Complexity Constraints in Two-Armed Bandit Problems: An Example (Tilman Börgers and Antonio J Morales) Extended Abstract

The two-armed bandit problem is a classical model 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 Foisted (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 Piscine and Ruination (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 Piscine and Ruination'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.

Although our work is related to Piscine and Ruination (1997), it 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 Piscine and Ruination's work, which information will be stored, and which will be forgotten, is exogenously given.

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.

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.

Our work is closely related to Kalai and Solan (2003) who have presented a general study of optimal finite automata for Markov decision problems. 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. However, this is a minor point. The superiority of randomized strategies over deterministic strategies was already demonstrated by Kalai and Solan. What we present here is an application of Kalai and Solan's general framework to two-armed bandit problems.

Schlag (2003) has also studied several desirable properties of simple learning algorithms. 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.