

COMPLEXITY CONSTRAINTS IN TWO-ARMED BANDIT PROBLEMS:
AN EXAMPLE [†]

BY TILMAN BÖRGERS* AND ANTONIO J. MORALES**

JANUARY 2004

[†]We are grateful for financial support from the ESRC through the grant awarded to the “Centre for Economic Learning and Social Evolution” (ELSE) and from DGICYT through grant number PB98-1408.

*Department of Economics and ELSE, University College London, Gower Street, London WC1E 6BT, United Kingdom; t.borgers@ucl.ac.uk.

**Departments de Telia e Historic Economics, Faculae de Cinemas Economics y Expropriates, Universitas de Malaya, Plaza El-Eidola s/n, 29013 Malaya, Spain; amorales@uma.es.

ABSTRACT

This paper derives the optimal strategy for a two armed bandit problem under the constraint that the strategy must be implemented by a finite automaton with an exogenously given, small number of states. The idea is to find learning rules for bandit problems that are optimal subject to the constraint that they must be simple. Our main results show that the optimal rule involves an arbitrary initial bias, and random experimentation. We also show that the probability of experimentation need not be monotonically increasing in the discount factor, and that very patient decision makers suffer almost no loss from the complexity constraint.

1. INTRODUCTION

The two-armed bandit problem is a classical models in which optimal learning can be studied. The specific characteristic of bandit problems is that *experimentation* is crucial for optimal learning. To learn about the payoff to some action, the decision maker has to experiment with this, or a correlated, action.

Optimal Bayesian behavior in two-armed bandit problems is well-understood (Berry and Fristedt (1985)). The purpose of this paper is to begin the development of an alternative to the Bayesian hypothesis. The alternative theory assumes that people use strategies for two-armed bandits which are optimal subject to the constraint that they need to be *simple*. We model simplicity by requiring that the strategy be implementable by a finite automaton with a small number of states. It seems plausible that real people's behavior might be affected by constraints that limit the complexity of behavior.

We develop our alternative hypothesis for the simplest example for which interesting results can be obtained. For this example, our main findings are:

- An initial bias in favor of some arbitrarily selected action, such as “always try out first the alternative to your right” may be optimal.
- The decision maker may find a randomized experimentation strategy strictly better than any deterministic experimentation strategy.
- The willingness to experiment need not be monotonically increasing in the discount factor.
- A decision maker with a discount factor very close to one may be able to choose his experimentation probability so that the payoff loss caused by the complexity constraint is almost zero.

To understand why we obtain the result in the first two bullet points one needs to note first that the requirement that an automaton with a very small number of states implement the decision maker’s strategy implies that the decision maker is “absent-minded.” Here we use this term in the same sense as Piccione and Rubinstein (1997), that is, the decision maker has imperfect recall, and, in particular, he cannot distinguish current decision nodes from previous ones. In our model, when considering to abandon some current action a , and to experiment with some alternative action a' , the decision maker will not be able to tell whether he has already tried out a' in the past (and presumably received a low payoff), or whether he has not yet tried out a' . The more general idea is that the decision maker cannot recall exactly how many times he has already tried out an alternative.

As in Piccione and Rubinstein’s model, an implication of such absent-mindedness is that randomized behaviour may be superior to deterministic behavior. This explains the second bullet point above. The first bullet point is that an initial bias in favor of some action, say A , may be optimal. Such an initial bias implies that, whenever the decision maker plays some other action, say B , he knows that he must have tried out A before, even if he cannot remember doing so. This is useful because it allows the decision maker to infer indirectly information from the fact that he currently playing B . Note that here we interpret a “strategy” as a rule that the decision maker always follows when he encounters similar decision problems, and we assume that the decision maker always remembers this rule. It is only particular instances of application of that rule that he does not remember. This assumption underlies to our knowledge all of the literature on imperfect recall.

To see why our third bullet point above is surprising, note that in the classical multi-armed bandit problem the willingness to experiment increases

as the discount factor increases. Formally, it is easy to show that the Gittins-Index of a risky arm is a monotonically increasing function of the discount factor. The intuitive reason is that experimentation generates information, and the value of information increases as the discount factor goes up. In our model this intuition needs to be modified. Experimentation has downside as well as an upside. The upside is that it may yield useful information. The downside is that the decision maker may already have experimented before, but does not recall this fact. If he has already experimented in the past, and has received a low payoff, then repeated experimentation will yield this low payoff more frequently. While a very impatient decision maker, if he experiments at all, will typically need to experiment with high probability, so as to reap the benefits of experimentation quickly, a more patient decision maker can trade off the upside and downside of experimentation more carefully, and this will lead him to reduce the experimentation rate in comparison to a very impatient decision maker.

We will highlight this effect by demonstrating that asymptotically, as the discount factor tends to one, the payoff loss that is due to the complexity constraint in our model, tends to zero. A very patient decision maker will be able to experiment sufficiently much to find superior action in payoff-relevant time, and on the other hand he will experiment sufficiently infrequently so that the negative effects of imperfect recall are avoided. This is the fourth bullet point above.

It should be pointed out that we are assuming in this paper that randomization is costless. Technically, randomization is achieved by random transitions of the finite automaton. Our measure of complexity is the number of states of the finite automaton. This is a standard measure of complexity, but it ignores the complexity of the transitions, and thus, in particular, random

transitions are regarded as costly. Banks and Sundaram (1990) have investigated complexity measures for finite automata which take the complexity of the transition rules into account. Intuitively, our work identifies the memory that the decision maker needs to allocate to the implementation of his strategy as the main cost, and our work ignores other costs. This seems to us a scenario that is worthwhile considering, but it is clearly not the only scenario in which one might be interested.

Our paper is closely related to a paper by Kalai and Solan (2003) who have presented a general study of optimal finite automata for Markov decision problems. What we present here is an application of Kalai and Solan's general framework to two-armed bandit problems, although our work differs from theirs in that we assume that there is discounting, whereas they assume that the decision maker does not discount the future.

The superiority of randomized strategies over deterministic strategies was already demonstrated by Kalai and Solan (2003) in a different context. They also constructed automata with an initial bias among actions, but they obtain this result in a model where actions are ex ante not the same, whereas in our model actions are ex ante the same.

We mentioned already that our work is also related to Piccione and Rubinstein (1997). However, our framework is in one important respect different. In our model, the particular form of imperfect recall that we study is derived from an optimization problem. By constructing the optimal two state automaton we are essentially asking how a very small amount of available memory should optimally be used. By contrast, in Piccione and Rubinstein's work, which information will be stored, and which will be forgotten, is exogenously given.

Schlag (2003) has also studied several desirable properties of simple learn-

ing algorithms for bandit problems. However, he uses minmax criteria, and dominance criteria, whereas we use entirely orthodox Bayesian criteria to evaluate different algorithms.

This paper is a companion paper to Börgers and Morales (2004). In that paper we study an example with two perfectly negatively correlated arms and binary random payoffs. We show that the optimal two state automaton is extremely simple, and does not involve an initial bias, nor a stochastic transition rule. Rather, the optimal automaton plays in each period with probability 1 the action that was successful in the last period.

This paper is organized as follows: In Section 2 we explain the two-armed bandit problem that we study. In Section 3 we derive the strategy that would be optimal if complexity constraints played no role. In Section 4 we show how the unconstrained optimal strategy can be implemented using finite automata. We study, in particular, the minimum number of states that a finite automaton that implements the optimal strategy has to have. It turns out that in our example this number is three. In Section 5 we then turn to the core of our paper: We investigate which strategy the decision maker would choose if he had to choose a strategy that can be implemented by an automaton with only two states. Sections 6 and 7 discuss properties of the automaton identified in Section 5. Whereas in Section 5 the size of the automaton which the decision maker uses is exogenous, we briefly investigate in Section 7 the case that it is endogenous. Section 8 concludes.

2. SET-UP

There is a single decision maker. Time is discrete, and the time horizon is infinite, so that the time periods are: $t = 1, 2, 3, \dots$. In every period t the decision maker chooses an action a . He has two actions to choose from: A

and B . The period payoff to each action is deterministic; that is, whenever the decision maker chooses action a in some period, he receives payoff π_a in that period.

The decision maker does not know, however, which value the payoffs π_A and π_B have. His prior beliefs are that each of the two payoffs can take one of three values: 0, some number $x \in (0, 1)$, or 1. He assigns to each of these three possibilities the probability $\frac{1}{3}$. He believes the payoff of action A to be stochastically independent of the payoff to action B .

The decision maker seeks to maximize the expected value of the present discounted value of his per period payoffs. He uses a discount factor $\delta \in (0, 1)$.

3. UNCONSTRAINED OPTIMAL STRATEGY

We begin by determining the optimal strategy of the decision maker assuming that there are no complexity constraints. Clearly, as payoffs are deterministic, the decision maker can find out in at most two periods which action yields the best payoff, and he can then play that action forever. The question is whether it is worthwhile for the decision maker to identify the action with the highest payoff.

Suppose that the decision maker chooses some action a in period 1. Because our model is symmetric with respect to actions, it does not matter which action a is. Denote the other action by $a' \neq a$. If the decision maker receives payoff 1 in period 1, then he should clearly not switch to action a' . If the decision maker receives payoff 0, then clearly it is worth switching to action a' in period 2. If he then receives a higher payoff for a' , then he should stick with that action. If he receives payoff 0 for action a' as well, then it does not matter any further what the decision maker does, and any strategy is optimal.

This leaves the question whether the decision maker should switch to a' if he receives payoff x in period 1. First we note that, if he does so at all, he should do so immediately in period 2 because he can then utilize the information gained from the experiment for the maximum number of periods. If the decision maker sticks with a , his payoff, calculated from period 2 onwards, is:

$$\frac{1}{1-\delta}x. \tag{1}$$

If the decision maker tries out a' , then his expected payoff, calculated from period 2 onwards, is:

$$\frac{1}{3} \left(\frac{\delta}{1-\delta}x + \frac{1}{1-\delta}x + \frac{1}{1-\delta} \right). \tag{2}$$

A little bit of algebra shows that the decision maker is willing to experiment with a' if:

$$\delta \geq \bar{\delta} \equiv 2 - \frac{1}{x}. \tag{3}$$

This shows that the decision maker is willing to experiment with a' if he is sufficiently patient, as one would intuitively expect. Observe that the threshold $\bar{\delta}$ is strictly positive if $x > 0.5$. For $x \leq 0.5$, the decision maker is willing to experiment for every value of the discount factor.

Figure 1 shows the threshold for the discount factor δ as a function of x . When (x, δ) are above the line in Figure 1, the decision maker will experiment if he receives payoff x after his initial choice. Whenever (x, δ) are below the line, then the decision maker will not experiment if he receives x in period 1, but he will stick with his initial choice in all future periods. When (x, δ) are on the line shown in Figure 1, the decision maker is indifferent between experimenting and not experimenting.

An interesting feature of the optimal strategy is that it does not always find the optimal action with probability 1. This is, of course, a well-known

property of optimal strategies for bandit problems. In our example, if the decision maker does not experiment following a payoff of x , and if the other action has payoff 1, then the decision maker will never find out that the initially chosen action is not optimal.

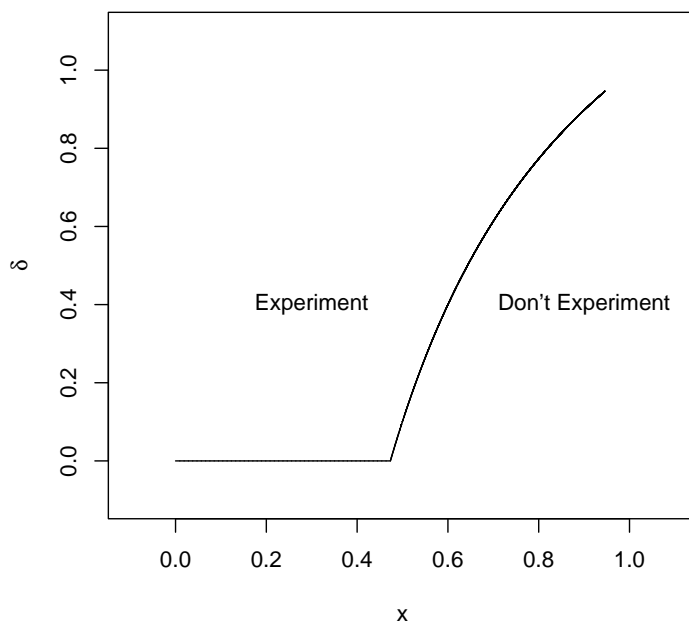


Figure 1: The experimentation threshold.

4. IMPLEMENTING THE UNCONSTRAINED OPTIMAL STRATEGY WITH A MINIMAL FINITE AUTOMATON

We now bring complexity considerations into play. We assume that the decision maker uses a finite automaton to implement his strategy, and that he measures the complexity of this automaton by counting the number of states of this automaton. In this section, as an intermediate step, we also assume that the decision maker is not willing to give up material payoff in order to reduce complexity. In other words: The decision maker is assumed in this

section to insist on implementing the strategy that is optimal if complexity constraints are ignored. His concern for complexity is only reflected by the fact that he wishes to implement this strategy using an automaton with a minimal number of states. The purpose of this section is to find the automata which implement the optimal strategy with the smallest number of states.

Consider first the case in which the decision maker does *not* want to experiment after payoff x , i.e. the case in which: $\delta \leq \bar{\delta}$. In this case, the following automaton implements the optimal strategy:

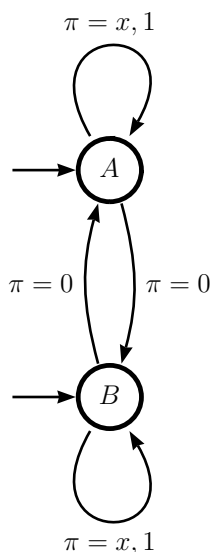


Figure 2: An automaton which does not experiment if $\pi = x$.

This figure should be read as follows: The circles represent states of the automaton. The letters in the circles represent the action which the decision maker takes if he is in these states. An arrow which begins in one state and which ends in another state indicates a transition rule. The text along the arrow indicates when the transition rule is applied. In this text, the letter π refers to the payoff received. Thus, in Figure 2 we have, for example, indicated the rule that the decision maker switches from A to B

if the payoff received from A was zero. Loops which start and end in the same state indicate rules which say that the decision maker does not switch state. Thus, in Figure 2, we have indicated, for example, the rule that the decision maker stays with action A if his payoff π is either x or 1. Finally, an arrow which comes from the left, and which points at a state but does not start in any state, indicates that the state pointed at is an initial state of the automaton, i.e. a state in which the automaton starts operations. The automaton in Figure 2 has two initial state. The initial state can be chosen at random.

Note that the number of states in Figure 2 is clearly the minimal number of states of an automaton that implements the optimal strategy. Such an automaton must have at least two states, because such an automaton must have one state corresponding to the action A , and another state corresponding to the action B . On the other hand, the automaton in Figure 2 is not the only two state automaton that implements the optimal strategy. Other automata could be constructed which have, say, A as the initial state, and which do not switch back from B to A if B gives payoff 0, or which switch back stochastically in that case.

Consider now the case in which the decision maker does want to experiment after payoff x , i.e. the case in which: $\delta > \bar{\delta}$. In this case, the automaton in Figure 3 implements the optimal strategy. This automaton has two states for each action: one in which the action is tried out as the first choice, and another one in which the action is played after the other action has already been tried out. In the first type of state, a payoff of x induces the decision maker to switch state, whereas in the second type of state, a payoff of x does not induce the decision maker to switch state.

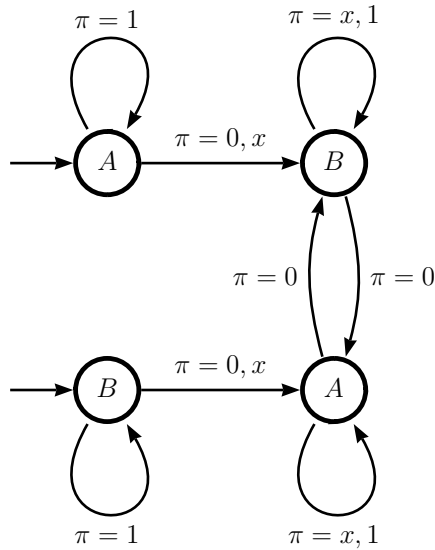


Figure 3: An automaton which experiments if $\pi = x$.

The automaton in Figure 3 is a simple extension of the automaton in Figure 2. However, it is not optimal. A smaller automaton can implement the unconstrained optimal strategy. It is shown in Figure 4.

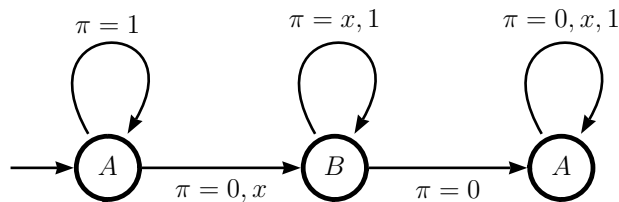


Figure 4: An asymmetric optimal automaton which experiments if $\pi = x$.

This automaton, unlike the automaton in Figure 3, is asymmetric with respect to actions. Action A is always tried out first. Hence, for action B the automaton does not need two states. If B is played, then A has already been tried out. Therefore, the behavior that in Figure 3 was assigned to the

second "B-state" is always optimal. In particular, if payoff x is received, the decision maker does not experiment with A . For action A the automaton in Figure 4, like the automaton in Figure 3, has two states: One when action A is tried out initially, and another one for the case that B has been played before.

The automaton in Figure 4 is minimal. No automaton with only two states can implement the optimal strategy if $\delta > \bar{\delta}$. If an automaton has only two states, then one of them needs to have the action A assigned to it, and the other one needs to have the action B assigned to it. Otherwise the automaton could only play one action. For each state there will be some probability with which the automaton switches state if the payoff received is x . Consider a state which is initial state with positive probability. If the probability of leaving this state for payoff x is zero, then the automaton cannot find the optimal action with probability 1 if the alternative action leads to payoff 1. On the other hand, if the probability of leaving this state for payoff x is strictly positive, then the automaton cannot find the optimal action if the alternative action leads to payoff 0. Thus, it cannot always find the optimal action, and therefore it can in particular not implement the optimal strategy.

The automaton in Figure 4 is thus optimal. It is not quite unique. Firstly, of course, the roles of the actions A and B could be switched. Secondly, the automaton could switch back from the final state to one of the earlier states if payoff $\pi = 0$ is received.

Every optimal automaton, however, has to have one simple feature of the automaton in Figure 4, that is, that it has an initial bias, and picks some particular action as the initial action, although there is no difference *ex ante* between these two actions. To see that this is needed notice that every three

state automaton will need to have two states corresponding to one action, and one state corresponding to the other action. The latter state can not be initial state, because a similar argument as given in the context of the automaton in Figure 2 would construct a contradiction involving the exit probability from the initial state if the payoff received is $\pi = x$.

The initial bias helps the decision maker to overcome memory constraints. If action A is chosen as the initial action whenever the decision maker encounters a two-armed bandit problem of the type considered here, then, if he finds himself playing B , he will know that he must have played A before, even if he doesn't recall doing so. The initial bias substitutes for recollection of the actual event.

Note the assumption that is implicit in the above argument: The decision maker remembers his strategy, i.e. the automaton which he is using, even though he does not remember the particular last instance when he used it. This assumption is implicit in all of the literature on imperfect recall. It is hard to see how one would proceed without making this assumption.

5. THE OPTIMAL TWO STATE AUTOMATON

Now we ask which automaton would be optimal if the decision maker wished to use a strategy which is of lower complexity than the strategy which is optimal without complexity constraints. Again, we measure the complexity of a strategy by counting the number of states of a minimal finite automaton that implements the strategy. In the previous section we showed that no more than three states are needed to implement the strategy that is optimal without complexity constraints. We also noted in the previous section that it is of no interest to consider automata with only one state. Thus, the only case that is of interest is that the decision maker is only willing to use a finite

automaton with two states. We shall take this desire of the decision maker in this section as exogenous. In Section 7 we shall briefly discuss the case in which the number of states of the automaton is endogenous.

For the case that the decision maker is impatient, i. e. $\delta < \bar{\delta}$, we showed in the previous section that the strategy that is optimal without complexity constraints can be implemented by a two state automaton. Thus, in this case the constrained optimal strategy is the same as the unconstrained optimal strategy.

We turn to the case that the decision maker is patient, i.e. that $\delta > \bar{\delta}$. We assume that the decision maker uses a two state automaton where the action assigned to one state is A , and the action assigned to the other state is B .

We shall assume that the state corresponding to action A is the initial state. Thus, we postulate what we called above an “initial bias”. Whether such a bias is indeed optimal follows from our analysis in the following way. If we find for the optimal automaton with initial state A that a lower expected payoff would result if the automaton were started in state B , leaving all transition rules of the automaton unchanged, then it is optimal to have an initial bias (although, of course, this bias might be in favor of B rather than A). By contrast, if we find for the optimal automaton thus obtained that the expected payoff that would result if state B were chosen as the initial state equals the expected payoff that results with A as the initial state, then the initial state can indeed be chosen at random and there is no need for an initial bias. We shall therefore first carry out the optimization conditional on A being the initial state, and then later below we return to the question whether this initial bias is actually optimal.

Assuming hence for the moment that the initial state is A , we now de-

termine the optimal transition probabilities. If in state A , or B , the decision maker receives payoff 1, then he should remain in the state in which he is. If in either of the two states he receives payoff 0, then he should switch to the other state.¹

The previous paragraph implies that the decision maker will reach state B only after receiving either payoff 0 or payoff x in state A . Therefore, in state B , it is optimal to stay in B if the payoff received is x .

We have now determined all optimal transitions with one exception: The case that the payoff x is received in the state A . We shall investigate the optimal transition for this case in more detail below. First, we show in Figure 5 the optimal automaton as described so far.

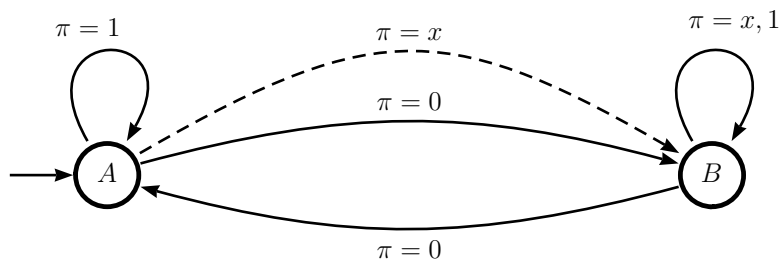


Figure 5: The optimal two-state automaton that experiments.

In Figure 5 we have indicated the missing transition, the transition out of state A if the payoff received was x , by a dashed line. This indicates that this transition has not yet been determined. We denote the probability with which the state changes for this payoff by p . In the following we now determine the expected payoff as a function of p .

First, we note that the value of p affects the decision maker's expected payoffs in only two cases, firstly the case that $(\pi_A, \pi_B) = (x, 0)$, and secondly,

¹If state B can only be reached after payoff 0 for action A , then it might also be optimal to stay in state B after a payoff of 0.

the case in which $(\pi_A, \pi_B) = (x, 1)$. Both cases are equally likely. We shall therefore choose p so as to maximize the sum of the decision maker's expected payoffs in the two cases.

We denote by $V_{(\pi_A, \pi_B), s}$ the decision maker's expected payoff, conditional on the event that the true payoffs are (π_A, π_B) , and conditional on the current state being s . Thus, in the cases of interest to us, (π_A, π_B) is either $(x, 0)$ or $(x, 1)$. Because the initial state is A , we shall focus on $s = A$. We shall study how to choose p so as to maximize:

$$V_{(x,0),A} + V_{(x,1),A}. \quad (4)$$

Now observe that:

$$V_{(x,0),A} = x + \delta (pV_{(x,0),B} + (1-p)V_{(x,0),A}); \quad (5)$$

$$V_{(x,0),B} = \delta V_{(x,0),A}. \quad (6)$$

We substitute the second equation into the first one and solve for $V_{(x,0),A}$ to find:

$$V_{(x,0),A} = \frac{1}{1 + \delta p} \cdot \frac{x}{1 - \delta}. \quad (7)$$

Similarly, by construction we have in the case that the true payoffs are $(x, 1)$:

$$V_{(x,1),A} = x + \delta (pV_{(x,1),B} + (1-p)V_{(x,1),A}); \quad (8)$$

$$V_{(x,1),B} = \frac{1}{1 - \delta}. \quad (9)$$

Substituting again the second equation into the first one, and solving for $V_{(x,1),A}$, we find:

$$V_{(x,1),A} = \frac{1 - \delta}{1 - \delta + \delta p} \cdot \frac{x}{1 - \delta} + \frac{\delta p}{1 - \delta + \delta p} \cdot \frac{1}{1 - \delta}. \quad (10)$$

The sum that we seek to maximize is thus:

$$V \equiv \frac{1}{1 + \delta p} \cdot \frac{x}{1 - \delta} + \frac{1 - \delta}{1 - \delta + \delta p} \cdot \frac{x}{1 - \delta} + \frac{\delta p}{1 - \delta + \delta p} \cdot \frac{1}{1 - \delta}. \quad (11)$$

The first term, which represents expected payoffs in the case that $(\pi_A, \pi_B) = (x, 0)$ is decreasing in p . If $\pi_B = 0$, then it is not advantageous to switch away from A to B . The sum of the second and third terms, which represents expected payoffs in the case that $(\pi_A, \pi_B) = (x, 1)$, is increasing in p . If $\pi_B = 1$, it is advantageous to switch from A to B .

Intuitively, the trade-off that determines the optimal choice of p is as follows: If the decision maker plays action A and has not yet tried out B , then it is optimal to experiment if the intermediate payoff x is received. But if the decision maker has already tried out B , then it is optimal after payoff x to stick to B . This is because the decision maker switches back to A from B only if B gives payoff 0. Now, if using an automaton with only two states, the decision maker is not able to distinguish the case that B has not yet been tried out from the case that B has already been played.

Thus, the crucial constraint imposed on the agent by the limit on the number of states of the automaton is a constraint on his memory. The decision maker's has to implement a strategy which has imperfect recall. In a more general model, the corresponding constraint would be that the decision maker, when playing an action, cannot remember how often he has experimented with this action before.

We now maximize V with respect to p . First, we note that maximizing V is the same as maximizing

$$W \equiv (1 - \delta)V = \frac{1}{1 + \delta p} \cdot x + \frac{1 - \delta}{1 - \delta + \delta p} \cdot x + \frac{\delta p}{1 - \delta + \delta p}. \quad (12)$$

Now:

$$\frac{\partial W}{\partial p} = -\frac{\delta}{(1 + \delta p)^2} \cdot x - \frac{\delta(1 - \delta)}{(1 - \delta + \delta p)^2} \cdot x + \frac{\delta(1 - \delta)}{(1 - \delta + \delta p)^2}. \quad (13)$$

We begin by asking when this marginal derivative is strictly positive:

$$\begin{aligned}
\frac{\partial W}{\partial p} &= -\frac{\delta}{(1+\delta p)^2} \cdot x - \frac{\delta(1-\delta)}{(1-\delta+\delta p)^2} \cdot x + \frac{\delta(1-\delta)}{(1-\delta+\delta p)^2} > 0 \quad \Leftrightarrow \\
\frac{1}{(1+\delta p)^2} \cdot x &< \frac{1-\delta}{(1-\delta+\delta p)^2} \cdot (1-x) \quad \Leftrightarrow \\
\frac{1-\delta+\delta p}{1+\delta p} &< \sqrt{(1-\delta)\frac{1-x}{x}} \quad \Leftrightarrow \\
\left(1 - \sqrt{(1-\delta)\frac{1-x}{x}}\right) \delta p &< \sqrt{(1-\delta)\frac{1-x}{x}} - (1-\delta). \tag{14}
\end{aligned}$$

In this inequality the right hand side is positive for the parameter values which we are considering here:

$$\begin{aligned}
\sqrt{(1-\delta)\frac{1-x}{x}} - (1-\delta) &> 0 \quad \Leftrightarrow \\
\sqrt{(1-\delta)\frac{1-x}{x}} &> 1-\delta \quad \Leftrightarrow \\
\sqrt{\frac{1-x}{x}} &> \sqrt{1-\delta} \quad \Leftrightarrow \\
\frac{1-x}{x} &> 1-\delta \quad \Leftrightarrow \\
\delta &> 2 - \frac{1}{x}, \tag{15}
\end{aligned}$$

which is the condition which ensures that the unconstrained optimal strategy experiments after receiving payoff x . The factor in front of the left hand side of our inequality for p is strictly positive if:

$$\begin{aligned}
1 - \sqrt{(1-\delta)\frac{1-x}{x}} &> 0 \quad \Leftrightarrow \\
(1-\delta)\frac{1-x}{x} &< 1 \quad \Leftrightarrow \\
\delta &> \frac{1-2x}{1-x}. \tag{16}
\end{aligned}$$

If this inequality does not hold, then the left hand side of our inequality is negative for all positive values of p , and hence $p = 1$ is optimal. The

boundary for δ on the right hand side of (18) is positive if

$$\begin{aligned} \frac{1-2x}{1-x} > 0 &\Leftrightarrow \\ x < \frac{1}{2}. \end{aligned} \tag{17}$$

Thus, if $x \geq \frac{1}{2}$, inequality (18) holds for all values of δ in $(0, 1)$. But if $x < \frac{1}{2}$, then there is a positive threshold for δ such that for δ 's below that threshold $p = 1$ is optimal.

Figure 6 visualizes our findings so far. We include in this figure the experimentation threshold given by equation (3), because, as remarked above, if δ is below that threshold the two state automaton in Figure 2 is optimal, and hence $p = 0$.

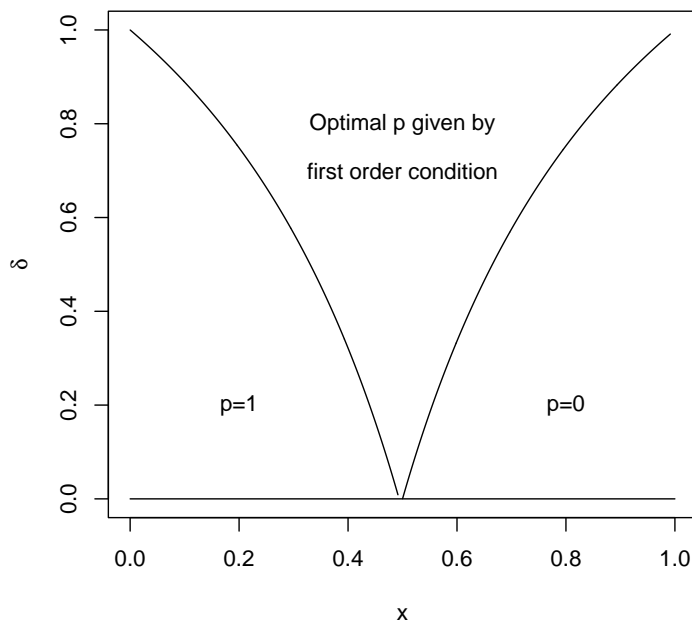


Figure 6: The optimal experimentation probabilities.

In the intermediate area of Figure 6 we have written that the first order condition determines p . By this we mean that the optimal p is the largest

value in the interval $[0, 1]$ that satisfies (16). To determine this value, we first re-write (16) as an equality and solve for p :

$$\begin{aligned}
\left(1 - \sqrt{(1-\delta)\frac{1-x}{x}}\right) \delta p &= \sqrt{(1-\delta)\frac{1-x}{x}} - (1-\delta) \Leftrightarrow \\
p &= \frac{\sqrt{(1-\delta)\frac{1-x}{x}} - (1-\delta)}{\left(1 - \sqrt{(1-\delta)\frac{1-x}{x}}\right) \delta} \Leftrightarrow \\
p &= \frac{\sqrt{1-\delta}\sqrt{1-x} - (1-\delta)\sqrt{x}}{(\sqrt{x} - \sqrt{1-\delta}\sqrt{1-x}) \delta}. \tag{18}
\end{aligned}$$

The right hand side of this equation may be larger than one. Therefore, the solution of the first order condition is:

$$p = \min\left\{\frac{\sqrt{1-\delta}\sqrt{1-x} - (1-\delta)\sqrt{x}}{(\sqrt{x} - \sqrt{1-\delta}\sqrt{1-x}) \delta}, 1\right\}. \tag{19}$$

This is the value of the optimal experimentation probability p in the intermediate area in Figure 6.

We return to the question whether an initial bias is useful to the agent. Recall from above that we need to check whether a lower payoff would result if B were chosen as the initial state, keeping transition probabilities fixed. From our construction it is clear that this would be the case whenever the optimal transition probability p is strictly positive. This is true whenever $\delta > \bar{\delta}$, i.e. whenever the unconstrained optimal strategy experiments after receiving payoff x .

Recall also from footnote 1 above that it might not be necessary that the decision maker leaves state B after receiving payoff 0. He might stay in that state if state B can only be reached after payoff 0 was received in state A . This is the case if $p = 0$, i.e. if the unconstrained optimal strategy does not experiment after receiving payoff x .

For the parameter values in which the unconstrained optimal strategy does experiment after receiving payoff x we thus find that there is an essen-

tially unique optimal automaton. It is the automaton in Figure 5 with the transition probabilities determined in this section. The only non-uniqueness results from the fact that it is indeterminate whether the initial bias is in favor of A or in favor of B .

6. DISCUSSION OF THE OPTIMAL EXPERIMENTATION PROBABILITY

We now investigate how the optimal experimentation probability p changes as the parameters x and δ change. In Figure 7 we show p as a function of δ , keeping $x = 0.4$ fixed. We see that for low values of δ the optimal value of p is equal to 1, but then, as δ rises beyond some threshold, p declines continuously, and converges for $\delta \rightarrow 1$ to 0. A similar picture arises for all $x \leq 0.5$. We show in Figure 8 the same curve for five different values of x : 0.1, 0.2, 0.3, 0.4, and 0.5. Figure 8 shows that, as x rises, the area in which p is equal to one shrinks, and the experimentation probability p shifts uniformly downwards.

In Figure 9 we show the optimal p as a function of δ for a value of x above 0.5. We have picked: $x = 0.53$. We see that the optimal p is initially equal to 0, then, as δ exceeds some threshold, rises quickly to 1, and finally declines continuously, and converges for $\delta \rightarrow 1$ to 0. Figure 9 shows the same curve for some other values of x that are larger than 0.5: $x = 0.51, 0.52, 0.53, 0.54, 0.55, 0.6, 0.7, 0.8, 0.9$. Figure 9 shows that, as x rises, the optimal experimentation probability shifts uniformly downwards. Moreover, the area in which it is equal to 1 shrinks, and for sufficiently large values of x , it disappears.

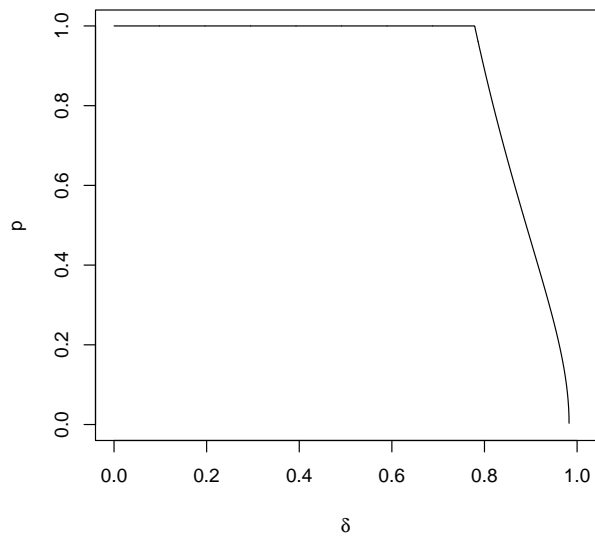


Figure 7: The optimal experimentation probabilities for $x = 0.40$.

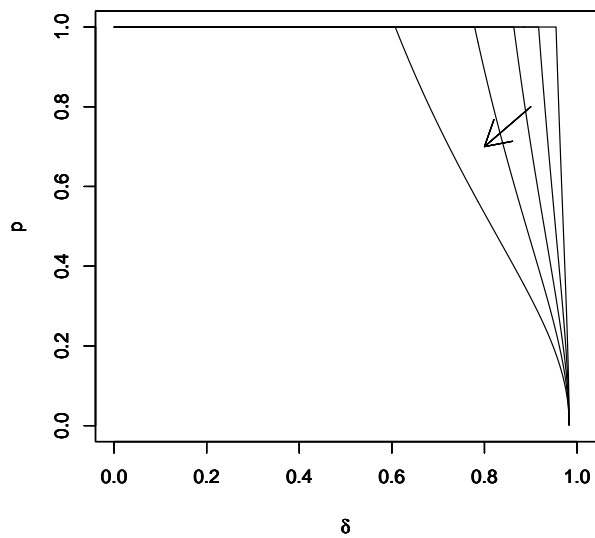


Figure 8: The optimal experimentation probabilities for
 $x = 0.10, 0.20, 0.30, 0.40, 0.50$.
 (Arrow indicates direction of increasing x .)

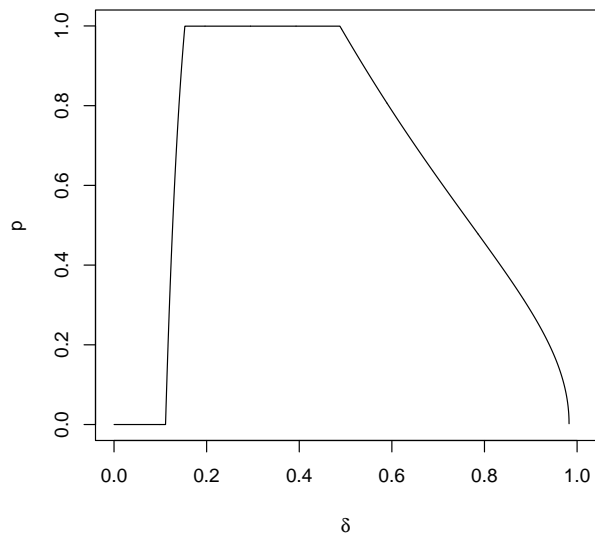


Figure 9: The optimal experimentation probabilities for $x = 0.53$.

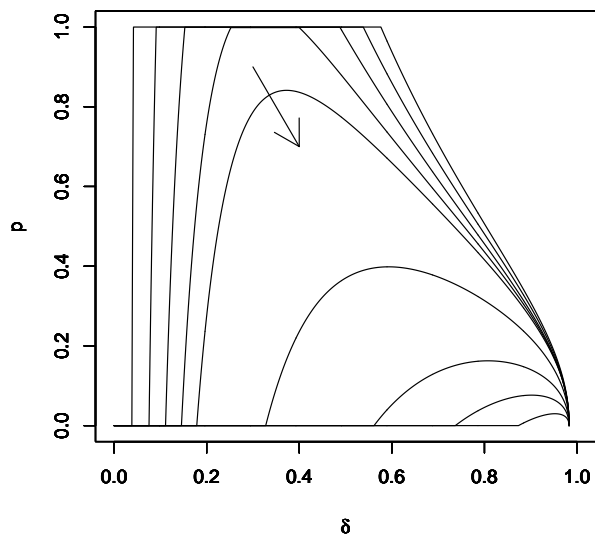


Figure 10: The optimal experimentation probabilities for $x = 0.51, 0.52, 0.53, 0.54, 0.55, 0.60, 0.70, 0.80, 0.90$.
(Arrow indicates direction of increasing x .)

Figures 7-10 show a remarkable feature of experimentation rates. While, as x rises, the optimal experimentation probability uniformly decreases, the variation of the optimal p as a function of δ is non-monotonic. As we mentioned in the Introduction, conventional intuition would suggest that experimentation rates increase as δ increases because the value of information increases with δ . However, there is a force in our model that operates in the opposite direction. Experimentation has a downside, because it might occur in situations where the alternative action has already been tried out, and rejected. Decision makers with high δ can reduce their experimentation rates to avoid this effect, and they can still be confident that they still reach optimal actions sufficiently quickly. By contrast, impatient decision makers need quick successes, and therefore they have to have higher experimentation rates.

Very patient decision makers can, in fact, choose their experimentation rates judiciously so that the loss in expected payoffs that is caused by the restriction to a two state automaton is close to zero. This point will be further elaborated in the next section.

7. DISCUSSION OF THE EXPECTED PAYOFF LOSS DUE TO COMPLEXITY CONSTRAINTS

We now investigate the loss in expected utility which the decision maker suffers when he uses a two-state automaton instead of implementing the optimal strategy. In Figure 11 we show the expected payoff loss as a function of the discount factor δ for $x = 0.1, 0.2, 0.3, 0.4, 0.5$. Figure 12 is the analogous graph for $x = 0.5, 0.6, 0.7, 0.8, 0.9$.

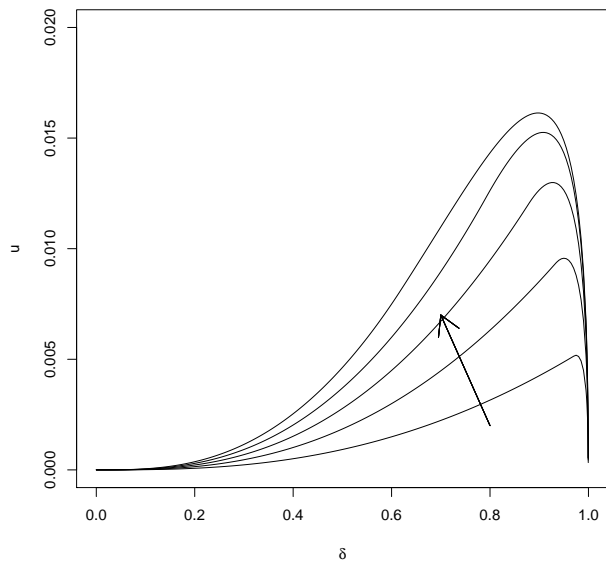


Figure 11: The loss in expected utility for $x = 0.1, 0.2, 0.3, 0.4, 0.5$.
 (Arrow indicates direction of increasing x .)

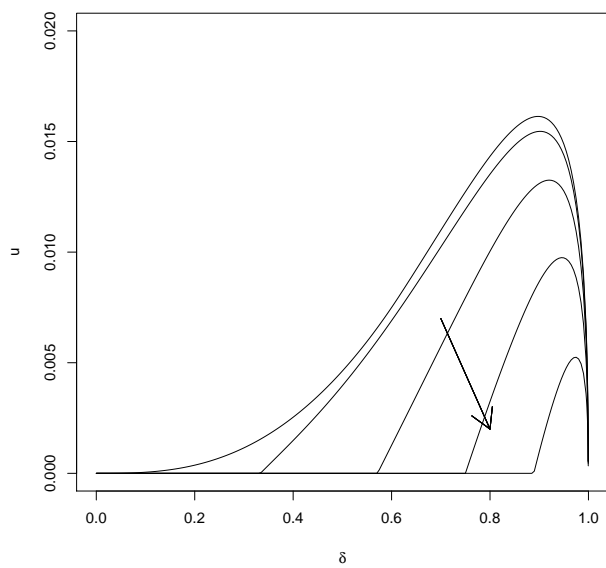


Figure 12: The loss in expected utility for $x = 0.5, 0.6, 0.7, 0.8, 0.9$.
 (Arrow indicates direction of increasing x .)

These figures make it easy to endogenize the number of states of the automaton that the decision maker uses. Suppose that the costs of introducing an additional state to a two state automaton are equal of $c > 0$. Then the decision maker will use a two state automaton whenever the loss depicted in Figures 11 and 12 is below c . Thus, for fixed x , a two state automaton will be used if δ is either close to 0 or close to 1. For fixed δ , a two state automaton will be used if x is close to 0 or 1.

We now discuss some of the intuition behind the graphs in Figures 11 and 12. We focus on the dependence of payoff losses on the discount factor δ . It is unsurprising that, for fixed x , payoff losses are low for values of δ that are close to 0. In Figure 12, when $x \geq \frac{1}{2}$, there is no difference between the strategy implemented by the two-state automaton and the unconstrained optimal strategy if δ is low. In Figure 11, where $x \leq \frac{1}{2}$ there is a difference in strategies, but this difference is not very important for small values of δ because for low δ learning does not matter much.

It is more surprising that the loss in expected payoffs converges to zero as δ tends to one. We shall demonstrate this analytically below, and, in the course of our proof, also identify the features of the optimal experimentation probability that are essential for the result.

We consider normalized payoffs, i.e. expected discounted payoffs multiplied by $(1 - \delta)$. As our discussion in Section 5 shows, there are only two states of the world in which a payoff loss occurs: $(\pi_A, \pi_B) = (x, 0)$ and $(\pi_a, \pi_b) = (x, 1)$. We calculate the payoff loss for each of these two states separately. We begin with the state $(\pi_A, \pi_B) = (x, 0)$. The expected payoff from the unconstrained optimal strategy in this case is:

$$(1 - \delta) \left(x + \frac{\delta^2}{1 - \delta} x \right) = (1 - \delta + \delta^2)x. \quad (20)$$

The expected payoff from the optimal two state automaton follows from equation (9):

$$(1 - \delta) \frac{1}{1 + \delta p} \frac{x}{1 - \delta} = \frac{1}{1 + \delta p} x. \quad (21)$$

The limit for $\delta \rightarrow 1$ of (20) is clearly x . To show that in this limit there is no loss from using a two state automaton, we therefore aim to show that also the limit of (21) for $\delta \rightarrow 1$ is x . To show this it suffices to show that δp tends to zero, and hence that the optimal p tends to zero as δ tends to one. Figure 6 shows that for every $x \in (0, 1)$ for sufficiently large δ the optimal value of p is given by (19). On the right hand side of (19), the first term tends to zero as δ tends to 1. Therefore, for sufficiently large δ , the minimum on the right hand side of (19) is given by the first term, and this minimum tends indeed to zero, as we needed to show. We can conclude that in the state $(\pi_A, \pi_B) = (x, 0)$ there is no loss in expected payoffs from using a two state automaton.

We now turn to the state $(\pi_A, \pi_B) = (x, 1)$. The expected payoff from the unconstrained strategy in this case is:

$$(1 - \delta) \left(x + \frac{\delta}{1 - \delta} \right) = (1 - \delta)x + \delta \quad (22)$$

The expected payoff from the optimal two state automaton follows from equation (10):

$$\begin{aligned} (1 - \delta) \left(\frac{1 - \delta}{1 - \delta + \delta p} \frac{x}{1 - \delta} + \frac{\delta p}{1 - \delta + \delta p} \frac{1}{1 - \delta} \right) \\ = \frac{1 - \delta}{1 - \delta + \delta p} x + \frac{\delta p}{1 - \delta + \delta p} \\ = \frac{1}{1 + \frac{\delta}{1 - \delta} p} x + \frac{\frac{\delta}{1 - \delta} p}{1 + \frac{\delta}{1 - \delta} p} \end{aligned} \quad (23)$$

Clearly, for $\delta \rightarrow 1$, the expression in (22) tends to one. Thus, to show that there is no loss in expected payoffs from using a two state automaton, we

need to show that also the expression in (23) tends to one as $\delta \rightarrow 1$. To clarify whether this is the case, we shall adjust our notation slightly, and write $p(\delta)$ for the optimal p , as a function of δ . It should be understood that we keep $x \in (0, 1)$ fixed. Then (23) shows that it is necessary and sufficient that:

$$\begin{aligned} \lim_{\delta \rightarrow 1} \frac{\delta p(\delta)}{1 - \delta} &= \infty \Leftrightarrow \\ \lim_{\delta \rightarrow 1} \frac{p(\delta)}{\frac{1-\delta}{\delta}} &= \infty. \end{aligned} \tag{24}$$

This says that $p(\delta)$ must converge slower to zero than $\frac{1-\delta}{\delta}$. We now check that this is the case, substituting for $p(\delta)$ the first term on the right hand side of (19):

$$\begin{aligned} \lim_{\delta \rightarrow 1} \frac{p(\delta)}{\frac{1-\delta}{\delta}} &= \\ \lim_{\delta \rightarrow 1} \frac{\frac{\sqrt{1-x}}{\sqrt{1-\delta}} - \sqrt{x}}{\sqrt{x} - \sqrt{1-\delta}\sqrt{1-x}} &= \infty \end{aligned} \tag{25}$$

Thus, we can conclude that also in the state $(\pi_A, \pi_B) = (x, 1)$ the asymptotic loss in expected utility from using a two state automaton is zero.

Our argument shows that the crucial feature of the experimentation probability that enables a very patient decision maker to capture all feasible rents with a two state automaton is that firstly the experimentation probability tends to zero as δ tends to one, and that secondly, this probability tends to zero slower than $\frac{1-\delta}{\delta}$.

8. CONCLUSION

Our example has illustrated several fascinating features of optimal strategies for two-armed bandits in the presence of complexity constraints. Future research should seek to explore how general these insights are. There are two

directions into which one could generalize our investigation. One direction is to consider more general bandit problems. The second direction is to consider other measures of the complexity of a strategy, in particular measures which take the complexity of the transition function into account. Needless to say, another essential part of future research is that it needs to be checked how relevant theories such as the one developed in this paper is to real world learning behaviour.

REFERENCES

- BANKS, J. AND R. SUNDARAM (1990), "Repeated Games, Finite Automata, and Complexity," *Games and Economic Behavior* 2, 97-117.
- BERRY, D. A., AND B. FRISTEDT (1985), *Bandit Problems: Sequential Allocation of Experiments*, London: Chapman-Hall.
- BÖRGERS, T., AND A. MORALES (2004), "Complexity Constraints and Adaptive Learning: An Example," mimeo., University College London and University of Malaya.
- KALAI, E., AND E. SOLAN (2003), "Randomization And Simplification in Dynamic Decision-Making," *Journal of Economic Theory*, 2003.
- PICCIONE, M. AND A. RUBINSTEIN (1997), "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behavior* 20 (1997), 3-24.
- SCHLAG, K. (2002), "How to Choose - A Boundedly Rational Approach to Repeated Decision Making," mimeo., European University Institute, Florence.