CIRJE-F-7

Learning about Stochastic Payoff Structures

Hitoshi Matsushima
Faculty of Economics, University of Tokyo
Institute of Economic Research, University of Kyoto

June 1998

Learning about Stochastic Payoff Structures*

Hitoshi Matsushima+

Faculty of Economics, University of Tokyo
Institute of Economic Research, University of Kyoto

June 4, 1998
(The First Version: May 17, 1997)

---

* This paper is the revised version of a part of the article entitled "Procedural Rationality and Inductive Learning I: Towards a Theory of Subjective Games", which was written on the basis of the seminar at the Tokyo Center of Economic Research (TCER), Tokyo, January 1997. This paper has a lot of new insights which have never been discussed in the elder versions. The first version of this paper was presented at the TCER International Summer Conference, Tokyo, May 1997, and the Eastern Meeting of the Japan Association of Economics and Econometrics, Shiga, May 1997.

+ Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113, Japan. Fax: (03) 3818-7082. E-mail: hitoshi@e.u-tokyo.ac.jp

Abstract

The situation in which a decision maker is confronted with decision making problems infinitely many times is considered. She does *not* know the state-dependent stochastic payoffs, and learns from past experiences according to some *adaptive learning rule*. She is motivated by the maximization of the subjective expected payoff, and *never* experiments with actions. We show that the decision maker comes to choose *only one action* in the long run, irrespective of which states she anticipates are likely to occur. This result holds even though she can *almost perfectly* monitor the true state. We give a characterization and argue that the action chosen in the long run may be objectively *maximin*.

JEL Classification Numbers: D80, D81, D83.

1. Introduction

The situation in which a decision maker is confronted with decision making problems infinitely many times is considered. The decision maker has little knowledge about the relevant aspects of this repeated situation, and therefore, has to learn inductively from past experiences according to some *adaptive learning rule*. The decision maker is mainly motivated by the maximization of the subjective expected payoff in a myopic way, and *never* experiments with any action which neither maximizes the subjective expected payoff nor is equal to the action chosen in the preceding period.

In the middle run, the decision maker may frequently change the choices of action. At the end of every period, the decision maker observes some signal which has information as to which state of the world has actually occurred. The payoff obtained in this period relies, not only on the choice of action, but also on the state of the world. Hence, the decision maker positively learns to evaluate the probability of occurrence of the informative signal, and then learns to evaluate the probability of occurrence of the state of the world. Especially in the non-stationary situation, she may frequently change her evaluations as to which signals are likely to occur, and therefore, the action maximizing the subjective expected payoff may change as time goes on.

This paper, however, will propose the opposite scenario *in the long run*: That is, the decision maker becomes gradually insensible to the change of her subjective evaluation as to which states are likely to occur, and eventually comes to choose *only one action* forever.

We assume that the decision maker does *not* know how the payoff depends on the state of the world, and therefore, has to learn, not only the probability of observation of the signal, but also the state-dependent payoff structure. With this assumption, the decision maker comes to reach the following rigid *misperception*: That is, the decision maker comes to believe incorrectly that there exists an action which always maximizes her subjective expected payoff *irrespective of* which states she anticipates are likely to occur. Hence, the decision maker comes to stop choosing the other actions, and stop gathering information relevant to the payoffs for these actions. Of particular importance is that this long run rigidity holds even though the decision maker can, not perfectly, but *almost perfectly*, monitor the state of the world.

In the literature of dynamic decision making problems, Bayesian models called the multi-armed bandit models have long been regarded as being standard, where the decision maker updates her probability belief according to the Bayes rule (Rothschild

(1974), Gittins (1989)). Bayesian models have also been applied to the situation in which the decision maker has little objective knowledge and updates her subjective belief (Savage (1954), Kalai and Lehrer (1995)). Despite its mathematical tractability, the application of Bayesian models to choice under *uncertainty* is criticized because the underlying assumption that the decision maker is ideally rational is inappropriate (Simon (1955, 1976), Binmore (1987, 1988), Dekel and Gul (1996), Matsushima (1997a)).

An alternative approach to Bayesian models in game theory is the study of *adaptive learning* (Brown (1951), Marimon (1997), Fudenberg and Tirole (1997)), where a player may not be fully rational, may not be well-informed, and learns inductively from past experiences about how the opponents will play according to a well-specified learning rule. Most relevant works assume either that a player has the complete knowledge about her own payoff function, or, in an evolutionary context, that a player can observe the other player's choices of action and their resultant payoffs. In contrast with these works, this paper assumes that the decision maker does not know her own payoff function, and also assumes that she can not observe the other decision maker's choices of action and their resultant payoffs.

In some works relevant to adaptive learning called the study of stimulus-response learning, the decision maker experiments with multiple actions, i.e., she plays a totally mixed action according to the suggestions of randomly selected stimuli (Cross (1983), Börgers and Sarin (1996)). In contrast with these works, this paper assumes that in every period the decision maker *never* experiments with any action which neither maximizes the subjective expected payoff nor is equal to the action chosen in the preceding period.

Most of the works relevant to adaptive learning have studied the convergence to the rational choice of the objective best response, with the constraints of limited prior knowledge and bounded rationality. This paper, however, investigates the possibility of permanent *dissonance* between the *subjective* evaluations and the *objective* structure. In the cognitive-dissonance literature of psychology, the tendency of people in a laboratory to manipulate subjective belief by avoiding unappealing information has long been discussed (Cotton (1985)). Recently, the possibility of the decision maker's misperception induced by the weak incentive to gather unbiased information becomes a growing concern among economists (Akerlof and Dickens (1982), Rabin (1995), Carrillo and Mariotti (1997)).

We also characterizes which actions the decision maker comes to choose in the long run. We will show that, with positive probability, the decision maker comes to choose

only an action, if and only if there exists no other action whose objective minimal payoff among possible signals and whose payoff evaluation in the initial period for some signal are more than the objective minimal payoff for this action. Especially, we will give a necessary and sufficient condition under which the objective *maximin* action, which maximizes the objective minimal payoff among possible actions, is the unique action which the decision maker comes to choose in the long run. That is, the decision maker comes to choose only the objective maximin action, if and only if there exists at least one signal for which the payoff evaluation for this action in the initial period is larger than the objective minimal payoff for any other action.

Readers may be reminded of one of the classical decision criteria of rationality under uncertainty called "the maximin criterion", according to which the decision maker by intention chooses the maximin action (Luce and Raiffa (1957), Simon (1955)). We must note that the decision maker in this paper does *not* accept this maximin criterion. What this paper says is that the action which the decision maker chooses by maximizing the subjective expected payoff happens to be objectively maximin.

After completing the previous version of this paper (Matsushima (1997)), I have become aware of the work by Sarin and Vahid (1997), which has independently showed the similar maximin property. That is, Sarin and Vahid showed, by investigating a different class of learning rules, that the decision maker comes to choose only the objective maximin action among the actions which she has ever chosen. Though points of departure are similar, I focus on different issues and the differences from their work might be even more important than the similarities. In particular, Sarin and Vahid assumed that the decision maker can not observe any signal relevant to the state of the world and never evaluates the probability of occurrence of the states of the world. In contrast with their work, one of the main theorem of this paper is that the rigidity of the decision maker's choice of action holds true even though she can *almost perfectly* monitor the states of the world.

Throughout this paper, we assume that the decision maker is myopic. Needless to say, this assumption is restrictive. Most papers in the Bayesian study of bandit models have assumed in fact that the decision maker maximizes the discounted sum of expected future payoffs with *positive* discount factor. I have not examined in this paper, but I have a conjecture that the tendency of the decision maker to misperceive the situations is robust even though we allow the decision maker to behave more patiently.

The organization of this paper is as follows. Section 2 gives an example which explains the intuition about the long run rigidity. It is shown that the decision maker succeeds to choose the objective best response in the perfect monitoring case, whereas

she fails to do so in the imperfect monitoring case. Section 3 gives the formal model, several assumptions, and one of the main theorem (Theorem 1) which says that the decision maker comes to choose only one action in the long run.

Section 4 characterizes the actions chosen in the long run (Proposition 2). Section 4 also gives a necessary and sufficient condition under which the decision maker comes to choose only the objective maximin action in the long run (Theorem 5).

Assumption 3 presented in Section 3 is the driving force for our results, which says that whenever the decision maker choose an action then she never get information relevant to the payoffs for the other actions. Section 5 weakens Assumption 3 and discuss the robustness of the rigidity of the decision maker's choices.

Section 6 presents an algorithm to determine the action which the decision maker comes to choose in the long run, where we assume that, in the initial period, the decision maker has the *state-independent* payoff evaluations and maximizes the subjective expected payoff.

Section 7 gives several discussions. Subsection 7.1 compares different signal structures. Subsection 7.2 gives a remark on the effect of the change of the minimal payoff vector. Finally, Subsection 7.3 gives a concluding remark.

2. An Example

Consider an entrepreneur (a decision maker) who chooses between the *uncertain* action and the *safe* action infinitely many times. There are two possible states of the world, i.e., state "*boom*" and state "*recession*". In every period, state "boom" occurs with positive probability $p > 0$, whereas state "recession" occurs with positive probability $1 - p > 0$. If the entrepreneur chooses the uncertain action and state "boom" occurs, she obtains 100 dollars. If she chooses the uncertain action and state "recession" occurs, she loses 100 dollars. If she chooses the safe action, she obtains 0 dollar, irrespective of which state actually occurs (See Figure 1).

[Figure 1]

At the end of every period t, the entrepreneur can *not* observe the realization of state of the world $\omega(t) \in \{boom, recession\}$, but instead can observe the realization of some *random signal* $\phi(t) \in \{B, R\}$. If state "boom" occurs, the entrepreneur observes signal "B" with probability $1 - \varepsilon \geq 0$ and signal "R" with probability $\varepsilon \geq 0$. If state "recession" occurs, she observes signal "B" with probability ε and signal "R" with probability $1 - \varepsilon$ (See Figure 2).

[Figure 2]

The entrepreneur a priori knows that the state-independent payoff for the safe action is zero. The entrepreneur does *not* know the stochastic payoffs for the uncertain action. She also does *not* know this signal structure.

She learns *inductively* from the past experiences: In every period t, the entrepreneur evaluates the payoff for state "boom" by $v^{(B)}(t) \in R$, and evaluates the payoff for state "recession" by $v^{(R)}(t) \in R$. She evaluates the probability for signal "B"

by $\delta(t) \in [0,1]$. In every period $t > 1$, the entrepreneur evaluates the payoffs and the probability by

$$v^{(\phi)}(t) = \begin{cases} v(t-1) & \text{if she chooses the uncertain action in period } t-1 \text{ and } \phi(t-1) = \phi, \\ v^{(\phi)}(t-1) & \text{otherwise,} \end{cases}$$

and

$$\delta(t) = \begin{cases} \theta + (1-\theta)\delta(t-1) & \text{if } \phi(t-1) = B, \\ (1-\theta)\delta(t-1) & \text{otherwise,} \end{cases}$$

respectively, where $\phi \in \{B, R\}$ and $\theta \in (0,1)$. When the entrepreneur has chosen the uncertain action and observed signal ϕ in the last period, she anticipates in this period that, by choosing the uncertain action and observing the same signal ϕ, she obtains the same payoff as the one which she has obtained in the last period. Otherwise, the subjective payoff evaluation is unchanged. We assume that θ is so close to 1 that for every large enough t, $\delta(t)$ is in the neighborhood of the true probability for signal "B", i.e., $p(1-\varepsilon) + (1-p)\varepsilon$. For convenience of the arguments in this section, we assume that $v^{(B)}(1) > 100$ and $v^{(R)}(1) > 100$.

In every period t, the entrepreneur chooses the *same* action as the one chosen in the last period with positive probability $\eta > 0$, and also chooses the action which *maximizes* her subjective expected payoff with positive probability $1 - \eta > \eta$. We must note that the safe action maximizes her subjective expected payoff if

$$\delta(t)v^{(B)}(t) + (1-\delta(t))v^{(R)}(t) \leq 0,$$

whereas the uncertain action maximizes it if

$$\delta(t)v^{(B)}(t) + (1-\delta(t))v^{(R)}(t) \geq 0.$$

Of particular importance is that the entrepreneur *never* experiments with the action which neither maximizes her subjective expected payoff nor is the same as the action chosen in the last period.

First, consider the case that $\varepsilon = 0$. In this case, the entrepreneur can *perfectly* monitor what the true state of the world is, that is, she observes signal "B" if and only if state "boom" occurs. By choosing the uncertain action and observing signal "B" in a period t, the entrepreneur certainly obtains 100 dollars and makes $v^{(B)}(t+1)$ equal to 100. By choosing the uncertain action and observing signal "R", she certainly loses 100 dollars and makes $v^{(R)}(t+1)$ equal to -100. Moreover, in the long run, the probability evaluation $\delta(t)$ for signal "B" is in the neighborhood of the true probability p for state "boom", because $\varepsilon = 0$ and θ is close enough to unity. Hence, the entrepreneur comes

to choose only the *objective* best response, that is, comes to choose only the safe action if $p < 1/2$, and choose only the uncertain action if $p > 1/2$.

Our main concern is the case that $\varepsilon > 0$. In this case, the entrepreneur can *not* perfectly monitor the state of the world. We must note that the entrepreneur can not perfectly, but *almost* perfectly, monitor the state of the world if ε is close enough to zero. This imperfect monitoring case substantially *differs* from the perfect monitoring case: The entrepreneur makes both $v^{(B)}(t)$ and $v^{(R)}(t)$ equal to -100, and stops choosing the uncertain action forever, whether it is the objective best response or not (See Figure 3).

[Figure 3]

We give the brief explanation below. For every $t \geq 1$, consider the event that:

i) the entrepreneur chooses the uncertain action, observes signal "B", and obtains payoff -100 in period $t+1$,

ii) she chooses the uncertain action, observes signal "R", and obtains payoff -100 in period $t+2$,

and

iii) she chooses the safe action in period $t+3$.

We must note that if this event occurs, it holds that $v^{(B)}(t+3) = -100$ and $v^{(R)}(t+3) = -100$. Hence, the entrepreneur regards the safe action as the *subjective* best response, and therefore, never chooses the uncertain action. Because the entrepreneur never experiments with the uncertain action, she never changes its payoff evaluations. Given that the entrepreneur has chosen the uncertain action in period t, this event occurs in period $t+3$ at least with positive probability

$$(1-p)^2 \varepsilon(1-\varepsilon)\eta^2(1-\eta) > 0.$$

Suppose that the entrepreneur chooses the uncertain action infinitely many times. Then, this event almost surely occurs at least at once within the finite time horizon. This is a contradiction, because once this event occurs she never chooses the uncertain action. Hence, she chooses the uncertain action at most finitely many times, and therefore, she eventually comes to stop choosing it.

We can also check that the entrepreneur comes to make both $v^{(B)}(t)$ and $v^{(R)}(t)$

equal to -100: Suppose that the entrepreneur stops choosing the uncertain action but $v^{(B)}(t) \neq -100$, that is, either $v^{(B)}(t) = v^{(B)}(1) > 100$ or $v^{(B)}(t) = 100$. Choose $\xi > 0$ such that

$$\xi v^{(B)}(t) + (1-\xi)v^{(R)}(t) > 0,$$

and choose a positive integer s' large enough to satisfy that

$$(1-\theta)^{s'} < \xi.$$

For every period $t \geq 1$, consider the event that the entrepreneur observes signal "B" from period $t+1$ through period $t+s'$. If this event occurs, the subjective expected payoff for the uncertain action in period $t+s'+1$ becomes more than zero. Since this event occurs with positive probability $\{p(1-\varepsilon)+(1-p)\varepsilon\}^{s'} > 0$ in every period, it occurs at least at once within the finite time horizon. This is a contradiction, because once this event occurs the entrepreneur has incentive to choose the uncertain action again. Hence, it must hold that $v^{(B)}(t) = -100$ in the long run. Similarly we can also check that $v^{(R)}(t) = -100$ in the long run.

From these observations, we have concluded that $v^{(B)}(t) = v^{(R)}(t) = -100$ and the entrepreneur comes to choose only the safe action in the long run, whether it is the objective best response or not.

The above example is unsatisfactory for the following reasons. First, we assumed that the probability that state "boom" occurs, p, is constant. It is more realistic that it depends on time and history, and therefore, which action is the objective best response crucially depends on time and history.

Second, we assumed that there exists the unique uncertain action. It is more realistic that there are multiple uncertain actions the payoffs for which are a priori unknown to the decision maker.

Third, we confined our attention to a very limited class of learning procedures. It might be much more satisfactory if we can derive general results in a wider class of learning procedures.

Fourth, it is needless to say that the assumption that the payoff evaluations for the uncertain action in the initial period are more than 100, is very restrictive.

These drawbacks urges us to investigate the more general models in the next sections.

3. Assumptions and the Basic Theorem

A long-run single-person decision making problem $D \equiv (A, \Omega, \Phi, H, u^{(\cdot)}, p^{(\cdot)}, q^{(\cdot)})$ is defined as follows: A is the finite set of *actions*. Ω is the finite set of *states*. Φ is the finite set of *signals*. A decision maker chooses actions among A infinitely many times. In each period $t \geq 1$, the decision maker chooses an action $a(t) \in A$ and a state $\omega(t) \in \Omega$ is realized. The decision maker can *not* directly observe this realized state $\omega(t)$. At the end of period t, the decision maker obtains a payoff $v(t) \in R$ and observes the realization of a *random* signal $\phi(t) \in \Phi$.

Let h^0 be *the null history* and $H^0 \equiv \{h^0\}$. Let $u^{(h^0)}: A \times \Omega \to R$ be the payoff function in period 1. Recursively, for every $t \geq 1$, let $h^t \equiv (a(\tau), \omega(\tau), v(\tau), \phi(\tau))_{\tau=1}^{t}$ be *a history up to period* t, and let $u^{(h^t)}: A \times \Omega \to R$ be the payoff function in period $t+1$ provided that h^t is realized, where $v(\tau) = u(a(\tau), \omega(\tau))$ for all $\tau \in \{1, \ldots, t\}$. Let H^t be the set of all histories h^t up to period t, and $H \equiv \bigcup_{t=0}^{\infty} H^t$. The payoff function $u^{(h^t)}$ may be history-dependent. In every period t, the decision maker obtains the payoff $v(t) = u^{(h^{t-1})}(a(t), \omega(t))$ according to this history-dependent payoff function $u^{(h^{t-1})}$.

Let $p^{(h^{t-1})} \equiv p^{(h^{t-1})}(\cdot | a(t)): \Omega \to R_+$ be a probability function on Ω, and let $q^{(h^{t-1})} \equiv q^{(h^{t-1})}(\cdot | a(t), \omega(t)): \Phi \to R_+$ be a probability function on Φ. When h^{t-1} was realized and the decision maker chose $a(t) \in A$ in period t, a state $\omega(t) \in \Omega$ is realized with probability $p^{(h^{t-1})}(\omega(t) | a(t))$, and the decision maker observes a signal $\phi(t) \in \Phi$ with probability $q^{(h^{t-1})}(\phi(t) | a(t), \omega(t))$. We must note that both of these probability functions, $p^{(h^{t-1})}$ and $q^{(h^{t-1})}$, may be history-dependent.

A decision maker is modeled by *a learning rule* (d, Γ). $d: H \to \Delta(A)$ is *a decision rule*, where $\Delta(A)$ is the set of mixed actions. For every $t \geq 1$, every $h^{t-1} \in H^{t-1}$ and every $a \in A$, the decision maker chooses a with probability $d(h^{t-1})(a)$. We must note that $d(h^{t-1})$ is *independent* of $(\omega(\tau))_{\tau=1}^{t-1}$, because the decision maker can not observe the true states. $\Gamma \equiv (\Gamma^{(a)})_{a \in A}$ is *an evaluation rule*, where $\Gamma^{(a)} \equiv ((v^{(a,\phi)})_{\phi \in \Phi}, \delta^{(a)})$ is *an evaluation rule for action* a, $v^{(a,\phi)}: H \to R$ is *a payoff evaluation rule for* (a, ϕ), $\delta^{(a)}: H \to \Delta(\Phi)$ is *a probability evaluation rule for action* a, and $\Delta(\Phi)$ is the set of probability functions on Φ. The decision maker anticipates that she obtains the payoff, or the expected payoff, $v^{(a,\phi)}(h^{t-1})$, in period t, when h^{t-1} was realized, she chooses action $a(t) = a$, and she observes signal $\phi(t) = \phi$. The decision maker also anticipates that she observes signal $\phi(t) = \phi$ with probability $\delta^{(a)}(h^{t-1})(\phi)$ in period t, when h^{t-1} was realized and she chooses action $a(t) = a$. For every $a \in A$, every $t \geq 1$, and every

$h^{t-1} \in H^{t-1}$, *the subjective expected payoff* is defined by

$$V^{(a)}(h^{t-1}) \equiv \sum_{\phi \in \Phi} \delta^{(a)}(h^{t-1})(\phi) v^{(a,\phi)}(h^{t-1}).$$

We assume that $v^{(a,\phi)}(\cdot)$ is *bounded*.

We present six assumptions on D and (d, Γ) as follows.

Assumption 1: For every $t > 1$ and every $h^{t-1} \in H^{t-1}$,

$$d(h^{t-1})(a(t-1)) > 0,$$

$$[V^{(a)}(h^{t-1}) \geq V^{(a')}(h^{t-1}) \text{ for all } a' \in A] \Rightarrow [d(h^{t-1})(a) > 0],$$

and

$$[a \neq a(t-1) \text{ and } V^{(a)}(h^{t-1}) < V^{(a')}(h^{t-1}) \text{ for some } a' \in A]$$
$$\Rightarrow [d(h^{t-1})(a) = 0].$$

The first inequality in Assumption 1 implies that the decision maker maximizes the subjective expected payoff with positive probability. The second inequality implies that she maximizes the subjective expected payoff with positive probability. The third inequality implies that she never experiments with any action which neither maximizes the subjective expected payoff nor is the action chosen in the last period.

Assumption 2: For every $t \geq 1$, every $h^t \in H^t$, and every $(a, \phi) \in A \times \Phi$,

$$v^{(a,\phi)}(h^t) \geq \min\left[v(t), v^{(a,\phi)}(h^{t-1})\right] \text{ whenever } a(t) = a.$$

Assumption 2 implies that the decision maker never makes the payoff evaluations for an action less than either the obtained payoff or the payoff evaluations in the preceding period, provided that she has chosen this action.

Assumption 3: For every $t \geq 1$, every $h^t \in H^t$, and every $(a, \phi) \in A \times \Phi$,

$$[(a(t), \phi(t)) \neq (a, \phi)] \Rightarrow [v^{(a,\phi)}(h^t) = v^{(a,\phi)}(h^{t-1})].$$

Assumption 3 implies that the payoff evaluation for an action and a signal is influenced *only* by the decision maker's experiences when she *actually* choosing this action and observing this signal. Assumption 3 plays the crucial role in deriving Theorem 1 presented in this section. Later in Section 5, we will weaken Assumption 3.

Assumption 4: For every $\mu > 0$, there exists a positive integer s^* such that for every $(a, v, \phi) \in A \times V \times \Phi$, every $t > s^*$ and every $h^{t-1} \in H^{t-1}$, if

$$(a(\tau), v(\tau), \phi(\tau)) = (a, v, \phi) \text{ for all } \tau = t - s^*, \ldots, t-1,$$

then

$$|v^{(a,\phi)}(h^{t-1}) - v| < \mu,$$

and

$$\delta^{(a')}(h^{t-1})(\phi) > 1 - \mu \text{ for all } a' \in A.$$

Assumption 4 says that if the decision maker has chosen action a, obtained payoff v, and observed signal ϕ for a large number of periods, then she believes that, by choosing action a, she almost certainly observes signal ϕ and obtains payoff v approximately. Assumption 4 is satisfied if the decision maker always gives a fixed positive weight to the current experience.

Assumption 5: For every $a \in A$, there exists a finite set of possible payoffs $V(a) \subset R$ such that for every $t \geq 0$ and every $h^t \in H^t$,

$$\{v \in R: v = u^{(h^t)}(a, \omega) \textit{ for some } \omega \in \Omega\} = V(a).$$

Assumption 5 automatically holds if the payoff functions $u^{(h^t)}$ are independent of time and history. Clearly, the example in Section 2 satisfies Assumption 5.

We define

$$\underline{v}(a) \equiv \min V(a),$$

which is the *minimal* payoff for action a. We must note from Assumptions 2, 3 and 5 that for every $t \geq 1$, every $h^t \in H^t$, and every $(a, \phi) \in A \times \Phi$,

$$v^{(a,\phi)}(h^t) \geq \min\left[\underline{v}(a),\ v^{(a,\phi)}(h^0)\right], \tag{1}$$

which will be presented again as Assumption 9 in Section 5.

Assumption 6: There exists a positive real number $\varepsilon > 0$ such that for every $t \geq 1$, every $h^{t-1} \in H^{t-1}$, and every $(\omega, \phi) \in \Omega \times \Phi$,

$$p^{(h^{t-1})}(\omega | a) > \varepsilon,$$

$$q^{(h^{t-1})}(\phi | a, \omega) > \varepsilon,$$

$$[d(h^{t-1})(a) > 0] \Leftrightarrow [d(h^{t-1})(a) > \varepsilon] \text{ for all } a \in A,$$

and

$$|v^{(a,\phi)}(h^t) - \underline{v}(a')| > \varepsilon \text{ for all } a \in A \text{ and all } a' \neq a.$$

The first two inequalities imply that the probability functions $p^{(h^{t-1})}(\cdot | a)$ and $q^{(h^{t-1})}(\cdot | a, \omega)$ have the *full supports* with uniform positive lower bounds. The third

inequalities imply that the positive probabilities on the choice of action have also the uniform positive lower bound.

We must note that these inequalities automatically hold if this dynamic stochastic process is described by a *finite Markov chain*. The fourth inequalities imply that the payoff evaluation for an action never converges to the minimal payoff for another action. We must note that the fourth inequalities and Assumption 4 imply that for every $a \in A$ and all $a' \neq a$,

$$\underline{v}(a) \neq \underline{v}(a'),$$

and

$$v^{(a,\phi)}(h^t) \neq \underline{v}(a') \text{ for all } t \geq 0 \text{ and all } h^t \in H^t.$$

Let the real number $\varepsilon > 0$ in Assumption 6 be chosen close to zero.

Assumption 6 is required for simplicity of our arguments. The essence of the theorem presented below will be unchanged even though Assumption 6 is dropped.

For every $t' \geq 1$ and every $t > t'$, a history up to period t, $h^t \in H^t$, is said to be *reachable from* a history up to period t', $h^{t'} \in H^{t'}$, if for every $\tau \in \{t'+1, \ldots, t\}$,

$$d(h^{\tau-1})(a(\tau)) > 0.$$

The basic theorem in this paper is presented as follows.

Theorem 1: *Suppose that Assumptions 1 through 6 hold. Then, for every* $\xi \in (0,1]$, *there exists a positive integer* s *such that for every* $t \geq s$, *the following property holds at least with probability* $1-\xi$: *There exists* $a^* \in A$ *such that*

$$d(h^t)(a^*) = 1,$$

$$v^{(a^*,\phi)}(h^t) \geq \underline{v}(a^*) - \varepsilon > v^{(a,\phi)}(h^t) \text{ for all } a \neq a^* \text{ and all } \phi \in \Phi,$$

and for every $t' \geq 1$ *and every* $h^{t'} \in H^{t'}$ *that is reachable from* h^t,

$$d(h^{t'})(a^*) = 1,$$

$$v^{(a^*,\phi)}(h^{t'}) \geq \underline{v}(a^*) - \varepsilon,$$

and

$$v^{(a,\phi)}(h^{t'}) = v^{(a,\phi)}(h^t) \text{ for all } a \neq a^* \text{ and all } \phi \in \Phi.$$

Theorem 1 implies that, in the long run, the decision maker comes to choose only one action a^*, never choose any other action, and never change the payoff evaluations for any action other than action a^*, irrespective of which signals and payoffs she experiences in the future. This long run rigidity of the decision maker's choices and payoff evaluations holds true even though the decision maker can *almost perfectly* monitor the true state of the world. Since which action is the objective best response crucially depends on the distributions of state of the world $p^{(h^{t-1})}(\omega|a)$, one gets from

Theorem 1 that the decision maker in general fails to choose the objective best response in the long run.

Proof of Theorem 1: Let $m^* \equiv |\Phi|$, $k^* \equiv |A|$, and denote $\Phi = \{\phi^1, \ldots, \phi^{m^*}\}$ and $A = \{a^1, \ldots, a^{k^*}\}$. Let the real number $\mu > 0$ in Assumption 4 be less than ε. For every $t \geq 1$, consider the event that:

1) from period $t+1$ through period $t+s^*$, the decision maker chooses action $a = a(t)$, observes signal ϕ^1 and obtains payoff $\underline{v}(a)$,

recursively, for every $k \in \{1, \ldots, k^*\}$ and every $m \in \{1, .., m^*\}$,

2) from period $\tau(k,m)+1$ through period $\tau(k,m)+m^*s^*$, the decision maker chooses the action a which maximizes the subjective expected payoff in period $\tau(k,m)+1$ and obtains payoff $\underline{v}(a)$, where

$$\tau(k,m) \equiv t + s^* + (k-1)(m^*)^2 s^* + (m-1)m^* s^*,$$

for every $m' \in \{1, \ldots, m^*\}$,

3) from period $\tau(k,m)+(m'-1)s^*+1$ through period $\tau(k,m)+m's^*$, the decision maker observes signal $\phi^{m+m'}$, where

$$\phi^{m+m'} = \phi^{m+m'-m^*} \text{ if } m+m' > m^*,$$

and

4) in period $\tau(k^*+1,1)+1$, the decision maker chooses the action a which maximizes the subjective expected payoff, obtains payoff $\underline{v}(a)$, and observes signal ϕ^1.

Here s^* is the integer introduced in Assumption 4.

Assumption 4 says that for every $m \in \{1, \ldots, m^*\}$, every $k \in \{1, \ldots, k^*\}$, every $m' \in \{1, \ldots, m^*\}$, and for $a = a(\tau(k,m)+m's^*)$,

$$\underline{v}(a) - \mu < v^{(a,\phi^{m+m'})}(h^{\tau(k,m)+m's^*}) < \underline{v}(a) + \mu.$$

Let $A^* \subset A$ be the set of actions a such that

$$a = a(t') \text{ for some } t' \in \{1 + s^* + 1, \ldots, \tau(k^*+1,1)\}.$$

Let $a^* \in A^*$ be the action such that

$$\underline{v}(a^*) > \underline{v}(a) \text{ for all } a \in A^* / \{a^*\}.$$

Assumptions 1 and 3 say that if this event occurs, for every $\phi \in \Phi$,

$$|v^{(a,\phi)}(h^{\tau(k^*+1,1)+1}) - \underline{v}(a)| < \mu \text{ for all } a \in A^*,$$

and

$$v^{(a,\phi)}(h^{\tau(k^*+1,1)+1}) < \underline{v}(a^*) + \mu \text{ for all } a \notin A^*.$$

From the fourth inequalities in Assumption 6 and $\mu < \varepsilon$, the last inequality means

$$v^{(a,\phi)}(h^{\tau(k^*+1,1)+1}) < \underline{v}(a^*) - \varepsilon.$$

Since we can choose μ and ε so close to zero that

$$\underline{v}(a) + \mu < \underline{v}(a^*) - \varepsilon \text{ for all } a \in A^* / \{a^*\},$$

one gets that for every $\phi \in \Phi$ and every $a \neq a^*$,

$$v^{(a^*,\phi)}(h^{\tau(k^*+1,1)+1}) \geq \underline{v}(a^*) - \varepsilon > v^{(a,\phi)}(h^{\tau(k^*+1,1)+1}).$$

This, together with Assumption 1, implies that after this event occurs the decision maker never chooses any action other than a^*. Assumption 3 implies that the decision maker never changes the payoff evaluations for any action other than a^*. Assumption 2 implies that the decision maker never makes the payoff evaluations for action a^* less than $\underline{v}(a^*) - \varepsilon$. Moreover, Assumption 6 implies that in every period this event occurs at least with positive probability $\varepsilon^{3\{s^*+k^*(m^*)^2 s^*+1\}} > 0$, and therefore, it is almost certain that this event occurs at least at once within the finite time horizon. From these observations, we have completed the proof of this theorem.

Q.E.D.

4. Characterization

In this section, we characterizes the action a^* presented in Theorem 1 which the decision maker sticks to choose in the long run.

Proposition 2: *Suppose that Assumptions 1 through 6 hold. Then, the action* $a^* \in A$ *presented in Theorem 1 satisfies the following inequalities for* $a = a^*$,

$$\underline{v}(a) \geq \min\left[\underline{v}(a'), \max_{\phi \in \Phi} v^{(a',\phi)}(h^0)\right] \text{ for all } a' \in A. \tag{2}$$

Moreover, if an action $a \in A$ *satisfies inequalities (2) and* $d(h^0)(a) > 0$, *then* $a^* = a$ *holds with positive probability.*

Proof: Suppose that a^* does not satisfy inequalities (2), that is,

$$\underline{v}(a^*) < \underline{v}(a) \text{ and } \underline{v}(a^*) < v^{(a,\phi)}(h^0) \text{ for some } a \neq a^* \text{ and some } \phi \in \Phi.$$

This, together with inequalities (1), implies that for every $t \geq 0$ and every $h^t \in H^t$,

$$v^{(a,\phi)}(h^t) > \underline{v}(a^*),$$

which is a contradiction because Assumption 1 implies that the decision maker chooses $a \neq a^*$ with positive probability. Hence, a^* must satisfy inequalities (2).

Next, suppose that a satisfies inequalities (2) and $d(h^0)(a) > 0$. Consider the event presented in the proof of Theorem 1 for $t = 1$ with $a(1) = a$. Since $d(h^0)(a) > \varepsilon$, this event occurs at least with positive probability $\varepsilon^{3\{s^* + k^*(m^*)^2 s^* + 1} > 0$, and after period $2 + s^* + k^*(m^*)^2 s^*$, action $a^* = a$ satisfies the properties presented in Theorem 1.

Q.E.D.

Inequalities (2) say that there exists no action whose minimal payoff and whose payoff evaluation in the initial period for some signal are more than the minimal payoff for action a. Proposition 2 says that inequalities (2) are, in some sense, not only necessary, but also sufficient.

Let $\hat{a} \in A$ be the *maximin* action in the sense that

$$\underline{v}(\hat{a}) \geq \underline{v}(a) \text{ for all } a \in A.$$

The maximin action maximizes the minimal payoff $\underline{v}(a)$ with respect to pure action $a \in A$.[1] Since $\underline{v}(a) \neq \underline{v}(a')$ for all $a \in A$ and all $a' \neq a$, the maximin action $\hat{a}$

---

[1] The definition of maximin action in this paper is different from the definition in the textbooks for game theory. The latter is in terms of mixed action, whereas the former is in terms of pure action.

uniquely exists.

Lemma 3: *Inequalities (2) for* $a = \hat{a}$ *always holds.*

Proof: From the definition of maximin action, one gets that for every $a \in A$,

$$\underline{v}(\hat{a}) \geq \underline{v}(a) \geq \min\left[\underline{v}(a), \max_{\phi \in \Phi} v^{(a,\phi)}(h^0)\right].$$

Q.E.D.

From Proposition 2 and Lemma 3, one gets that whenever the learning rule assigns the maximin action a positive probability in the initial period, i.e., $d(h^0)(\hat{a}) > 0$, then, with positive probability, the decision maker comes to choose only the maximin action $\hat{a}$ in the long run.

The main purpose of this section is to clarify a necessary and sufficient condition under which the maximin action $\hat{a}$ is the *unique* action which the decision maker sticks to choose in the long run.

Lemma 4: *The maximin action* $\hat{a}$ *is the unique action* a *which satisfies inequalities (2) if and only if there exists* $\phi \in \Phi$ *such that*

$$v^{(\hat{a},\phi)}(h^0) \geq \underline{v}(a) \text{ for all } a \neq \hat{a}. \tag{3}$$

Proof: The proof of the "if" part is as follows. Inequalities (3) says that

$$\min\left[\underline{v}(\hat{a}), \max_{\phi \in \Phi} v^{(\hat{a},\phi)}(h^0)\right] > \underline{v}(a) \text{ for all } a \neq \hat{a},$$

which implies the violation of inequalities (2) for all $a \neq \hat{a}$. Lemma 3 says that $\hat{a}$ is the unique action a which satisfies inequalities (2).

Next, the proof of the "only if" part is as follows. Choose an action $\tilde{a} \neq \hat{a}$ so as to satisfy

$$\underline{v}(\tilde{a}) \geq \underline{v}(a') \text{ for all } a' \neq \hat{a}.$$

Suppose that inequalities (3) do not hold. Then,

$$\underline{v}(\tilde{a}) > v^{(\hat{a},\phi)}(h^0) \text{ for all } \phi \in \Phi,$$

and therefore, one gets

$$\underline{v}(\tilde{a}) \geq \min\left[\underline{v}(a'), \max_{\phi \in \Phi} v^{(a',\phi)}(h^0)\right] \text{ for all } a' \in A,$$

which is equivalent to inequalities (2) for $a = \tilde{a}$.

Q.E.D.

Inequalities (3) say that there exists at least one signal for which the payoff evaluation for the maximin action in the initial period is larger than the minimal payoff for any other action.

The main theorem of this section is presented as follows.

Theorem 5: *Suppose that Assumptions 1 through 6 hold. If there exists* $\phi \in \Phi$ *which satisfies inequalities (3), then* $a^* = \hat{a}$ *always holds. If there exists no such* ϕ *and an action* $a \neq \hat{a}$ *satisfies that* $d(h^0)(a) > 0$ *and*

$$\underline{v}(a) \geq \underline{v}(a') \text{ for all } a' \neq \hat{a},$$

then $a^* = a \neq \hat{a}$ *holds with positive probability.*

Proof: The first part of this theorem is straightforward from Proposition 2, Lemma 3 and lemma 4.

We will prove the latter part of this theorem as follows. Since there exists no ϕ which satisfies inequalities (3), the proof of the "only if" part of Lemma 4 says that the action $a \neq \hat{a}$ such that $\underline{v}(a) \geq \underline{v}(a')$ for all $a' \neq \hat{a}$ satisfies inequalities (2). Hence, the latter part of this theorem is straightforward from the latter part of Proposition 2.

Q.E.D.

Theorem 5 implies that the decision maker comes to choose only the maximin action $\hat{a}$ in the long run, if and only if inequalities (3) holds, that is, if and only if there exists at least one signal for which the payoff evaluation for the maximin action $\hat{a}$ in the initial period is larger than the minimal payoff for any action other than $\hat{a}$.

5. Some Generalization

Assumption 3, which is one of the most crucial assumption for deriving these results, may be restrictive in a class of situations. A real decision maker may learn something relevant to an action when she chooses any other action. In this section, we will weaken Assumption 3 and present necessary and sufficient conditions under which these properties hold.

The following two assumptions are weaker than Assumption 3.

Assumption 7: For every $t \geq 1$, every $h^t \in H^t$, and every $(a, \phi) \in A \times \Phi$,

$$v^{(a,\phi)}(h^t) \leq \max\left[v(t),\ v^{(a,\phi)}(h^{t-1})\right] \text{ whenever } a = a(t) \text{ and } \phi \neq \phi(t).$$

Assumption 8: For every $\tau \geq 1$, every $t \geq \tau + 1$, every $h^t \in H^t$, and every $a \in A$, if $a(\tau) = a$, $a(t) = a$, and $a(\tau') \neq a$ for $\tau' = \tau + 1, \ldots, t-1$, then

$$v^{(a,\phi)}(h^{t-1}) \leq v^{(a,\phi)}(h^{\tau}) \text{ for all } \phi \in \Phi.$$

Assumption 7 excludes the case in which the decision maker makes the payoff evaluation for an action and a signal larger than both the obtained payoff and the payoff evaluation in the preceding period, provided that she has chosen this action and observed this signal. Assumption 8 implies that the payoff evaluations for an action when the decision maker starts choosing it is always less than or equal to the evaluations in the preceding period in which she has just stopped choosing it.

Theorem 1 in Section 3 explained the possibility that the decision maker comes to choose only one action in the long run. The following proposition says that the property in Theorem 1 still holds even if we replace Assumption 3 with weaker assumptions such as Assumptions 7 and 8.

Proposition 6: *Suppose that Assumptions 1, 2, and 4 through 8 hold. Then, for every* $\xi \in (0, 1]$, *there exists a positive integer* s *such that for every* $t \geq s$, *the following property holds at least with probability* $1 - \xi$*: There exists* $a^* \in A$ *such that*

$$d(h^t)(a^*) = 1,$$

and for every $t' > t$ *and every* $h^{t'} \in H^{t'}$ *that is reachable from* h^t,

$$d(h^{t'})(a^*) = 1.$$

Proof: Let the real number $\mu > 0$ in Assumption 4 be less than ε. For every $t \geq 1$, consider the event presented in the proof of Theorem 1. Assumption 4 says that for

every $m \in \{1,\ldots,m^*\}$, every $k \in \{1,\ldots,k^*\}$, every $m' \in \{1,\ldots,m^*\}$, and for $a = a(\tau(k,m)+m's^*)$,

$$\underline{v}(a) - \mu < v^{(a,\phi^{m+m'})}(h^{\tau(k,m)+m's^*}) < \underline{v}(a) + \mu .$$

Assumptions 1, 7 and 8 say that if this event occurs, for every $\phi \in \Phi$,

$$|v^{(a,\phi)}(h^{\lambda(a)}) - \underline{v}(a)| < \mu \text{ for all } a \in A^*,$$

and

$$v^{(a,\phi)}(h^{\lambda(a)}) < \underline{v}(a^*) + \mu \text{ for all } a \notin A^*,$$

where A^* and $a^* \in A^*$ are the set of actions and the action defined in the proof of Theorem 1, and $\lambda(a) \in \{1,\ldots,\tau(k^*+1,1)+1\}$ is the last period in which the decision maker has chosen action a. Let $\lambda(a) = 0$ if $a(\tau) \neq a$ for all $\tau \in \{1,\ldots,\tau(k^*+1,1)+1\}$.

We can check that the decision maker never chooses any action other than a^* in the following way: If this is not true, then there exist an action $a \neq a^*$ and a period $\tau > \tau(k^*+1,1)+1$ such that $a = a(\tau)$ and $a(\tau') = a^* \neq a$ for all $\tau' \in \{\lambda(a)+1,\ldots,\tau-1\}$. Assumption 1 says that this action a maximizes the subjective expected payoff in period τ. But this is a contradiction, because Assumption 8 says that for every $\phi \in \Phi$,

$$v^{(a,\phi)}(h^{\tau-1}) \leq v^{(a,\phi)}(h^{\lambda(a)}) < \underline{v}(a^*) - \varepsilon ,$$

whereas Assumption 2 says that for every $\phi \in \Phi$,

$$v^{(a,\phi)}(h^{\tau-1}) \geq \underline{v}(a^*) - \varepsilon .$$

Similarly to the proof of Theorem 1, one gets that it is almost certain that this event occurs at once within the finite time horizon. Hence, we have completed the proof of this proposition.

Q.E.D.

We also introduce another assumption which automatically holds if Assumptions 2, 3, and 5 hold.

Assumption 9: For every $t \geq 1$ and every $h^t \in H^t$, inequalities (1) hold.

Proposition 2 in Section 4 characterized the class of actions which the decision maker sticks to choose in the long run. The following proposition says that the characterization in Proposition 2 still holds even if we replace Assumption 3 with weaker assumptions such as Assumptions 7, 8 and 9.

Proposition 7: *Suppose that Assumptions 1, 2, and 4 through 9 hold. Then, the action $a^* \in A$ presented in Proposition 6 satisfies inequalities (2) for $a = a^*$. Moreover, if an action $a \in A$ satisfies inequalities (2) and $d(h^0)(a) > 0$, then $a^* = a$ holds with*

positive probability.

Proof: The former part of this proposition is proved in the same way as the proof of the former part of Proposition 2.

Next, suppose that a satisfies inequalities (2) and $d(h^0)(a) > 0$. Consider the event presented in the proof of Proposition 6 for $t = 1$ with $a(1) = a$. Since $d(h^0)(a) > \varepsilon$, this event occurs at least with positive probability $\varepsilon^{3\{s^*+k^*(m^*)^2 s^*+1}} > 0$, and after period $2 + s^* + k^*(m^*)^2 s^*$, $a^* = a$ satisfies the properties presented in Proposition 5.

Q.E.D.

Theorem 5 in Section 4 explained the possibility that the decision maker comes to choose only the unique maximin action in the long run. The following proposition says that the property presented in Theorem 5 still holds even if we replace Assumption 3 with weaker assumptions such as Assumptions 7, 8 and 9.

Proposition 8: *Suppose that Assumptions 1, 2, and 4 through 9 hold. If there exists* $\phi \in \Phi$ *such that inequalities (3) hold, then* $a^* = \hat{a}$ *always holds. If there exists no such* ϕ *and an action* $a \neq \hat{a}$ *satisfies that* $d(h^0)(a) > 0$ *and*

$$\underline{v}(a) \geq \underline{v}(a') \text{ for all } a' \neq \hat{a},$$

then $a^* = a \neq \hat{a}$ *holds with positive probability.*

Proof: The first part of this proposition is straightforward from Proposition 7, Lemma 3 and lemma 4.

We will prove the latter part of this proposition as follows. Since there exists no ϕ which satisfies inequalities (3), the proof of the "only if" part of Lemma 4 says that the action $a \neq \hat{a}$ such that $\underline{v}(a) \geq \underline{v}(a')$ for all $a' \neq \hat{a}$ satisfies inequalities (2). Hence, the latter part of this proposition is straightforward from the latter part of Proposition 7.

Q.E.D.

In spite of the above arguments on the robustness of our results presented in the previous sections, we must admit that Assumption 8 is still restrictive when the observed signal ϕ includes information as to the other decision makers' choices of action and resultant payoffs in similar decision making problems. If the decision maker observed the fact that the other decision makers have chosen an action other than action a^* and obtained high payoffs, it might be reasonable that she increases the payoff evaluations for this action and choose it again. Apparently this contradicts the

requirement of Assumption 8. A related point will be also discussed in the companion paper Matsushima (1998a).

6. State-Independent Initial Evaluations

In this section, by requiring additional assumptions on learning rules, we give an *algorithm* to determine the action a^* which the decision maker comes to choose in the long run .

At the very beginning of the long-run decision making problem, the decision maker may have serious lack of reasons why and how her payoff evaluations condition on different signals. In this case it might be reasonable for the decision maker to start with the *state-independent* payoff evaluations. Hence, we assume:

Assumption 10: For every $a \in A$,

$$v^{(a,\phi)}(h^0) = v^{(a,\phi')}(h^0) \text{ for all } \phi \in \Phi \text{ and all } \phi' \in \Phi.$$

Needless to say, a special case which satisfies Assumption 10 is the case in which the decision maker can observe no signal, i.e., the set of signals Φ is a singleton.[2]

On Assumption 10, inequalities (2) is equivalent to the following inequalities:

$$\underline{v}(a) \geq \min\left[\underline{v}(a'), V^{(h^0)}(a')\right] \text{ for all } a' \in A. \tag{4}$$

We denote $A = \{a_1, \ldots, a_k\}$, where $k \equiv |A|$, and

$$V^{(h^0)}(a_1) > V^{(h^0)}(a_2) > \cdots > V^{(h^0)}(a_k).$$

For every $q \in \{1, \ldots, k\}$, the action a_q is the $q-th$ subjective best response in the initial period. Moreover, we assume that the decision maker chooses the action which maximizes the subjective expected payoff in the initial period *with certainty*. That is,

Assumption 11: $d(h^0)(a_1) = 1$.

We will specify an action, action $\tilde{a}$, according to the following algorithm. For every $q \in \{1, \ldots, k\}$, we define

$$A_q \equiv \{a_1, \ldots, a_q\},$$

and let $a^{(q)} \in A_q$ be the maximin action among A_q, i.e.,

$$\underline{v}(a^{(q)}) \geq \underline{v}(a) \text{ for all } a \in A_q.^3$$

Let $\tilde{q} \in \{1, \ldots, k\}$ be the minimal integer q such that

$$\underline{v}(a^{(q)}) \geq V^{(h^0)}(a_{q+1}).$$

---

[2] This is the case that Sarin and Vahid (1997) investigated.

[3] We must note that $a^{(q)}$ may not be equal to a_q.

This inequality means that the maximin payoff among A_q is more than or equal to the subjective expected payoff for action a_{q+1} in the initial period. Finally, we specify action $\tilde{a}$ by

$$\tilde{a} \equiv a^{(\tilde{q})}.$$

The following proposition says that the action $\tilde{a}$ specified above is the unique action which the decision maker comes to choose in the long run, provided that the decision maker has the state-independent payoff evaluations and maximizes the subjective expected payoff in the initial period with certainty.

Proposition 9: *Suppose that Assumptions 1 through 6, 10 and 11 hold. Then,*

$$a^* = \tilde{a}.$$

Proof: Let $q^* \in \{1,\ldots,k\}$ be the integer such that $a^* = a_{q^*}$. Let $q'' \in \{1,\ldots,k\}$ be the integer such that $\tilde{a} = a_{q''}$. All we have to do is to show $q^* = q''$.

Suppose that $q^* > \tilde{q}$. Then, from inequalities (1) and the definition of $\tilde{a}$,

$$V^{(h^t)}(\tilde{a}) \geq \min[\underline{v}(\tilde{a}), V^{(h^0)}(\tilde{a})] > V^{(h^0)}(a_{q^*}).$$

Assumption 11 says that the decision maker never starts to choose action a_{q^*} in any period t as long as

$$V^{(h^{t-1})}(a_{q^*}) = V^{(h^0)}(a_{q^*}) < V^{(h^{t-1})}(\tilde{a}),$$

but this is a contradiction.

Suppose that $q'' < q^* \leq \tilde{q}$. Since $\tilde{a}$ is the maximin action among A_{q^*} also, one gets from the definitions of $\tilde{q}$ that

$$\underline{v}(\tilde{a}) < V^{(h^0)}(a_{q^*}).$$

Since $V^{(h^0)}(a_{q^*}) < V^{(h^0)}(\tilde{a})$, one gets

$$\underline{v}(a_{q^*}) < \underline{v}(\tilde{a}) = \min[\underline{v}(\tilde{a}), V^{(h^0)}(\tilde{a})],$$

which contradicts inequalities (4).

Finally, suppose that $q^* < q''$. The definition of $\tilde{q}$ says that

$$\underline{v}(a^{(q''-1)}) < V^{(h^0)}(a_{q''}) = V^{(h^0)}(\tilde{a}),$$

which, together with the definition of $a^{(q''-1)}$, says that

$$\underline{v}(a_{q^*}) \leq \underline{v}(a^{(q''-1)}) < V^{(h^0)}(\tilde{a}).$$

Moreover, since $a^{(q'')} = \tilde{a}$, one gets

$$\underline{v}(a_{q^*}) < \underline{v}(\tilde{a}).$$

These inequalities implies

$$\underline{v}(a_{q^*}) < \min\left[\underline{v}(a^{(\tilde{q})}), V^{(h^0)}(a^{(\tilde{q})})\right],$$

which contradicts inequalities (4).

Q.E.D.

7. Discussions

In this section, we present several points which have not been discussed yet in the previous sections.

7.1. Change of the Minimal Payoff Vector

We must admit that Assumption 5, which says that for every action $a \in A$ there exists the minimal payoff $\underline{v}(a)$ common to all histories, may be restrictive: In general, the minimal payoff for an action a with respect to the state of the world, $\min_{\omega \in \Omega} u^{(h^t)}(a, \omega)$, may be *history-dependent*, and therefore, the maximin action may be *history-dependent*. Unfortunately, we have not examined a generalization in this direction, and it seems to be beyond the purpose of this paper.

A relevant question is whether the decision maker in the *stationary* position will change the choice of action again, if the vector of the minimal payoffs suddenly changes. Suppose that the decision maker have already reached the stationary position described in Theorem 1, and suppose that the long run single-person decision making problem suddenly changes, and its constant minimal payoff vector is changed from $(\underline{v}(a))$ to $(\underline{v}'(a))$. Then, we can easily check from Theorem 1 that the decision maker will change the choice of action from action a^* to any other action, only if the minimal payoff for action a^* decreases, that is, only if

$$\underline{v}'(a^*) < \underline{v}(a^*).$$

Otherwise, the decision maker never stops choosing action a^*, no matter how drastically the minimal payoffs for the other actions are increased.

7.2. Comparison of Signal Structures

We have shown in this paper that the rigidity of choice of action in the long run holds true irrespective of the degree to which the random signal is informative about the true state of the world. However, the set of possible actions which the decision maker comes to choose in the long run, i.e., the set of all actions which satisfy inequalities (2) in Proposition 2, relies significantly on whether the decision maker can observe the signal or not. Suppose that the decision maker can not observe this signal. Then, inequalities (2), i.e., the necessary and sufficient condition for the actions chosen in the

long run, is replaced by the following inequalities:

$$\underline{v}(a) \geq \min\left[\underline{v}(a'), V^{(h^0)}(a')\right] \text{ for all } a' \in A. \tag{5}$$

Because

$$V^{(h^0)}(a') \leq \max_{\phi \in \Phi} v^{(a',\phi)}(h^0),$$

that is,

$$\min\left[\underline{v}(a'), V^{(h^0)}(a')\right] \leq \min\left[\underline{v}(a'), \max_{\phi \in \Phi} v^{(a',\phi)}(h^0)\right],$$

one gets that inequalities (5) are weaker than inequalities (2). Hence, the set of possible actions chosen in the long run in the case that the decision maker can not observe the signal, is bigger than, or equal to, the one in the case that she can observe it.

7.3. Further Remarks

The analysis of this paper can be applied to the game theory where the set of states is interpreted as the set of the opponent's actions and the decision maker at best imperfectly monitors the opponent's choice of action. The companion paper Matsushima (1988a) investigated the game-theoretic situation in which an individual is randomly matched with an opponent in every period and plays a symmetric component game together the payoff function of which is unknown. Matsushima (1988a) showed in a wide class of component games that the individual comes to believe incorrectly that there exists the *strictly dominant* action vector which is *uniquely* Pareto-efficient, and therefore, she comes to believe that there is no substantial strategic conflict with respect to *fairness* as well as *efficiency*.

Throughout this paper, we have assumed that at the beginning of every period, the decision maker never notices any characteristic of the current situation which distinguishes from the past situations. If the decision maker can notice such a characteristic, then she may regard the current situation as being exceptionally fit to choose any action different form action a^*, even though its payoff evaluations are much lower than the other actions'. This might put the decision maker's experimentation with multiple actions consistent with the maximization hypothesis of the subjective expected payoffs. The other companion paper Matsushima (1988b) investigated the repeated situation in which the decision maker is infinitely many times confronted with decision making problems which are contextually different each other, and learns from past experiences according to a more complex learning rule based on "*induction by analogy*".

Matsushima (1998b) characterized Markovian learning rules according to which the decision maker succeeds to choose the objective best response in the long run.

References

Akerlof, G. and W. Dickens (1982): "The Economic Consequences of Cognitive Dissonance," *American Economic Review* 72, 307-319.

Binmore, K. (1987, 1988): "Modeling Rational Players I, II," *Economics and Philosophy* 3, 4, 179-214, 9-55.

Börgers, T. and R. Sarin (1996): "Learning through Reinforcement and Replicator Dynamics," mimeo.

Brown, G. (1951): "Iterated Solution of Games by Fictitious Play," in *Activity Analysis of Production and Allocation*, edited by T. Koopmans, New York: Wiley.

Carrillo, J. and T. Mariotti (1997): "Wishful Thinking and Strategic Ignorance," mimeo.

Cotton, J. (1985): "Cognitive Dissonance in Selective Exposure," in *Selective Exposure in Communication*, edited by D. Zillman and J. Bryant, Lawrence Erlbaum Associates.

Cross, J. (1983): *A Theory of Adaptive Economic Behavior*, Cambridge University Press.

Dekel, E. and F. Gul (1997): "Rationality and Knowledge in Game Theory," in *Advances in Economics and Econometrics: Theory and Applications Vol. 1*, edited by D. Kreps and K. Wallis, Cambridge University Press.

Fudenberg, D. and J. Tirole (1997): *Theory of Learning in Games*, forthcoming, MIT Press.

Gittins , J. (1989): *Multi-Armed Bandit Allocation Indices*, New York: Wiley.

Harsanyi, J. (1967, 1968): "Games with Incomplete Information Played by Bayesian Players," *Management Science* 14, 159-182, 320-334, 486-502.

Kalai, E. and E. Lehrer (1995): "Subjective Games and Equilibria," *Games and Economic Behavior* 15, 1-26.

Luce, D. and H. Raiffa (1957): *Games and Decisions*, New York: Wiley.

Marimon, R. (1997): "Learning from Learning in Economics," *Advances in Economics and Econometrics: Theory and Applications Vol. 1*, edited by D. Kreps and K. Wallis, Cambridge University Press.

Matsushima, H. (1997a): "Bounded Rationality in Economics: A Game Theorist's View," *Japanese Economic Review* Vol. 48, No. 3.

Matsushima, H. (1997b): "Procedural Rationality and Inductive Learning I: Towards a Theory of Subjective Games," Discussion Paper 97-F-21, Faculty of Economics, University of Tokyo.

Matsushima, H. (1998a): "Towards a Theory of Subjective Games," mimeo.

Matsushima, H. (1998b): "Efficient Entrepreneurship," mimeo.

Rabin, M. (1995): "Moral Preference, Moral Constraints, and Self-Serving Biases," mimeo.

Rothschild, M. (1974): "A Two-Armed Bandit Theory of Market Pricing," *Journal of Economic Theory* 9, 185-202.

Sarin, R. and F. Vahid (1997): "Payoff Assessments without Probabilities: A Simple Dynamic Model of Choice," mimeo.

Savage, L. (1954): *The Foundation of Statistics*, New York: Wiley.

Simon, H. (1955): "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics* 69, 99-118.

Simon, H. (1976): "From Substantive to Procedural Rationality," in S. J. Latsis ed. *Methods and Appraisal in Economics*, Cambridge University Press.

	boom p	recession $1-p$
uncertain action	100	-100
safe action	0	0

Figure 1: The Stochastic Payoff Structure

	Signal "B"	Signal "R"
state "boom"	$p(1-\varepsilon)$	$p\varepsilon$
state "recession"	$(1-p)\varepsilon$	$(1-p)(1-\varepsilon)$

Figure 2: The Signal Structure

	boom (signal "B")	recession (signal "R")
uncertain action	-100	-100
safe action	0	0

Figure 3: The Objective Payoff Function