pi_b}L = -\delta\frac{(1+\delta\lambda_0-\delta)\left(\delta\lambda_0 w+\delta^2\lambda_0 w-\delta+1-w\delta^2+\delta w\right)^2(\delta\lambda_0 w+1-\delta w)^2}{\left(1-(1+\pi_b\lambda_0-\lambda_0-\lambda_0 w)\,\delta-w(1-\lambda_0)(1-\pi_b)\,\delta^2(1+\lambda_0-(1-\lambda_0)\,\delta)\right)^3} \leq 0$$

which completes the proof that $\frac{d}{d\pi_b}L \leq 0$ holds for $\delta \leq \delta_0$.

If $\pi_b = 0$ then

$$\frac{d}{dw}L = -\frac{1}{2}\frac{\begin{array}{c}\left(1-2\delta-\delta^2\lambda_0-\delta^3\lambda_0+\delta^3\right)+2\delta(1+\lambda_0+\delta\lambda_0-\delta)\,w \\ +\delta^2(1+\lambda_0+\delta\lambda_0-\delta)(\lambda_0+\delta\lambda_0-\delta)\,w^2\end{array}}{\left(w\delta^2\lambda_0+\delta\lambda_0 w-w\delta^2+1\right)^2}$$

Since $(1+\lambda_0+\delta\lambda_0-\delta) \geq 0$ we obtain $\frac{d}{dw}L|_{(w,\pi_b)=(0,0)} \leq 0$ implies $\frac{d}{dw}L|_{\pi_b=0} \leq 0$ which completes the proof of the fact that $(\pi_a, \pi_b) = (1, 0)$ maximizes L. ■

References

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

[2] Börgers, T., Morales, A.J., and R. Sarin (2001), "Expedient and Monotone Learning Rules," Mimeo, University College London, http://www.ucl.ac.uk/~uctpa01/Papers.htm.

[3] Isbell, J.R. (1959), On a Problem of Robbins," *Ann. Math. Statist.* **30**, 606-10.

[4] Kakigi, R. (1983), "A Note on Discounted Future Two-Armed Bandits," *Ann. Statist.* **11(2)**, 707-11.

[5] Robbins, H. (1952), "Some Aspects of the Sequential Design of Experiments," *Bull. Amer. Math. Soc.* **58(5)**, 527-35.

[6] Robbins, H. (1956), "A Sequential Decision Problem with a Finite Memory," *Proc. Nat. Acad. Sci.* **42**, 920-3.

[7] Samaranayake, K. (1992) "Stay-With-A-Winnter Rule for Dependent Bernoulli Bandits", *Ann. Statist.* **20(4)**, 2111-23.

[8] Savage (1972), *The Foundation of Statistics*, Dover, New York

[9] Simon (1982), *Models of Bounded Rationality*, MIT Press.

[10] Tsetlin, M.L. (1961), "On the Behaviour of Finite Automata in Random Media," *Automation and Remote Control* **22**, 1210-19.

[11] von Neumann, J. and O. Morgenstern (1944), *Theory of Games and Economic Behavior*, Princeton Univ. Press.

[12] Wald, A. (1950), *Statistical decision functions*, Chelsea: Bronx.