

# Non-probabilistic Decision Making with Memory Constraints

Alexander Vostroknutov\*

Department of Economics  
University of Minnesota

July 2007

## Abstract

In the model of choice, studied in this paper, the decision maker chooses the actions non-probabilistically in each period (Sarin and Vahid, 1999; Sarin, 2000). The action is chosen if it yields the biggest payoff according to the decision maker's subjective assessment. Decision maker knows nothing about the process that generates the payoffs. If the decision maker remembers only recent payoffs, she converges to the maximin action. If she remembers all past payoffs, the maximal expected payoff action is chosen. These results hold for any possible dynamics of weights and are robust against the mistakes. The estimates of the rate of convergence reveal that in some important cases the convergence to the asymptotic behavior can take extremely long time. The model suggests simple experimental test of the way people memorize past experiences: if any weighted procedure is actually involved, it can possibly generate only two distinct modes of behavior.

*JEL classification: D83, D81, C02.*

*Keywords: Adaptive learning, constrained memory, bandit problem.*

---

\*I would like to thank Tilman Börgers for invaluable discussions and Andrew Chesher, Beth Allen, and Aldo Rustichini for helpful comments.

# 1 Introduction

The environments in which economic agents make decisions can be very complex. This accounts for the tendency of the agents to simplify their decisions. I study the behavior of a simple decision maker who does not randomize while choosing among actions. Instead, she chooses the action with the biggest subjective assessment, which is represented by the weighted average of the past payoffs. The decision maker has no information about the environment apart from the payoffs she receives. I analyze the long run behavior of the decision maker and the rate of convergence for very general weight structure.

The model of the kind was first introduced in Sarin and Vahid (1999). In this paper the decision maker does not randomize among actions as well, but the weights given to the past payoffs are fixed. The long run behavior of the decision maker depends on various assumptions and might be stochastic (i.e. in the long run the agent is randomizing among several actions).

In this paper I consider a model which makes sharp predictions about the long run behavior for all possible weight configurations. Depending on the “size” of the memory, the decision maker can follow only two modes of behavior. Moreover, the result is robust against the mistakes by the decision maker. This suggests an easy way to test if the actual mechanism of memorizing payoffs in humans works in weighted average fashion. An experimental study in Sarin and Vahid (2001) already shows that the models of this type can explain existing data exceptionally well. The only drawback that becomes clear is that the rate of convergence to the long run action might be tremendously low in some cases. This can potentially create problems with testing the hypotheses.

In the model the decision maker faces the same decision problem repeatedly. In each period she chooses an action according to her subjective assessments, which are the weighted averages of the past payoffs. After the choice is made, the state of the world is realized and the payoff is revealed. The decision maker has no information about the process that generates the payoffs. After the payoff is known, she updates the subjective assessment of the action which generated the payoff. The subjective assessments of all other actions stay the same. We assume that the decision maker can make mistakes

with  $\varepsilon$  probability in which case she chooses an action with non-maximal assessment.

Similar model is studied extensively in computer science literature, where it is called  $\varepsilon$ -greedy learning procedure (Sutton and Barto, 1998). Most results in computer science, however, are simulations that compare this learning model to others. Very close learning model is used by Young (2007) to study the behavior in large games. When the action sets and the number of players are very big it becomes computationally impossible for either humans or computer agents to use learning procedures that try to take into account the information about the game. Thus, it is important to understand how learning models that use only payoff information behave.

The paper is organized as follows. In section 2 the model with finite memory is formalized and the maximin result is proved. General model is introduced in section 3 as well as the long run behavior results for different cases. Section 4 deals with the estimates of the rate of convergence. And Section 5 concludes. Proofs and definitions can be found in the Appendix 6.

## 2 The Model with Finite Memory

The decision maker faces the same decision problem repeatedly. Each period she is choosing one of the  $J$  actions from a finite set  $A = \{a_1, a_2, \dots, a_J\}$ . After the action is chosen, the state of the world  $\omega_t \in \Omega$  is realized ( $t$  stands for the time period). The set of states of the world  $\Omega$  is assumed finite. Suppose that there is some probability measure defined on  $\Omega$ . For each  $t$ ,  $\omega_t$  is identically and independently distributed in each time period. After the state of the world is chosen decision maker receives her payoff according to the utility function  $u : A \times \Omega \rightarrow \mathbb{R}_+$  (in what follows I will write  $u_i(t)$  for the payoff received by the decision maker from choosing  $i$ th action in period  $t$ ). Decision maker does not have any information about the process that generates the payoffs (even if it is random or not).

To choose an action the decision maker uses the following information. There exist real-valued subjective assessments of the payoffs, which the decision maker attaches to each action in  $A$ . Denote these assessments by  $\alpha_i(t)$  for  $i$ th action in period  $t$  and write the vector of assessments as  $\alpha(t) = (\alpha_1(t), \alpha_2(t), \dots, \alpha_J(t))$ . Each period the decision maker chooses the action with *the maximal subjective assessment*. For simplicity, assume that at any time  $t$  all the subjective assessments are different from each other. This assumption does not create any loss of generality but helps to avoid uninteresting ties.

Subjective assessments are updated each period. The assessment for each action  $i$  at period  $t$  is a time and action invariant function of the payoffs, received by the decision

maker after playing  $i$  in the past. These payoffs are stored in the memory. Assume that the memory of the decision maker is finite and she remembers  $m$  values of the payoffs received in the past from each action. Denote by  $m_i(t)$  the “memory vectors” for each action  $i$  at time  $t$ . These vectors contain the values of last  $m$  payoffs, which were obtained from playing each action. So  $m'_i(t) = (u_i(t_1), u_i(t_2), \dots, u_i(t_m))$ , where  $t_1 < t_2 < \dots < t_m \leq t$  are the periods when action  $i$  was played last  $m$  times in the past (the prime means transposition). The memory vector for an action changes whenever the action is played. The first (the oldest) value of the payoff in the memory vector disappears (the decision maker forgets it), all other values in the vector move one position leftward and new payoff takes its place in the rightmost position. For example, if the decision maker has played action  $i$  and received some payoff  $u_i(t)$  her memory for action  $i$  changes as follows:  $m'_i(t+1) = (u_i(t_2), u_i(t_3), \dots, u_i(t_m), u_i(t))$ , where  $u_i(t_2), \dots, u_i(t_m)$  are the  $m-1$  values taken from the vector  $m_i(t)$ . Call this transformation of memory vector  $T(m_i(t), u_i(t))$ . Think of  $m_i(0)$  as given for each action  $i$  and that all elements in  $m_i(0)$  are equal to one of the payoffs, say  $u_i(0)$ , which action  $i$  can possibly yield. Let us then assume that  $\alpha_i(0) = u_i(0)$ .

Now consider how the subjective assessments are calculated. Suppose that there exist the constant vector of non-negative weights  $\lambda' = (\lambda_1, \lambda_2, \dots, \lambda_m)$  with the property  $\sum_{i=1}^m \lambda_i = 1$ . The assessment of the payoff from playing  $i$ th action in period  $t$  is the  $\lambda$ -weighted average of the values of previous payoffs received from playing action  $i$ . More formally, in the end of the period  $t$ , when the decision maker obtains the payoff  $u_i(t)$  from choosing  $i$ th action, the memory and the assessments change as follows:

$$\begin{aligned} m_i(t+1) &= T(m_i(t), u_i(t)) \\ m_j(t+1) &= m_j(t), \quad \forall j \neq i \\ \alpha_i(t+1) &= \lambda' m_i(t+1) \\ \alpha_j(t+1) &= \alpha_j(t), \quad \forall j \neq i \end{aligned}$$

Only the memory and the assessment of the action that was played changes. All other memory vectors and assessments remain the same.

In addition, assume that the decision maker can make mistakes while choosing the action to play. Each period she chooses the action with maximal subjective assessment with probability  $1 - \varepsilon$ , where  $\varepsilon$  is some small positive number. With probability  $\varepsilon$  she makes a mistake and some other action is chosen instead. So, with probability  $1 - \varepsilon$  the decision maker chooses her subjectively maximal action and all other actions in  $A$  are chosen with probability  $\varepsilon/(J-1)$  each.

Say that the decision maker has *finite memory* if she remembers only  $m$  values of the payoffs received in the past (for each action). This is equivalent to saying that the memory can be in only finite number of states. Indeed, denote the set, which contains all possible values of the memory vector  $m_i$  for action  $i$  by  $M_i = \times_{j=1}^m u(i, \Omega)$ , where  $u(i, \Omega)$  is the finite set of possible payoffs from playing  $i$ . Denote the set of all possible “memories” of the decision maker by  $M = M_1 \times M_2 \times \dots \times M_J$ . Each  $M_i$  contains  $|\Omega|^m$  elements and thus is a finite set. Therefore,  $M$  is a finite set as well.

## 2.1 The Maximin Result

In this section I will prove the result concerning long-run properties of the model above. In particular, I will show that in the long run the decision maker will choose the maximin action (i.e. the action with maximal minimum payoff) almost surely. The proof will proceed in two stages. First, the model without mistakes, i.e. with  $\varepsilon = 0$ , will be considered and it will be shown that it can be described as a discrete-time Markov process with finite state space. Second, I will show that the original model with  $\varepsilon > 0$  is a *regular perturbed Markov process* which converges in some sense to the case  $\varepsilon = 0$  as  $\varepsilon$  gets small.

Call the model with  $\varepsilon = 0$  *the unperturbed system*. The state of the system is fully described by the state of the memory of the decision maker. Indeed, given some state of the memory  $\mu \in M$  decision maker’s subjective assessments are determined unambiguously. So, the probabilities of getting to other states are determined without ambiguity and depend only on the probability measure over  $\Omega$ .

**Definition 2.1** *Denote by  $P^{M, \lambda, 0}$  the discrete-time Markov process with finite state space  $M$ , which fully describes the unperturbed system.*

**Proposition 2.2** *In the unperturbed system the decision maker chooses only one action asymptotically. This action is the maximin action in  $A$  (the one with maximal smallest payoff).*

**Proof.** Given our assumption that the assessments in period 0 are in between maximal and minimal payoff for each action the proof of Proposition 1 in Sarin and Vahid (1999) gives the result. ■

The idea of the proof is as follows. The assessments of the payoffs from each action can only lie in between action specific maximal and minimal payoffs. Thus, whenever the assessment of some not maximin action falls below maximin that action is never chosen again. With time the assessments of all actions go down. This happens since

the decision maker switches to some other action only when the assessment of currently played action falls below some other assessment, then the process repeats itself. As time passes all assessments will become lower than maximin.

Now as we found the action, which is chosen by the decision maker in the unperturbed system, let us make some statements about Markov process  $P^{M,\lambda,0}$ . In the following Proposition assume without loss of generality that the maximin action has index 1 ( $a_1$ ). Also denote the minimal payoff from action  $a_i$  by  $u_{\min}^i = \min_{\omega \in \Omega} u(i, \omega)$ . Denote the maximin payoff by  $u_{\max \min} = \max_{1 \leq i \leq J} u_{\min}^i$ .

**Proposition 2.3** *The Markov process  $P^{M,\lambda,0}$  has finitely many recurrent classes. All of these recurrent classes consist only of the states in which the decision maker chooses maximin action. The set of states in any recurrent class can be described as*

$$C_Q = \{(m_1, m_2, \dots, m_J) : m_1 \in M_1, (m_2, \dots, m_J) = Q\}$$

where  $Q \in \{(m_2, \dots, m_J) \in \times_{i=2}^J M_i : \lambda^i m_i = \alpha_i < u_{\max \min} \forall i = 2, \dots, J\}$  is some constant value of the collection of  $J - 1$  memory vectors corresponding to not maximin actions with the property that the assessments of these actions (also constants) are less than maximin payoff.

**Proof.** See Appendix 6.1.

Now the analysis of the behavior of the model with mistakes is possible. Recall that in each period the decision maker chooses the action with maximal subjective assessment with probability  $1 - \varepsilon$  and any other action is chosen with probability  $\varepsilon/(J - 1)$ .

It is easy to see that the behavior of the model with mistakes is described by the Markov process similar to the model without mistakes. The difference is that from any given state  $\mu \in M$  the process can get to the states previously unreachable. The probability of these events is proportional to  $\varepsilon$ . Say that the Markov process  $P^{M,\lambda,\varepsilon}$  is a *perturbation* of the process  $P^{M,\lambda,0}$  if the transition matrix of  $P^{M,\lambda,\varepsilon}$  is slightly distorted version of the transition matrix of  $P^{M,\lambda,0}$ . Suppose that  $P^{M,\lambda,\varepsilon}$  fully describes the behavior of the model with mistakes.

**Definition 2.4 (Young, 1998)** *The perturbed Markov process  $P^{M,\lambda,\varepsilon}$  is called regular perturbed Markov process if it satisfies the following conditions:<sup>1</sup>*

1.  $P^{M,\lambda,\varepsilon}$  is irreducible for every  $\varepsilon \in (0, \varepsilon^*]$ ;

---

<sup>1</sup>In what follows  $P_{zz'}^{M,\lambda,\varepsilon}$  means the probability of transition from state  $z$  to the state  $z'$ .

2. for every two states  $z, z' \in M$  it is true that

$$\lim_{\varepsilon \rightarrow 0} P_{zz'}^{M,\lambda,\varepsilon} = P_{zz'}^{M,\lambda,0}$$

3. if  $P_{zz'}^{M,\lambda,\varepsilon} > 0$  for some  $\varepsilon > 0$ , then  $0 < \lim_{\varepsilon \rightarrow 0} P_{zz'}^{M,\lambda,\varepsilon} / \varepsilon^{r(z,z')} < \infty$  for some  $r(z, z') \geq 0$ .

**Proposition 2.5** *The perturbed Markov process  $P^{M,\lambda,\varepsilon}$  is a regular perturbed Markov process.*

**Proof.** See Appendix 6.1.

Now, according to the Theorem 3.1 of Young (1998) any regular perturbed Markov process  $P^{M,\lambda,\varepsilon}$  has unique stationary distribution  $\mu^\varepsilon$  for each  $\varepsilon > 0$ . More than that,  $\lim_{\varepsilon \rightarrow 0} \mu^\varepsilon = \mu^0$ , where  $\mu^0$  is some stationary distribution of the unperturbed process  $P^{M,\lambda,0}$ . So as  $\varepsilon \rightarrow 0$  the perturbed system converges to some stationary distribution of the unperturbed system. The stationary distribution of the unperturbed system necessarily lies in one of the recurrent classes. Since we have shown that in all the states inside any recurrent class the decision maker chooses maximin action, the same maximin action will be chosen in the perturbed system with probability close to one. When  $\varepsilon > 0$  is small enough the decision maker will “circulate” around maximin action, but still choose it most of the time.

### 3 General Model

In this section I consider the case when the memory of the decision maker is infinite. This will be the only difference of the model in this section from the model with finite memory. In each period the decision maker chooses the action with maximal assessment. At the same time the decision maker can make mistakes and choose some different action instead. There are three different modes of behavior which can emerge when the memory is infinite. Difference arises because of the way of formation of weights  $\lambda_i$  in the infinity.

All the assumptions of the model with finite memory remain the same, only the assessments updating rule changes. Here I do not assume that before choosing an action for the first time the decision maker has already infinite number of payoff experiences. Instead, suppose that in the beginning the decision maker has some given vector of assessments  $\alpha(0) = (\alpha_1(0), \dots, \alpha_J(0))$ , with each component being between the minimum and the maximum possible payoff from each action. After receiving the payoff from

playing action  $a_i$  the decision maker updates her assessments in the following way:

$$\begin{aligned}\alpha_i(t) &= \lambda'_t m_i^\infty(t) \\ \alpha_j(t) &= \alpha_j(t-1) \quad \forall j \neq i\end{aligned}$$

Here  $\lambda_t = (\lambda_{tt}, \lambda_{t(t-1)}, \dots, \lambda_{t0})'$  denotes the column vector of length  $t + 1$  of the weights attached to the payoff just received and the payoffs from playing the same action in the past. The column vector  $m_i^\infty(t)$  of length  $t + 1$  contains the payoff just received in the period  $t$ , all the payoffs received by the decision maker from playing action  $a_i$  in the past, and the value of  $\alpha_i(0)$ :

$$m_i^\infty(t) = (\alpha_i(0), u_i(1), u_i(2), \dots, u_i(t))'.$$

As before assume that the decision maker finds weighted average of the payoffs, so  $\sum_{i=0}^t \lambda_{ti} = 1 \quad \forall t \in \mathbb{N}$ . As time passes the vectors become longer as more and more payoffs are put to the memory of the decision maker. There are no assumptions on the particular way of evolution of weights as more payoffs are received. The only restriction is that there is some procedure, the same for all actions, which generates the weights each period.<sup>2</sup>

Now as the model is described, let us define the three different ways the weights  $\lambda$  can behave in infinity. Consider the triangular array of weights

$$\begin{array}{cccc} 1 & & & \\ \lambda_{11} & \lambda_{10} & & \\ \lambda_{22} & \lambda_{21} & \lambda_{20} & \\ \vdots & & & \\ \lambda_{tt} & \lambda_{t(t-1)} & \dots & \lambda_{t0} \\ \vdots & & & \end{array}$$

where  $\sum_{i=0}^t \lambda_{ti} = 1, \quad \forall t \in \mathbb{N}$ .

In each period  $t$  the most recent payoff received by the decision maker is given the weight  $\lambda_{t0}$ . The oldest payoff ( $\alpha_i(0)$ ) is given the weight  $\lambda_{tt}$ . Consider three types of

---

<sup>2</sup>As time passes the memory of the decision maker for different actions will contain different number of payoffs. The assumption means that whenever the number of payoffs for two different actions in any point in time is the same, the weighted average is calculated using the same weights.

arrays.

**Case I.** This case is close to the finite memory model as no conditions are imposed on weights to stay bounded away from zero. For any triangular array of weights  $\lambda$  define a function  $N_\lambda : \mathbb{N} \times \mathbb{R} \rightarrow \mathbb{N}$  by

$$N_\lambda(n, \delta) = \min\{k : \sum_{i=0}^k \lambda_{ni} \geq 1 - \delta\}$$

Say that the triangular array of weights  $\lambda$  belongs to Case I whenever

$$\forall \delta > 0 \ N_\lambda(\cdot, \delta) \text{ is bounded}$$

**Case II.** In this case the decision maker takes into account infinitely many past payoffs. As  $t \rightarrow \infty$  all the weights approach zero, however the weight does not “escape” to infinity.

**Definition 3.1 (No Escape Condition)** *The triangular array of weights  $\lambda$  satisfies No Escape Condition whenever*

$$\forall k \in \mathbb{N} \ \lim_{t \rightarrow \infty} \lambda_{tk} = 0$$

The array of weights belongs to Case II if No Escape Condition is satisfied and

$$\lim_{t \rightarrow \infty} \max_{k \leq t} \lambda_{tk} = 0.$$

**Case III.** In this case the weight “escapes” to infinity. Any array of weights which does not satisfy No Escape Condition belongs to Case III.

### 3.1 Long Run Behavior in the General Model

In this section I prove results concerning long run behavior of the decision maker as the number of periods played grows to infinity. Depending on the way the weights are formed, the decision maker’s behavior differs in the long run.

**Proposition 3.2** *In the model with no mistakes and Case I weights as  $t \rightarrow \infty$  the decision maker converges to the maximin action.*

**Proof.** See Appendix 6.1.

It can be easily seen that when we introduce mistakes to the model with Case I weights we still get robust convergence to maximin action (in the sense of Section 2.1).

In the special case when the array of weights converges row-wise to some element of the space  $\ell_1$  we can get as good approximation of infinite memory by finite memory models as we desire since the convergence to maximin result holds for any finite memory model.<sup>3</sup> If the array of weights does not converge in  $\ell_1$  we still can approximate by finding the finite sequence of weights  $(\lambda_1, \dots, \lambda_n)$  with the function  $N(n, \delta)$  “close” to the function of the array as  $n \rightarrow \infty$ .

**Proposition 3.3** *In the model with mistakes and Case II weights the decision maker asymptotically plays the action with maximal expected payoff with probability  $1 - \varepsilon$ .*

**Proof.** See Appendix 6.1.

In Case III nothing specific can be said about the long run behavior of the decision maker since asymptotically she puts positive weights only on the payoffs received infinitely long time ago. Depending on the realizations of payoffs anything can happen.

However, this case can still have some meaning to it. For example, if the decision maker has very strong “first impression”: after playing some action for the first time she remembers only the first payoff and never changes her assessment afterwards. This behavior corresponds to the array of weights with

$$\lambda_{nn} = 1 \quad \forall n \in \mathbb{N} \quad \text{and} \quad \lambda_{nk} = 0 \quad \forall k < n$$

## 4 Rate of Convergence

In this subsection I will state the results concerning the rate of convergence to maximin in the finite memory model and to maximal expected payoff in Case II model.

### 4.1 Finite Memory Model

It seems intuitively true that bigger memory slows down the convergence to maximin. Recall the proof of Proposition 2.2. There it was stated that if the decision maker chooses non-maximin action then in finite number of steps, the subjective assessment of this action would fall to some level below the value of maximin payoff. The subsequent occurrence of  $m$  lowest payoffs, for example, is enough. If  $m$  is very large, then the subjective assessment will be an average of a big number of random variables and it will take longer for it to fall lower than the maximin payoff. Thus, as  $m$  grows it will take longer and longer for the decision maker to switch to the maximin action.

---

<sup>3</sup> $\ell_1$  is the space of all infinite real sequences  $(x_1, x_2, \dots)$  with the norm  $\sum_{i=1}^{\infty} |x_i|$ .

I find an estimate of the *minimum* number of periods needed for the decision maker to converge to maximin action. Say that the decision maker *has converged* to maximin action once the subjective assessments of all non-maximin actions become less than the value of maximin payoff *for the first time*. This means that “the convergence takes place” whenever the decision maker chooses maximin action *not erroneously*. Of course, since the decision maker makes mistakes, it can happen that she will switch to some non-maximin action in the future. However, since the assessments of all other non-maximin actions will still be less than the value of maximin payoff, the decision maker will switch back to the maximin relatively quickly (especially if the number of actions is big).

Denote by  $F_i$  the distribution function of the subjective assessment  $\alpha_i$ , which is a weighted average of  $m$  discrete random variables. Without loss of generality let us assume that maximin action is the action  $a_1$  with corresponding subjective assessment  $\alpha_1$ .

Consider the sequence of periods  $t_1^j, t_2^j, \dots$  when action  $a_j$  is chosen. We can treat the values of the subjective assessment  $\alpha_j$  in periods  $t_1^j, t_2^j, \dots$  as a sequence of independent realizations of random variable with distribution function  $F_j$ .<sup>4</sup> Each period the action  $a_j$  is chosen the probability that the subjective assessment of this action will become less than  $u_{\max \min}$  equals to  $P_j = F_j(u_{\max \min})$ . Simple calculation shows that the expected number of periods needed for the subjective assessment  $\alpha_j$  to become less than  $u_{\max \min}$  is

$$\sum_{i=1}^{\infty} iP_j(1 - P_j)^{i-1} = \frac{1}{P_j}.$$

Let us think about the sequences of periods when each action is played as a separate sequence. We need to find the minimum expected time needed for all non-maximin assessments to become less than  $u_{\max \min}$ . The expected number of periods, which is necessary for this, is the sum of the expected number of periods for each action:

$$N = \sum_{i=2}^J \frac{1}{P_i} = \sum_{i=2}^J \frac{1}{F_i(u_{\max \min})}$$

here we use our assumption that the maximin action has index 1.

In general,  $N$  depends on 1) distributions of all  $u_i$  except the distribution of maximin

---

<sup>4</sup>In reality the realizations of the subjective assessment are not independent, since the components of one realization are present in the next realization. However, for our purposes this fact is not very significant, since we are looking for expected *minimum* number of periods needed for convergence. If we take into account this correlation between subsequent realizations the time of convergence will become only longer.

action; 2) the value of  $u_{\max \min}$ ; 3) number of actions  $J$ ; 4) length of memory  $m$  ( $F_i$  depend on length of memory, the longer is the memory, the smaller is the variance of the distributions). It is straightforward that  $N$  is an increasing function of  $m$  and  $J$  and a decreasing function of  $u_{\max \min}$ .

## 4.2 Case II Model

Another approach is used here to estimate the number of periods needed for convergence to maximal expected payoff. Denote by  $V_i$  the variance of the random variable  $u(i, \cdot)$ , by  $E_i$  its expectation and let  $\sigma_i = \sqrt{V_i}$ . Without loss of generality assume that action 1 has the biggest maximal payoff. Let  $\Delta_i = E_1 - E_i$ . I am interested in how the variance of the assessments behaves. Notice that given any assessment which was updated  $t$  times the variance of the assessment satisfies

$$V_i \min_k \lambda_{tk} \leq \text{Var}[\alpha_i(t)] = V_i \sum_{j=0}^t \lambda_{tj}^2 \leq V_i \max_k \lambda_{tk}$$

Denote by  $\sigma_{it}$  the standard deviation of assessment  $\alpha_i$  after  $t$  updates, then

$$\sqrt{V_i \min_k \lambda_{tk}} \leq \sigma_{it} \leq \sqrt{V_i \max_k \lambda_{tk}}$$

The question now is what is the number  $n$  of updates such that for some  $\kappa$  the intervals  $[E_i, E_i + \kappa\sigma_{in}]$  and  $[E_1 - \kappa\sigma_{1n}, E_1]$  do not intersect. The intuition is that if these two intervals are disjoint then the assessment  $i$  is less than assessment 1 with probability at least  $1 - 1/\kappa^2$  (Chebyshev's inequality). Clearly the minimal number of updates solves

$$\kappa(\sigma_{1n} + \sigma_{in}) = \Delta_i$$

Suppose that  $n_i^*$  solves this equation for each  $i$ . Then the number of updates needed for the assessments of all actions to be less than the assessment of action 1 with probability  $1 - 1/\kappa^2$  is

$$N = \max_{i>1} n_i^* + \sum_{i=2}^J n_i^*.$$

The  $\max_{i>1} n_i^*$  term is needed as action 1 should be chosen enough times as well.

It is not easy to explicitly find  $n_i^*$  in general, however we can find an estimate of  $N$  which involves only  $\min_k \lambda_{nk}$  and  $\max_k \lambda_{nk}$  by using the inequality above. True  $n_i^*$  lies

in between the solutions to

$$\begin{aligned}\kappa\sqrt{\max_k \lambda_{nk}}(\sigma_1 + \sigma_i) &= \Delta_i \\ \kappa\sqrt{\min_k \lambda_{nk}}(\sigma_1 + \sigma_i) &= \Delta_i\end{aligned}$$

Suppose that  $n_i^U$  solves the first equation and  $n_i^L$  the second. Then

$$N^L = \max_{i>1} n_i^L + \sum_{i=2}^J n_i^L \leq N \leq \max_{i>1} n_i^U + \sum_{i=2}^J n_i^U = N^U.$$

For example, assume that  $\min_k \lambda_{nk} = 1/n^\mu$  where  $0 < \mu \leq 1$  then the minimal number of periods needed for the action 1 to be played with probability  $1 - 1/\kappa^2$  is

$$\max_{i>1} \left( \frac{\kappa(\sigma_1 + \sigma_i)}{\Delta_i} \right)^{2/\mu} + \sum_{i=2}^J \left( \frac{\kappa(\sigma_1 + \sigma_i)}{\Delta_i} \right)^{2/\mu}$$

One can see that the rate of convergence can be tremendously low if  $\mu$  is small (say,  $\mu = 0.1$ ). This means that if the decision maker uses the weights that go to zero not too fast, but still fast enough for the assessments to converge to expectation, then we should not expect to see the maximal expectation action to be chosen.

## 5 Conclusion

Economic agents tend to simplify their decisions when facing very complex environment. The model in this paper studies the behavior of a simple decision maker, who does not possess any information about the process that generates payoffs. The decision maker simplifies her choice by choosing actions non-probabilistically.

The results differ depending on the weights the decision maker attaches to the past payoffs. If she cares only about recent payoffs then the maximin action is chosen in the long run. This result holds even if we introduce mistakes into the model. If the decision maker cares about payoffs received long time ago then she converges to the maximal expected payoff action. In the maximin case, the rate of convergence to the long run behavior depends on the number of actions, length of the memory and the process that generates the payoffs. In the maximal expectation case, the rate of convergence depends on how fast the biggest weight goes to zero. Remarkably, this rate can be very low in some cases when the weights decrease slow enough.

The model can be extended in several ways. At first, the technique used to prove

the convergence result in the general model can be applied almost without changes to the case of infinite state space  $\Omega$ . The only condition that matters is bounded support of the distributions of payoffs. At second, it could be interesting to investigate the case when the action space is big and the decision maker experiments only in the vicinity of the action she plays. At third, one might check how this model performs in games (see Huck and Sarin (2004); Young (2007)).

## 6 Appendix

### 6.1 Proofs

**Proof of Proposition 2.3.** The recurrent class of the Markov process is a set of the states, which satisfy two properties: 1) any pair of states in the recurrent class communicate; and 2) none of the states outside the recurrent class can be accessed from any of the states inside it. To prove property 1 of this statement notice that once the decision maker chooses maximin strategy and  $\alpha_i < u_{\max \min}$  for all  $i = 2, \dots, J$  she chooses only it forever after. This means that, after it happens, the memories (and the assessments) of all not maximin actions are not updated anymore and are constant forever (this constant is denoted  $Q$  in the Proposition). However, the decision maker continues to choose maximin action and so its memory vector is changing all the time. Given that all other memory vectors are constant, it is easy to see that the decision maker can reach any state of the maximin memory vector.  $2m$  steps is enough to reach any memory configuration from any other. Any of these  $2m$  steps has positive probability of occurrence. Thus any set of states described in the Proposition satisfies property 1. Property 2 follows from the proof of the Proposition 2.2. So we conclude that any set of states of the form stated in the Proposition constitutes a recurrent class of the Markov process  $P^{M,\lambda,0}$  and all of these recurrent classes consist only of the states in which the decision maker chooses maximin action. There is no other set of states, which do not belong to some of the recurrent classes described in the Proposition, and, at the same time, constitute a recurrent class. This again directly follows from the proof of the Proposition 2.2. ■

**Proof of Proposition 2.5.** We first prove that  $P^{M,\lambda,\varepsilon}$  is irreducible. It is enough to show that any two states  $z, z' \in M$  communicate. Consider any two such states  $z$  and  $z'$ . There exists a sequence of mistakes together with particular states of the world occurring at each transition such that with positive probability the system will move from state  $z$  to  $z'$ . It is easy to imagine such sequence which gradually makes the same all memory vectors in  $z$  and  $z'$ . Now we show that conditions 2 and 3 of Definition 2.4 are satisfied. Indeed, in all possible perturbations in the transition matrix,  $\varepsilon$  enters as a linear term  $c\varepsilon$  where  $c$  is some constant. This is enough for condition 2 to be true. For the same reason the condition 3 is satisfied when  $r(z, z') = 1$  for any  $z$  and  $z'$  such that

$P_{zz'}^{M,\lambda,0} = 0$  and when  $r(z, z') = 0$  for any  $z$  and  $z'$  such that  $P_{zz'}^{M,\lambda,0} > 0$ . ■

**Proof of Proposition 3.2.** Given the definition of the Case I array of weights we can be sure that for any given  $\delta > 0$  we can find a number  $n_\delta$  such that

$$\forall t \in \mathbb{N} \quad \sum_{i=1}^{n_\delta} \lambda_{ti} \geq 1 - \delta.$$

Since all  $u_{\min}^i$  are different, this guarantees that there exists  $\delta$  small enough for any assessment to fall below  $u_{\max \min}$  with positive probability ( $n_\delta$  occurrences of  $u_{\min}^i$  will suffice). Now we get the result by placing this instead of the analogous argument in Proposition 1 of Sarin and Vahid (1999). ■

**Proof of Proposition 3.3.** Let us show that the assessments of all actions converge to the expected value of payoffs. Consider some action  $i$  and time periods  $t_1, t_2, \dots$  when this action is played ( $t_1 < t_2 < \dots$ ). We are interested in the behavior of the assessment  $\alpha_i(t_n)$  as  $n \rightarrow \infty$  (since the decision maker makes mistakes we can be sure that any action is played infinitely often). For each update period  $t_n$  we know that

$$\alpha_i(t_n) = \lambda_{nn}\alpha_i(0) + \lambda_{n(n-1)}u(i, \omega_{t_1}) + \dots + \lambda_{n0}u(i, \omega_{t_n})$$

where  $u(i, \omega_t)$  is the realization of the payoff in period  $t$ . To prove the statement use the theorem from Fristedt and Gray (1997) (Theorem 25, p.311).<sup>5</sup> First we show that the assumptions of the theorem are satisfied and then explicitly find the limiting distribution of  $\alpha_i(t)$ .

Consider the triangular array of random variables:

$$\begin{array}{ccccccc} \alpha_i(0) & & & & & & \\ \lambda_{11}\alpha_i(0) & \lambda_{10}u(i, \omega_{t_1}) & & & & & \\ \lambda_{22}\alpha_i(0) & \lambda_{21}u(i, \omega_{t_1}) & \lambda_{20}u(i, \omega_{t_2}) & & & & \\ \vdots & & & & & & \\ \lambda_{nn}\alpha_i(0) & \lambda_{n(n-1)}u(i, \omega_{t_1}) & \dots & \lambda_{n0}u(i, \omega_{t_n}) & & & \\ \vdots & & & & & & \end{array}$$

This array is row-wise independent.<sup>6</sup> Indeed, each time the decision maker plays  $a_i$  she receives independent realization of  $u(i, \cdot)$ . Moreover this array is uniformly asymptotically negligible: this is easy to see since, by the definition of Case II weights, for any fixed  $\delta > 0$  we can always find the row of weights small enough for  $\sup_k P[|\lambda_{nk}u(i, \omega_{t_{n-k}})| > \delta] = 0$  to be true for some  $n$  and any row that follows. This is the consequence of the

---

<sup>5</sup>See Appendix 6.2

<sup>6</sup>For the definitions of the terms used in this proof see Appendix 6.2.

fact that  $u(i, \cdot)$  can take on values in only bounded interval of  $\mathbb{R}_+$ .

Now let us verify the conditions of the theorem. We claim that the Lévy measure  $\nu(x, \infty] = 0, \forall x > 0$  satisfies the first condition. For any  $x > 0$  we can find  $n$  big enough so that for all  $k \leq n$  we have  $P[\lambda_{nk}u(i, \omega_{t_{n-k}}) > x] = 0$ , hence the probability measure corresponding to any random variable  $\lambda_{\ell k}u(i, \omega_{t_{\ell-k}})$  where  $\ell \geq n$  and  $k \leq \ell$  is zero on the interval  $[x, \infty)$ . Thus, the limit of sums of these measures in each row is zero.

Denote by  $Q_{nk}^i$  the distribution of  $\lambda_{nk}u(i, \omega_{t_{n-k}})$  and consider the integral

$$\int_{(0, \delta]} x Q_{nk}^i(dx)$$

Since  $\lim_{n \rightarrow \infty} \lambda_{nk} = 0$  there exists  $n$  big enough so that  $Q_{nk}^i[\delta, \infty) = 0$ . Therefore for all  $\ell \geq n$

$$\int_{(0, \delta]} x Q_{\ell k}^i(dx) = \lambda_{\ell k} E[u(i, \cdot)]$$

and

$$\sum_{k=1}^{\ell} \int_{(0, \delta]} x Q_{\ell k}^i(dx) = E[u(i, \cdot)]$$

So second condition of the theorem is clearly satisfied:

$$\lim_{\delta \searrow 0} \limsup_{n \rightarrow \infty} \sum_{k=1}^n \int_{(0, \delta]} x Q_{nk}^i(dx) = \lim_{\delta \searrow 0} \liminf_{n \rightarrow \infty} \sum_{k=1}^n \int_{(0, \delta]} x Q_{nk}^i(dx) = E[u(i, \cdot)]$$

Now theorem tells us that the assessment converges to the random variable which corresponds to the pair  $(E[u(i, \cdot)], 0)$  via the Lévy-Khinchin Representation Theorem. This random variable has moment generating function  $\exp(-E[u(i, \cdot)])$  which obviously corresponds to delta distribution at  $E[u(i, \cdot)]$ . This finishes the proof of the statement above.

In the model the decision maker makes mistakes, so all the actions are played infinitely often. Therefore, as it was shown, assessments of all actions converge to the expected value. Since the decision maker chooses the action with the maximal assessment, she will eventually choose the one with maximal expected payoff. ■

## 6.2 Triangular array problem

Consider the triangular array of random variables<sup>7</sup>

$$\begin{array}{c} X_{11} \\ X_{22} \ X_{21} \\ X_{33} \ X_{32} \ X_{31} \\ \vdots \\ X_{nn} \ X_{n(n-1)} \ \dots \ X_{n1} \\ \vdots \end{array}$$

For each  $n$  the vector  $(X_{nk} : k = 1, \dots, n)$  is independent. So call such an array row-wise independent.

**Definition 6.1 (uan arrays)** *Triangular array which satisfies*

$$\lim_{n \rightarrow \infty} \sup_k P[|X_{nk}| > \delta] = 0 \quad \text{for all } \delta > 0$$

*is called uniformly asymptotically negligible.*

How do the row sums  $S_n = \sum_{k=1}^n X_{nk}$  which come from the uan arrays behave? The following theorems characterize this.

A measure  $\nu$  on  $(0, \infty)$  is a *Lévy measure* for  $\mathbb{R}_+$  if

$$\int_{(0, \infty)} \min\{y, 1\} \nu(dy) < \infty$$

**Theorem 6.2 (Lévy-Khinchin Representation for  $\mathbb{R}_+$ )** *The pair  $(\xi, \nu)$  where  $\xi \in \mathbb{R}_+$  and  $\nu$  is a Lévy measure for  $\mathbb{R}_+$  corresponds to a unique infinitely divisible distribution with moment generating function given by  $\exp(-\theta(v))$  where*

$$\theta(v) = \xi v + \int_{(0, \infty)} (1 - e^{-vy}) \nu(dy)$$

If  $X_1$  and  $X_2$  are random variables with distributions  $Q_1$  and  $Q_2$  let  $Q_1 * Q_2$  mean the distribution of  $X_1 + X_2$  and  $\sum_i Q_i$  be the sum of distributions *as measures*, so that  $Q_1 + Q_2$  is a measure that can take on values up to 2.

**Theorem 6.3 (Fristedt, Theorem 25, p.311)** *Let  $(Q_{nk} : 1 \leq k \leq n, n = 1, 2, \dots)$  be a uan triangular array of distributions on  $\mathbb{R}_+$ . For each  $n$ , let*

$$Q_n = Q_{n1} * Q_{n2} * \dots * Q_{nn}$$

---

<sup>7</sup>The material in this subsection is taken from Fristedt and Gray (1997). See also Loève (1978) (p.329) for analogous result.

In order that the sequence  $(Q_n : n = 1, 2, \dots)$  converge to a distribution on  $\mathbb{R}_+$  it is necessary and sufficient that there exist a nonnegative number  $\xi$  and a Lévy measure  $\nu$  for  $\mathbb{R}_+$  satisfying the following two conditions:

$$\nu[x, \infty) = \lim_{n \rightarrow \infty} \sum_{k=1}^n Q_{nk}[x, \infty) \quad \text{if } 0 < x \text{ and } \nu\{x\} = 0$$

$$\begin{aligned} \xi &= \lim_{\delta \searrow 0} \limsup_{n \rightarrow \infty} \sum_{k=1}^n \int_{(0, \delta]} x Q_{nk}(dx) \\ &= \lim_{\delta \searrow 0} \liminf_{n \rightarrow \infty} \sum_{k=1}^n \int_{(0, \delta]} x Q_{nk}(dx) \end{aligned}$$

In case these conditions are satisfied, the sequence  $(Q_n : n = 1, 2, \dots)$  converges to the infinitely divisible distribution on  $\mathbb{R}_+$  corresponding to  $(\xi, \nu)$  via the Lévy-Khinchin Representation Theorem for  $\mathbb{R}_+$ .

## References

- FRISTEDT, B., AND L. GRAY (1997): *A modern approach to probability theory*. Boston: Birkhäuser.
- HUCK, S., AND R. SARIN (2004): “Players with limited memory,” *Contributions to Theoretical Economics*, 4(1).
- LOÈVE, M. M. (1978): *Probability Theory*. New York, NY: Springer-Verlag.
- SARIN, R. (2000): “Decision Rules with Bounded Memory,” *Journal of Economic Theory*, 90(1), 151–160.
- SARIN, R., AND F. VAHID (1999): “Payoff Assessments without Probabilities: A Simple Dynamic Model of Choice,” *Games and Economic Behavior*, 28(2), 294–309.
- (2001): “Predicting How People Play Games: A Simple Dynamic Model of Choice,” *Games and Economic Behavior*, 34(1), 104–122.
- SUTTON, R. S., AND A. G. BARTO (1998): *Reinforcement Learning: An Introduction*. Cambridge, Mass.: MIT Press.
- YOUNG, H. P. (1998): *Individual strategy and social structure: an evolutionary theory of institutions*. Princeton, N.J.: Princeton University Press.
- (2007): “Learning by Trial and Error,” mimeo, University of Oxford, The Brookings Institution.