How to Minimize Maximum Regret in Repeated Decision-Making

Karl H. Schlag[1]

July 13 2003

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

Abstract

Consider repeated decision making in a stationary noisy environment given a finite set of actions in each round. Payoffs belong to a known bounded interval. A rule or strategy attains minimax regret if it minimizes over all rules the maximum over all payoff distributions of the difference between achievable and achieved discounted expected payoffs.

Linear rules that attain minimax regret are shown to exist and are optimal for a Bayesian decision-maker endowed with the prior where learning is most difficult.

Minimax regret behavior for choosing between two actions given small or intermediate discount factors is derived and only requires two rounds of memory.

JEL classification: D81, D83.

Keywords: Two-armed bandit, Bernoulli, bounded rationality, minimax regret, limited memory.

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. Rational decision making as we call it today (von Neumann-Morgenstern 1944, Savage 1972) proceeds as follows. 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 his initial prior after any new information arrives. We will refer to a decision-maker as *Bayesian* if he behaves according to this procedure as probability updating follows Bayes' rule. The underlying behavioral rule will be called *Bayesian optimal*. Rational decision-making has been criticized from the beginning. In particular it has been questioned whether individuals are able to form priors and whether they have the ability and time to perform the necessary calculations when making their choices and updating their prior. These objections are particularly relevant when stakes are low, time is scarce and priors are diffuse (cf Simon, 1982).

We follow an alternative approach and investigate behavior of a decision-maker who makes choices that attain 'minimax regret' (Wald, 1950, Robbins, 1952, cf. Savage, 1951). This is a distribution-free approach in the sense that priors for the specific decision problem are not specified by the decision-maker. The advantage of a distribution-free approach is that the decision-maker does not have to determine a new prior and compute a new behavior each time he faces a similar decision problem. Instead he behaves the same at each first encounter and adapts behavior over time through experience and learning to the specific environment. Numerous different rules are suggested in the literature to describe learning without priors. This paper adds to the few papers (e.g. Börgers et al., 2001, Schlag, 1998) that formally select among such rules.

The environment of this paper is the same as in the classic *multi-armed bandit* problem which can be described as follows. An individual must repeatedly choose from a finite set of actions or *arms*. Each choice yields a random payoff which is drawn from an action dependent distribution that is stationary and independent of previous choices or payoffs achieved. All payoffs are assumed to belong to the interval $[0, 1]$. The specification of a set of actions and

of a payoff distribution for each action will be called a *decision problem*. So the individual repeatedly and independently faces the same decision problem. Finally, the classic multi-armed bandit specification includes a *prior* which is a probability measure over the set of decision problems. In the alternative setting where payoffs only belong to $\{0,1\}$ we call the decision problem a *Bernoulli decision problem*. The payoff distribution underlying a choice of an action in a given decision problem the individual actually faces should not be confused with the prior distribution over the decision problems the individual might face.

A rule or strategy is a description of which action the individual chooses next given his previous observations. We distinguish between *deterministic rules* that do not involve randomizing between actions and *(randomized) rules* that are probability measures over deterministic rules.

We assume that the individual is risk neutral and that future payoffs are discounted with a given discount factor δ where $\delta \in (0,1)$. An action that maximizes expected payoffs in a given decision problem as a *best action*. For a given rule and a given decision problem *regret* is defined as the difference between the maximal discounted expected payoff obtainable (i.e. the payoff to choosing a best action forever) and the discounted expected payoff achieved by this rule in this decision problem. Regret is strictly positive whenever the decision maker is a priori uncertain (or ignorant) of which action is best. This results from the fact that regret is never negative, that regret can be interpreted as the discounted sum of regret per round and that zero regret will not be attained in round one if the decision-maker is uncertain about which action is best.

Before introducing our methodology in more detail it is useful to explain how selecting behavior according maxmin (Wald, 1950, Gilboa and Schmeidler, 1989) fails in our setting. We allow for any prior over decision problems that yield payoffs in $[0,1]$. For any given rule expected payoffs are minimized when each action yields the payoff 0 for sure. So all rules yield the same minimal expected payoff and hence a maxmin decision maker should be indifferent among all rules.

There is little value to learning about the returns to the different actions if all actions yield similar expected payoffs. We will ignore such decision problems and focus on an individual who wishes to perform well when there is an incentive to learn. Performance of a rule will be measured by the maximal regret it achieves over the set of all decision problems. Accordingly we search for rules that attain *minimax regret* which means that the decision maker minimizes over

all rules the maximum regret over all decision problems of this rule. In other words, minimax regret behavior minimizes the maximum loss due to ignorance of the true state of affairs.[1]

In our main characterization of minimax regret behavior we extend results obtained by Berry and Fristedt (1985) for Bernoulli two-armed bandits to our setting in which payoffs belong to $[0, 1]$ and where more than two actions are allowed. Accordingly, a rule attains minimax regret if and only if it is an equilibrium strategy of the decision-maker in the zero sum game with nature where the decision-maker minimizes, and nature maximizes regret. A rule that attains minimax regret is shown to exist within the set of *Bernoulli equivalent rules* that are *symmetric*. A Bernoulli equivalent rule is a rule that is linear in payoffs and that behaves in any decision problem as in the Bernoulli decision problem in which actions receive the same expected payoff as in the original decision problem. A rule (or a prior) is called *symmetric* if its description does not depend on how the actions are labelled.

As in Berry and Fristedt (1985) we are able to show the relationship between minimax regret and Bayesian decision-making. Any minimax regret rule is Bayesian optimal under a so-called *worst case prior* that has the interpretation that it is the environment in which learning is most difficult for a Bayesian decision maker. More formally, a worst case prior maximizes over all priors the regret of a Bayesian decision-maker and is an equilibrium strategy of nature in the fictitious zero sum game mentioned above. Confirming our intuition we find that worst case priors exist within the set of symmetric priors that put weight on Bernoulli decision problems only.

In the rest of the paper we investigate minimax regret behavior in more detail when there are two actions only. Special focus is on whether minimax regret can be attained by a rule with finite memory. Specifically, a rule has n *round memory* for some natural number n if the next choice only depends on choices or payoffs obtained in the previous n rounds. The minimal size of memory needed to describe a rule is a candidate measure of a rule's complexity.

Our results build on understanding under which circumstances worst case priors are 'simple' where simple refers here to the fact that their support only contains two Bernoulli decision

[1] See French (1986) for a discussion of minimax regret along with alternative distribution free measures of behavior. Other studies on minimax regret include Chamberlain (2000) and, in terms of relative regret, Neeman (2001).

problems. Let Q_0 be the symmetric prior that puts equal weight on the two deterministic (two-action) decision problems in which one action yields payoff 1 and the other payoff 0. When minimax regret can be attained with a rule with finite memory then Q_0 is the only candidate for a symmetric worst case prior that is simple in the above sense. This is proven using results by Kakigi (1983) and Samaranayake (1992) on Bayesian optimal decision-making under simple priors. Furthermore we show that Q_0 can only be a worst case prior when $\delta \leq 0.62$, here proofs rely on investigating Taylor expansions of the regret of a Bernoulli equivalent rule near Q_0. This means for $\delta > 0.62$ that either the worst case prior is not simple or minimax regret cannot be attained by a rule with finite memory. Further results below complete the picture as they show that Q_0 is in fact a worst case prior for all $\delta \leq 0.62$. It is intuitive that Q_0 is a worst case prior when δ is sufficiently small as Q_0 maximizes the minimum regret in the first round. Aside we obtain that minimax regret behavior is never deterministic as long as $\delta \leq 0.62$.

There is an obvious candidate for a simple symmetric Bernoulli equivalent rule that attains minimax regret when Q_0 is a worst case prior as such a rule must be Bayesian optimal against Q_0. It is the single round memory rule that specifies in the first round to choose each action equally likely and in any later round to repeat the previous action with probability equal to the payoff obtained in the previous round. Our calculations show that this rule attains minimax regret if and only if $\delta \leq 0.41$. We also find that there is no single round memory rule that attains minimax regret when $\delta > 0.41$.

Typical for Bernoulli equivalent rules, the single round memory rule selected above prescribes random behavior whenever a payoff is realized in $(0, 1)$. We find that there exists a single round memory rule that is deterministic apart from the choice in the first round if and only if $\delta \leq 1/3$. A rule with this property for all $\delta \leq 1/3$ specifies to choose each action equally likely in the first round and in later rounds to repeat the previous action if and only if the payoff obtained in the previous round was greater than $1/3$.

We then present a symmetric Bernoulli equivalent two round memory rule that attains minimax regret if and only if $\delta \leq 0.62$. Should payoffs be only realized in $\{0, 1\}$ then this rule specifies to choose each action equally likely in the first round, to choose the same action again in the next two rounds whenever receiving payoff 1 and to switch actions otherwise. Behavior after receiving interior payoffs is more intricate and essentially involves a four state stochastic

automaton. We also show that there is no two round memory rule that attains minimax regret for δ larger than 0.62 if it attains minimax regret for all $\delta \leq 0.62$.

Finally we investigate for which values of δ between 0.41 and 0.62 minimax regret can be attained with two round memory of actions but only on a single round memory of previous payoffs - rules we call *two round action memory rules*. We find that minimax regret is attainable with such a rule if and only if $\delta \leq 0.54$. The rule presented with this property is Bernoulli equivalent, symmetric, specifies to choose the same action again after receiving payoff 1 and to sometimes choose it again after receiving payoff 0 if the same action has been chosen twice in a row.

This is the first paper in which minimax regret behavior has been explicitly derived for two-armed bandits. Partial results existed previously only for the scenario in which all payoffs are contained in $\{0,1\}$. Berry and Fristedt (1985) provided upper and lower bounds on minimax regret when δ is close to 1. A series of papers in the statistics and in the machine learning literature present specific examples of rules to be used when the decision maker is infinitely patient, i.e. $\delta = 1$ (e.g. Robbins, 1952, 1956, Samuels, 1968, Narendra and Thathachar, 1989). In particular, two rules suggested by Robbins (1952, 1956) coincide to the rules selected by us for small ($\delta \leq 0.41$) and for intermediate ($\delta \leq 0.62$) discount factors when payoffs are limited to $\{0,1\}$.

The presentation of the material proceeds as follows. Section two introduces the basic setting. In Section three we supply the main characterization result of minimax regret behavior and worst case priors. In Section four 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, Rules and Selection

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 or arms $W = \left\{a_1,..,a_{|W|}\right\}$ with $|W| \geq 2$ and for each action $c \in W$ a measurable payoff distribution $P_c \in \Delta[0,1]$.[2] Sometimes we will index parameters by the decision problem D they refer to, e.g. write $P_c(D)$ instead of P_c. The set

[2] Our results can be applied to payoff distributions over a known bounded interval $[\alpha,\omega]$ by first rescaling payoffs into $[0,1]$ using the linear transformation $x \longmapsto \frac{x-\alpha}{\omega-\alpha}$.

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 $\{0,1\}$. The set of all Bernoulli multi-action decision problems will be denoted by $\mathcal{D}_0$. Payoffs 0 and 1 are sometimes referred to as *failure* and *success* respectively.[3]

Consider an individual who repeatedly faces the same multi-armed bandit (W,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-action decision problem $\left(W,\tilde{P}\right)$ the individual will be facing according to the prior Q. Choice of action c in round t yields a payoff realized according to $\tilde{P}_c$ that is drawn independently of previous choices and payoff realizations.

A rule (or strategy) is the formal description of how the individual makes his choice as a function of his previous experience. A *deterministic rule* is a mapping $f : \emptyset \cup_{m=1}^{\infty} \{\times_{k=1}^{m} \{W \times [0,1]\}\} \to W$ where $f(\emptyset)$ is the action chosen in the first round and $f(a_1, x_1, .., a_m, x_m)$ is the action choosen in round $m+1$ after choosing action a_k and receiving payoff x_k in round k for $k = 1, .., m$. The set of deterministic rules will be denoted by $\mathcal{F}$. A *(randomized) (behavioral) rule* ϕ is a probability measure over the set deterministic rules and hence an element of $\Delta\mathcal{F}$. We identify $c \in W$ with the probability distribution in ΔW that selects c with probability one so that $\mathcal{F} \subset \Delta\mathcal{F}$. We will also write $\phi(\emptyset)_c$ as the probability of choosing action c in the first round and $\phi(a_1, x_1, .., a_m, x_m)_c$ as the probability of choosing action c in round $m+1$ after the history $(a_1, x_1, .., a_m, x_m)$. Notice that these probabilities need not be independent across rounds.

Assume throughout that the individual decision-maker is risk neutral and discounts future payoffs with a given discount factor $\delta \in (0,1)$.[4] For a given rule ϕ and a given decision problem D let $p_c^{(n)} = p_c^{(n)}(\phi, D)$ be the probability of choosing action $c \in W$ in round n unconditional on previous choices. Let $\pi_c(D) = \int x dP_c(x, D)$ denote the expected payoff of choosing action c when facing the multi-action decision problem (W, D). Then $\pi^{\delta}(\phi, D) := (1-\delta)\sum_{n=1}^{\infty} \delta^{n-1} \sum_{c \in W} p_c^{(n)}(\phi, D)\, \pi_c(D)$ is the discounted value of future payoffs. The *regret*

[3] The machine learning literature (cf Naremdra and Thathachar, 1989) refers to the Bernoulli case as the P-model and to our setting with payoffs in $[0,1]$ as the S-model. In the Q-model the support of the payoff distribution is finite.

[4] Our analysis also applies to agents that are not risk neutral by replacing each payoff x with a von Neumann-Morgenstern utility $u(x)$ where $u(0) = 0$ and $u(1) = 1$.

(or opportunity loss) of a rule ϕ when facing the multi-action decision problem D is defined as $L_\phi(D) := \max_{c\in W}\{\pi_c(D)\} - \pi^\delta(\phi, D)$. Regret is a measure of the loss due to ignorance of the true state of affairs where the state of affairs is identified with a decision problem.[5] Elements of $\arg\max_{c\in W}\{\pi_c(D)\}$ will sometimes be referred to as *best actions*.

A Bayesian decision-maker is an individual who chooses a rule $\hat{\phi} \in \arg\max \int \pi^\delta_\phi(D)\, d\tilde{Q}(D)$. His choice $\hat{\phi} = \hat{\phi}\left(\tilde{Q}\right)$ is called a *Bayesian optimal rule under* $\tilde{Q}$. We will call $\bar{Q}$ a *worst case prior* if it maximizes the expected regret of a Bayesian decision-maker over all priors, i.e. if $\bar{Q} \in \arg\max_{Q\in\Delta\mathcal{D}} \int L_{\hat{\phi}(Q)}(D)\, dQ(D)$. Simplifying we obtain that $\bar{Q}$ is a worst case prior if and only if $\bar{Q} \in \arg\max_{Q\in\Delta\mathcal{D}} \min_{\phi\in\Delta\mathcal{F}} \int L_\phi(D)\, dQ(D)$.

If the prior $\tilde{Q}$ is unknown (while W is known) then according to Savage (1972) the individual specifies a subjective prior $\hat{Q}$ and chooses a Bayesian optimal rule under $\hat{Q}$. We follow an alternative approach (Wald, 1950, Gibbons, 1952) that is distribution-free as the individual does not invoke a specific prior to select a rule. We assume that the individual selects a rule that minimizes among all rules the maximal regret among all decision problems (W, D). More specifically, we say that ϕ^* *attains minimax regret* if $\phi^* \in \arg\min_{\phi\in\Delta\mathcal{F}} \sup_{D\in\mathcal{D}} L_\phi(D)$.

3 A General Characterization

Some definitions are need before we present our characterization of minimax regret behavior and worst case priors.

3.1 Symmetry

As the various actions belonging to W cannot be distinguished (apart from their labels), symmetry will play an important role in our investigation.

Given $D \in \mathcal{D}$ and a permutation ι of the elements 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 such that $P_c(D_\iota) = P_{\iota(c)}(D)$ for $c \in W$. For a given multi-armed bandit (W, Q) with $Q \in \Delta\mathcal{D}$ let Q_ι be the distribution defined by exchanging each decision problem D in the support of Q by D_ι. A prior

[5] Notice how we thus differ from the approach of Savage (1951, cf. French, 1986) that is based on a set of states, each being without uncertainty and where regret is considered in each state separately.

Q is called *symmetric* if $Q = Q_\iota$ holds for any permutations ι of the elements of W. The set of symmetric priors over a subset $\mathcal{Z}$ of $\mathcal{D}$ will be denoted by $\Delta_p\mathcal{Z}$.

Given a deterministic rule f and a permutation ι of the elements of W let f_ι be the deterministic rule that is derived from f by permuting actions with ι such that $f_\iota(\emptyset)_c = f(\emptyset)_{\iota(c)}$ and $f_\iota(a_1, x_1, .., a_m, x_m)_c = \phi(\iota(a_1), x_1, .., \iota(a_m), x_m)_{\iota(c)}$. A randomized rule ϕ is called *symmetric* if $\phi(T) = \phi(\{f_\iota \text{ s.t. } f \in T\})$ holds for all permutations ι of W and for all measurable sets of deterministic rules T. The set of symmetric randomized rules will be denoted by $\Delta_p\mathcal{F}$. Notice that if ϕ is symmetric then $\phi(\emptyset)_c = \frac{1}{|W|}$ for all $c \in W$.

3.2 Linearity and Bernoulli Equivalence

In our setting there are no restrictions on how the action prescribed by a given rule in a given round depends on previous payoffs obtained. We will find that simple rules in the sense that behavior is a linear function of previous payoffs will play an important role for attaining minimax regret behavior. More specifically, a subset of the linear rules called Bernoulli equivalent rules will play this important role.

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

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

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

A linear rule ϕ is called *Bernoulli equivalent* if in any decision problem it behaves as it does in the Bernoulli decision problem in which actions have the same expected payoff as in the original decision problem. More formally, given $D \in \mathcal{D}$ let $D_0(D) \in \mathcal{D}_0$ be defined by the fact that $\pi_c(D) = \pi_c(D_0(D))$ holds for all $c \in W$. Then we require for all $D \in \mathcal{D}$ and for any sequence of actions $a_1, .., a_m$ that the probability that action a_i is chosen in round i for all $i = 1, .., m$ is

the same under D as it is under $D_0(D)$. Formally,

$$\begin{aligned}
&\int \phi(\emptyset)_{a_1} * \prod_{k=1}^{m-1} \phi(a_1, x_1, .., a_k, x_k)_{a_{k+1}} \, dP_{a_1}(x_1) \, ..dP_{a_{m-1}}(x_{m-1}) \\
= \quad & \sum_{j=1}^{m} \sum_{y_j=0}^{1} \prod_{j=1}^{m} \left(y_j \pi_{a_j} + (1-y_j)\left(1-\pi_{a_j}\right)\right) * \phi(\emptyset)_{a_1} * \prod_{k=1}^{m-1} \phi(a_1, y_1, .., y_k, x_k)_{a_{k+1}}
\end{aligned}$$

The set of Bernoulli equivalent rules with support $M \subseteq \mathcal{F}$ will be denoted by $\Delta^B M$.

Next we illustrate why not all linear rules are Bernoulli equivalent. Consider a linear rule f. It is easily checked that f satisfies the conditions imposed on a Bernoulli equivalent rule in the first two rounds. However this is not necessarily true in round three. For instance, the probability of obtaining the sequence of actions a,b,b in the first three rounds equals

$$\begin{aligned}
& f(\emptyset)_a \int f(a,x)_b \, f(a,x,b,y)_b \, dP_a(x) \, dP_b(y) \\
= \quad & f(\emptyset)_a \left(\pi_b \int f(a,x)_b \, f(a,x,b,1)_b \, dP_a(x) + (1-\pi_b) \int f(a,x)_b \, f(a,x,b,0)_b \, dP_a(x) \right)
\end{aligned}$$

If f prescribes to randomize independently in each round then

$$\int f(a,x)_b \, f(a,x,b,1)_b \, dP_a(x) = \left(\pi_a f(a,1)_b + (1-\pi_a) f(a,0)_b\right)\left(\pi_a f(a,1,b,1)_b + (1-\pi_a) f(a,0,b,1)_b\right).$$

On the other hand, if f is Bernoulli equivalent then

$$\int f(a,x)_b \, f(a,x,b,1)_b \, dP_a(x) = \pi_a f(a,1)_b \, f(a,1,b,1)_b + (1-\pi_a) f(a,0)_b \, f(a,0,b,1)_b .$$

So if the linear rule f is Bernoulli equivalent and $f(\emptyset)_a > 0$ then one of the following three statements is true: (i) $f(a,0)_b = f(a,1)_b$, (ii) $f(a,0,b,y)_b = f(a,1,b,y)_b$ for all $y \in [0,1]$, or (iii) randomization under f in rounds two and three is not independent. So linearity can only coincide with Bernoulli equivalence if payoffs obtained more than one round ago have only a limited impact on present behavior (see Sections 4.2.1 and 4.2.3 below).

Finally we show how to extend a rule ϕ defined on the set of Bernoulli decision problems to a Bernoulli equivalent rule. We present two ways to generate the same behavior. (A) When a payoff, say $x_i \in [0,1]$, is obtained in round i then realize an independent random variable that yields 1 with probability x_i and 0 otherwise. Remember the realization $d_{x_i} \in \{0,1\}$ of this random variable and forget the payoff x_i itself. Apply the rule in all later rounds as if d_{x_i} was the payoff received in round i. (B) Given a sequence $z = (z_i)_{i=1}^{\infty}$ with $z_i \in [0,1]$ for all i define the rule

ϕ_z by setting $\phi_z(\emptyset) = \phi(\emptyset)$ and setting $\phi_z(a_1, x_1, .., a_m, x_m) = \phi\left(a_1, 1_{\{x_1 \geq z_1\}}, .., a_m, 1_{\{x_m \geq z_m\}}\right)$ for any history $(a_1, x_1, .., a_m, x_m)$ where $1_{\{x_i \geq z_i\}}$ is the indicator function that takes value 1 if $x_i \geq z_i$ and value 0 otherwise. The Bernoulli equivalent extention of the rule ϕ is then obtained by randomizing over the set of rules ϕ_z by choosing z_i iid from a uniform distribution on $[0, 1]$ for all i.

Under the behavior defined in (A), each sequence of actions has the same probability of occurring in the decision problem D as it does in the Bernoulli decision problem $D_0(D)$. Nonethe-less, the construction in (A) does not formally define a rule as the memory of the decision-maker is changed which is not allowed in our definition of a randomized rule. Our second alternative (B) leads to a formal definition of a rule. It is easily shown that the rule defined in (B) is behaviorally equivalent to the one described in (A) and hence that it is Bernoulli equivalent.

Remark 1 *Linear rules, in particular Bernoulli equivalent rules, typically involve randomizing when receiving payoffs in* $(0, 1)$*. More specifically, it is easily deduced from (1) for a linear rule* f *that either* $f(a_1, x_1, .., a_n, x_n)$ *is independent of* $x_1, .., x_n$ *or* $f(a_1, x_1, .., a_n, x_n) \notin W$ *for all* $x_1, .., x_n \in (0, 1)$*. In contrast we show in the appendix that Bayesian optimal rules typically do not involve randomizing behavior.*

3.3 The Result

The following characterization will be very useful as it reduces the search for a rule that attains minimax regret to the search for an equilibrium of a zero-sum game. At the same time it reveals a close connection between minimax regret behavior and Bayesian decision making.

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

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

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

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

$$\int L_{\phi^*}(D)\, dQ(D) \leq \int L_{\phi^*}(D)\, dQ^*(D) \leq \int L_{\phi}(D)\, dQ^*(D) \;\; \forall \phi \in \Delta\mathcal{F} \; \forall Q \in \Delta\mathcal{D}. \quad (2)$$

In particular, any rule that attains minimax regret is Bayesian optimal under any worst case prior.

The above generalizes findings that Berry and Fristedt (1985) have obtained for Bernoulli two-armed bandits.

Proof. We first review the results Berry and Fristedt (1985) obtained for Bernoulli two-armed bandits which is statement (i) and the 'if' statements of (ii) and (iii). They introduce a topology on the set of strategies and then show for the zero sum game where the individual chooses a rule to minimize regret and nature chooses a prior to maximize regret that a Nash equilibrium (ϕ^*, Q^*) exists. If (ϕ^*, Q^*) is a such a Nash equilibrium (i.e. (2) 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 the above also holds when payoffs are not restricted to $\{0,1\}$. Let $(\phi^*, Q^*) \in \Delta^B\mathcal{F} \times \Delta\mathcal{D}_0$ be a Nash equilibrium of the zero-sum game when restricting attention to $\mathcal{D}_0$. Since ϕ^* is Bernoulli equivalent, $\max_{Q\in\Delta\mathcal{D}_0} \int L_{\phi^*}(D)\, dQ(D) = \max_{Q\in\Delta\mathcal{D}} \int L_{\phi^*}(D)\, dQ(D)$ 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 (2) 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

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

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. ■

4 Two-Armed Bandits

In the following we investigate minimax regret behavior when there are two actions only. Let $W=\{a,b\}$. Special attention will focus on whether minimax regret behavior can be implemented with a rule that has finite memory. An important ingredient will be to understand Bayesian optimal behavior under specific very simple priors.

We say that the rule ϕ has n *round memory* if $\phi(a_1,x_1,..,a_m,x_m)_c$ is independent of (a_k,x_k) for $k\leq m-n$. ϕ has *finite memory* if there exists n such that ϕ has n round memory. ϕ has n *round action memory* if ϕ has n round memory and if $\phi(a_1,x_1,..,a_m,x_m)_c$ is independent of x_k for $k\leq m-1$. The amount of memory needed to implement a rule can be considered a measure of its complexity.

4.1 Necessary Conditions

Following Proposition 1, rules that attain minimax regret are Bayesian optimal under some prior over Bernoulli decision problems. Insights into Bayesian optimal behavior can thus teach us about minimax regret behavior. Unfortunately many results on Bayesian optimal decision making deal only with independent arms - we do not expect a worst case prior ever to have this property. We will use results on dependent arms due to Kakigi (1983) and Samaranayake (1992) who consider priors that put weight on two Bernoulli decision problems.

$Q\in\Delta_p\mathcal{D}_0$ will be called a *symmetric two point Bernoulli prior* if it only has two elements in its support, formally if there exist v and w with $0\leq v<w\leq 1$ such that $Q\left(\tilde{D}\right)=1/2$ for $\tilde{D}\in\mathcal{D}_0$ with $\pi_1\left(\tilde{D}\right)=v$ and $\pi_2\left(\tilde{D}\right)=w$. We will write $Q=Q(v,w)$ and also write

Q_0 instead of $Q(0,1)$. Kakigi (1983) derives a particular Bayesian optimal rule for such priors. Results found in Samaranayake (1992) can be used to show that a Bayesian optimal rule under such a prior will have the 'stay with a winner' property. The rule ϕ is said to have the *stay with a winner* property if it specifies to choose the same action again after any success, i.e. if $\phi(a_1, x_1, .., a_m, 1)_{a_m} = 1$ for all a_k, x_k, $k = 1, .., m-1$, all a_m and all m.[6]

Proposition 2 *Consider* $|W| = 2$. *If the finite memory rule* ϕ^* *attains minimax regret and* $Q(v,w)$ *is a worst case prior (for instance when* $\arg\max_{D \in \mathcal{D}_0 : \pi_a(D) > \pi_b(D)} L_{\phi^*}(D)$ *is single valued) then* $Q(v,w) = Q_0$.

Proof. First assume $0 < v < w < 1$. Kakigi (1983) shows that the following symmetric rule is Bayesian optimal under such a prior $Q(v,w)$. Choose action a in round one. Choose action a in round n if and only if the updated belief that a yields a higher expected payoff than b is at least 0.5. Notice that this rule cannot be implemented with a finite memory as $0 < v < w < 1$. What we have to show in the following is that no Bayesian optimal rule will have finite memory.

It is shown in Kakigi (Proof of Theorem 2, 1983) that the difference between the value of choosing a and the value of choosing b when continuing thereafter optimally is non decreasing in the belief that action a yields a higher expected payoff than action b. Let $r(s)$ denote this difference where s is the corresponding belief.

Samaranayake (Example 2.2, 1992) shows that actions a and b are negatively correlated after any history. Since $v < w$, the support of the marginal distributions of choosing a (or of choosing b) has two elements. So the results in Proposition 5.2(b) in Samaranayake (1992) holds with strict inequalities. This means that a Bayesian decision maker strictly prefers action c over action d after action c yielded a success.

Thus $r(s^+) > 0$ if s^+ is the updated belief after having belief $1/2$ and receiving a success by choosing a. Together with the fact that r is non decreasing we obtain that r is strictly increasing in s. In other words, the rule from Kakigi (1983) described above prescribes the unique Bayesian optimal behavior whenever the belief does not equal $1/2$. Hence no Bayesian optimal rule under $Q(v,w)$ with $0 < v < w < 1$ has finite memory.

[6] Note that the result proven by Berry and Fristedt (1985) for independent arms is weaker as it only states that there always exists a Bayesian optimal rule with the stay with a winner property.

Now assume $v = 0$ and $w \in (0,1)$. Consider a Bayesian optimal behavior ϕ^* under $Q(0,w)$. As behavior when $s = 1/2$ does not matter we can assume that ϕ^* is symmetric and specifies to switch after any failure. Of course ϕ^* locks in on the same action whenever he obtains the first success. We now calculate regret of ϕ^* when facing $Q(0,\bar{w})$ for some $\bar{w} \in (0,1)$. Let z be the future value after only failures obtained previously. Then $z = (1-\delta)\frac{1}{2}\bar{w} + \frac{1}{2}\bar{w}\delta\bar{w} + \left(1 - \frac{1}{2}\bar{w}\right)\delta z$ so $z = \bar{w}\frac{1-\delta+\bar{w}\delta}{2-2\delta+\bar{w}\delta}$ and hence $L_{\phi^*}(Q(0,\bar{w})) = \bar{w} - z = \frac{(1-\delta)\bar{w}}{2-2\delta+\bar{w}\delta}$. $L_{\phi^*}(Q(0,\bar{w}))$ as a function of $\bar{w}$ obtains its unique maximum when $\bar{w} = 1$ and hence $Q(0,w)$ is never a worst case prior if $w < 1$.

Finally, assume $v > 0$ and $w = 1$. Let ϕ^* be the symmetric Bayesian optimal rule under $Q(v,1)$ that switches after a failure and has the stay with a winner property. Consider regret under $Q(\bar{v},1)$ for some $\bar{v} \in (0,1)$. Let x be the future value of payoffs after only achieving successes in the previous rounds with the worse action. We obtain $x = (1-\delta)\bar{v} + \bar{v}\delta x + (1-\bar{v})\delta$ so $x = \frac{\delta+\bar{v}-2\delta\bar{v}}{1-\delta\bar{v}}$ and hence $L_{\phi^*}(Q(\bar{v},1)) = 1 - \frac{1}{2} - \frac{1}{2}\frac{\delta+\bar{v}-2\delta\bar{v}}{1-\delta\bar{v}} = \frac{1}{2}\frac{(1-\delta)(1-\bar{v})}{1-\delta\bar{v}}$. $L_{\phi^*}(Q(\bar{v},1))$ as a function of $\bar{v}$ obtains its unique maximum when $\bar{v} = 0$ and hence $Q(v,1)$ is never a worst case prior if $v > 0$. ■

So Q_0 is the only candidate for a simple worst case prior. Using Taylor expansions of regret we derive an upper bound on the set of discount factors under which Q_0 can be a worst case prior.

Proposition 3 *Consider $|W| = 2$. Then Q_0 is not a worst case prior for $\delta > \frac{1}{2}\sqrt{5} - \frac{1}{2} \approx 0.62$.*

Proof. Consider a symmetric Bernoulli equivalent rule ϕ^* that attains minimax regret with Q_0 being a worst case prior. Since ϕ^* is symmetric, $\phi^*(\emptyset)_a = 1/2$. Since ϕ^* is a Bayesian optimal rule under Q_0, ϕ^* has the stay with a winner property and $\phi^*(c,0)_c = 0$ for $c \in \{a,b\}$. So all we have to check in order for ϕ^* to attain minimax regret is that Q_0 maximizes regret of the rule ϕ^*.

In the following we will derive necessary conditions such that Q_0 maximizes regret $L_{\phi^*}(Q)$ among priors Q contained in $\{Q(v,1) : 0 \leq v < 1\} \cup \{Q(0,w) : 0 < w \leq 1\}$. In fact we will only be considering priors close to Q_0 which means that v is close to 0 and w is close to 1. Looking at first order effects only means that when facing $Q(v,1)$ we can ignore events in which the 'bad' action yields two successes. Similarly, when facing $Q(0,w)$ we can ignore events when the best action yields two failures.

Below we alter the behavior of ϕ^* to obtain a rule ϕ' that chooses if possible a best response to both $Q(v,1)$ and $Q(0,w)$. ϕ' retains the properties of ϕ^* that ϕ' is a best response to Q_0 and that Q_0 maximizes regret under ϕ'. The latter follows since $L_{\phi^*}(Q_0) = L_{\phi'}(Q_0)$ and $L_{\phi^*}(Q) \geq L_{\phi'}(Q)$ implies $L_{\phi'}(Q_0) \geq L_{\phi'}(Q)$.

Let ϕ' choose action a forever after observing a failure from action b and a success from action a in the first two rounds. Here ϕ' chooses a best response to both $Q(0,w)$ and to $Q(v,1)$. Let ϕ' choose action a forever after observing $(a,1,a,1)$ or $(a,1,a,0,a,1)$. As we are only interested in first order approximation, we ignore the possibility that we could be facing $(\pi_a, \pi_b) = (v,1)$. Similarly, based on first order approximation ϕ' chooses action a forever after observing two failures of b and one failure of a in the first three rounds.

Let $x = \phi^*(c,1,c,0)_d$ and $y = \phi^*(d,0,c,0)_d$ for $c \neq d$. Then

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

where the expressions refer in the order of their appearance 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 $(b,0,a,0,b,0)$. Consequently,

$$L_{\phi'} = \frac{1}{2}(1-\delta) - \frac{1}{2}(1-\delta)\left(1-\delta-\delta^2-x\delta^2\right)(1-w) + o\left((1-w)^2\right).$$

Since Q_0 maximizes $L_{\phi'}(Q)$ we obtain $1-\delta-\delta^2-x\delta^2 \geq 0$ which implies $1-\delta-\delta^2 \geq 0$ which implies $\delta \leq \frac{1}{2}\sqrt{5} - \frac{1}{2}$. ■

We combine Propositions 2 and 3 to obtain the following.

Corollary 4 *Consider $|W| = 2$ and $\delta > \frac{1}{2}\sqrt{5} - \frac{1}{2}$. Then either there is no finite memory rule that attains minimax regret or $\arg\max_{D \in \mathcal{D}_0 : \pi_a(D) > \pi_b(D)} L_{\phi^*}(D)$ is not single valued for any ϕ^* that attains minimax regret.*

At this point of our analysis we have no evidence for which values (if any) of $\delta \leq \frac{1}{2}\sqrt{5} - \frac{1}{2}$ that Q_0 is a worst case prior. However, if Q_0 is a worst case prior at $\delta = \frac{1}{2}\sqrt{5} - \frac{1}{2}$ then the expansion technique used in the proof of Proposition 3 reveals properties of a minimax regret rule.

Lemma 5 *Consider* $|W| = 2$ *and* $\delta = \frac{1}{2}\sqrt{5} - \frac{1}{2}$. *Assume that* ϕ^* *is a symmetric rule that attains minimax regret and assume that* Q_0 *is a worst case prior. Then* $\phi^*(c,0)_c = 0$, $\phi^*(c,1,c,0)_c = 1$, $\phi^*(c,1,c,1,c,0)_c = 1$, $\phi^*(c,1,c,0,c,0)_c = 0$, $\phi^*(c,0,d,1,d,0)_d = 1$, $\phi^*(c,0,d,0,c,0)_c = 0$ *if* $\phi^*(c,0,d,0)_d < 1$, $\phi^*(c,0,d,0,d,0)_d = 0$ *if* $\phi^*(c,0,d,0)_d > 0$ *and in the first three rounds* ϕ^* *does not switch after a success.*

Consequently, no single round memory nor any n round action memory for some n attains minimax regret under this critical value of δ. Nor does one of the rules suggested by Robbins (1956) or Isbell (1959) for $n > 2$ (and $\delta = 1$) have this property.

Proof. First we provide the analogous calculations as in the proof of Proposition 3 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 action 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 action 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

$$L_{\phi'} = \frac{1}{2}(1-\delta) - \frac{1}{2}(1-\delta)\left(1-\delta-(1-x)\delta^2\right)v + o\left(v^2\right)$$

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 3 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. ■

Finally we show that any rule that attains minimax regret when Q_0 is a worst case prior behaves in round one like a symmetric rule.

Proposition 6 *Consider* $|W| = 2$. *If* Q_0 *is a worst case prior and* ϕ^* *attains minimax regret then* $\phi^*(\emptyset)_a = 1/2$.

Proof. Consider a rule ϕ^* that attains minimax regret when Q_0 is a worst case prior. Then ϕ^* is Bayesian optimal under Q_0. Let D_c be the Bernoulli two-action decision problem with $P_c(1) = P_d(0) = 1$ where $d \neq c$. Then $L_{\phi^*}(D_a) = 1 - \phi^*(\emptyset)_a - \phi^*(\emptyset)_b \delta$ and $L_{\phi^*}(D_b) = 1 - \phi^*(\emptyset)_b - \phi^*(\emptyset)_a \delta$. Since Q_0 is a worst case we obtain $L_{\phi^*}(D_a) = L_{\phi^*}(D_b)$ and hence $\phi^*(\emptyset)_a = 1/2$. ■

4.2 Sufficient conditions

Above we show that Q_0 is the only candidate for a simple worst case prior. For any symmetric rule ϕ regret equals $L_\phi(D) = (1-\delta)\frac{1}{2}|\pi_a - \pi_b| + (1-\delta)o(\delta)$ so we actually expect Q_0 (which maximizes $|\pi_a - \pi_b|$) to be a worst case prior for sufficiently small δ. Interestingly we find below that Q_0 does not have to be "that small" for this to be true.

4.2.1 Single round memory

In the following we search for single round memory rules that attain minimax regret. Note that for single round memory rules there is no difference between linearity and Bernoulli equivalence.

Proposition 7 *Consider* $|W| = 2$.

(i) The symmetric linear single round memory rule ϕ^* *that has the stay with a winner property and that satisfies* $\phi^*(a,0)_a = 0$ *attains minimax regret if and only if* $\delta \leq \sqrt{2}-1 \approx 0.41$. *This rule yields*

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

(ii) For any $\delta \in (0,1)$ *there is no other symmetric linear single round memory rule that attains minimax regret.*

(iii) There is no single round memory rule that attains minimax regret for some $\delta > \sqrt{2}-1$.

Notice that Bayesian optimal rules generally do not have finite memory even when δ is small. For instance, as pointed out in the proof of Proposition 2, any Bayesian optimal rule under the two point distribution $Q(v,w)$ with $0 < v < w < 1$ does not have finite round memory.

Proof. It follows immediately that the rule ϕ^* described above is the unique symmetric linear single round memory Bayesian optimal rule under Q_0. 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 the two equations in the two unknowns z_a and z_b yields the expression for $\pi^\delta = 0.5\,(z_a + z_b)$ given above.

For $\pi_a > \pi_b$ we obtain

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

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

Similarly we obtain for $\pi_a > \pi_b$

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

Thus, L is maximized at $(\pi_a, \pi_b) = (1, 0)$ and Q_0 is a worst case prior if and only if $\delta \leq \sqrt{2}-1$. Since ϕ^* is the unique symmetric linear single round memory rule that is Bayesian optimal under Q_0 there is no alternative symmetric linear single round memory rule that attains minimax regret when $\delta \leq \sqrt{2} - 1$.

It is easily verified 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 4 ϕ^* does not attain minimax regret when $\delta > \sqrt{2} - 1$.

Note that $p_b = \frac{1}{2}\frac{1+\delta-2\delta\pi_a}{1+\delta-\delta\pi_a-\delta\pi_b}$ is the sum of the discounted probabilities of choosing action b under ϕ^* where p_b can be derived as the solution to $\pi^\delta = (1 - p_b)\,\pi_a + p_b\pi_b$.

Consider an alternative symmetric linear single round memory rule ϕ. Let q_c be the probability of choosing action c in the next round given c is chosen in the present round, then $q_c = \pi_c y + (1 - \pi_c)\,z$ where $y = \phi(c, 1)_c$ and $z = \phi(c, 0)_c$. Consequently $\pi^\delta(\phi) = \pi_a + \frac{1}{2}\frac{1+\delta-2\delta q_a}{1+\delta-\delta q_a-\delta q_b}(\pi_b - \pi_a)$ and $L_\phi = \frac{1}{2}\frac{1+\delta-2\delta q_a}{1+\delta-\delta q_a-\delta q_b}(\pi_a - \pi_b)$ when $\pi_a > \pi_b$. It is easily verified that $\frac{d}{dy}L_\phi < 0 < \frac{d}{dz}L_\phi$ when $\pi_a > \pi_b$. Thus ϕ^* is among linear symmetric single round memory rules the only candidate for a Bayesian optimal rule and hence the only candidate for a symmetric rule that attains minimax regret. ■

Following Propositions 6 and 7, any single round memory rule that attains minimax regret randomizes in round one, choosing each action with probability 0.5. However, the rule selected in Proposition 7 also randomizes in later rounds whenever receiving a payoff in $(0, 1)$. In the

following we investigate when and whether this sort of randomizing is also necessary for attaining minimax regret.

Proposition 8 *Consider $|W| = 2$. Consider a single round memory rule ϕ with $\phi(c,x)_c \in \{0,1\}$ for all $x \in [0,1]$. Then*

(i) ϕ attains minimax regret for all $\delta \leq 1/3$ if and only if $\phi(\emptyset)_a = 1/2$, $\phi(c,x)_c = 0$ if $x < 1/3$ and $\phi(c,x)_c = 1$ if $x > 1/3$.

(ii) ϕ does not attain minimax regret for $\delta > 1/3$.

Proof. Consider a symmetric single round memory rule ϕ^o with $\phi^o(c,x)_c \in \{0,1\}$ for all $x \in [0,1]$. Consider $D^* \in \arg\max_{D \in \mathcal{D}_0 : \pi_a(D) > \pi_b(D)} L_{\phi^o}(D)$. Then ϕ^o behaves (in terms of sequences of actions chosen) when facing D^* as the rule ϕ^* defined in Proposition 7 does when facting $D_0 \in \mathcal{D}_0$ defined by $P_c(1, D_0) = P_c(\{x : \phi^o(c,x)_c = 1\}, D^*)$. Setting $q_c = P_c(1, D_0)$ and using the expression p_b from the proof of Proposition 7 we obtain $L_{\phi^o}(D^*) = \frac{1}{2}\frac{1+\delta-2\delta q_a}{1+\delta-\delta q_a-\delta q_b}(\pi_a - \pi_b)$.

Now assume that ϕ^o attains minimax regret. Let $\rho_o = \sup_x \{\phi^o(c,x)_c = 0\}$ and $\rho_u = \inf_x \{\phi^o(c,x)_c = 1\}$. Since D^* maximizes the regret of ϕ^o we derive that $\pi_a(D^*) = q_a + (1-q_a)\rho_o$ and $\pi_b(D^*) = q_b \rho_u$. Since $\rho_o \geq \rho_u$, the decision-maker can lower this maximal regret by choosing a rule with $\rho_o = \rho_u =: \rho$. In other words, the decision-maker chooses ρ and nature chooses q_a and q_b and regret is given by

$$L_{f^o}(D^*) = \frac{1}{2}\frac{1+\delta-2\delta q_a}{1+\delta-\delta q_a-\delta q_b}\left(q_a + (1-q_a)\rho - q_b\rho\right) .$$

Following Proposition 7, if f attains minimax regret then Q_0 is a worst case prior. Hence we need to verify that

$$\begin{aligned}
\frac{d}{dq_a} L_{f^o}|_{(q_a,q_b)=(1,0)} &= \frac{1}{2}(1-\delta)(1-\rho) - \frac{1}{2}\delta(1+\delta) \geq 0 \\
\frac{d}{dq_b} L_{f^o}|_{(q_a,q_b)=(1,0)} &= -\frac{1}{2}(1-\delta)(\rho-\delta) \leq 0
\end{aligned}$$

which simplifies to $\delta \leq \rho \leq 1/3$.

Finally, if $\rho = 1/3$ then it is easily verified that $\frac{d}{dq_2} L \leq 0$ holds $\delta \leq 1/3$ and that $q_2 = 0$ implies $\frac{d}{dq_1} L \geq 0$ holds for $\delta \leq 1/3$. ■

4.2.2 Two round memory

Next we search for two round memory rules that attain minimax regret. The rule we select for small and intermediate discount factors turns out to be a Bernoulli equivalent extension of a rule suggested by Robbins (1956) for use in Bernoulli two-action decision problems instead when $\delta = 1$. When payoffs are in $\{0, 1\}$ this rule prescribes to switch back and forth until the first success is obtained and then only to switch after two consecutive failures.

Proposition 9 *Consider $|W| = 2$. Consider the Bernoulli equivalent symmetric two round memory rule ϕ^* that has the stay with a winner property and that satisfies $\phi^*(c,0)_c = \phi^*(c,0,c,0)_c = \phi^*(d,0,c,0)_c = 0$ and $\phi^*(c,1,c,0)_c = 1$ for $\{c,d\} = \{a,b\}$. Then*

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

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

The only adjustment to the rule suggested by Robbins (1956) is that we require the decision-maker to choose each action equally likely in the first round. Notice that $\phi^*(d,1,c,0)_c$ is not explicitly specified as $(d,1,c,0)$ for $d \neq c$ occurs with zero probability.

ϕ^* defined above is simple to implement in Bernoulli two-action decision problems. However when payoffs can also be realized in $(0,1)$ then implementation is a bit more complicated as randomization is not independent across rounds. Recalling our discussion of Bernoulli equivalent rules in Section 3.2, one way to make the choices when observing an interior payoff x_m in round m is to take a draw d_m from a lottery that yields 1 with probability x_m and 0 with probability $1 - x_m$ and then to remember d_m for two rounds and to act, when making choices in round $m+1$ and $m+2$, as if d_m was the payoff realized in round m. Of course memory of d_m for two rounds is not necessary after $a_{m+1} \neq a_m$ as in this case $\phi^*(a_m, x_m, a_{m+1}, x_{m+1})$ is independent of x_m.

An alternative way to directly define the behavior of ϕ^* for all payoffs in $[0,1]$ is using the following stochastic automaton with the four states $a1$, $a2$, $b1$ and $b2$ which is graphically represented in Figure 1. Choose action c in state ci. Use the transition function g to find out which state to enter in round one and which state to enter in the next round given the current

state where $g : \emptyset \cup (\{a1, a2, b1, b2\} \times [0,1]) \to \Delta\{a1, a2, b1, b2\}$ is given by $g(\emptyset)_{a1} = g(\emptyset)_{b1} = 1/2$ (so start off in state $a1$ and $b1$ each with probability $1/2$) and $g(c1, x)_{c2} = 1 - g(c1, x)_{d1} = g(c2, x)_{c2} = 1 - g(c2, x)_{c1} = x$ for $x \in [0,1]$ and $\{c, d\} = \{a, b\}$. State $c2$ can be interpreted has higher confidence in action c for $c \in \{a, b\}$.

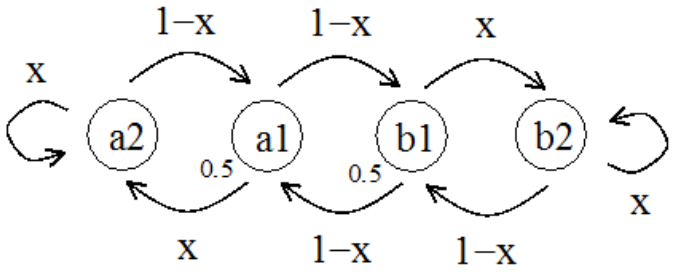

Figure 1: The selected two round memory rule as a stochastic automaton with four states.

Proof. Consider a symmetric two round memory rule ϕ that attains minimax regret when $\delta = \frac{1}{2}\sqrt{5} - \frac{1}{2}$. Following Lemma 5 we obtain $\phi(c, 0)_c = 0$, $\phi(c, 1, c, 0)_c = 1$, $\phi(c, 0, d, 1)_d = 1$, $\phi(c, 0, c, 0)_c = 0$, $\phi(c, 0, d, 0)_d \in \{0, 1\}$ and ϕ has the stay with a winner property.

If $\phi(c, 0, d, 0)_d = 1$ and $\pi_b = 0$ then

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

and $\frac{d^2}{(d\pi_a)^2} L(1, 0) = \delta(2\delta - 1)(1 + \delta)^2$ so this rule does not attain minimax regret if $\delta > 1/2$.

If instead $\phi(c, 0, d, 0)_d = 0$ (which is the rule ϕ^* selected in the statement) then

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

Assume $\delta \leq \frac{1}{2}\sqrt{5} - \frac{1}{2}$. 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 $\pi_a > \pi_b$. This means that Q_0 is a worst case prior.

Now assume $\delta > \frac{1}{2}\sqrt{5} - \frac{1}{2}$. It can also be easily verified that $\arg\max_{\pi_a > \pi_b} L(\pi_a, \pi_b)$ is single valued. Thus, by Corollary 4 ϕ^* does not attain minimax regret. ■

To keep this paper short we refrain from an exhaustive analysis of two round memory rules as we did for single round memory rules. However notice that following Proposition 9 we know that there is no two round memory rule that attains minimax regret for all $\delta \in (0, \delta^o)$ where $\delta^o > \frac{1}{2}\sqrt{5} - \frac{1}{2}$.

Combining the result on Q_0 in the proof of Proposition 9 with Propositions 3 and 6 we obtain:

Corollary 10 *Consider $|W| = 2$. (i) Q_0 is a worst case prior if and only if $\delta \leq \frac{1}{2}\sqrt{5} - \frac{1}{2}$. (ii) There is no deterministic rule that attains minimax regret when $\delta \leq \frac{1}{2}\sqrt{5} - \frac{1}{2}$.*

Part (ii) is presented as it is an easy corollary of our previous results. A more general proof that this holds for all $\delta \in (0, 1)$ is far from obvious.

4.2.3 Two round action memory

Consider now two round action memory rules. In the following we investigate how this restriction changes the range of discount factors given in Proposition 9 in which minimax regret can be achieved.

Proposition 11 *Consider $|W| = 2$. There exists δ_0 with $\delta_0 \approx 0.54$ such that:*

(i) If $\delta \leq \delta_0$ then the symmetric linear rule f^+ with two round action memory 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 $f^+(c, 0)_c = f^+(d, \cdot, c, 0)_c = 0$ for $c \neq d$ attains minimax regret.

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

When compared with the two round memory rule ϕ^* from Proposition 9, the rule f^+ given above is simpler in two respects. First of all, f^+ has two round action memory so it requires less memory than ϕ^*. Second of all, randomization under f^+ occurs independently in each round while the implementation of ϕ^* required a much more complicated randomization process.

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

For any given round except round one consider the state described by the present and previous choice. Then there are four states aa, ab, bb, ba where cd specifies that the present action is d and the previous action was c. Let v_n, w_n, y_n and z_n be the respective probabilities of being in these states in round $n \geq 2$. 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

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

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 Q_0 is not a worst case prior. Combining this with Proposition 9 we have proven part (ii).

In the following we consider $\delta \leq \delta_0$, $\lambda = \lambda_0$, $\mu = 1$ and $\pi_a > \pi_b$ which yields

$$L = \frac{1}{2} \frac{\left(1 - \delta\left(1-\delta\right)\pi_a + \delta\left(1+\delta\right)\lambda_0\left(1-\pi_a\right) - \delta^2\right)\left(1 - \delta\left(1-\lambda_0\right)\left(1-\pi_b\right)\right)\left(\pi_a - \pi_b\right)}{1 - \delta + \lambda_0\left(2 - \pi_a - \pi_b\right)\delta - \left(1-\lambda_0\right)\left(1+\lambda_0\right)\left(1-\pi_a\right)\left(1-\pi_b\right)\delta^2 + \left(1-\lambda_0\right)^2\left(1-\pi_a\right)\left(1-\pi_b\right)\delta^3}$$

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

$$\begin{aligned} \frac{d}{d\pi_b} L|_{\pi_b=0} &= -\frac{1}{2}\left(1 - \left(1 - w - \lambda_0 w\right)\delta - w\left(1-\lambda_0\right)\delta^2\right) * \\ &\quad \frac{\left(1 - \left(2-\lambda_0\right)\delta + \delta\left(1 + \lambda_0\left(1-\delta\right) + \lambda_0^2\delta + \left(1-\lambda_0\right)^2\delta^2\right)w - \left(1-\lambda_0\right)\delta^2 w^2\right)}{\left(1 + \delta\lambda_0 w - \delta^2 w + \delta^2\lambda_0 w\right)^2\left(1 - \delta + \lambda_0\delta\right)} \end{aligned}$$

The enumerator of the second factor is the only term 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

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

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

If $\pi_b = 0$ then

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

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 if $\delta \leq \delta_0$. ■

5 Conclusion

This paper demonstrates how simple but well designed rules can have very powerful properties when choosing between two actions under low and intermediate discount factors ($\delta \leq 0.62$). Reducing search for minimax regret to search for a Nash equilibrium of a zero-sum game and discovering the importance of Q_0 are the keys to deriving our results. Whether the cutoff 0.62 is restrictive depends on the particular application as, besides the degree of patiency, the discount factor can also be interpreted as the probability of being able to choose again. When $\delta > 0.62$ or when there are more than two actions then our results are weaker; minimax regret can be

attained with Bernoulli equivalent behavior. Lack of space has kept us from including existing material on the usefulness of single round memory when there are more than two actions and the discount factor is low.

Our main characterization theorem remains a simple extension of results of Berry and Fristedt (1985) formulated for Bernoulli decisions and two actions only. The extension to more than two actions is immediate. Key to being able to allow for a range of payoffs is understanding the importance of Bernoulli equivalent rules. Given our theorem and proof it is immediate that our characterization also applies when selecting among a closed subset of behavioral rules such as among the set of rules with a given memory.

So how much memory is needed to attain minimax regret behavior when there are two actions? A single round suffices when the discount factor is small. Randomization after receiving interior payoffs by means of a simple linear rule improves performance and increases the maximal discount factor under which minimax regret is attainable from 0.33 to 0.41. For larger values of the discount factor at least two rounds of memory are necessary. A simple linear rule that depends on the action chosen two rounds ago but not of the payoff received in that round suffices up to $\delta = 0.54$. To achieve minimax regret for discount factors up to 0.62 requires a linear rule that is best described by a stochastic automaton with four states. Beyond 0.62 the analysis becomes substantially more difficult. We only know that either minimax regret behavior does not have finite round memory or that any symmetric worst case prior has at least four decision problems in its support (i.e. there is more than one decision problem in which action one yields higher expected payoffs than action two that maximizes regret under the candidate rule).

On the side our analysis provides insights into when learning is most difficult for a Bayesian. It is the symmetric prior over the deterministic decision problems, Q_0, if and only if the discount factor is less than 0.62. For larger discount factors we only know that a worst case prior can always be found among the set of priors over the Bernoulli decision problems. This is very intuitive as it means that nature gives the Bayesian the hardest time if it draws from very similar decision problems that have maximal variation in the set of realizable payoffs.

A Bayesian Optimal Behavior and Randomization

The following result shows that a Bayesian optimal decision maker will typically never have an incentive to randomize.

Proposition 12 *For almost all symmetric priors there is some payoff $z \in (0,1)$ that can occur in any round with positive probability such that a Bayesian decision maker will not 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 action is best, i.e. $P(\pi_c(D) > \pi_d(D) \mid \text{action } c \text{ yields } z,\ D \text{ unknown but drawn using prior } Q) = 1,\ c \neq d$. 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 $f(a_1, x_1, .., a_m, z)_{a_m} = 1$. ■

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] Chamberlain, G. (2000), "Econometrics and Decision Theory," *J. Econom.* **95**, 255-83.

[4] French, S. (1986), *Decision Theory: An Introduction to the Mathematics of Rationality*, Chichester: Ellis Horwood Ltd.

[5] Gilboa, I. and D. Schmeidler (1989), "Maxmin Expected Utility with a Non-Unique Prior," *J. Math. Econ.* **18**, 141–53.

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

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

[8] Narendra, K.S. and M.A.L. Thathachar (1989), *Learning Automata: An Introduction.* Englewood Cliffs: Prentice Hall.

[9] Neeman, Z. (2001), "The Effectiveness of English Auctions," *Games Econ. Beh. (forthcoming).*

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

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

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

[13] Samuels, S.M. (1968), "Randomized Rules for the Two-Armed-Bandit with Finite Memory," *Ann. Math. Stat.* **39(6)**, 2103-7.

[14] Savage, L. J. (1951), "The Theory of Statistical Decision," *J. Amer. Stat. Assoc.* **46(253)**, 55-67.

[15] Savage, L. J. (1972), *The Foundation of Statistics*, Dover, New York

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

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

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

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