ESSAYS ON ADAPTIVE LEARNING

by

NAOKI FUNAI

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Birmingham Business School
Department of Economics
The University of Birmingham
August 2013

Abstract

This thesis consists of three interrelated chapters on adaptive learning. In each chapter, I investigate the way in which adaptive decision makers/players behave in the long run. In particular, I consider subjective assessment maximizers; each player assigns a subjective assessment to each of his actions based on its past performance and chooses the action which has the highest assessment. They update their assessments adaptively using realized payoffs. I mainly focus on the following three cases; (1) an adaptive decision maker takes into account not only direct payoff information, but also foregone payoff information; (2) adaptive players face a normal form game with strict Nash equilibrium in each of infinitely many periods; and (3) adaptive players face a finitely repeated game in each of infinitely iterated periods. Then I show the conditions under which (1) adaptive decision maker chooses the optimal action, (2) adaptive players end up choosing Nash equilibrium strategies, and (3) adaptive players' behavioural strategies converge to an agent quantal response equilibrium, which is a quantal response equilibrium for extensive form games.

ACKNOWLEDGEMENTS

I am greatly indebted to my supervisor, Prof. Rajiv Sarin. Without his help, I would not have been able to complete my thesis. I am thankful to Prof. Guoqiang Tian, Dr. Martin Jensen, Dr. Rodrigo Velez, Dr. Antonella Ianni, Prof. Indrajit Ray for many helpful comments on my chapters. I also appreciate Dr. Eve Richards and Mr. Charles Rahal for proofreading my thesis. I would like to give many thanks to my family and Ahran Song, who have supported my PhD studies in Texas and Birmingham.

CONTENTS

CHAPTER 1

INTRODUCTION

Adaptive learning has attracted many theoretical and experimental researchers. Their research helps us to understand how people learn and how they behave in the long run when they face the same situation repeatedly. Yet there exist many aspects which are important and have not been investigated thoroughly. The purpose of this thesis is to provide theoretical investigation and insights on such aspects.

In this PhD thesis, I investigate the long run behaviour of adaptive decision makers in a decision problem and games. I consider the situation where they play the same game or decision problem repeatedly but have limited information about their environment; they do not know the payoff functions or the opponent players. They assess each of their available actions based on past payoff information and pick the action which has the highest assessment.

In Chapter 2, I provide a theoretical prediction on the way in which an adaptive agent with foregone payoff information behaves in the long run. In the model, when the agent updates his assessments of actions in an adaptive manner, he uses not only the objective payoff information, but also the foregone payoff information, which may be distorted.

The distortion may arise from pessimism/optimism or envy/gloating; which depends on how the agent views the source of the information. This chapter shows the conditions in which the assessment of each action converges, where the limit assessment is the average of the expected objective and distorted payoffs. It is also shown that the agent chooses the optimal action most frequently in the long run if the expected distorted payoff of the optimal action is greater than the expected distorted payoffs of the other actions. The relations of this model to experience-weighted attraction learning, stochastic fictitious play, and quantal response equilibrium are also considered.

In Chapter 3, I provide a theoretical prediction of the way in which the adaptive players in games with strict Nash equilibrium behave in the long run. In this model, each player updates his assessment of the chosen action only in an adaptive manner. Almost sure convergence to a Nash equilibrium is shown under one of the following conditions: (i) that, at any non-Nash equilibrium action profile, there exists a player who receives a payoff which is less than his maximin payoff, (ii) that all non-Nash equilibrium action profiles give the same payoff. I show almost sure convergence to a Nash equilibrium in the following games: pure coordination games; the battle of the sexes game; the stag hunt game; and the first order statistic game. While in the game of chicken and market entry games, players may end up playing the maximin action profile.

In Chapter 4, I investigate the way in which adaptive players who play a finitely repeated game in each of infinitely iterated periods behave in the long run. In this model, each player assigns subjective assessments on his actions after any history. After receiving payoffs, they update their assessments of chosen actions using the realized payoffs in an adaptive manner; in particular, I consider the Q-learning updating rule introduced by

Watkins and Dayan (1992) and Sarin and Vahid (1999) updating rule. When players experience emotional shocks on their assessments, players' behaviour strategies converge to the agent quantal response equilibrium introduced by McKelvey and Palfrey (1998) if the stage game has perfect information. For the general case, I provide an additional condition to guarantee convergence. When players do not experience the shocks, in the long run, I show the following results; (1) when they play the finitely repeated prisoner's dilemma, both players may end up cooperating in each stage game, and (2) when they play a finitely repeated coordination game, both players coordinate in each stage game.

In Chapter 5, I summarize the work and discuss some potential extensions of this work.

CHAPTER 2

AN ADAPTIVE LEARNING MODEL WITH FOREGONE PAYOFF INFORMATION

2.1 Introduction

We often learn not only from our own experiences but also from others. For example, consider the adoption of new agricultural technologies by farmers, where they face new fertilizers. Since they are not familiar with the new technologies, they do not know the possible outcomes from each fertilizer or the way in which the outcomes are realized. It is common that farmers have neighboring farmers who face the same decision problem, which may facilitate him to find other farmers who have chosen the different fertilizers. Then the farmers can learn from other farmers about the outcomes from the other fertilizers.

This chapter studies the behaviour of an agent who has limited information about his decision-making environment; he knows the available actions, however he may not know all the possible outcomes of each action or how the outcome is realized. When we face such a complicated problem, it is natural for us to simplify the problem. In this chapter, I consider an agent who simplifies the problem in the following way; based on past payoff information, he assigns a subjective assessment to each action, where the assessment of

each action represents the payoff which he expects to receive from the action, and picks the action which he thinks gives the best payoff. Before choosing an action, he experiences temporary random shocks on his assessments; therefore, he picks an action which has the highest shock-affected assessment. The shocks may be interpreted as emotional noise on his evaluations.

When updating his assessments adaptively, the agent uses the payoff information not only from his own experience but also from others. However, the information from others need not be treated in the same way as the direct payoff information. For instance, when a farmer cannot directly observe what the other farmers have received, he may believe that his neighbors exaggerate (depreciate) the effect of other fertilizers. Then he discounts (increases) the effectiveness and takes the discounted (increased) payoff information into account. In another example, when a farmer observes others' payoffs, he may envy his neighbors' outcomes which are better than his outcome, while he gloats over their outcomes when his outcome is better than theirs. In these cases, he takes into account increased foregone payoff information when he envies and discounted foregone payoff information when he gloats. Hence it is reasonable to assume that when the agent processes the foregone payoff information, the information may be distorted, depending on the way in which he views the source of the information.

I first show the conditions in which the assessments of all actions converge. I use a stochastic approximation method and approximate a trajectory of assessments by the solution of ordinary differential equations (ODEs). Then I show the conditions under which the ODEs have a unique rest point to which the assessments converge, where the limit assessment of each action is an average of the expected objective and distorted

payoffs. Next, given the limit assessments, I provide conditions in which the agent chooses the optimal action most frequently in the long run. I show that if the expected distorted payoff of the optimal action is higher than the ones of other actions, then he chooses the optimal action most frequently in the long run. In particular, if (1) the distortion function for each action is an affine map and (2) the agent envies and/or gloats over other agents' payoffs, then the agent chooses the optimal action most frequently. However, I also show the case in which he picks a non-optimal action most frequently; it happens when the agent distorts his forgone payoff information of the non-optimal action more upwardly than the one of the optimal action.

In addition, I show the necessary and sufficient condition for convergence to the quantal response equilibrium proposed by McKelvey and Palfrey (1995). At the quantal response equilibria, given payoff perturbations, players pick the action which has the highest expected perturbed payoff, where the expectation is derived from the equilibrium strategy. Since the decision problem here can be considered as a 2-player game in which a player plays against nature, if all the subjective assessments converge to the expected objective payoffs, then the agent's choice probabilities converge to the quantal response equilibrium of this model. I show that this happens if and only if the expected distorted payoff is equal to the expected objective payoff for each action; there is no distortion on average for each action, that is, the distortion function is mean preserving.

Camerer and Ho (1999) provide a model of agents' behaviour with foregone payoff information, where the model is called the experience-weighted attraction (EWA, hereafter) learning model. The differences between the EWA learning model and this model are the way in which decision makers treat the foregone payoff information and their method of

updating their assessments. In Camerer and Ho (1999), decision makers observe or can infer the foregone payoff and they take into account the discounted payoff. In this chapter, I investigate the relationship between this model and the EWA learning model of Camerer and Ho (1999) and show that this model becomes equivalent to the EWA learning model when (1) the agent in this model discounts the foregone payoff information by a discount factor δ and (2) the discount factor for the previous experiences, ρ, and the discount factor for the previous attraction, ϕ, in the EWA learning model are equal: $\rho = \phi = 1$. In addition, if $\delta = 1$, then this model incorporates the stochastic fictitious play model of Fudenberg and Kreps (1993). For both cases, I show that the agent chooses the optimal action most frequently in the long run.

The model of the adaptive agent in this chapter is proposed by Heller and Sarin (2001), which is based on Sarin and Vahid (1999) (SV, hereafter)[1]. The differences of the model in this chapter and theirs are that the agent in SV does not have access to payoff information about actions which are not chosen by the agent, while the agent in Heller and Sarin model does not experience stochastic emotional shocks in the assessments. Therefore, this model can be considered as an extension and complement of the analysis of Heller and Sarin (2001). It is worth to note that since the decision of the agent in this model is affected by emotional shocks, he may not pick the action which has the highest assessment; it happens when the emotional noise of the chosen action is so big that the noise-affected assessment of the action is greater than the ones of the other actions, even though the assessment of the chosen action is not the highest. It is shown that there exists a case in which the noise makes the agent to choose the optimal action most frequently;

[1]The use of SV model, rather than other learning models such as Erev and Roth (1998), is supported by some empirical work (See, Sarin and Vahid, 2001; Yechiam and Busemeyer, 2008; Chen and Khoroshilov, 2003).

without the noise, as in Heller and Sarin (2001), the agent may not end up choosing the optimal action.

There also exist other related models with emotional shocks in games, but the following authors do not consider the effect of foregone payoff information. Leslie and Collins (2005) investigate games played by the same type of agents in this model, payoff-assessment maximizers, with a slightly different assessment updating rule. They show that in 2-player partnership games and 2-player zero sum games, strategies converge to Nash distribution, which is a Nash equilibrium under stochastic payoff perturbations. Cominetti, Melo and Sorin (2010) investigate the model with stochastic perturbations which have the extreme value distribution. They show that with the parameters in specific ranges, the choice probabilities of players converge to a Nash distribution in general games.

Some experimental and empirical literature shows the importance of foregone payoff information for decision makings. Conley and Udry (2010) investigate learning by farmers in Ghana about fertilizer use and show that their decisions are influenced by their neighbors' decisions. Duffy and Feltovich (1999) investigate the effect of foregone payoffs in the ultimatum game and the best-shot game and show that players show different behaviors with and without foregone payoff information. Grosskopf et al. (2006) investigate the effect in decision problems with foregone payoffs, showing that foregone payoff information affects decision maker's behaviour but the difference of his behaviour with and without foregone payoff disappears as he gains experience, except in cases where alternatives are correlated to each other.

There are some evidence showing that foregone payoffs are in fact distorted. Camerer and Ho (1999) show that in median-action games and beauty contests, foregone payoffs

are discounted from actual payoffs. Using their experimental data, Grosskopf et al. (2006) estimate parameters in generalized fictitious play with discount factor on foregone payoff and show that the foregone payoff is actually discounted. It is worth to note that discounting is not the only way that people distort the foregone payoff information. Grygolec et al. (2012) show that in a lab experiment, agents' evaluations on lotteries are affected by the outcomes of other agents; their envy and gloating affect their evaluations.

2.2 The Model

I consider an agent who faces the same decision problem repeatedly. In each period, $n \in \mathbb{N}$, he picks an action from the set $A = \{1, 2, ..., M\}$. After picking an action, the agent receives a payoff; let $(\Omega, \mathcal{F}, P)$ be the probability space on which all random variables are defined and $\pi_n^i : \Omega \to \mathbb{R}$ be the payoff function of action i in period n. I assume that the environment is stationary and thus $E[\pi_n^i] = E[\pi_m^i] =: E[\pi^i]$ for any $n, m \in \mathbb{N}$ and $i \in A$, where $E[\pi^i]$ denotes the expected payoff of action i. I assume that the payoff function for each action is bounded. In this model, he knows the action set and observes his own realized payoffs but he does not know the state space, realized state or his payoff function. In period n, he assigns a subjective assessment on each action: $Q_n = (Q_n^1, ..., Q_n^M)$ denotes the assessments of actions in period n, where Q_n^i is the assessment of action i in period n. Before choosing an action, he receives stochastic emotional shocks on the assessments; the random vector of the emotional shocks for all actions, $\eta = (\eta^1, ..., \eta^M)$, takes a value in $\mathbb{R}^M$ and the distribution of η does not depend on his payoff or his assessments. After the stochastic emotional shocks on the assessments of all actions are realized, the agent chooses the action which has the highest total value

of the assessment and the realized shock. Therefore, the probability of choosing action i given his assessments is as follows:

$$C^i(Q) = \Pr\big(\arg\max_{j\in A}(Q^j + \eta^j) = i\big),$$

where $C^i : \mathbb{R}^M \to [0,1]$ is a mapping which specifies a probability of choosing action i for each assessment Q and can be a correspondence. I assume that the distribution of the stochastic noise η has a strictly positive density on the domain, so that C^i becomes a continuous function [1]. One example of this type of choice probability is the logit choice rule, which is derived by the i.i.d. shocks with the extreme value distribution of $F(\eta_i) = \exp(-\exp(-\frac{1}{\tau}\eta_i))$; the logit choice rule has the following form:

$$C^i(Q) = \frac{\exp(\frac{1}{\tau}Q^i)}{\sum_{j\in A}\exp(\frac{1}{\tau}Q^j)},$$

where τ is sometimes called "noise term" [2]. Note that (i) if τ approaches infinity, then the choice probability becomes the uniform distribution and (ii) if τ approaches 0, then the choice probability approaches the degenerate probability; the probability of the action with the highest assessment becomes 1 and 0 for the other actions [3].

After receiving a payoff in each period, the agent updates his assessments using the payoff information; he updates the assessment of chosen action using the realized payoff, while he updates the assessments of the other actions using the foregone payoff informa-

[1] For example, consider the case where $\eta^i = 0$ with probability one for all $i \in \{1,2\}$. Then if $Q^1 = Q^2$, then the choice probability that the agent chooses i depends on his tie break rule and C^i may become discontinuous or a correspondence.

[2] See Hofbauer and Sandholm (2002)

[3] If there exist more than two actions which have the highest assessment, then those actions are chosen equally.

tion, which may be distorted. Assume that the agent has chosen an action $j \neq i$ and the payoff of action i, π^i, is realized, where he cannot observe the payoff. Let $D^i : \mathbb{R} \to \mathbb{R}$ be the distortion function of the payoff from action i. Therefore, $D^i(\pi^i)$ is the distorted payoff information of action i, which has not been chosen by the agent.

I assume that the distortion function is bounded. Another reasonable assumption on distortion function is that the function is monotonically non-decreasing[1], so that it weakly preserves the order of objective payoffs; if $x > y$ then $D^i(x) \geq D^i(y)$. If the inequality holds with equality, then the agent cannot distinguish precisely the foregone payoff from the objective payoff realization x and y. Since the distortion function cannot distort the order, the agent can still receive meaningful information. An additional assumption on action i 's distortion function is that the other actions' payoff realizations do not affect the extent to which the payoff information of the action i is distorted[2].

Using the objective payoff information and foregone payoff information, the agent updates his assessment on each action in the following manner:

$$Q^i_{n+1} = \begin{cases} (1-\lambda_{n+1})Q^i_n + \lambda_{n+1}\pi^i_n & if\ action\ i\ is\ chosen\ in\ period\ n \\ (1-\lambda_{n+1})Q^i_n + \lambda_{n+1}D^i(\pi^i_n) & otherwise \end{cases}$$

where $\{\lambda_n\}_{n\geq 1}$is a deterministic sequence of weighting parameters satisfying the following condition:

$$\sum_{n\geq 1}\lambda_n = \infty,\ \sum_{n\geq 1}(\lambda_n)^2 < \infty.$$

These assumptions on weighting parameters indicate that the effect by which new

[1]However, the result in the analysis here does not depend upon this assumption.

[2]This assumption is relaxed in the later section by introducing an envy-and-gloating distortion function.

information affects next period's assessment (i) decreases over time, which captures the power law of practice, but (ii) does not disappear in the later periods. Note that the sum of the sequence of frequentist's weighting parameters, $\lambda_n = \frac{1}{n}$ for each $n \in \mathbb{N}$, satisfy the conditions. To clarify this argument, it may be helpful to rewrite the updating rule as follows:

$$\begin{aligned} Q^i_{n+1} &= Q^i_n + \lambda_{n+1}(\mathbf{1}_{i,n}(\pi^i_n - Q^i_n) + (1 - \mathbf{1}_{i,n})(D^i(\pi^i_n) - Q^i_n)) \\ &= Q^i_n + \lambda_{n+1}(\mathbf{1}_{i,n}\pi^i_n + (1 - \mathbf{1}_{i,n})D^i(\pi^i_n) - Q^i_n) \end{aligned} \tag{2.1}$$

for each $i \in A$, where $\mathbf{1}_{i,n} = 1$ if action i is chosen in period n, and 0 otherwise.

It is worth noting the relation between this model and the experience-weighted attraction (EWA hereafter) model introduced by Camerer and Ho (1999). Under some parameters, the updating rule of the assessments in this model coincides with the updating rule of the attractions in EWA learning model. Since choice rules of this model and EWA learning model coincide under the logistic choice rule, the two models become equivalent when the assessments correspond to the attractions. To find such parameters, the general updating rule of EWA model should be described here.

Let N_n be the discounted number of past experiences and let A^i_n be the agent's attraction to action i. The updating rules of both variables are described as follows. For the variable N_n,

$$N_{n+1} = \rho N_n + 1,$$

where the parameter ρ is the discount factor of previous experience. For A_n^i,

$$A_{n+1}^i = \frac{\phi N_n A_n^i + \mathbf{1}_{i,n}\pi_n^i + (1 - \mathbf{1}_{i,n})\delta\pi_n^i}{N_{n+1}},$$

where the factor ϕ is the discount factor for previous attraction, while the factor δ is the discount factor for foregone payoff.

I consider the case where $\rho = \phi = 1$. Then the updating rule of A_n^i is expressed as follows:

$$A_n^i = A_n^i + \frac{1}{N_{n+1}}(\mathbf{1}_{i,n}\pi_n^i + (1 - \mathbf{1}_{i,n})\delta\pi_n^i - A_n^i).$$

If we have $A_n^i = Q_n^i, \lambda_n = \frac{1}{N_n}$ and $D^i(\pi^i) = \delta\pi_n^i$, then this EWA updating rule is equivalent to the updating rule in this model. Since the choice rule given attractions in EWA learning model is equivalent to the logit choice rule given assessments in this model, EWA learning model coincides with this model. In addition, if $\delta = 1$, then this model is parallel to the stochastic fictitious play model by Fudenberg and Kreps (1993).

2.3 Limit Assessments

To investigate the agent's behaviour in the long run, I show that with probability one, the assessment of each action converges to an average of expected objective and distorted payoffs, which is the sample mean of directly and indirectly observed payoffs of the action in the limit[1].

The proof of this claim is motivated by the argument in Cominetti, Melo and Sorin (2010); while notice again that players in their model do not observe foregone payoff

[1]The formal statement is given in Theorem 1, which is located just before the next section.

information and therefore the dynamics of players' assessments are different.

To prove this claim, I use a stochastic approximation method and show that the ordinary difference equations (2.1) can be approximated in the long run by the solution of the following ordinary differential equations (ODEs, hereafter);

$$\dot{Q}_t^i = C^i(Q_t)E[\pi^i] + (1 - C^i(Q_t))E[D^i(\pi^i)] - Q_t^i,\ i \in A, \tag{2.2}$$

where $\dot{Q}_t^i := \frac{\mathrm{d}}{\mathrm{d}t}Q_t^i$.

Let $\dot{Q}_t^i = F_i(Q_t)$ and $F = (F_1, ..., F_M)$. In Lemma 1, I provide the condition by which (i) F is Lipschitz continuous and (ii) there exists a unique rest point for ODEs (2.2).

Lemma 1. *(i)$F = (F_1, ..., F_M)$ is Lipschitz continuous and (ii) there exists a unique rest point of ODEs (2.2) if the following condition holds: for any Q and Q'*

$$\mid C^i(Q) - C^i(Q^{'}) \mid\mid E[D^i(\pi^i)] - E[\pi^i] \mid \leq \delta_i \|Q - Q^{'}\|_\infty, \tag{2.3}$$

for all i and some $\delta_i \in [0, 1)$, where $\|\cdot\|_\infty$ is the infinity norm.

Proof. (i) Let i be the action such that the following condition satisfies:

$$\|F(Q) - F(Q^{'})\|_\infty = \left|(Q^i - Q^{i\prime}) + (C^i(Q) - C^i(Q'))(E[D^i(\pi^i)] - E[\pi^i])\right|.$$

Then we have

$$\begin{aligned}\|F(Q) - F(Q^{'})\|_{\infty} &= \left|(Q^{i} - Q^{i\prime}) + (C^{i}(Q) - C^{i}(Q'))(E[D^{i}(\pi^{i})] - E[\pi^{i}])\right| \\ &\leq \left|Q^{i} - Q^{i\prime}\right| + \delta_{i}\|Q - Q^{'}\|_{\infty} \\ &\leq (1 + \max\delta_{i})\|Q - Q^{'}\|_{\infty},\end{aligned}$$

where $\delta_{max} := \max_{j}\delta_{j} \in [0, 1)$. Therefore, F is Lipschitz continuous.

(ii) Consider the function $f = (f^{1}, f^{2}, ..., f^{M})$ such that

$$f^{i}(Q) = E[\pi^{i}] + (1 - C^{i}(Q))(E[D^{i}(\pi^{i})] - E[\pi^{i}])$$

for each $i \in A$. Notice that $f(Q) - Q = F(Q) = \dot{Q}$. Let i be an action such that $\|f(Q) - f(Q^{'})\|_{\infty} =\mid f^{i}(Q) - f^{i}(Q^{'}) \mid$. Then

$$\begin{aligned}\|f(Q) - f(Q^{'})\|_{\infty} &= \mid f^{i}(Q) - f^{i}(Q^{'}) \mid \\ &= \mid (-C^{i}(Q) + C^{i}(Q^{'}))(E[D^{i}(\pi^{i})] - E[\pi^{i}]) \mid \\ &= \mid C^{i}(Q) - C^{i}(Q^{'}) \mid\mid E[D^{i}(\pi^{i})] - E[\pi^{i}] \mid\end{aligned}$$

Now by the hypothesis, there exists $\delta_{i} \in [0, 1)$ such that

$$\mid C^{i}(Q) - C^{i}(Q^{'}) \mid\mid E[D^{i}(\pi^{i})] - E[\pi^{i}] \mid\leq \delta_{i}\|Q - Q^{'}\|_{\infty}$$

for all i. Then we have

$$\begin{aligned}\|f(Q) - f(Q^{'})\|_{\infty} &= \mid C^{i}(Q) - C^{i}(Q^{'}) \mid\mid E[D^{i}(\pi^{i})] - E[\pi^{i}] \mid \\ &\leq \delta_{i}\|Q - Q^{'}\|_{\infty} \\ &\leq \max_{j} \delta_{j}\|Q - Q^{'}\|_{\infty}.\end{aligned}$$

Since $\delta_{max} := \max_{j} \delta_{j} \in [0, 1)$, f is a contraction mapping and by the contraction mapping theorem, $Q = f(Q)$ has a unique solution. Hence, if condition (2.3) is satisfied for all actions, then ODEs $\dot{Q}$ have a unique rest point. □

In the following arguments in this section, I assume that condition (2.3) holds. If $E[D^{i}(\pi^{i})] \neq E[\pi^{i}]$ for all i, then condition (2.3) tells us that the choice probability function $C = (C^{1}, ..., C^{M}) : \mathbb{R}^{M} \to [0, 1]^{M}$ is also Lipschitz continuous. It is helpful for understanding condition (2.3) to consider the case where emotional shocks have the extreme value distribution, so that the choice probability becomes the logistic choice rule:

$$C^{i}(Q) = \frac{\exp(\frac{1}{\tau}Q^{i})}{\sum_{j} \exp(\frac{1}{\tau}Q^{j})}.$$

Proposition 1. *In the logistic choice rule case, condition (2.3) holds if the following condition holds;*

$$\mid E[D^{i}(\pi^{i})] - E[\pi^{i}] \mid \cdot M < \tau \tag{2.4}$$

for each i.

Proof. Consider the difference of the choice probabilities of action i for two different

assessment vectors Q and Q'. We have

$$\mid C^i(Q) - C^i(Q^{'}) \mid = \mid \frac{\exp(\frac{1}{\tau}Q^i)}{\sum_j \exp(\frac{1}{\tau}Q^j))} - \frac{\exp(\frac{1}{\tau}Q^{i\prime})}{\sum_j \exp(\frac{1}{\tau}Q^{j\prime})} \mid .$$

By the mean value theorem, there exists Q^* such that

$$\begin{aligned}
\mid C^i(Q) - C^i(Q^{'}) \mid &= \mid \sum_j \frac{\partial}{\partial Q_j} C^i(Q^*)(Q^j - Q^{j\prime}) \mid \\
&\leq \sum_j \mid \frac{\partial}{\partial Q_j} C^i(Q^*) \mid\mid (Q^j - Q^{j\prime}) \mid \\
&\leq \sum_j \frac{1}{\tau} \mid (Q^j - Q^{j\prime}) \mid \\
&\leq \frac{M}{\tau} \| Q - Q^{'} \|_\infty .
\end{aligned}$$

From the second line to the third line, I use the fact that $\mid \frac{\partial}{\partial Q_j} C^i(Q) \mid \leq \frac{1}{\tau}$ for any Q and j. Hence if $\mid E[D^i(\pi^i)] - E[\pi^i] \mid < \frac{\tau}{M}$, then

$$\begin{aligned}
\mid & \quad C^i(Q) - C^i(Q^{'}) \mid\mid E[D^i(\pi^i)] - E[\pi^i] \mid \\
\leq & \quad \mid E[D^i(\pi^i)] - E[\pi^i] \mid \frac{M}{\tau} \| Q - Q^{'} \|_\infty \\
= & \quad \delta_i \| Q - Q^{'} \|_\infty ,
\end{aligned}$$

where $\delta_i = \mid E[D^i(\pi^i)] - E[\pi^i] \mid \frac{M}{\tau} \in [0, 1)$. □

Condition (2.4) says that the noise term τ should be great enough to cover the product of (i) the difference between the expected objective and distorted payoffs and (ii) the size of the action set. Therefore, if one of them becomes greater, then τ should also be greater, that is, his choice probabilities approach the uniform distribution.

Next, I show that a "continuous interpolated trajectory" of the assessments almost surely approaches the solution of ODEs (2.2). In the following argument, I mostly follow the notation and methods in Borkar (2008). Let $t_0 = 0$, $t_n = \sum_{i=1}^{n} \lambda_i$ and $I_n := [t_n, t_{n+1}], n \geq 0$. Then a continuous interpolated trajectory $\overline{Q}_t$ is expressed as follows:

$$\overline{Q}_t = Q_n + (Q_{n+1} - Q_n)\frac{t - t_n}{t_{n+1} - t_n}, t \in I_n.$$

Let $Q_t^{\geq s}, t \geq s$, denote the unique solution of (2.2) starting at s;

$$\dot{Q}_t^{\geq s} = F(Q_t^{\geq s}), t \geq s,$$

with $Q_s^{\geq s} = \overline{Q}_s, s \in \mathbb{R}$. Similarly, let $Q_t^{\leq s}, t \leq s$ denote the unique solution to (2.2) ending at s;

$$\dot{Q}_t^{\leq s} = F(Q_t^{\leq s}), t \leq s,$$

with $Q_s^{\leq s} = \overline{Q}_s, s \in \mathbb{R}$. Then, we have the following result:

Lemma 2. *For any $T > 0$,*

$$\begin{aligned} \lim_{s\to\infty} \sup_{t\in[s,s+T]} \|\overline{Q}_t - Q_t^{\geq s}\|_\infty &= 0 \ a.s., \ \textit{and} \\ \lim_{s\to\infty} \sup_{t\in[s-T,s]} \|\overline{Q}_t - Q_t^{\geq s}\|_\infty &= 0 \ a.s.. \end{aligned}$$

Proof. First I rewrite the updating rule above as follows;

$$Q_{n+1}^i - Q_n^i = \lambda_{n+1}\big(F_i(Q_n) + U_{n+1,i}\big), \text{for each } i \in A,$$

where

$$F_i(Q_n) = E[\pi^i] - Q_n^i + (1 - C^i(Q_n))(E[D^i(\pi^i)] - E[\pi^i]),$$

and

$$U_{n+1,i} = \Big(\pi_n^i - Q_n^i + (1 - \mathbf{1}_{i,n})(D^i(\pi^i) - \pi_n^i)\Big) - F_i(Q_n).$$

Now $F = (F_1, ..., F_M)$ is a Lipschitz continuous map from $\mathbb{R}^M$ to $\mathbb{R}^M$, and $U_n = (U_{n,1}, ..., U_{n,M})$ are random perturbations such that

$$E[U_{n+1} \mid Q_n] = 0$$

and

$$\sup_n E[\| U_{n+1} \|_\infty^2 | Q_n] \leq K,$$

where $K < \infty$ is a constant. Note that $\sup_n \| Q_n \|_\infty < \infty$, since the initial assessment for each action takes a finite value and the payoff function and distorted payoff function are also bounded. The second condition is also true since the choice probability, payoff function and distorted payoff function are bounded. Therefore, by Lemma 1 in Borkar (2008), the solution of ordinary difference equations (2.1) for the assessments is approximated in the long run by the solution of ODEs (2.2) almost surely. □

Note that the rest point of ODEs (2.2), $Q^* = (Q^{1*}, ..., Q^{M*})$, is the vector of the assessments where each action's assessment is an average of the expected objective and distorted payoffs of the action;

$$Q^{i*} = C^i(Q^*)E[\pi^i] + (1 - C^i(Q^*))E[D^i(\pi^i)],\ i \in A,$$

where the average is taken by the limit choice probability of the action.

In Lemma 3, I show the global convergence of assessments to the unique rest point of ODEs (2.2). To show convergence, I use the Lyapunov function method:

Lemma 3. *Given any initial assessment Q_1, the assessments of actions converge to the unique rest point of ODEs (2.2) almost surely: for any Q_1, $Q_n \overset{a.s.}{\to} Q^*$*

Proof. Consider the function

$$V_0(Q) = \| Q - Q^* \|_\infty,$$

where Q^* is the unique rest point of (2.2). This function is 0 when $Q^i = Q^{i*}$ for all i and strictly positive otherwise. Note that the derivative of V_0 coincides with the derivative of $\mid Q^i - Q^{i*} \mid$ of the action i which takes the highest value among all actions. First, I assume that $Q^i > Q^{i*}$. Then

$$\begin{aligned}
\frac{d}{dt}V_0 &= \frac{d}{dt}(Q^i - Q^{i*}) \\
&= (f(Q^i) - f(Q^{i*}) + Q^{i*} - Q^i) \\
&\leq \delta_{max} \| Q - Q^* \|_\infty - (Q^i - Q^{i*}) \\
&= (\delta_{max} - 1) \| Q - Q^* \|_\infty \\
&< 0.
\end{aligned}$$

Next, I assume that $Q^i \leq Q^{i*}$. Then

$$\begin{aligned}
\frac{d}{dt}V_0 &= \frac{d}{dt}(Q^{i*} - Q^i) \\
&= -(f(Q^i) - f(Q^{i*}) + Q^{i*} - Q^i) \\
&\leq \delta_{max} \parallel Q - Q^* \parallel_\infty -(Q^{i*} - Q^i) \\
&= (\delta_{max} - 1) \parallel Q - Q^* \parallel_\infty \\
&< 0.
\end{aligned}$$

Thus we have $\frac{d}{dt}V_0 < 0$ for all $Q \neq Q^*$. Hence V_0 is a Lyapunov function for ODEs (2.2) and, by Proposition 6.4 and Corollary 6.6 of Benaïm (1999) or Theorem 2 and Corollary 3 of Borkar (2008), assessments converge to the unique Q^* almost surely. □

Since there exists a density function for each shock and choice function, $\mathbf{1}_{i,n}$, is bounded, the choice probability function for each action is continuous with respect to assessments. Hence the convergence of assessments implies the convergence of choice probabilities. In addition, by the strong law of large numbers for dependent variables[1], the empirical frequency of each action, $\frac{1}{n}\sum_{m=1}^{n}\mathbf{1}_{i,m}$, converges to the choice probability given the limit assessments, $C^i(Q^*)$, almost surely.

Lemma 4. *With probability 1, the empirical frequency of each action converges to the choice probability of the action in the limit; for each i,*

$$\frac{1}{n}\sum_{m=1}^{n}\boldsymbol{1}_{i,m} \overset{a.s.}{\to} C^i(Q^*).$$

Proof. Let $\{\mathbf{1}_{i,n}, n \geq 1\}$ be a sequence of choice functions such that $\mathbf{1}_{i,n} = 1$ if i is chosen

[1]For example, see p36 in Hall and Heyde (1980)

at period n and 0 otherwise. Let $\{\mathcal{F}_n, n \geq 1\}$ be a sequence of $\sigma\text{-}fields$, where $\mathcal{F}_n$ is generated by random variables for all actions up to period n. Thus $\mathbf{1}_{i,n}$ is measurable with respect to $\mathcal{F}_n$ and $E[\mathbf{1}_{i,n} \mid \mathcal{F}_{n-1}] = C^i(Q_n)$. Now for each x and $n \geq 1, E[|1|] = 1 < \infty$ and $P(|\mathbf{1}_{i,n}| > x) \leq P(|1| > x)$. Notice also that $E[|1| \log^+ |1|] = 0 < \infty$. Thus by Theorem 2.19 in Hall and Heyde (1980), we have

$$\frac{1}{n} \sum_{m=1}^{n} [\mathbf{1}_{i,m} - C^i(Q_m)] \overset{a.s.}{\to} 0.$$

Since $C^i(Q_n)$ converges to $C^i(Q^*)$ almost surely, the empirical frequency of action i converges to $C^i(Q^*)$ almost surely. Since I pick i randomly, this argument is true for any $i \in A$. □

Now we are ready to state one of the main results;

Theorem 1. *Given condition (2.3), with probability one, the assessment of each action converges to the average of the expected objective and distorted payoffs where the average is taken by the limit frequency of the action; for any* Q_1,

$$Q_n \overset{a.s.}{\to} Q^*$$

and

$$Q^{i*} = \alpha^{i*} E[\pi^i] + (1 - \alpha^{i*}) E[D^i(\pi^i)], \quad \forall i \in A,$$

where $\alpha^{i*} := \lim_{n\to\infty} \frac{1}{n} \sum_{m=1}^{n} \boldsymbol{1}_{i,m}$.

In the following sections, using the results in this section, I show the conditions under which the decision maker chooses the optimal action most frequently in the long run.

2.4 Empirical Frequencies in the Long Run

In this section, I focus on the properties of the limit empirical frequencies of actions. I show that if the expected distorted payoff of the optimal action, which has the highest expected objective payoff, is greater than the expected distorted payoffs of the other actions, then the agent chooses the optimal action most frequently in the long run. I also show a case where the agent chooses a suboptimal action most frequently in the long run; it happens when, on average, the foregone payoffs of the suboptimal action are distorted more upward than the distorted foregone payoffs of the optimal action.

By the results in the last section, we know that $\alpha^{i*} \geq \alpha^{j*}$ if $C^i(Q^*) \geq C^j(Q^*)$. In addition, we assume that stochastic shocks are i.i.d., such as the case for the logit choice rule, and then $C^i(Q^*) \geq C^j(Q^*)$ if $Q^{i*} \geq Q^{j*}$. Therefore, in the following statements, I analyze conditions under which $Q^{i*} \geq Q^{j*}$ holds; given the conditions, we have $\alpha^{i*} \geq \alpha^{j*}$ as well.

First, I assume that one action gives better payoff information than another action on average: the expected values of directly and indirectly observed payoffs from one action are greater than the ones of another action. As shown, each assessment in the long run can be expressed as the average value of expected objective and distorted payoffs. Hence, if those two values of one action are bound to be higher than those values of another action, then the former action's limit assessment is higher than another action's limit assessment, which means that the former action are chosen more frequently than the latter action in the long run:

Lemma 5. *If* $\min\{E[D^i(\pi^i)], E[\pi^i]\} \geq \max\{E[D^j(\pi^j)], E[\pi^j]\}$ *holds, then* $\alpha^{i*} \geq \alpha^{j*}$.

Proof. See Appendix. □

The following corollaries are immediate consequences of Lemma 5.

Corollary 1. *If* $\min\{E[D^i(\pi^i)], E[\pi^i]\} \geq \max\{E[D^j(\pi^j)], E[\pi^j]\}$ *holds for all* $j \in A$*, then action* i *is chosen most frequently in the long run almost surely.*

Corollary 2. *If* $E[\pi^i] = E[D^i(\pi^i)]$ *for all* $i \in A$*, then* $\alpha^{i*} \geq \alpha^{j*}$ *if* $E[\pi^i] \geq E[\pi^j]$.

Corollary 2 says that if distortion functions for all actions are mean-preserving, then the action which has higher expected objective payoff will be chosen more frequently in the long run.

Second, I consider the case in which foregone payoffs of some actions are distorted upward and their expected distorted payoffs are greater than the expected objective payoffs. In this case, the action which has a greater expected objective payoff is chosen more frequently in the long run.

Lemma 6. *Suppose that* $E[D^i(\pi^i)] \geq E[D^j(\pi^j)] \geq E[\pi^i] \geq E[\pi^j]$*. Then* $\alpha^{i*} \geq \alpha^{j*}$.

Proof. See Appendix. □

Third, I consider the case where foregone payoff information is distorted downward for some actions. It is also shown that the action with higher expected objective payoff is chosen more frequently in the long run.

Lemma 7. *If* $E[\pi^i] \geq E[\pi^j] \geq E[D^i(\pi^i)] \geq E[D^j(\pi^j)]$*, then* $\alpha^{i*} \geq \alpha^{j*}$.

Proof. See Appendix. □

From the preceding results, it can be shown that the optimal action is chosen most frequently if the expected distorted payoff of the action is greater than the ones of the other actions;

Theorem 2. *Suppose that* $E[\pi^i] \geq E[\pi^j]$ *for some* i *and* j. *If* $E[D^i(\pi^i)] \geq E[D^j(\pi^j)]$ *then* $\alpha^{i*} \geq \alpha^{j*}$. *Therefore, the optimal action is chosen most frequently in the long run,* $\alpha^{i*} \geq \max_j \alpha^{j*}$, *with probability one if* $E[D^i(\pi^i)] \geq \max_j E[D^j(\pi^j)]$.

It is worth to note that it is not the case if the agent does not experience the emotional shocks on his assessments as in Heller and Sarin (2001). As an example, consider the case where there are only two actions, i and j. Also, for simplicity, I assume that there is no uncertainty for payoffs; the agent receives a constant payoff π^h from action $h \in \{i, j\}$. In particular, I assume that $\pi^i > \pi^j$. Then if he chooses the suboptimal action, j, then he receives the payoff π^j while he observes the foregone payoff information of $D^i(\pi^i)$. If the initial assessments of action i and j are such that $\pi^i > \pi^j > Q^j > Q^i > D^i(\pi^i) > D^j(\pi^j)$ then he always chooses action j.

Last, I consider the case where one action with lower expected objective payoff is chosen more often than another action with higher expected objective payoff. It happens when the foregone payoff of the worse action is distorted more upwardly, so that the expected objective and distorted payoffs of the better action is lower than the expected distorted payoff of the worse action. I show that if the arithmetic average of the expected distorted and objective payoffs of the worse action is greater than the maximum value of the expected objective and distorted payoffs of the better action, then in the long run the worse action is chosen more frequently with probability one.

Proposition 2. *If*

$$E[D^i(\pi^i)] > \max\{E[D^j(\pi^j)], E[\pi^j]\} > \min\{E[D^j(\pi^j)], E[\pi^j]\} > E[\pi^i]$$

and

$$\frac{E[D^i(\pi^i)] + E[\pi^i]}{2} \geq \max\{E[D^j(\pi^j)], E[\pi^j]\},$$

then $\alpha^{i*} \geq \alpha^{j*}$.

Proof. See Appendix. □

2.4.1 Affine Transformation

In this subsection, I consider the following distortion function; for each $i \in A$

$$D^i(\pi^i) = \beta\pi^i + \gamma,$$

where $\beta \geq 0$ and $\gamma \in \mathbb{R}$. This distortion function is an affine map and includes the case where the agent distorts the foregone payoff information by $\beta \geq 0$ and $\gamma = 0$, as in the EWA learning model[1]. Note that if $E[\pi^i] \geq E[\pi^j]$ then $E[D^i(\pi^i)] \geq E[D^j(\pi^j)]$. Therefore, if the distortion function for each action is an affine map, then the optimal action is chosen most frequently in the long run;

Corollary 3. *If* $D^i(\pi^i) = \beta\pi^i + \gamma$ *for each* $i \in A$, *then* $\alpha^{i*} \geq \alpha^{j*}$ *if* $E[\pi^i] \geq E[\pi^j]$.

Note that for the case where $\gamma = 0$, as in the EWA learning model, the action with the highest expected objective payoff is chosen most frequently.

2.4.2 Envy and Gloating

Next, I consider a distortion function which captures the concept of the decision maker's envy and gloating. In this subsection, I assume that the agent directly observes not only

[1]They use δ for the discount factor.

the payoff from chosen action, but also the payoffs from the other actions which are chosen by others. Then the decision maker envies other people's payoffs when the decision maker observes others' payoffs which are better than his, while he gloats over others' payoffs which are worse than his. One way to express this idea is that he distorts the foregone payoffs upward when he envies, while he distorts the foregone payoffs downward when he gloats. It is worthwhile noting that increasing the foregone payoff means that the action is more likely to be chosen, while discounting the foregone payoff means that the action is less likely to be chosen in the next period.

I first consider a situation in which the agent receives a payoff of an action, π^i, while he observes a foregone payoff of another action, π^j. Let $D^j(\pi^j, \pi^i)$ be the distortion function of action j when action i is chosen. Then I consider the following distortion function[1]:

$$D^j(\pi^j, \pi^i) = \pi^j + G_i^j(\pi^j - \pi^i),$$

where the function G_i^j, which expresses envy and gloating, is increasing and bounded and satisfies the following condition: $G_i^j(\pi^j - \pi^i) \geq 0$ if $\pi^j - \pi^i > 0$, $G_i^j(\pi^j - \pi^i) = 0$ if $\pi^j - \pi^i = 0$, and $G_i^j(\pi^j - \pi^i) \leq 0$ if $\pi^j - \pi^i < 0$. For example, a linear function $D^j(\pi^j, \pi^i) = \beta(\pi^j - \pi^i)$ with a slope $\beta > 0$ satisfies the conditions[2].

Then the distortion function $D^j(\pi^j)$ has the following form:

$$\begin{aligned} D^j(\pi^j) &= \sum_{i \in A \setminus \{j\}} \mathbf{1}_{i,n} D^j(\pi^i, \pi^j) \\ &= \sum_{i \in A \setminus \{j\}} \mathbf{1}_{i,n} (\pi^j + G_i^j(\pi^j - \pi^i)). \end{aligned}$$

[1]See equation (1) of Gygolec, Coricelli and Rustichini (2012)
[2]Boundedness is also satisfied, since I assume that payoffs are bounded

Now consider the updating rule of the assessment of an action, given the envy-and-gloat distortion function. For simplicity, and as assumed in Grygolec et al. (2012), I assume that there are only two actions available[1]. Then the updating rule of the assessment of action i is as follows:

$$\begin{aligned} Q^i_{n+1} &= Q^i_n + \lambda_{n+1}(\mathbf{1}_{i,n}\pi^i_n + (1-\mathbf{1}_{i,n})D^i(\pi^i) - Q^i_n) \\ &= Q^i_n + \lambda_{n+1}(\mathbf{1}_{i,n}\pi^i_n + (1-\mathbf{1}_{i,n})(\pi^i + G^i_j(\pi^i - \pi^j)) - Q^i_n) \end{aligned}$$

Notice that given the information in period n, payoff functions in period n and choice functions are independent. Thus, this extension of the distortion function does not change the dynamics of assessments.

Therefore I investigate the long run behaviour of this agent by comparing the limit assessments. From Theorem 2, we know that assuming $E[\pi^i] \geq E[\pi^j]$, action i is chosen more frequently if $E[D^i(\pi^i)] \geq E[D^j(\pi^j)]$. Given the envy-and-gloating distortion function, the expected distorted payoff of action i is as follows;

$$E[D^i(\pi^i)] = E[\pi^i] + E[G^i_j(\pi^i - \pi^j)].$$

In particular, as assumed in Grygolec et al. (2012), I consider the envy-and-gloating distortion function with

$$G^i_j(\pi^i - \pi^j) = \beta(\pi^i - \pi^j)$$

with a slope $\beta > 0$ for any i and j. Then the expected distorted payoff is expressed as

[1]If there are more than two actions, then we will have different dynamics for assessments. The case is left to be pursued in future work.

follows;

$$E[D^j(\pi^j)] = E[\pi^j] + \beta(E[\pi^j] - E[\pi^i]).$$

It is easy to see that assuming $E[\pi^i] \geq E[\pi^j]$, the inequality

$$E[\pi^i] + \beta(E[\pi^i] - E[\pi^j]) \geq E[\pi^j] + \beta(E[\pi^j] - E[\pi^i])$$

holds when $\beta \geq 0$. Hence the agent with the envy-and-gloating distortion function chooses the optimal action most frequently in the long run.

Next, I consider a class of envy-and-gloating distortion functions, which includes the envy-and-gloating distortion function above. I first assume that the function G^i_j does not depend on which action is chosen;

$$G^i_j(x) = G(x)$$

for actions $i, j \in A$. This means that the degree of decision maker's envy and gloating does not depend on which action is chosen, but the distance of his and other's payoffs. I next assume that the distortion function G is an odd function;

$$G(-x) = -G(x).$$

This means that given two payoffs, the degree of envy is equivalent to the degree of gloating, but envy and gloating have the opposite effect. Note that the envy-and-gloating function above satisfies both conditions. Then the agent chooses the optimal action most frequently in the long run if the function G is convex;

Proposition 3. *If there exist two actions and the function G is convex, then the agent chooses the optimal action most frequently in the long run:*

$$\alpha^{i*} \geq \alpha^{j*} \quad if \quad E[\pi^i] \geq E[\pi^j].$$

Proof. See Appendix. □

2.5 Quantal Response Equilibrium

I investigate the relationship of this model with the quantal response equilibrium model[1] introduced by McKelvey and Palfrey (1995). For the purpose, the quantal response equilibrium model is introduced here. Let A_ι be a finite set of actions for player $\iota \in \{1, 2\}$: $A_\iota = \{1, 2, ..., M_\iota\}$. Let $\pi_\iota : A_1 \times A_2 \to \mathbb{R}$ be a payoff function for player ι. Let $\Delta_\iota = \{p_\iota = (p_{\iota 1}, ..., p_{\iota M_\iota}) : \sum_j p_{\iota j} = 1, p_{\iota j} \geq 0\}$ be the set of probability measures of player ι, where $p_{\iota j}$ is the probability of player ι playing action j. The domain of the payoff function can be extended to the set of probability measures $\Delta = \Delta_1 \times \Delta_2$ and let $p = (p_1, p_2)$ be an element of Δ. When choosing action j, player ι receives a stochastic payoff, $\eta_{\iota j}$. The random vector $\eta_\iota = (\eta_{\iota 1}, ..., \eta_{\iota M_\iota})$ takes a value in $\mathbb{R}^{M_\iota}$. The assumptions of stochastic payoffs here are equivalent to the assumption of the stochastic shocks in the model of this chapter except that $E[\eta^j] = 0$ *for all* $j \in A_\iota$ and $\iota \in \{1, 2\}$. Given the opponent's choice probability $p_{-\iota}$, the player ι who uses j receives the following total payoff:

$$\overline{\pi}_\iota(j, p_{-\iota}) = \pi_\iota(j, p_{-\iota}) + \eta_{\iota j}.$$

[1]The decision problem can be considered a game in which the decision maker plays against nature, which chooses its action, which corresponds to the states, with a fixed probability. Therefore, it is possible to compare this model with the quantal response equilibrium model, which is introduced for games.

Given the total payoffs for actions, player ι picks action j such that $\overline{\pi_\iota}(j, p_{-\iota}) \geq \overline{\pi_\iota}(k, p_{-\iota})$ for all $k \in A_\iota$. Therefore, given the opponent's choice probability $p_{-\iota}$, the probability of choosing j is as follows:

$$C'^j(\pi_\iota, p_{-\iota}) = P(\arg\max_{k \in A}(\pi_\iota(k, p_{-\iota}) + \eta_{\iota k}) = j)$$

Then quantal response equilibrium is a choice probability profile $p^* = (p_1^*, p_2^*)$ such that for any ι and j,

$$p_{\iota j}^* = C'^j(\pi_\iota, p_{-\iota}^*).$$

If I set player 2 as nature and his actions as states, then this model can be considered as a decision problem for player 1. In this case, the nature picks a state ω randomly from $A_2 = \Omega$, where the distribution corresponds to p_2, and $\pi_1(j, p_2)$ corresponds to the expected payoff of action j, $E[\pi^j]$. Therefore, the choice probability of action j is expressed as follows;

$$C'^j(\pi_\iota, p_{-\iota}) = C^j(E[\pi]) = P(\arg\max_{k \in A}(E[\pi^k] + \eta_k) = j).$$

Since the choice probability for each action in this model is continuous with respect to assessments, if Q_n^j converges to $E[\pi^j]$ for each j, then the choice probability of each action converges to a quantal response equilibrium. Since the choice probability is always positive, by ODEs (2.2), the necessary and sufficient condition of the convergence point being the quantal response equilibrium is $E[D^j(\pi^j)] = E[\pi^j]$ for all j, which means that each distortion function should preserve the mean of each payoff function. Therefore, this proves the following result:

Proposition 4. *If the choice probabilities and the empirical frequencies of actions con-*

verge, then the convergence point is a quantal response equilibrium if and only if $E[D^j(\pi^j)] = E[\pi^j]$ *and* $E[\eta^j] = 0$ *for all* $j \in A$.

2.6 Discussion and Conclusion

This chapter has investigated theoretically the behaviour of a myopic and adaptive agent who simplifies his decision problem; to each action, he assigns a subjective assessment which is a weighted average of past realized payoffs, experiences an emotional shock on the assessment and picks the action which has the highest noise-affected assessment. Payoff information which is used to update his assessments is not only directly observed payoff information, but also foregone payoff information, which may be distorted. It is shown by using a stochastic approximation method that the limit assessment of each action is an average of expected objective and distorted payoffs. Given the limit assessments, the tendencies of the decision maker's behaviour in the limit are studied; the agent chooses the optimal action most frequently if the expected distorted payoff of the action is greater than the ones of the other actions. It is also shown that seeing the decision problem as a game against nature, the choice probabilities of the agent almost surely converge to a quantal response equilibrium if and only if he distorts his forgone payoff in a mean-preserving way.

2.6.1 Population Interpretation

This model can also be interpreted as a population model. Consider the situation in which there exists a continuum population of agents, who have limited information about the decision problem they face; they are to choose an action, but do not know the state space, realized state, and payoff function. In each period, an agent is picked from the

population randomly, chooses an action from the set A and receives a payoff. Once he has been picked, he never makes a decision again. In each period, the population forms public assessments on actions, $Q^i = (Q^1, ..., Q^M)$, which are based on the payoff information informed by agents. The public assessments are observed by all agents. In addition, each agent has his own assessments on those actions. The individual assessments are independently drawn by the common distribution which is equivalent to the distribution of emotional stochastic shocks in the model. Given the agent's assessment and public assessments, he chooses an action which has the highest total value of public assessment and his own assessment. Therefore, the probability that action i is chosen by the agent in period n is as follows;

$$C^i(Q_n) = \Pr\big(\arg\max_{j \in A}(Q^j + \eta^j) = i\big),$$

where $\eta = (\eta^1, ..., \eta^M)$ is the sequence of the individual assessments of the agent in a period. After making a decision, each agent informs the population of his received payoff. Moreover, the population can acquire the foregone payoff information from the other information sources. The population may incorporate the foregone payoff from others in the distorted way, as defined in the model. It is also possible that the foregone payoff information itself is distorted. Then the population updates the public assessments in the adaptive way which I have defined in this model;

$$Q^i_{n+1} = \begin{cases} (1 - \lambda_{n+1})Q^i_n + \lambda^i_{n+1}\pi^i_n & if\ action\ i\ is\ chosen \\ (1 - \lambda_{n+1})Q^i_n + \lambda^i_{n+1}D^i(\pi^i_n) & otherwise, \end{cases}$$

where π_n^i is the payoff which is realized in period n and $D^i(\pi_n^i)$ is the distorted payoff information of action i in period n. Therefore, the public assessments consist of the payoff information from its own agents and from others. For example, the public assessment of each action is the sample average of realized payoffs if $\lambda_n = \frac{1}{n}, \forall n$ and $D^i(x) = x, \forall i$.

To understand the argument further, consider a specific situation where there exist similar goods from different brands and consumers need to decide from which brand they will purchase. Those consumers may visit a web page which collects payoff information or reviews from its viewers. The page may also collects reviews from other web pages, however, the editor of the page may believe that the reviews from the other web pages may be distorted and he distorts the information from other pages. Then given the payoff information from its own consumers and from others, the page shows the public assessment which is an (weighted or sample) average of undistorted and distorted payoff information. Each consumer also has his own assessment and he therefore picks the brand which has the best value of public assessment and his own assessment. This chapter shows that the public assessment of each brand converges and consumers will end up picking the right brand most frequently when the distortion is non-discriminatory among brands.

2.7 Appendix

2.7.1 Proof of Lemma 5

For each action, i, the following equation holds;

$$Q^{i*} = C^i(Q^*)E[\pi^i] + (1 - C^i(Q^*))E[D^i(\pi^i)].$$

Hence if $\min\{E[D^i(\pi^i)], E[\pi^i]\} \geq \max\{E[D^j(\pi^j)], E[\pi^j]\}$ holds, then

$$\begin{aligned} Q^{i*} &= C^i(Q^*)E[\pi^i] + (1 - C^i(Q^*))E[D^i(\pi^i)] \\ &\geq C^j(Q^*)E[\pi^j] + (1 - C^j(Q^*))E[D^j(\pi^j)] \\ &= Q^{j*}. \end{aligned}$$

□

2.7.2 Proof of Lemma 6

Here, I prove by contradiction. First, I consider the case where $E[D^i(\pi^i)] > E[D^j(\pi^j)] \geq E[\pi^i] \geq E[\pi^j]$ holds. Assume that $Q^{i*} - Q^{j*} < 0$. Since $E[\pi^i] \geq E[\pi^j]$, we have

$$\begin{aligned} E[\pi^i] &= Q^{i*} - (1 - C^i(Q^*))(E[D^i(\pi^i)] - E[\pi^i]) \\ &\geq Q^{j*} - (1 - C^j(Q^*))(E[D^j(\pi^j)] - E[\pi^j]) = E[\pi^j]. \end{aligned}$$

Note that since $Q^{i*} < Q^{j*}$, we have

$$E[D^j(\pi^j)] - E[\pi^j] \geq E[D^i(\pi^i)] - E[\pi^i]. \tag{2.5}$$

Now, since $E[D^i(\pi^i)] > E[D^j(\pi^j)]$, we have

$$\begin{aligned} E[D^i(\pi^i)] &= Q^{i*} + C^i(Q^*)(E[D^i(\pi^i)] - E[\pi^i]) \\ &> Q^{j*} + C^j(Q^*)(E[D^j(\pi^j)] - E[\pi^j]) = E[D^j(\pi^j)]. \end{aligned}$$

And by the hypothesis $Q^{i*} < Q^{j*}$, we have

$$E[D^i(\pi^i)] - E[\pi^i] > E[D^j(\pi^j)] - E[\pi^j]. \tag{2.6}$$

However, the inequalities (2.5) and (2.6) contradict each other.

Next, I consider the case where $E[D^i(\pi^i)] = E[D^j(\pi^j)] = E[\pi^i] \geq E[\pi^j]$ holds. Since the limit assessment of each action takes a value between the expected objective and distorted payoffs, we should have that $Q^{i*} \geq Q^{j*}$.

Last, I consider the case where $E[D^i(\pi^i)] = E[D^j(\pi^j)] > E[\pi^i] \geq E[\pi^j]$ holds. Again, I assume that $Q^{i*} < Q^{j*}$. Since $E[D^i(\pi^i)] = E[D^j(\pi^j)]$, we have

$$\begin{aligned} & Q^{i*} + C^i(Q^*)(E[D^i(\pi^i)] - E[\pi^i]) \\ = \; & Q^{j*} + C^j(Q^*)(E[D^j(\pi^j)] - E[\pi^j]). \end{aligned}$$

However, this equation does not hold, since $Q^{i*} < Q^{j*}$, $C^i(Q^*) < C^j(Q^*)$ and $0 < E[D^i(\pi^i)] - E[\pi^i] \leq E[D^j(\pi^j)] - E[\pi^j]$. □

2.7.3 Proof of Lemma 7

I assume that one of the following inequalities,

$$E[\pi^i] \geq E[\pi^j] \geq E[D^i(\pi^i)] \geq E[D^j(\pi^j)],$$

holds strictly. Also I assume that $E[D^j(\pi^j)] > 0$. Consider some Q_1 such that $Q_1^i \geq Q_1^j$ for any $j \neq i$. Then

$$C^i(Q_1)E[\pi^i] + (1 - C^i(Q_1))E[D^i(\pi^i)] \geq C^j(Q_1)E[\pi^j] + (1 - C^j(Q_1))E[D^j(\pi^j)].$$

What I show here is that the trajectories of ODEs starting from the points with $Q_1^i = Q_1^j$ never enter the area of Q with $Q^i < Q^j$ so that at the unique rest point Q^*, which is globally asymptotically stable, we should have that $Q^{i*} \geq Q^{j*}$. First, consider the initial point Q_1 such that

$$\begin{aligned} Q_1^i = Q_1^j &\leq C^j(Q_1)E[\pi^j] + (1 - C^j(Q_1))E[D^j(\pi^j)] \\ &\leq C^i(Q_1)E[\pi^i] + (1 - C^i(Q_1))E[D^i(\pi^i)]. \end{aligned}$$

Note that $\dot{Q}^i \geq 0$ and $\dot{Q}^j \geq 0$ and if $\dot{Q}^i = 0$, then $\dot{Q}^j = 0$. Otherwise, $\dot{Q}^i > 0$ and $\dot{Q}^j \geq 0$,

$$0 \leq \frac{C^j(Q_1)E[\pi^j] + (1 - C^j(Q_1))E[D^j(\pi^j)] - Q^j}{C^i(Q_1)E[\pi^i] + (1 - C^i(Q_1))E[D^i(\pi^i)] - Q^i} \leq 1.$$

Therefore, the trajectories starting from Q_1 do not enter the area with $Q^i < Q^j$. Next, I assume that

$$\begin{aligned} &C^j(Q)E[\pi^j] + (1 - C^j(Q))E[D^j(\pi^j)] < Q^i = Q^j \\ \leq\ &C^i(Q)E[\pi^i] + (1 - C^i(Q))E[D^i(\pi^i)]. \end{aligned}$$

Then $\dot{Q}^j < 0$ and $\dot{Q}^i \geq 0$ and it is obvious that the trajectories of the ODEs do not enter the area with $Q^i < Q^j$. Finally, I assume that

$$\begin{aligned} C^j(Q)E[\pi^j] + (1 - C^j(Q))E[D^j(\pi^j)] &\leq C^i(Q)E[\pi^i] + (1 - C^i(Q))E[D^i(\pi^i)] \\ &< Q^i = Q^j. \end{aligned}$$

Then $\dot{Q}^i < 0$, $\dot{Q}^j < 0$ and

$$1 < \frac{C^j(Q)E[\pi^j] + (1 - C^j(Q))E[D^j(\pi^j)] - Q^j}{C^i(Q)E[\pi^i] + (1 - C^i(Q))E[D^i(\pi^i)] - Q^i}.$$

And again, the trajectories of the ODEs also do not enter the area with $Q^i < Q^j$. In sum, the trajectories which start from the points on the line with $Q^i = Q^j$ never enter the area of Q with $Q^i < Q^j$ and thus $Q^{i*} \geq Q^{j*}$. This argument can be applied to the other cases where $E[D^j(\pi^j)] \leq 0$. □

2.7.4 Proof of Proposition 2

Assume that $Q^{i*} < Q^{j*}$. By the property of choice rules, we have that $C^i(Q^*) < C^j(Q^*)$ and hence $C^i(Q^*) < (1 - C^i(Q^*))$. Since $C^i(Q^*) < \frac{1}{2}$, we have

$$\begin{aligned} Q^{i*} &= E[\pi^i] + (1 - C^i(Q^*))(E[D^i(\pi^i)] - E[\pi^i]) \\ &\geq E[\pi^i] + \frac{1}{2}(E[D^i(\pi^i)] - E[\pi^i]) \\ &= \frac{E[D^i(\pi^i)] + E[\pi^i]}{2}. \end{aligned}$$

Since $Q^{j*} \leq \max\{E[D^j(\pi^j)], E[\pi^j]\}$, we have $Q^{i*} \geq Q^{j*}$. However, this condition contradicts the original hypothesis. □

2.7.5 Proof of Proposition 3

Consider the difference of Q^{i*} and Q^{j*}. Then

$$\begin{aligned} Q^{i*} - Q^{j*} &= E[\pi^i] - E[\pi^j] + (1-\alpha^{i*})E[G(\pi^i - \pi^j)] - (1-\alpha^{j*})E[G(\pi^j - \pi^i)] \\ &= (E[\pi^i] - E[\pi^j]) + (1-\alpha^{i*})E[G(\pi^i - \pi^j)] + \alpha^{i*}E[G(\pi^i - \pi^j)] \\ &= (E[\pi^i] - E[\pi^j]) + E[G(\pi^i - \pi^j)]. \end{aligned}$$

Since $E[G(\pi^i - \pi^j)] \geq G(E[\pi^i] - E[\pi^j]) \geq 0$ if $E[\pi^i] \geq E[\pi^j]$, we have that $Q^{i*} \geq Q^{i*}$ if $E[\pi^i] \geq E[\pi^j]$. This means that $\alpha^{i*} \geq \alpha^{j*}$. □

CHAPTER 3

AN ADAPTIVE LEARNING MODEL IN COORDINATION GAMES

3.1 Introduction

Over the past few decades, learning models have received much attention in the theoretical and experimental literature of cognitive science. One such model is fictitious play, where players form beliefs about their opponents' play and best respond to these beliefs. In fictitious play model, players know the payoff structure and their opponents' strategy sets.

Whereas there are other learning models where players have limited information; players may not have information about payoff structure, opponents' strategy sets, or they may not even know whether they are playing against other players. In the situation, they may not be able to form beliefs about the way that the opponents play or all possible outcomes. What they do know is their own available actions and the results from the previous play, that is, the realized payoffs from chosen actions. Instead of forming beliefs about all possible outcomes, each player makes a subjective assessment on each of his actions based on the realized payoffs from the action and tends to pick the action which

has achieved better results than the others in the past.

One such model with limited information is the reinforcement learning model introduced by Erev and Roth (1998) (ER, hereafter), where they model the observed behaviour of agents in the lab[1]. In their model, agent chooses an action randomly, where the choice probability of the action is the fraction of the payoffs realized from the action over the total payoffs realized for all available actions[2].

In another learning model which is introduced by Sarin and Vahid (1999) (SV, hereafter), players make a subjective payoff assessment of each of his actions, where the assessment is a weighted average of past payoffs, and they choose the action which has the highest assessment. After receiving a payoff, each player updates the assessment of chosen action adaptively; the assessment of chosen action is adjusted toward the received payoff[3].

In this chapter, I provide a theoretical prediction of the way in which myopic players in the SV model[4] behave in the long run in general games, mostly in coordination games, which are of interest to a wide range of researchers[5]. In this model, the initial assessment of each action is assumed to take a value between the maximum and the minimum payoff that the action can provide[6]. For instance, players may have experienced the game in advance so they may use their knowledge of previous payoffs to form an initial assessment

[1]There also exists work which has invesitgated their model theoretically. For instance, see Beggs (2005) and Laslier et al. (2001).

[2]Since the payoffs are assumed to be positive, each player increases the probability of choosing an action whenever the action is chosen.

[3]It is worth to note that if the realized payoff of an action is lower than the assessment of the action, then the chance of the action being chosen in the next period becomes less likely.

[4]Note that the players do not observe the foregone payoff information and do not update the assessments of unchosen actions. Therefore, the learning dynamics here is different from the one introduced in Chapter 2.

[5]As examples of experimental works on coordination games, Cooper, DeJoung, Forsythe and Ross (1992) and Van Huyck, Battalio and Beil (1990) have investigated which among multiple Nash equilibria, is the one played in the lab.

[6]See also Sarin (1999) for the justification of the assumption.

of each action. Given those initial assessments, each player picks the action which has the highest assessment; in this chapter, each player does not experience any stochastic perturbations on his own assessments.

After players have played a game and received the payoffs, each player updates his assessment using the realized payoff; the new assessment of a chosen action is a convex combination of the current assessment and the realized payoff. In the present chapter, the weights on the realized payoffs are assumed to be random variables, meaning that players are not sure how much they incorporate the new payoff information into their assessments, which may be also affected by their mood. As a special instance, I also consider some cases where those weights are non-random variables. For example, I consider players who believe that the situation they are involved in is stationary so that each action's assessment is the arithmetic mean of its past payoffs. I also consider the case where players believe that the environment is non-stationary and put the same weight on all new payoff information.

Since the initial assessment of each action is smaller than the best payoff that the action can give, each player increases his assessment of the action when he receives the best payoff. If one action profile gives the best payoff to all players and they play it in some period, then players will keep choosing the action profile in all subsequent periods. I call an action profile absorbing state if once players play the action profile in some period, then they play it in all subsequent periods.

Furthermore, there exist other cases where players stick to one action profile. One such case is that their assessments of other actions become so low that the actions are never tried again. Another case is that payoffs from the action profile are greater than the

other assessments and players keep playing the action profile, even though it does not give them the best payoffs. It is shown that each pure Nash equilibrium is always a candidate of the convergence point, that is, for each strict Nash equilibrium there exists a range of assessments for all players and actions such that players stick to the Nash equilibrium forever. In addition, if (i) at any non-Nash equilibrium action profile, at least one player receives the payoff which is less than his maximin payoff, or (ii) all non-Nash equilibrium action profiles give the same payoff, then players end up playing a strict Nash equilibrium with probability one.

To see this in detail, I consider 2×2 coordination games and one non-2×2 coordination game. In 2×2 coordination games, since only two actions are available for each player, I can divide them into three categories according to the numbers of action profiles at which each player receives the payoff which is strictly greather than the other possible payoffs from his current action. Since each player receives the best payoff from his current action, he never changes his action; note that such an action profile is absorbing. Notice also that the number of such action profiles ranges from zero to two in 2×2 coordination games. The class of coordination games with two absorbing states includes the battle of the sexes and pure coordination games, where two absorbing states correspond to pure Nash equilibria. The class of coordination games with one absorbing state can be subdivided into the following cases: (1) the absorbing state corresponds to a Nash equilibrium; and (2) the absorbing state corresponds to one non-Nash equilibrium action profile. The class of coordination games in case (1) includes the stag hunt game, while the class of coordination games in case (2) includes the game of chicken and market entry games. Then I show the following results. In coordination games with two absorbing states, (i) if

the maximin actions of both players coincide, then they end up playing a Nash equilibrium with probability one, (ii) if maximin actions do not coincide for both players, then players end up playing a Nash equilibrium or the maximin action profile with probability one. In coordination games in case (1), players end up playing a Nash equilibrium with probability one if the maximin actions of both players coincide. In coordination games in case (2), players end up playing a strict Nash equilibrium or a maximin action profile.

In a non-2×2 coordination game introduced by Van Huyck, Battalio and Beil (1990) (VHBB, hereafter), each player is asked to pick a number from a finite set. If players fail to coordinate, the player who picks the smallest number among players' choices receives the highest payoff. In addition, each number gives a better payoff when the choice is closer to the smallest number among all the players' choices. I show that each Nash equilibrium, in which players coordinate to pick the same number, is absorbing[1]. It is also shown that the smallest number of the players' choices weakly decreases over time. Next, I consider the case where the second best payoff from each action is lower than the payoff from the maximin action, which is the smallest number of their choice set. Hence, players are better off if they choose the smallest number of their choice set when they fail to pick the smallest number among the players' choices. In this case, I show that players end up playing a Nash equilibrium with probability one, which can be also observed in the experimental results by VHBB.

[1]It is absorbing if the minimum number gives different payoffs for opponents' choices. If it gives the same payoff for any opponents' choice, then I have to assume an inertia condition for players' tie break rule for the corresponding Nash equilibrium to be absorbing. See the following argument.

3.1.1 Literature review

In this chapter, I investigate convergence properties of SV learning model in mainly coordination games. In the prisoner's dilemma game, Sarin (1999) shows that players end up playing for mutual cooperation or mutual defection. In this chapter, players do not experience any stochastic emotional noise on their assessments, whereas SV also investigate a decision problem in which a decision maker experiences the stochastic shocks on his assessments in each decision period. Then SV show that (1) assessment of the action which is played infinitely often converges in distribution to a random variable whose expected value is the expected objective payoff and (2) if one action first-order stochastically dominates the other, then the former action is played more often than the other on average. In the context of SV leanring with the shocks in games, Leslie and Collins (2006) investigate the model with slightly different updating rules and show convergence of strategies to a Nash distribution[1] in the partnership game and the zero-sum games. With the SV updating rule, Cominetti, Melo and Sorin (2010) show the general convergence result when each player's choice rule is the logistic choice rule. They show that players' choice probabilities converge to a unique Nash distribution if the noise term of the logistic choice rule for each player is big enough. By a property of the logistic choice rule if its noise term becomes large then the choice probability approaches a uniform distribution. Hence, players in their model are more likely to choose an action which does not have the highest assessment each time. However, players in the SV model without emotional shocks do not choose the actions; they always pick the action which they think is the best based on past payoff realizations. In this chapter, even the lack of exploration,

[1]Nash distribution is Nash equilibrium under stochastic perturbations on payoffs. If the expected values of the perturbations are 0, then Nash distribution coincides with the quantal response equilibrium proposed by McKelvey and Palfrey (1995)

it is shown that players end up playing a Nash equilibrium in several coordination games.

Lastly, some authors have provided empirical supports of this model. For instance, Sarin and Vahid (2001) show that the SV model can explain the data by ER at least as well as the ER model does. Chen and Khoroshilov (2003) show that among learning models comprising the ER model, the SV model, and the experience-weighted attraction learning model by Camerer and Ho (1999), the SV model can best explain the data in coordination games and cost sharing games.

3.2 General Games

There are M players who play the same game repeatedly over periods. Let $N = \{1, ..., M\}$ be the set of players. In each period, $n \in \mathbb{N}$, each player chooses an action from his own action set simultaneously. Let S^i be the finite set of actions for player $i \in N$. After all the players choose actions, each player receives a payoff. If players play $(s^i)_{i \in N} \in \Pi_{i \in N} S^i$, then player i's realized payoff is denoted by $u^i(s^i, s^{-i})$, where $s^{-i} = (s^1, ..., s^{i-1}, s^{i+1}, ..., s^M)$. When choosing an action, each player does not know the payoff functions or the environment in which he is involved.

In each period, each player assigns subjective payoff assessments on his actions; let $Q_n^i(s^i) \in \mathbb{R}$ denote player i's assessment on action s^i in period n. Let $Q_n^i = (Q_n^i(s^i))_{s^i \in S^i}$ be the vector of assessments for all actions for player i. I assume that the initial assessment for each action and each player takes a value between the maximum and the minimum value that the action gives; thus,

$$Q_1^i(s^i) \in (\min_{s^{-i}} u^i(s^i, s^{-i}), \max_{s^{-i}} u^i(s^i, s^{-i})),$$

for all $i \in N$ and $s^i \in S^i$. If $\min_{s^{-i}} u^i(s^i, s^{-i}) = \max_{s^{-i}} u^i(s^i, s^{-i})$, then I assume that $Q_1^i(s^i) = \min_{s^{-i}} u^i(s^i, s^{-i}) = \max_{s^{-i}} u^i(s^i, s^{-i})$.

In each period, each player chooses the action which he believes will give the highest payoff; given his assessments, he chooses the action which has the highest assessment in the period. Therefore, if s_n^{i*} is the action that player i chooses in period n, then

$$s_n^{i*} = \arg\max_{s^i} Q_n^i(s^i).$$

For a tie break situation, which arises when more than two actions have the highest assessment, I introduce two types of tie break rules. I say that a tie break rule satisfies the inertia condition if the rule picks the action which was chosen in the last period; if actions which have the highest assessment were not chosen in the last period, then the rule picks one of the actions randomly. As a comparison, I also introduce another tie break condition, the uniform condition, where the rule picks each of the actions which have the highest assessment with equal probability. In the following argument, I specify a tie break rule if the result depends on the tie break rule; otherwise, the results do not depend on the tie break rule assumption.

After playing the game in each period, each player observes only his own payoff; players observe neither their opponents' actions nor their payoffs. Given his own realized payoff, each player updates his assessment of the action chosen in the previous period. Specifically, if player i receives a payoff $u_n^i(s^i, s^{-i})$ when players play (s^i, s^{-i}), then he

updates Q_n^i as follows;

$$Q_{n+1}^i(s^i) = \begin{cases} (1-\lambda_n^i(s^i))Q_n^i(s^i) + \lambda_n^i(s^i)u_n^i(s^i, s^{-i}) & \textit{if } s^i \textit{ is chosen in period } n \\ Q_n^i(s^i) & \textit{otherwise} \end{cases}$$

where $\lambda_n^i(s^i)$ is player i's weighting parameter for action s^i in period n. I assume that $\lambda_n^i(s^i)$ is a random variable which takes a value between 0 and 1; $\lambda_n^i(s^i) \in (0,1)$. It reflects the idea that players are uncertain how far to incorporate the new payoff information into their new assessments. The uncertainty can also be interpreted as players' emotional shocks. How far they incorporate the new payoff information depends on their random mood. I also assume that the sequence of weighting parameters, $\{\lambda_n^i(s^i)\}_{i,n,s^i}$ is independent among periods, players and actions and it is identically distributed among periods. I assume that the weighting parameter $\lambda_n^i(s^i)$ has a density function which is strictly positive on the domain $(0,1)$ for all i and s^i.

3.3 Results

In this section, I investigate the convergence results in general games. In later sections, I focus on specific games, in particular coordination games. I say $(s^i)_{i\in I}$ is *absorbing* if once players play the action profile in a period then they play it in all subsequent periods.

Proposition 5. *If $(s^i)_{i\in I}$ is such that (i) for all i,*

$$u^i(s^i, s^{-i}) = \max_{t^{-i}\in S^{-i}} u^i(s^i, t^{-i})$$

and (ii) for all i there exists r^{-i} such that

$$\max_{t^{-i} \in S^{-i}} u^i(s^i, t^{-i}) > u^i(s^i, r^{-i})$$

then $(s^i)_{i \in I}$ is absorbing.

Proof. Consider the case where players pick the action profile $(s^i)_{i \in I}$ in some period n. In the case, player i receives the payoff $u^i(s^i, s^{-i})$. Note that the value $u^i(s^i, s^{-i})$ is the maximum value that action s^i can give; therefore, by condition (ii), player i inflates the assessment of the action s^i. Since the assessments of other actions do not change in the next period, player i plays action s^i in period $n+1$ again. Since this logic can be applied to other periods and I pick player i randomly, players play the same action profile in all the subsequent periods. □

If the inertia condition is always assumed for each player's tie break rule, then condition *(ii)* in Proposition 5 is not required. However, if the uniform condition is assumed, without condition *(ii)*, players may not converge to play one action profile. As an extreme example, if two actions give the same payoff for any opponents' actions and the payoff is higher than any other payoffs that any other action can give, then he plays those two actions with equal probability forever.

From Proposition 5, it is easy to see that even action profiles which consist of dominated strategies for all players can be absorbing. To see this, assume that two players play the prisoner's dilemma game which has the following payoff matrix;

where the strategy "C" is strictly dominated by the strategy "D" for both players. Notice that at (C,C), both players receive the highest payoffs from the action "C";

	C	D
C	1,1	-1,2
D	2,-1	0,0

$u^i(C,C) = \max_{s^{-i} \in \{C,D\}} u^i(C, s^{-i})$ for for both players. Hence, if players play (C,C) once, then they always play it afterwards[1].

In the next statement, I show that player i stops playing an action if the assessment of the action becomes smaller than the minimum payoff that another action can give;

Proposition 6. *If $Q_n^i(s^i) < \min_{s^{-i}} u(t^i, s^{-i})$ in some period n for $t^i \neq s^i$, then player i does not choose s^i after period n.*

Proof. From the fact that $Q_n^i(t^i) > \min_{s^{-i}} u(t^i, s^{-i})$, we have the fact that $Q_n^i(t^i) > Q_n^i(s^i)$. Notice that s^i is not chosen in period n. Since the assessment of the chosen action is a convex combination of realized payoff and the assessment of the previous period with $\lambda_n^i \in (0,1)$ for all i and n, we have $Q_{n+1}^i(t^i) > \min_{s^{-i}} u(t^i, s^{-i})$. Notice also that s^i is not chosen in period n and thus the assessment of the action is unchanged in the next period $n+1$. Therefore we have $Q_{n+1}^i(t^i) > \min_{s^{-i}} u(t^i, s^{-i}) > Q_{n+1}^i(s^i)$ and player i will not choose s^i in period $n+1$. The same logic can be applied in later periods and thus player i will not choose s^i in any subsequent periods. □

Once the assessment of one action becomes lower than the worst payoff from another action, then the action will not be chosen forever. Therefore, if the worst payoff from one action is greater than the best payoff from another action, then the latter action is never chosen at any time. One natural question is whether players end up playing a strict Nash equilibrium. In the following statement, I show that for any strict Nash equilibrium,

[1] See Sarin (1999) for the result.

there exist assessments for all players such that the players end up playing the strict Nash equilibrium:

Proposition 7. *For any strict Nash equilibrium, there exist assessments in period n for all players such that they play the Nash equilibrium in the period and all subsequent periods.*

Proof. Let $(s^{i*})_{i\in N}$ be a strict Nash equilibrium and s^{i*} be player i's strategy at the strict Nash equilibrium. Then, we have the following condition; for all $i \in N$,

$$u^i(s^{i*}, s^{-i*}) > u^i(t^i, s^{-i*}) \tag{3.1}$$

for all $t^i \neq s^{i*}$. I assume that in period n, the following conditions for assessments are satisfied; for all i,

$$Q_n^i(s^{i*}) > Q_n^i(t^i) \tag{3.2}$$

and

$$u^i(s^{i*}, s^{-i*}) > Q_n^i(t^i) \tag{3.3}$$

for all $t^i \neq s^{i*}$. Note that condition (3.3) holds, since by condition (3.1), the minimum value of the assessment of action t^i is less than or equal to $u^i(t^i, s^{-i*})$, which is strictly less than $u^i(s^{i*}, s^{-i*})$. Thus, players play the strict Nash equilibrium in period n. Note also that

$$Q_{n+1}^i(s^{i*}) \geq \min\{Q_n^i(s^{i*}), u^i(s^{i*}, s^{-i*})\} > Q_n^i(t^i) = Q_{n+1}^i(t^i)$$

for all $t^i \neq s^{i*}$ and players play the strict Nash equilibrium again in period n+1. □

Proposition 7 says that any strict Nash equilibrium is always a candidate of the conver-

gence point. However, it is possible that players end up playing a non-Nash equilibrium. Hence, it is natural to consider the case where if they converge to play one action profile, then it should be a strict Nash equilibrium. Notice that if one action profile $(s^i)_{i\in I}$ is played forever, then (1) each player receives a better payoff than the assessment of chosen action and he plays the action again, (2) each player receives a payoff which is not better than the assessment of chosen action, but the assessments of the other actions are less than the payoff, so that he plays the action again, or (3) the action gives the same payoff for any other players' actions, so that the assessment of the action is unchanged and the assessments of other actions are strictly less than the assessment of the action[1].

I say that players end up playing $(s^i)_{i\in I}$ if there exists n such that for all periods after n, players play $(s^i)_{i\in N}$. If the condition $Q^i_m(s^i) > Q^i_m(t^i)$ satisfies for all i, $m > n$, and $s^i \neq t^i$, then players end up playing $(s^i)_{i\in I}$ [2]. In the following statements, I focus on the cases where all pure Nash equilibria are strict. I also assume that there do not exist any redundant actions which always give the same constant payoff; for any $i \in N$ and actions $s^i, t^i \in S^i$, $s^i \neq t^i$, the following condition does not hold;

$$u^i(s^i, s^{-i}) = u^i(t^i, t^{-i}) \text{ for all } s^{-i}, t^{-i} \in S^{-i}.$$

Lemma 8. *For any initial assessments, players never end up playing* (s^i, s^{-i}) *if* $\exists i \in N$, $\exists t^i \in S^i$ *s.t.*

$$u^i(t^i, t^{-i}) \neq u^i(t^i, r^{-i}) \text{ for some } r^{-i} \neq t^{-i} \in S^{-i}$$

[1] If players' tie break rule satisfy the inertia condition, then the assessment of other actions need to be weakly less than the assessment of this action and the action is played in the previous period.

[2] This condition does not include some convergence case which happens when I assume the inertia condition to all players. In such a case, I can weaken the condition as follows; players converge to play $(s^i)_{i\in N}$ if there exist n and $(Q^i_n)_{i\in I}$ such that for all $m \geq n$, i, and $t^i \neq s^i$, $Q^i_m(s^i) \geq Q^i_m(t^i)$ where player i picks s^i in period n.

and

$$u^i(s^i, s^{-i}) \leq \min_{t^{-i} \in S^{-i}} u^i(t^i, t^{-i}). \tag{3.4}$$

Proof. I prove by contradiction; I assume that there exists a set of assessments such that players end up playing $(s^i)_{i \in I}$. Hence, there exists n such that for all $m > n$, $Q_m(s^i) > Q_m(t^i) > \min_{t^{-i} \in S^{-i}} u^i(t^i, t^{-i})$ for all $t^i \in S^i$ [1]. If $u^i(s^i, s^{-i}) \geq Q_m(s^i)$, then $u^i(s^i, s^{-i}) \geq Q_m(s^i) > Q_m(t^i) > \min_{t^{-i} \in S^{-i}} u^i(t^i, t^{-i})$, which contradicts the hypothesis. If $Q^i_m(s^i) > u^i(s^i, s^{-i})$, then it should be that $Q^i_m(s^i) > u^i(s^i, s^{-i}) \geq Q^i_m(t^i) > \min_{t^{-i} \in S^{-i}} u(t^i, t^{-i})$, if not, then $Q^i_m(s^i)$ becomes less than $Q^i_m(t^i)$. However, again the condition contradicts the hypothesis. □

If condition (3.4) is satisfied at non-Nash equilibrium action profiles, then players never end up playing one of them. It is also obvious that the condition is not satisfied at each strict Nash equilibrium. Condition (3.4) says that there exists at least one player who can find an action which always gives a better payoff than his current payoff from the action. It also means that there exists a player who receives a payoff which is less than his maximin payoff. Though the condition limits the class of games, still there exist interesting games which satisfy the condition. For example, the stag hunt game satisfies condition (3.4) at non-Nash equilibrium action profiles and has the following payoff matrix;

	Rabbit	Stag
Rabbit	1,1	2,0
Stag	0,2	5,5

At non-Nash equilibrium action profile, one player decides to hunt a stag while the other player decides to hunt a rabbit. The player who decides to hunt a stag fails and

[1] The following argument is also true if I assume that $Q_m(s^i) \geq Q_m(t^i)$ for all $m > n$, which is the condition for convergence when the inertia condition is assumed for each player's tie break rule.

receives nothing and the payoff is less than the minimum payoff from hunting a rabbit, 1, which is given when both players decide to hunt a rabbit together and share it.

Another coordination game which satisfies condition (3.4) is the first order statistic game where each player chooses a number from a finite set and coordination occurs when all of them pick the same number. In addition, if players succeed to coordinate at a higher number then they receive a better payoff. When they fail to coordinate on choosing the same number, the player who has chosen the smallest number receives the best payoff, the player who has chosen the second smallest number receives the second best payoff, and so on; the smaller number the player has chosen, the better payoff he receives. For example, I consider the case where each player picks a number from one to four and the payoff matrix of each player is expressed as follows:

	1	2	3	4
1	1	1.5	1.5	1.5
2	0	2	2.5	2.5
3	-1	0	3	3.5
4	-2	-1	0	4

The first column represents player i's choice while the first row represents the minimum value of his opponents' choices. It is easy to see that at each Nash equilibrium, all players pick the same number. Since action 1 gives at least 1 and players who fail to pick the smallest number receives at most 0, this game satisfies the condition (3.4).

In both games, condition (3.4) holds strictly. In other games, such as the battle of the sexes and pure coordination games, condition (3.4) holds weakly, in particular $u^i(s^i, s^{-i}) = u^i(t^i, t^{-i})$ for all i and $(s^i), (t^i) \notin E^*$, where E^* is the set of pure Nash

equilibria. For instance, the battle of the sexes game has the following payoff;

	s_1^2	s_2^2
s_1^1	1,2	0,0
s_2^1	0,0	2,1

In the following theorem, I show that players end up playing a Nash equilibrium almost surely if (i) condition (3.4) is satisfied strictly at non-Nash equilibrium profiles, or (ii) if each player's payoffs at non-Nash equilibrium action profiles are equal;

Theorem 3. *Players end up playing a strict Nash equilibrium almost surely if (i)* $\forall (s^i)_{i\in N} \notin E^*$, $\exists i \in N$, $\exists t^i \in S^i$ *s.t.*

$$u^i(s^i, s^{-i}) < \min_{t^{-i}\in S^{-i}} u^i(t^i, t^{-i}). \tag{3.5}$$

or (ii) $u^i(s^i, s^{-i}) = u^i(t^i, t^{-i})$ $\forall i \in N$ *and* $\forall (s^i)_{i\in N}, (t^i)_{i\in N} \notin E^*$.

Proof. See Appendix. □

3.4 VHBB Coordination Games

I first consider the coordination game proposed by Van Huyck, Battalio and Beil (1990), where there exist M players with $S^i = S = \{1, 2, ..., J\}$ for all $i \in N = \{1, ..., M\}$ and players have the following payoff function;

$$u^i(s^i, s^{-i}) = a(\min\{s^1, ..., s^M\}) - bs^i,$$

where $a > b > 0$ for all $i \in N$. If J=4, then player i's payoffs are shown by the following matrix;

	1	2	3	4
1	a-b	a-b	a-b	a-b
2	a-2b	2a-2b	2a-2b	2a-2b
3	a-3b	2a-3b	3a-3b	3a-3b
4	a-4b	2a-4b	3a-4b	4a-4b

where the numbers in the first column correspond to player i's action and the numbers in the first row correspond to the minimum values of the opponents' actions. It is easy to check that $(j, j, j, ..., j)$, $j \in S$, is a pure Nash equilibrium.

Notice that the pure Nash equilibria except $(1, 1, ..., 1)$ are absorbing. However, if I assume the inertia condition for each player's tie break rule, then $(1, 1, ..., 1)$ is also absorbing. In this section, I assume that each player's tie break rule satisfies the inertia condition.

Lemma 9. *For $j \in S$, the pure Nash equilibrium $(j, j, ..., j)$ is absorbing.*

When a player is choosing the smallest action among players' actions, he is receiving the best payoff that the action can give. Therefore, the player does not change his action when he is choosing the smallest action except when he chooses 1 and is facing a tie break situation. If the inertia condition is satisfied, then he chooses 1 forever and the minimum value of actions does not increase over time. Moreover, since the minimum value is bounded below, it converges.

Lemma 10. *The minimum value of actions among players is non-increasing over periods and converges almost surely.*

I additionally assume that each action's second best payoff, $a(j-1)-bj$ for $j \in S/\{1\}$,

is less than the secure payoff, $a - b$. That is,

$$a(j-1) - bj < a - b$$

for all $j \in S/\{1\}$. This means that each player receives a payoff better than the secure payoff only when his choice is the smallest among all players' choices. Given this assumption, players end up playing a Nash equilibrium.

Proposition 8. *If $a(j-1) - bj < a - b$ for all $j \in S\backslash\{1\}$ and players' tie break rules satisfy the inertia condition, then players end up playing a pure Nash equilibrium almost surely.*

Proof. If a player is choosing an action which is not the smallest action among players, then the payoff which the action gives is less than $a - b$. Let $\underline{j}(n)$ be the minimum value of actions in period n. From Lemma 10, $\underline{j}(n) \geq \underline{j}(m)$ for $m \geq n$. Hence, actions which are strictly greater than $\underline{j}(n)$ always give a payoff less than $a - b$ after period n. Therefore, each player never plays $s > \underline{j}(n)$ infinitely often. If $s > \underline{j}(n)$ is played infinitely often, then the assessment of the action becomes lower than $a - b$ in some period $m > n$ with probability one; that is, the assessment of the action becomes lower than the assessment of action 1. Since the assessment of the action 1 never changes, he never plays action s afterwards, which contradicts the hypothesis. Thus, after some period $l > n$, he plays $\underline{j}(n)$ or some lower action. If all players play $j = \underline{j}(n)$, then players play $(j, j, ..., j)$ afterwards. If one player plays $k < \underline{j}(n)$ in period $m > n$ and $j(m) = k$, then I can apply the same logic. If $\underline{j}(n) = 1$, then there is no lower number that players can choose and they end up playing Nash equilibrium $(1, 1, ..., 1)$. Since there are finitely many players and actions,

players end up playing a Nash equilibrium almost surely. □

3.5 2×2 Coordination Games

In this section, I focus on 2×2 coordination games, which have the following payoff matrix;

	s_1^2	s_2^2
s_1^1	a_{11}, b_{11}	a_{12}, b_{12}
s_2^1	a_{21}, b_{21}	a_{22}, b_{22}

where $a_{11} > a_{21}$, $a_{22} > a_{12}$, $b_{11} > b_{12}$ and $b_{22} > b_{21}$ hold. Note that in these coordination games, the pure Nash equilibria are (s_1^1, s_1^2) and (s_2^1, s_2^2). For the purpose of analysis, I divide 2×2 coordination games into three categories according to the number of action profiles at each of which each player receives a payoff which is strictly better than another payoff that his current action gives; if (s_i^1, s_j^2) is such an action profile, then $a_{ij} > a_{ik}$ and $b_{ij} > b_{lj}$ for $j \neq k$ and $i \neq l$. Note that such action profile is absorbing. Therefore, the categorization also depends on the number of absorbing states under the tie break rule with the uniform condition. It is easy to check that there exist three possible cases for general 2×2 games: (1) both diagonal or both off-diagonal action profiles are absorbing states; (2) only one action profile is an absorbing state; or (3) there does not exist any absorbing state.

Since 2×2 coordination games have additional conditions, off-diagonal action profiles cannot be absorbing at the same time. Therefore, the condition for (1) is as follows;

$$(1)\ \min\{a_{11}, a_{22}\} > \max\{a_{21}, a_{12}\}\ \textit{and}\ \ \min\{b_{11}, b_{22}\} > \max\{b_{12}, b_{21}\}.$$

In the case of (2) and (3), the following condition should hold:

$$(2),(3)\ \min\{a_{11},a_{22}\} \leq \max\{a_{21},a_{12}\}\ \textit{or}\ \ \min\{b_{11},b_{22}\} \leq \max\{b_{12},b_{21}\}.$$

Without loss of generality, I assume for case (2) and (3) that $a_{22} \leq a_{21}$, that is, $a_{11} > a_{21} \geq a_{22} > a_{12}$ holds. Note that if an absorbing state exists, then it should be (s_1^1, s_1^2) or (s_2^1, s_1^2). Given the inequality of payoffs for player 1, (2-1) if $b_{11} > b_{21}$ holds, then (s_1^1, s_1^2) is the unique absorbing state; (2-2) if $b_{21} > b_{11}$ and $a_{21} > a_{22}$ hold, then (s_2^1, s_1^2) is the unique absorbing state; (3) if otherwise, then there does not exist an absorbing state.

In the following sections, I investigate games in categories (1), (2-1), (2-2) and (3). Specifically, the following games are considered: the battle of the sexes game and pure coordination games from category (1), the stag hunt game from category (2-1) and market entry games and the game of chicken from category (2-2) and (3).

3.5.1 The Battle of the Sexes Game and Pure Coordination Games

In this subsection I consider coordination games in category (1). Games in this category satisfy the following conditions; $\min\{a_{11},a_{22}\} > \max\{a_{21},a_{12}\}$ and $\min\{b_{11},b_{22}\} > \max\{b_{12},b_{21}\}$ and on-diagonal action profiles, pure Nash equilibria, are absorbing states. The condition says that for both players, coordinating one of the Nash equilibria always gives a better payoff than playing non-Nash equilibrium profiles. It is easy to see that the battle of the sexes game and pure coordination games satisfy the condition. For instance, the battle of the sexes game has the following payoff matrix;

In this game, the row player prefers going to a football game together to going to an

	Opera	Football
Opera	$1, 2$	$0, 0$
Football	$0, 0$	$2, 1$

opera together, while the column player enjoys going to the opera together more than going to a football game together. However, players are worse off when they fail to coordinate to go to one of them.

By Theorem 3, we know that players end up playing a pure Nash equilibrium almost surely;

Corollary 4. *In 2×2 coordination games in category (1), if* $u^1(s_k^1, s_l^2) \geq u^1(s_l^1, s_k^2)$ *and* $u^2(s_l^1, s_k^2) \geq u^2(s_k^1, s_l^2)$ *for* $k \neq l$*, then players end up playing a pure Nash equilibrium.*

Another case to be considered is that each player receives the worst payoff from the same action profile. Assume that players have the following payoff matrix;

	Opera	Football
Opera	1,2	0,0
Football	0.5,0.5	2,1

Notice that the row player enjoys going to a football game alone more than going to an opera alone. The column player is in the opposite situation - she enjoys going to the opera alone more than going to the football game alone. In this case, it is a possible outcome that players fail to coordinate and they end up playing their favored actions (Football, Opera).

Proposition 9. *In 2×2 coordination games in category (1), if* $u^1(s_k^1, s_l^2) > u^1(s_l^1, s_k^2)$ *and* $u^2(s_k^1, s_l^2) > u^2(s_l^1, s_k^2)$ *for* $k \neq l$*, then players end up playing a Nash equilibrium or* (s_k^1, s_l^2)*.*

Proof. Since (s_l^1, s_k^2) gives the worst payoff for both players, they never play (s_l^1, s_k^2) infinitely often. Notice, too, that if $Q^1(s_l^1) < a_{kl}$ and $Q^2(s_k^2) < b_{kl}$, then player 1 never plays s_l^1 and player 2 never plays s_k^2; they end up playing (s_k^1, s_l^2). In sum, players end up playing a Nash equilibrium or (s_k^1, s_l^2). □

3.5.2 The Stag Hunt Game

In this subsection, I consider coordination games in category (2-1), where the conditions $a_{11} > a_{21} \geq a_{22} > a_{12}$ and $b_{11} > b_{21}$ hold. For example, the stag hunt game satisfies this condition; the condition $b_{11} > b_{12} \geq b_{22} > b_{21}$ holds in the stag hunt game. For instance, the stag hunt game has the following payoff matrix;

	s_1^2	s_2^2
s_1^1	10,10	0,8
s_2^1	8,0	7,7

It is worth noting that in the stag hunt game, Nash equilibrium (s_2^1, s_2^2) is not absorbing. However, players end up playing one of pure Nash equilibria, including (s_2^1, s_2^2). In the stag hunt game, at each off-diagonal action profile, one player receives the worst payoff. Therefore, by Theorem 3, players end up playing a Nash equilibrium almost surely. In category (2-1), a slightly weaker condition on off-diagonal payoffs is required for the convergence to Nash equilibrium;

Proposition 10. *In 2×2 coordination games in category (2-1), players end up playing a pure Nash equilibrium almost surely if* $b_{12} \geq b_{21}$.

Proof. Note that if players play (s_1^1, s_1^2) once, they play it forever. Now I show that players never stick to (s_1^1, s_2^2) or (s_2^1, s_1^2). If players play (s_1^1, s_2^2) infinitely often, then whenever they

play it, there is a positive probability that the assessment of s_1^1 becomes lower than a_{22} and then player 1 stops playing s_1^1. Next I assume that $b_{12} > b_{21}$. By the same logic, players cannot play (s_2^1, s_1^2) infinitely often, since player 2 stops playing s_1^2 in some period in which $Q^2(s_1^2) < b_{12}$. Last, I assume that $b_{12} = b_{21}$. I assume that players never play (s_1^1, s_1^2). Therefore, players play only (s_1^1, s_2^2), (s_2^1, s_1^2) or (s_2^1, s_2^2). Note that when players (s_2^1, s_1^2), player 2 is receiving the worst payoff, while player 1 is receiving the best payoff from s_2^1. Therefore, player 2 changes his action to s_2^2 at some point. Note also that if $Q^1(s_1^1) > a_{22}$, then players change to play (s_1^1, s_2^2) at some point, since player 1 is receiving the worst payoff from s_2^1. At (s_1^1, s_2^2), both players receive the worst payoff so players change and play (1) (s_2^1, s_1^2) or (2) (s_2^1, s_2^2). Hence, players infinitely play (s_1^1, s_2^2). If so, then at some period, the assessment of action s_1^1 becomes lower than a_{22}. Therefore, player 1 stops playing s_1^1. Given this fact, players end up by playing (s_2^1, s_2^2) almost surely. This is because (1) at (s_2^1, s_1^2) player 2 receives the worst payoff and he changes to s_2^2, (2) at (s_2^1, s_2^2), player 1 receives the worst payoff from action s_2^1, though the assessment of s_1^1 is lower than a_{22}; player 1 never changes his action to s_1^1. □

If $b_{12} < b_{21}$ is satisfied, then there exists a possibility that players play (s_2^1, s_1^2) forever. This happens when $Q^2(s_2^2) < b_{21}$ and $Q^1(s_1^1) < a_{22}$.

3.5.3 The Game of Chicken and Market Entry Games

In this subsection, I first consider coordination games in category (2-2), where $b_{21} > b_{11}$ and $a_{21} > a_{22}$ hold. Since (s_2^1, s_1^2) is absorbing, the convergence to Nash equilibrium is not guaranteed in games in this category. For example, the game of chicken satisfies the condition, where it has the following payoff matrix:

	Swerve	Stay
Stay	1, -1	-10, -10
Swerve	0, 0	-1, 1

where each player shows his cowardice to the audiences when he swerves while his opponent stays. If both players swerve, then both of them are safe and receive nothing. However, the best outcome for each player is that he stays while the opponent swerves, so that he can gain reputation. The worst scenario is that both players stay and have a severe accident.

Note that when they play (Swerve, Swerve), the assessment for the action "Swerve" for both players does not deteriorate and they continue to play (Swerve, Swerve). Notice that they never end up at (Stay, Stay). If so, then one player's assessment of action "Stay" becomes lower than -1 at some point and the player stops playing the action.

In addition, when players play action profiles except (Swerve, Swerve), there exists a positive probability that the assessment of "Stay" becomes lower than -1. If so, the player stops playing "Stay" and they end up playing a Nash equilibrium or (Swerve, Swerve).

For this type of game, we have the following result:

Proposition 11. *In coordination games in category (2-2), players end up playing a Nash equilibrium or* (s_2^1, s_1^2) *almost surely.*

Proof. First of all, players cannot end up playing (s_1^1, s_2^2). If it does, one player's assessment of the action becomes lower than the minimum payoff of the other action and he stops playing the action. Notice also that (s_2^1, s_1^2) is absorbing and players end up playing (s_2^1, s_1^2) once players play it. In addition, for each pure Nash equilibrium, there exists an assessment for each player and each action such that players end up playing the

Nash equilibrium. Therefore, the other case to be considered is that players play Nash equilibria and (s_1^1, s_2^2) infinitely often without converging to either of them. However, this happens with probability zero. It is easy to show that (s_1^1, s_2^2) is played infinitely often in this case. The reason is that the player who is receiving the best payoff at a Nash equilibrium does not change his action while the other player receives the worst payoff from the action and changes his action. Thus players change to play from one Nash equilibrium to another action profile, (s_1^1, s_2^2). When (s_1^1, s_2^2) is played, with the positive probability that is bounded below, one player's assessment of the action becomes lower than the minimum payoff of the other action and he stops playing the action. In this case, players end up playing a Nash equilibrium or (s_2^1, s_1^2). Since (s_1^1, s_2^2) is played infinitely often, players end up playing a Nash equilibrium or (s_2^1, s_1^2) almost surely. □

Now consider a market entry game which has the following payoff matrix;

	Stay Out	Enter
Enter	100,0	-50,-50
Stay Out	0,0	0,100

where the action "Stay Out" always gives 0. Notice that this game satisfies the condition $b_{21} = b_{11}$ and $a_{21} = a_{22}$ and there does not exist any absorbing state. In this case, players end up playing (Enter, Stay Out), (Stay Out, Enter) or (Stay Out, Stay Out). For instance, once player 1's assessment of "Enter" becomes lower than 0, he does not play "Enter" any more. Then, players end up playing (Stay Out, Enter) if player 2's assessment of "Enter" is greater or equal to 0 and players end up playing (Stay Out, Stay Out) otherwise. Since (Enter, Enter) gives the worst payoff to both players, at some point, at least one player's assessment of "Enter" becomes lower than 0. Therefore, players end

up playing one of the action profiles, except for (Enter, Enter).

3.6 Non-Random Weighting Parameters

In this section, I assume that players' weighting parameters are not random variables. For example, players may believe that all past experiences equally represent the corresponding action's value, that is, players believe that the environments in which they are involved are stationary. Therefore, in each period, players put the same weight on all past experiences and players' assessments become arithmetic mean of past payoffs. Note that the weighting parameters for each player are as follows; $\lambda_n^i(s_j^i) = \frac{1}{\tau(n)+1}$ for all $i \in N$ and $s_j^i \in S^i$ where $\tau(n)$ is the number of times that the action s_j^i is played until period n.

I also consider the players who have the following weighting parameters; $\lambda_n^i(s_j^i) = \lambda$ for all i, s_j^i and n as in Sarin and Vahid (2001); all players have constant weighting parameters in all periods, that is, both players always put the same weight on the received payoff in each period. It is reasonable to assume this condition if players believe that the situation they are facing is non-stationary. If λ is close to 1, then players believe that only the most recent payoffs give information about the values of corresponding actions. If λ is close to 0, then players believe that initial assessments of actions mostly represent the actions' value.

In this section, I consider the battle of the sexes game, in which players may play off-diagonal action profiles alternately without ending up at a Nash equilibrium. In detail, I first consider the case where $\lambda_n^i(s_j^i) = \frac{1}{\tau(n)+1}$ for all i, s_j^i and n and off-diagonal payoffs for each player are all equivalent; $a_{12} = a_{21}, b_{12} = b_{21}$. In particular, I assume that $a_{12} = 0$ and $b_{12} = 0$.

As an example, consider the case where players' initial assessments are as follows: $Q_1^1(s_1^1) = 0.2$, $Q_1^1(s_2^1) = 0.2 + \epsilon, Q_1^2(s_1^2) = 0.2 + \epsilon, Q_1^2(s_2^2) = 0.2$, where $\epsilon \in (0, 0.2)$ is an irrational number. In this case, in the first period, they play (s_2^1, s_1^2) and both players receive payoff 0. In period 2, players' assessments are as follows: $Q_2^1(s_1^1) = 0.2, Q_2^1(s_2^1) = \frac{1}{2}(0.2 + \epsilon), Q_2^2(s_1^2) = \frac{1}{2}(0.2 + \epsilon), Q_2^2(s_2^2) = 0.2$. Notice that the assessments of s_1^1 and s_2^2 are greater than the assessments of s_2^1 and s_1^2. Hence, players play (s_1^1, s_2^2) and both players receive payoff 0. Using the payoff information in period 2, they update their assessments and they have the following assessments in period 3: $Q_3^1(s_1^1) = \frac{1}{2}(0.2)$, $Q_3^1(s_2^1) = \frac{1}{2}(0.2+\epsilon)$, $Q_3^2(s_1^2) = \frac{1}{2}(0.2 + \epsilon)$, $Q_3^2(s_2^2) = \frac{1}{2}(0.2)$. Then players play (s_2^1, s_1^2) in period 3. Notice that their assessments of action s_1^1 and s_2^2 never coincide with the assessments of action s_2^1 and s_1^2 at any period because of ϵ. After period 3, players play (s_2^1, s_1^2) until the corresponding assessments become lower than the assessments of (s_1^1, s_2^2). After the event, players again switch back to play (s_1^1, s_2^2), and so on.

When $\lambda_n^i(s_j^i) = \frac{1}{\tau(n)+1}$ for all i, s_j^i and n, the following statement shows the condition of initial assessments for coordination failures, which is the play on off-diagonal action profiles alternately. In this section, I assume that players' tie break rules satisfy the inertia condition.

Proposition 12. *In 2 × 2 coordination games with $a_{12} = a_{21} = b_{12} = b_{21} = 0$, under the inertia condition, if $\lambda_n^i(s_j^i) = \frac{1}{\tau(n)+1}$ for all i, s_j^i and n, then the necessary and sufficient condition for the coordination failure is as follows:*

$$\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} = \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)}$$

Proof. See Appendix. □

This result says that players will play non-Nash equilibria alternately forever if and only if players' ratios of initial assessments "coordinate".

Next, I consider the players who have the following weighting parameters; $\lambda_n^i(s_j^i) = \lambda$ for all i, s_j^i and n. Then the necessary and sufficient condition of initial assessments for the coordination failure is as follows:

Proposition 13. *In 2 × 2 coordination games with $a_{12} = a_{21} = b_{12} = b_{21} = 0$, under the inertia condition, if $\lambda_n^i(s_j^i) = \lambda$ for all i, s_j^i and n, then the necessary and sufficient condition for the coordination failure is as follows; for some $z \in \mathbb{Z}$,*

$$(1-\lambda)^{z-1} > \frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} \geq (1-\lambda)^z \text{ and } (1-\lambda)^{z-1} > \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)} \geq (1-\lambda)^z$$

or

$$(1-\lambda)^{z-1} \geq \frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} > (1-\lambda)^z \text{ and } (1-\lambda)^{z-1} \geq \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)} > (1-\lambda)^z$$

Proof. See Appendix. □

Since players play a Nash equilibrium forever if they coordinate once on the Nash equilibrium, for each case, the negation of the condition is the one for the success of coordination. For instance, if off-diagonal payoffs are all zero and players are frequentists, then they coordinate in some period and in all subsequent periods if and only if the initial assessments for both players and actions should satisfy the following condition: $\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} \neq \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)}$.

3.6.1 Coordinated Play on the Off-Diagonal Action Profiles

It is an interesting question whether the empirical frequency of play on the off-diagonal action profiles converges to the mixed Nash equilibrium. In fictitious play, Monderer and Shapley (1996) show that every 2×2 game with the diagonal property[1] has the fictitious play property; the empirical frequency of past play, which is a belief of players about an opponent player's behaviour, converges to a Nash equilibrium.

First note that 2×2 coordination games with $a_{21} = a_{12} = b_{12} = b_{21} = 0$ also have the diagonal property. In the case, under the condition of coordination failure, players forever play off-diagonal action profiles alternately. However, the frequency of the play need not converge to the mixed Nash equilibrium. I show this by an example. Consider the battle of the sexes game which has the following payoff matrix;

	s_1^2	s_2^2
s_1^1	1,2	0,0
s_2^1	0,0	2,1

I assume that weighting parameters and initial assessments for players are as follows: $\lambda_n^1(s_1^1) = \lambda_n^2(s_2^2) = \frac{1}{2}$, $\lambda_n^1(s_2^1) = \lambda_n^2(s_1^2) = \frac{1}{4}$, $Q_1^1(s_1^1) = Q_1^2(s_2^2) = \frac{1}{2}$, $Q_1^1(s_2^1) = Q_1^2(s_1^2) = \frac{1}{4}$. Under the inertia condition for both players, it is easy to see that players play action profiles in the following order; $(s_1^1, s_2^2) \to (s_1^1, s_2^2) \to (s_2^1, s_1^2) \to (s_1^1, s_2^2) \to (s_1^1, s_2^2) \to (s_2^1, s_1^2) \to ....$ In period 1, they play (s_1^1, s_2^2) and the assessments of s_1^1 and s_2^2 become $\frac{1}{4}$. Because of the inertia condition, they choose (s_1^1, s_2^2) again in period 2 and their assessments become $\frac{1}{8}$. Now players change to play (s_2^1, s_1^2) in period 3 and the assessments of

[1]The game has the diagonal property if $\alpha \neq 0$ and $\beta \neq 0$, where

$$\alpha = a_{11} + a_{22} - a_{12} - a_{21}$$

and

$$\beta = b_{11} + b_{22} - b_{12} - b_{21}.$$

s_2^1 and s_1^2 become $\frac{1}{16}$. In period 4, players return to play (s_1^1, s_2^2) and so on. Therefore, the empirical frequencies of play for both players converge to $((\frac{2}{3}, \frac{1}{3}), (\frac{1}{3}, \frac{2}{3}))$, while the mixed Nash equilibrium in this game is $((\frac{1}{2}, \frac{1}{2}), (\frac{1}{2}, \frac{1}{2}))$.

3.7 Discussion

This model can be also interpreted as a population model. Consider the situation in which there exist two large populations of naive players. In each period one player is picked from each population randomly and plays a 2×2 coordination game, but he can play the game only once[1]. After each player plays the game, he reports the payoff which he has received to each population. I assume that each population does not share information with the other population. Each population accumulates information as a public assessment, which consists of realized payoffs and the initial assessment. In each period, the public assessment of the action which is played is updated, using realized payoffs as defined above; the convex combination of the realized payoff and the public assessment in the previous period. Each player may not know whether he is playing a game, but he knows the public assessment. Using the public assessment, each player chooses an action which has the highest public assessment.

For example consider the battle of the sexes game. After the result of going to the opera or the football, both players report the realized payoff to the population which they belong to so that people in the population can make an assessment before they play the game themselves. The result above says that players from two different populations never coordinate when initial assessments satisfy the condition in Proposition 12 when they are

[1]Or each population is so large that the probability that a player plays a game again is almost 0.

frequentists. Otherwise, players coordinate to play one of the pure Nash equilibria.

3.8 Appendix

3.8.1 Proof of Theorem 3

It is a direct consequence from Lemma 8 that if players end up playing one action profile, then it should be a strict Nash equilibrium. Therefore, it should be shown that they actually end up playing a strict Nash equilibrium. The intuition of the proof is as follows. Since off-diagonal action profiles cannot be played infinitely often, there exists a period after which players only play Nash equilibria. Since I consider games with strict Nash equilibrium, players should change their actions at the same time when they move from one Nash equilibrium to another Nash equilibrium. Note also that weighting parameters are assumed to be independent, so that perfect correlated play on Nash equilibria is impossible. Now, the detailed proofs are given in the following arguments.

(i) At any non-Nash equilibrium action profile, there exists a positive probability such that one player who is receiving a worse payoff stops playing the action and plays another action. Note that at the non-Nash equilibrium action profile, $(s^i)_i$, the player who is suffering the worse payoff never plays his current action at least with the following probability:

$$\Pr(Q_n^i(s^i) \in (u^i(s^i, s^{-i}), \min_{t^{-i} \in S^{-i}} u^i(t^i, t^{-i})) \mid A),$$

where $A := \{Q_{n-1}^i(s^i) > Q_{n-1}^i(t^i)\ \forall i \in N, \forall t^i \neq s^i\}$ and this probability is bounded below

by the following probability:

$$\Pr(Q_n^i(s^i) \in (u^i(s^i, s^{-i}), \min_{t^{-i} \in S^{-i}} u^i(t^i, t^{-i}))) \mid A, B),$$

where $B = \{Q_{n-1}^i(s^i) = \max_{s^{-i}} u^i(s^i, s^{-i})\}$. Since the sets of players and actions are finite, if players play a non-Nash equilibrium action profile infinitely often, then the player who receives a worse payoff stops playing the action with probability one. Therefore, players do not play a non-Nash equilibrium action profile infinitely often. Hence, I assume that players only play some Nash equilibrium action profiles. The cases to be considered are that players play some Nash equilibria alternately without converging one of them. Since the game which I consider here has only strict Nash equilibria, all players should change their strategies at the same time when they change from one Nash equilibrium to another. Let $(s^{i*})_{i \in N}$ and $(s^{i**})_{i \in N}$ be two different strict Nash equilibrium action profiles which are played infinitely often. In this argument, I assume that players play only those two strict Nash equilibria alternately. The argument can be extended easily to the case where players play more than two Nash equilibria. Note that since players change one strict Nash equilibrium action profile to another strict Nash equilibrium action profile at the same time, all players should receive the payoffs which are strictly less than their current assessments. It should be true that $u^i(s^{i*}, s^{-i*}) = u^i(s^{i**}, s^{-i**})$ for all i and each player i's assessment never reaches the level $u^i(s^{i*}, s^{-i*})$ in a finite period. In the following argument, I show that players fail to play strict Nash equilibria alternately with probability one; to show that, I consider the periods in which players change from $(s^{i**})_{i \in N}$ to $(s^{i*})_{i \in N}$.

By the assumption on weighting parameters, we can ignore the case where $Q^i(s^{i*}) =$

$Q^i(s^{i**})$. Note also that we have the following results. Consider period n such that the condition $Q^i_{n-1}(s^{i*}) > Q^i_{n-1}(s^{i**}) > Q^i_n(s^{i*})$ holds. Then for any small $\varepsilon \in (0,1)$, there exist $0 < c^i, d^i < 1$ such that

$$\Pr(Q^i_n(s^{i*}) \in (\varepsilon u^i(s^{i**}, s^{-i**}) + (1-\varepsilon)Q^i_{n-1}(s^{i**}), Q^i_{n-1}(s^{i,**})) \mid C) \leq c^i$$

and

$$\Pr(Q^i_n(s^{i*}) \in (u^i(s^{i**}, s^{-i**}), (1-\varepsilon)u^i(s^{i**}, s^{-i**}) + \varepsilon Q^i_{n-1}(s^{i**})) \mid C) \leq d^i,$$

where $C := \{Q^i_{n-1}(s^{i*}) > Q^i_{n-1}(s^{i**}) > Q^i_n(s^{i*})\}$. Note that

$$\begin{aligned}
& \Pr(Q^i_n(s^{i*}) \in (\varepsilon u^i(s^{i**}, s^{-i**}) + (1-\varepsilon)Q^i_{n-1}(s^{i**}), Q^i_{n-1}(s^{i**})) \mid C) \\
= \; & \frac{\Pr(\varepsilon u^{i**} + (1-\varepsilon)Q^{i**}_{n-1} < \lambda u^{i**} + (1-\lambda)Q^{i*}_{n-1} < Q^{i**}_{n-1})}{\Pr(\lambda u^{i**} + (1-\lambda)Q^{i*}_{n-1} < Q^{i**}_{n-1})} \\
= \; & \frac{F(K) - F((1-\varepsilon)K)}{F(K)} \\
= \; & 1 - \frac{F((1-\varepsilon)K)}{F(K)}
\end{aligned}$$

and

$$\begin{aligned}
& \Pr(Q^i_n(s^{i*}) \in (u^i(s^{i**}, s^{-i**}), (1-\varepsilon)u^i(s^{i**}, s^{-i**}) + \varepsilon Q^i_{n-1}(s^{i**})) \mid C) \\
= \; & \frac{\Pr(u^{i**} < \lambda u^{i**} + (1-\lambda)Q^{i*}_{n-1} < (1-\varepsilon)u^{i**} + \varepsilon Q^{i**}_{n-1})}{\Pr(\lambda u^{i**} + (1-\lambda)Q^{i*}_{n-1} < Q^{i**}_{n-1})} \\
= \; & \frac{F(\varepsilon K)}{F(K)},
\end{aligned}$$

where $u^{i**} = u^i(s^{i**}, s^{-i**})$, $\lambda = \lambda^i_n(s^{i**})$, $Q^{i*}_{n-1} = Q^i_{n-1}(s^{i*})$, $Q^{i**}_{n-1} = Q^i_{n-1}(s^{i**})$, $F(x) =$

$\Pr((1-\lambda) \leq x)$ for $x \in (0,1)$ and $K = \frac{Q^{i**}_{n-1} - u^{i**}}{Q^{i*}_{n-1} - u^{i**}}$. Notice that for any $K \in (0,1]$, $\frac{F(cK)}{F(K)} \in (0,1)$ and $\lim_{K\to 0} \frac{F(cK)}{F(K)} = \lim_{K\to 0} \frac{cf(cK)}{f(K)} = c$ where $c \in \{\varepsilon, 1-\varepsilon\}$, f is the density function for the weighting parameter and $f(0) < \infty$.

Therefore for player i, with probability one, there exist infinitely many periods n such that

$$Q^i_n(s^{i*}) < \varepsilon u^i(s^{i**}, s^{-i**}) + (1-\varepsilon) Q^i_{n-1}(s^{i**}).$$

and

$$Q^i_n(s^{i*}) > (1-\varepsilon) u^i(s^{i**}, s^{-i**}) + \varepsilon Q^i_{n-1}(s^{i**}).$$

I focus on the cases where both conditions hold when player i changes his action from s^{i*} to s^{i**}.

Now I consider period n in which players are playing $(s^{i**})_{i\in N}$. For the case

$$Q^j_{n+1}(s^{j*}) \in [\varepsilon u^j(s^{j**}, s^{-j**}) + (1-\varepsilon) Q^j_n(s^{j**}), Q^j_n(s^{j**}))$$

for $j \neq i$, we have

$$\Pr(Q^i_{n+1}(s^{i**}) \geq Q^i_n(s^{i*})) \times \Pr(Q^j_{n+1}(s^{j**}) < Q^j_n(s^{j*})) \geq e_{1,ij},$$

and for the case

$$Q^j_n(s^{j*}) < \varepsilon u^j(s^{j**}, s^{-j**}) + (1-\varepsilon) Q^j_n(s^{j**})$$

for $j \neq i$, we have

$$\Pr(Q^i_{n+1}(s^{i**}) < Q^i_n(s^{i*})) \times \Pr(Q^j_{n+1}(s^{j**}) \geq Q^j_n(s^{j*})) \geq e_{2,ij}$$

for some $e_{1ij} > 0$ and $e_{2,ij} > 0$. In any cases, the probability that players fail to play the same strict Nash equilibrium in period $n+1$ has positive probability which has the lower bound $\min_{h\in\{1,2\}} \min_{ij,i\neq j}\{e_{h,ij}\} > 0$. Since players change from $(s^{i**})_{i\in N}$ to $(s^{i*})_{i\in N}$ infinitely many times, players fail to play strict Nash equilibrium with probability one, which contradicts the hypothesis. Therefore, the only possibility is that players play only one Nash equilibrium after some period.

(ii) Note that if the condition in (ii) satisfies, then the payoff from any Nash equilibrium should be greater than the payoff from non-Nash equilibrium; $u^i(s^{i*}, s^{-i*}) > u^i(s^i, s^{-i})$ for all $i \in N$, $(s^{i*}) \in E^*$, and $(s^i) \notin E^*$. Therefore, each pure Nash equilibrium is absorbing and players who play a Nash equilibrium once play it forever. By the same logic as the proof in (i), players cannot play only non-Nash equilibrium action profiles forever. That is, with probability one, players play a Nash equilibrium at some time and then play it in all subsequent periods. □

3.8.2 Proof of Proposition 12

I assume that each player's initial assessments of both actions are different. Then the condition of coordination failure under the inertia condition for each player's tie break rule is as follows; for $j \neq k$ and (1) for the initial assessment, $Q_1^i(s_j^i) > Q_1^i(s_k^i)$ and $Q_1^{-i}(s_k^{-i}) > Q_1^{-i}(s_j^{-i})$ and (2) for any n, $Q_n^i(s_j^i) \geq Q_n^i(s_k^i)$ and $Q_n^{-i}(s_k^{-i}) \geq Q_n^{-i}(s_j^{-i})$, where if one of the inequalities holds, then (i) $Q_{n-1}^i(s_j^i) > Q_{n-1}^i(s_k^i)$ and $Q_{n-1}^{-i}(s_k^{-i}) > Q_{n-1}^{-i}(s_j^{-i})$ and (ii) $Q_{n+1}^i(s_j^i) < Q_{n+1}^i(s_k^i)$ and $Q_{n+1}^{-i}(s_k^{-i}) < Q_{n+1}^{-i}(s_j^{-i})$. Let $\hat{Q}_t^i(s_j^i)$ be the assessment of action s_j^i when only (s_j^i, s_k^{-i}) is played t times where $j \neq k$. Then it can be easily verified that the condition for coordination failure is equivalent to the following condition;

for any $u, t \in \mathbb{N}$, $\hat{Q}^i_u(s^i_j) \geq \hat{Q}^i_t(s^i_k)$ and $\hat{Q}^{-i}_u(s^{-i}_k) \geq \hat{Q}^{-i}_t(s^{-i}_j)$ where if one of inequalities holds, then $\hat{Q}^i_{u-1}(s^i_j) > \hat{Q}^i_t(s^i_k)$ and $\hat{Q}^{-i}_{u-1}(s^{-i}_k) > \hat{Q}^{-i}_t(s^{-i}_j)$ and $\hat{Q}^i_{u+1}(s^i_j) < \hat{Q}^i_t(s^i_k)$ and $\hat{Q}^{-i}_{u+1}(s^{-i}_k) < \hat{Q}^{-i}_t(s^{-i}_j)$ for $j \neq k$. Therefore, in the following proofs, I use the latter condition.

The important factor of this argument is that the players change actions at the same time and they 'coordinate' at coordination failure. If the players coordinate on diagonal action profiles once, then they succeed in coordinating. Therefore if the following conditions are satisfied, players never coordinate; for any m and $n \in \mathbb{N}$,

$$\frac{1}{n}Q^1_1(s^1_j) \geq \frac{1}{m}Q^1_1(s^1_k) \textit{ and } \frac{1}{n}Q^2_1(s^2_k) \geq \frac{1}{m}Q^2_1(s^2_j)$$

holds, where equalities among them do not hold consecutively; if one of the equalities holds at m, n then both inequalities hold strictly at m, $n-1$ and m, $n+1$. In the following argument, I show that this condition is equivalent to the following condition:

$$\frac{Q^1_1(s^1_2)}{Q^1_1(s^1_1)} = \frac{Q^2_1(s^2_1)}{Q^2_1(s^2_2)}$$

To make this clear, I assume first that if one of inequalities holds with equality, then both inequalities should hold with equality. Then the original condition above can be expressed as follows:

$$Q^1_1(s^1_2) < \frac{m}{n}Q^1_1(s^1_1) \text{ and } Q^2_1(s^2_2) > \frac{n}{m}Q^2_1(s^2_1)$$

or

$$Q_1^1(s_2^1) > \frac{m}{n} Q_1^1(s_1^1) \text{ and } Q_1^2(s_2^2) < \frac{n}{m} Q_1^2(s_1^2)$$

or

$$Q_1^1(s_2^1) = \frac{m}{n} Q_1^1(s_1^1) \text{ and } Q_1^2(s_2^2) = \frac{n}{m} Q_1^2(s_1^2).$$

Note that if $\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)}$ is a rational number, then there exist m and n such that $\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} = \frac{m}{n}$. By the last condition, we should have $\frac{Q_1^2(s_2^2)}{Q_1^2(s_1^2)} = \frac{n}{m}$, that is, $\frac{Q_1^2(s_2^2)}{Q_1^2(s_1^2)}$ should be a rational number too. If $\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)}$ is an irrational number, then $\frac{Q_1^2(s_2^2)}{Q_1^2(s_1^2)}$ should be also an irrational number. If $\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} \neq \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)}$, say if $\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} > \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)}$, then there exists a rational number $\frac{m}{n}$ such that $\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} > \frac{m}{n} > \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)}$. This means that $Q_1^1(s_2^1) > \frac{m}{n} Q_1^1(s_1^1)$ and $Q_1^2(s_2^2) > \frac{n}{m} Q_1^2(s_1^2)$ and it contradicts the conditions above. Hence the following relation $\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} = \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)}$ is the only case which satisfies the condition above.

Now consider the other cases. There exist m and $n \in \mathbb{N}$ such that $\frac{1}{n} Q_1^i(s_1^i) = \frac{1}{m} Q_1^i(s_2^i)$ and $\frac{1}{n} Q_1^j(s_2^j) \neq \frac{1}{m} Q_1^j(s_1^j)$, say $\frac{1}{n} Q_1^j(s_2^j) > \frac{1}{m} Q_1^j(s_1^j)$,[1]. Then

$$\frac{1}{n-1} Q_1^i(s_1^i) > \frac{1}{m} Q_1^i(s_2^i) \text{ and } \frac{1}{n-1} Q_1^j(s_2^j) > \frac{1}{m} Q_1^j(s_1^j)$$

and

$$\frac{1}{n+1} Q_1^i(s_1^i) < \frac{1}{m} Q_1^i(s_2^i) \text{ and } \frac{1}{n+1} Q_1^j(s_2^j) < \frac{1}{m} Q_1^j(s_1^j)$$

[1] If $\frac{1}{n} Q_1^j(s_2^j) < \frac{1}{m} Q_1^j(s_1^j)$, then

$$\frac{1}{n} Q_1^i(s_1^i) < \frac{1}{m-1} Q_1^i(s_2^i) \text{ and } \frac{1}{n} Q_1^j(s_2^j) < \frac{1}{m-1} Q_1^j(s_1^j)$$

should hold. Notice that $\frac{Q_1^i(s_1^i)}{Q_1^i(s_2^i)}$ should be a rational number $\frac{n}{m}$. Moreover, by the conditions above, we have $\frac{n+1}{m} > \frac{Q_1^j(s_2^j)}{Q_1^j(s_1^j)} > \frac{n}{m}$. It is easy to see that at $2n$ and $2m$, the following conditions also satisfy; $\frac{1}{2n}Q_1^i(s_1^i) = \frac{1}{2m}Q_1^i(s_2^i)$ and $\frac{1}{2n}Q_1^j(s_2^j) > \frac{1}{2m}Q_1^j(s_1^j)$. Thus we have the following condition: $\frac{2n+1}{2m} > \frac{Q_1^j(s_2^j)}{Q_1^j(s_1^j)} > \frac{2n}{2m}$. Using the same logic, the condition should be satisfied for any kn and km where $k \in \mathbb{N}$. If $k \to \infty$, then the condition becomes as follows; $\frac{n}{m} \geq \frac{Q_1^j(s_2^j)}{Q_1^j(s_1^j)} > \frac{n}{m}$. However, there do not exist initial assessments which satisfy this condition[1]. Therefore the necessary and sufficient condition for initial assessments for the coordination failure in this case is equivalent to the following condition:

$$\frac{Q_1^1(s_2^1)}{Q_1^1(s_1^1)} = \frac{Q_1^2(s_1^2)}{Q_1^2(s_2^2)}.$$

□

3.8.3 Proof of Proposition 13

It can be shown that the following condition is equivalent to the condition for the coordination failure in the coordination game; for any t, there exists u such that

$$\hat{Q}_t^i(s_1^i) \in (\hat{Q}_{u+1}^i(s_2^i), \hat{Q}_u^i(s_2^i)] \text{ and } \hat{Q}_t^{-i}(s_2^{-i}) \in (\hat{Q}_{u+1}^{-i}(s_1^{-i}), \hat{Q}_u^{-i}(s_1^{-i})]$$

or

$$\hat{Q}_t^i(s_1^i) \in [\hat{Q}_{u+1}^i(s_2^i), \hat{Q}_u^i(s_2^i)) \text{ and } \hat{Q}_t^i(s_2^{-i}) \in [\hat{Q}_{u+1}^{-i}(s_1^{-i}), \hat{Q}_u^{-i}(s_1^{-i}))$$

for all i[2]. Since $\hat{Q}_t^i(s_j^i) = (1-\lambda)^t \hat{Q}_0^i(s_j^i)$, the condition in Proposition 13 can be easily

[1] If $\frac{1}{n}Q_1^j(s_2^j) < \frac{1}{m}Q_1^j(s_1^j)$ then it satisfies that $\frac{n+1}{m} < \frac{Q_1^j(s_2^j)}{Q_1^j(s_1^j)} < \frac{n}{m}$. By the same argument, There exist no initial assessments which satisfy the conditions.

[2] For example, if $\hat{Q}_m^i(s_1^i) > \hat{Q}_0^i(s_2^i)$, then I assume that $\hat{Q}_{-1}^i(s_2^i)$ is the maximum payoff which both

derived. □

actions give so that $\hat{Q}^i_m(s^i_1) \in (\hat{Q}^i_0(s^i_2), \hat{Q}^i_{-1}(s^i_2)]$. In addition, let $\hat{Q}^i_\infty(s^i_2)$ be the minimum payoff which both actions give and $\hat{Q}^i_{\infty+1}(s^i_2)$ be the minimum payoff of those which both actions give. Then if $\hat{Q}^i_m(s^i_1) \leq \hat{Q}^i_\infty(s^i_2)$, $\hat{Q}^i_m(s^i_1) \in (\hat{Q}^i_{\infty+1}(s^i_2), \hat{Q}^i_\infty(s^i_2)]$.

CHAPTER 4

ADAPTIVE LEARNING MODELS IN FINITELY REPEATED GAMES

4.1 Introduction

Many theoretical researchers have analyzed adaptive learning models in normal form games. In the literature, they investigate the behaviour of adaptive players who learn the opponents' behaviour or the values of their own actions over repeated plays of the game. One of their main interests is whether their behaviour in the long run corresponds to Nash equilibrium or perturbed Nash equilibrium, which is Nash equilibrium under payoff perturbations. Meanwhile, adaptive learning in extensive form games without payoff perturbations is also of interest among theoretical researchers, who focus on equilibrium concepts such as sequential equilibrium, subgame perfect equilibrium and self-confirming equilibrium, which is introduced by Fudenberg and Kreps (1995)[1]. While an equilibrium concept in extensive form games with payoff perturbations, agent quantal response equilibrium, is introduced by McKelvey and Palfrey (1998), to my best knowledge, there is no

[1]The self-confirming equilibrium may not be Nash equilibrium, since it does not require the correct belief about behavioural strategies of other players at relevant information sets at off the equilibrium path. In detail, see Fudenberg and Kreps (1995)

literature which investigates adaptive learning that leads to the equilibrium. Therefore, it is intriguing to investigate which type of adaptive learning leads to the equilibrium.

It is not only among theoretical researchers' interests but also experimental researchers to investigate learning in extensive form games. For example, the centipede game (McKelvey and Palfrey, 1992, Palacious-Huerta and Volij, 2009) and the finitely repeated prisoner's dilemma (Selten and Stoecker, 1986) are such examples. In the experimental literature, we observe that players learn from their past plays and adjust their behaviour. In addition, the experiments show deviations from the equilibrium predictions. It is also of interest to investigate which type of adaptive learning leads to such consequences.

In this chapter, I investigate the learning process of adaptive players who face a fixed extensive form game, in particular a finitely repeated game in each of infinitely many periods. In particular, I consider the case in which the players have limited information about their decision-making environment; they know their available actions in each period and observe realized payoffs but they may not know their own and opponents' payoff functions. Therefore, we need a model which does not require players to have knowledge about the payoff functions; I consider the case in which each player assigns his subjective assessments on his actions based on his past experience and picks the action which he thinks is the best. Using realized payoff information, each player updates the assessments of chosen actions adaptively; I consider players who follow the Q-learning updating rule and Sarin and Vahid (1999) updating rule.

I first assume that players experience random shocks on their assessments. If each stage game of the finitely repeated game consists of an extensive form game with perfect information, then I show that their behavioural strategies converge to the agent quantal

response equilibrium (AQRE hereafter,) introduced by McKelvey and Palfrey (1998). When a normal form game is played in each stage game, I provide an additional condition which guarantees convergence to the unique AQRE of the supergame. Next, I assume that players do not experience random shocks on their assessments. Then I show that (1) when they face the finitely repeated prisoner's dilemma, both players may end up cooperating at each stage game; and (2) when they play some finitely repeated coordination games, both players end up coordinating in each stage game.

The adaptive learning models considered here are introduced by Watkins and Dayan (1992) and Sarin and Vahid (1999). The models are developed to analyze decision problems, some authors have applied the models to investigate normal form games (Sarin, 1999, Leslie and Collins, 2005 and Cominetti et al., 2010) and extensive form games of perfect information with a unique subgame perfect equilibrium (Jehiel and Samet, 2005). In particular, Jehiel and Samet (2005) show that players' strategies, where they follow a specific Sarin and Vahid (1999) updating rule, approach the unique subgame perfect equilibrium. Note that in this chapter, I show a similar result, but the underlying games are allowed to have multiple subgame perfect equilibria. Another learning model in games, reinforcement learning model, is introduced by Erev and Roth (1998) and when an extensive form game of perfect information with a unique subgame perfect equilibrium is played by the players, convergence to the unique subgame perfect equilibrium is shown by Laslier and Walliser (2005). Lastly, in a learning model which requires players to have knowledge about the structures of the game, such as fictitious play model, convergence of beliefs to a unique sequential equilibrium is investigated by Hendon et al. (1993) and Groes et al. (1999).

The structure of this chapter is as follows. In Section 4.2, the notation for general extensive form games is introduced. In Section 4.3 the equilibrium concept under payoff perturbation, agent quantal response equilibrium, and its properties are provided. In Section 4.4, the learning rule and the decision rule of adaptive players are introduced. The results of the learning process are also shown in the chapter. In Section 4.5, the case without noise is analyzed, and Section 4.7 concludes. All proofs are placed in the Appendix.

4.2 Extensive Form Games

I first consider an extensive form game[1] Γ, which consists of the set of players N, histories H, player function P, and information sets $\mathcal{I}$. The set of players consists of M players; $N = \{1, 2, ..., M\}$. A history $h \in H$ is a sequence of actions taken by players; $h = (a_1, ..., a_K)$, $a_k \in \mathcal{A}$ for $1 \leq k \leq K$, is a history with the length of K, where $\mathcal{A}$ is the set of actions for all players,. The set of histories H includes the empty set $\emptyset =: h_0$, which corresponds to the initial node. A history $h = (a_1, ..., a_K)$ is terminal if there does not exist $a_{K+1} \in \mathcal{A}$ such that $(a_1, ..., a_{K+1}) \in H$. Given a history h, the partial history of length J is denoted by $h_J = (a_1, ..., a_J)$ where $J \leq K$. The set of actions which are available after a non-terminal history h is denoted by A_h. Thus $\mathcal{A} = \cup_{h \in H} A_h$. Let Z be the set of terminal histories. The player function P assigns a member in N to each non-terminal history; $P(h)$, $h \in H \diagdown Z$, is the player who chooses an action after the history h. Let $\mathcal{I}^i$ be a partition of the set $\{h : P(h) = i\}$ and I^i be a member of $\mathcal{I}^i$ with the property that $A_h = A_{h'} =: A_{I^i}$ for $h, h' \in I^i$. Thus I^i is an information set of player i and $\mathcal{I}^i$ is the

[1] I follow the notation of Osborne and Rubinstein (1994)

set of player i's information sets. Let $\mathcal{I} = \cup_{i \in N} \mathcal{I}^i$ denote the set of information sets for all players, which is a partition of non-terminal histories. Note that if any $I \in \mathcal{I}$ consists of a single history, then the extensive form game is the one with perfect information. Let $\pi : Z \to \mathbb{R}^M$ be a payoff function which assigns payoffs of all players to each terminal history and π^i be a payoff function for player i.

In this chapter, I restrict our attention to extensive form games with perfect recall. Let I_h be the information set which contains h. Let $\mathcal{I}_h^i$ denote the set of player i' information sets which are reached by $h = (a_1, ..., a_k)$: $\mathcal{I}_h^i = \{I_{h'} : h' \text{ is a partial history of } h \text{ and } P(h') = i\}$. Let $l_h^i : \{1, 2, ..., |\mathcal{I}_h^i|\} \to \mathcal{I}_h^i$ be a function which orders the information sets in $\mathcal{I}_h^i$ in the way in which the information sets are reached. Let $a_{l_h^i} : \{1, 2, ..., |\mathcal{I}_h^i|\} \to \mathcal{A}$ be the function such that player i's actions taken at information sets in $\mathcal{I}_h^i$ are ordered as they occur[1]. Then the extensive form game has perfect recall if for each $i \in N$, $\mathcal{I}_h^i = \mathcal{I}_{h'}^i$, $l_h^i = l_{h'}^i$ and $a_{l_h^i} = a_{l_{h'}^i}$ if $h \in I^i$ and $h' \in I^i$ for some $I^i \in \mathcal{I}^i$.

Letting $\triangle(A) = \{x \in \mathbb{R}^{|A|} : x_i \geq 0 \ \forall i \text{ and } \sum_{i \in A} x_i = 1\}$, a behavioural strategy of player i is a function β^i satisfying $\beta^i(I^i) \in \triangle(A_{I^i})$ for all $I^i \in \mathcal{I}^i$. Thus $\beta^i(I^i)$ assigns probabilities over available actions at the information set I^i and $\beta^i(I^i)(a)$ is the probability that player i assigns to action $a \in A_{I^i}$ at his information set I^i.

4.2.1 Finitely Repeated Games

I now consider a specific case of an extensive form game, a T times repeated game, in which a fixed game is played at each round $t \in \{1, 2, ..., T\}$, where the game is allowed to be an extensive form game with perfect information or a normal form game. Let

[1] If $I_h \in \mathcal{I}^i$, then $a_{l_h^i(|\mathcal{I}_h^i|)} := \varnothing$

h_t, $t \in T \cup \{0\}$, be a history of actions chosen by players until round t : $h_0 = \varnothing$ and $h_t = (a_{1,1}, ..., a_{1,K^1}, a_{2,1}, ..., a_{2,K^2}, a_{3,1}, ..., a_{t,K^t})$, where $a_{s,k}$ is k - th action taken at round s and a_{s,K^s} is the action taken at the end of s-th round. Let H_t be the set of histories until round t. Let I_t^i represent an information set of player i at round t. Let $h_{t,T} \in H$ be the partial history of $h_T = (a_{1,1}, ..., a_{T,K^T})$ until round t; $h_{t,T} = (a_{1,1}, ..., a_{t,K^t})$. Let $h_{t\backslash s-1,T}$ be a partial history of h_T from period s to period t; $h_{t\backslash s-1,T} = (a_{s,1}, a_{s,2}, ..., a_{t,K^t})$.

Let π^i_{s,h_T}, $s \leq T$, be the realized payoff of player i at round s given a terminal history $h_T \in Z$. Let $\pi^i(h_T) = \pi^i_{h_T} = \sum_{s=1}^{T} \pi^i_{s,h_T}$ be the total payoff of player i given the history h_T[1], while $\pi^i_{s\geq t,h_T} = \sum_{s=t}^{T} \pi^i_{s,h_T}$ be the partial payoff of player i from round t to round T.

4.3 Agent Quantal Response Equilibrium

The equilibrium concept for an extensive form game with payoff perturbations, the agent quantal response equilibrium, is introduced by McKelvey and Palfrey (1998). To state the concept here formally, I first consider the agent normal form $\Gamma' = (N', H, P, \mathcal{I}, \pi)$. $N' = N \times \mathcal{I}$ is the set of agents each of whom is assigned to a single information set. $j = (i, I^i) \in N'$ is the agent who serves for player i at information set I^i and $j_t = (i, I_t^i)$ is the agent who serves for player i at information set I_t^i at round t. After a terminal history h_T is realized, agent $j = (i, I^i)$, $I^i \in \mathcal{I}^i_{h_T}$, receives payoff $\pi^j_{h_T} = \sum_{s\leq T} \pi^i_{s,h_T}$. Let π^j_β be the expected payoff of agent j given behavioural strategies $\beta = (\beta^j)_j$, where for $j = (i, I^i)$, $\beta^j = \beta^i(I^i)$. Let Z_{I^i} be the set of terminal histories which pass the information set I^i: $Z_{I^i} = \{h \in Z$: there exists a partial history h' of h s.t. $h' \in I^i\}$.

[1]It can be also defined as follows; $\pi^i_{h_T} = \sum_{s=1}^{T} (\delta^i)^{s-1} \pi^i_{s,h_T}$, where δ^i is player i's discount factor for future payoffs. Now let $\tilde{\pi}^i_{s,h_T} = (\delta^i)^{s-1} \pi^i_{s,h_T}$. Then the total payoff that player i receives given history h_T is also defined by $\pi^i_{h_T} = \sum_{s\leq T} \tilde{\pi}^i_{s,h_T}$ and I use this expression in this chapter.

Let $\beta(h_T)$ represent the probability that h_T occurs under agents' behavioural strategies β and $\beta_I(h_T)$ be the conditional probability of h_T given that information set I is reached : $\beta_I(h_T) = \frac{\beta(h_T)}{\sum_{h \in Z_I} \beta(h)}$ if $\beta(Z_I) := \sum_{h \in Z_I} \beta(h) > 0$. Then $\pi^j_\beta = \sum_{h_T \in Z_{I^i}} \beta_{I^i}(h_T) \pi^j_{h_T} = \sum_{h_T \in Z_{I^i}} \beta_{I^i}(h_T) \sum_{s \leq T} \pi^i_{s,h_T}$. Let $\pi^j_{a,\beta^{-j}}$ be the expected payoff of agent j when he chooses action a and the others follow behavioural strategy β^{-j}.

Now each agent chooses an action which has the highest expected payoff given some payoff perturbations. Let C^j_a be a probability that agent j chooses $a \in A_{I^i}$. Then agent j's choice probability of action a given behavioural strategy β is as follows;

$$C^j_a(\pi^j_\beta) = \Pr\big(\arg\max_{b \in A_{I^i}} (\pi^j_{b,\beta^{-j}} + \epsilon^j_b) = a \big),$$

where the random vector $\epsilon^j = (\epsilon^j_a)_{a \in A_{I^i}}$ represents agent j's payoff perturbations and the following conditions are assumed: (i) ϵ^j takes a value in $\mathbb{R}^{|A_{I^i}|}$, (ii) the distribution of the stochastic perturbations has a density which is strictly positive on its domain, (iii) $(\epsilon^j)_j$ is independent, and (iv) the expected value of ϵ^j_a exists for each j and $a \in A_{I^i}$. Now I provide the equilibrium concept for extensive form games, which is introduced by McKelvey and Palfrey (1998);

Definition. The behavioural strategy profile $\beta^* = (\beta^{1*}, ..., \beta^{M*})$ in an extensive form game Γ is an agent quantal response equilibrium (AQRE) if it is a normal form quantal response equilibrium of the agent normal form $\Gamma' = (N', H, P, \mathcal{I}, \pi) : \beta^{i*}(I^i)(a) = \beta^{j*}(a) = C^j_a(\pi^j_{\beta^*})$ for $a \in A_{I^i}$ and $j = (i, I^i) \in N'$.

McKelvey and Palfrey (1998) show the existence of an AQRE;

Proposition (McKelvey and Palfrey 1998). *For any Γ, an AQRE exists*

One well-known choice probability form, which is derived by i.i.d. perturbations with the extreme value distribution $F(\epsilon_a^j) = \exp(-\exp(-\frac{1}{\tau}\epsilon_a^j))$, is the logit choice rule;

$$C_a^j(\pi_\beta^j) = \frac{\exp(\frac{1}{\tau}\pi_{a,\beta^{-j}}^j)}{\sum_{b\in A_{I^i}} \exp(\frac{1}{\tau}\pi_{b,\beta^{-j}}^j)},$$

where τ is called noise term[1]. If τ goes to infinity, then the choice probability becomes the uniform distribution, while if τ approaches 0, then the choice probability approaches the degenerate probability where the probability of the action which has the highest expected payoff is 1.

McKelvey and Palfrey (1998) have called the AQRE under the stochastic disturbance with the extreme value distribution logit-AQRE. They show that when the noise term τ goes to 0, then the logit-AQRE converges to a sequential equilibrium strategy profile.

Proposition (McKelvey and Palfrey 1998). *For every finite extensive form game, every limit point of a sequence of logit-AQRE with τ going to zero corresponds to the strategy of a sequential equilibrium assessment of the game.*

4.4 Assessment and Decision Rule

I now consider the decision rule of adaptive players in an extensive form game. I assume that the players assign assessments on their own actions available after each non-terminal history. Let $Q_{n,h}^i : A_h \to \mathbb{R}$ be player i's subjective assessment function in period n, where the function assigns a subjective assessment to each action available after history h; $Q_{n,h}^i(a)$, $a \in A_h$, is the assessment of player i's action a after the history h in period n.

[1]See Hofbauer and Sandholm (2002)

I assume that for any $I^i \in \mathcal{I}^i$ and $h, h' \in I^i, Q^i_{n,h}(a) = Q^i_{n,h'}(a) =: Q^i_{n,I^i}(a)$ for any period $n \in \mathbb{N}$ and $a \in A_{I^i}$.

Before a player makes a decision, the assessment of each action is affected by stochastic perturbation, which is interpreted as temporary emotional noise on the assessment; the random vector $\eta^i_{I^i} = (\eta^i_{I^i,a})_{a \in A_{I^i}}$ takes a value in $\mathbb{R}^{|A_{I^i}|}$ and the distribution of $\eta^i_{I^i}$ does not depend on the history, payoffs, or assessments of players. This emotional noise captures the idea that humans may not always pick the best-performed action; it also captures the probabilistic choice behaviour of humans.

Each player chooses the action which has the highest subjective assessment affected by the noise; the probability with which player i chooses action $a \in A_{I^i}$ given his assessments and noise is as follows: for $h \in I^i$,

$$C^i_{h,a}(Q^i_n) = C^i_{I^i,a}(Q^i_n) = \Pr\big(\arg\max_{b \in A_{I^i}}(Q^i_{n,I^i}(b) + \eta^i_{I^i,b}) = a\big),$$

where $C^i_{I^i,a} : \mathbb{R}^{|A_h|} \to [0,1]$ is the probability of choosing action $a \in A_{I^i}$. I assume that for $i \in N$ and I^i, the distribution of the stochastic noise has a density which is strictly positive on its domain and thus $C^i_{I^i,a}$ becomes a continuous function almost surely[1].

If I assume i.i.d. emotional noise with the extreme value distribution then again we have the logit choice rule;

$$C^i_{I^i,a}(Q^i_n) = \frac{\exp(\frac{1}{\tau}Q^i_{n,I^i}(a))}{\sum_{b \in A_{I^i}} \exp(\frac{1}{\tau}Q^i_{n,I^i}(b))}.$$

[1]For example, consider the case where $\eta^i = 0$ with probability one for all $i \in \{1,2\}$. Then if $Q^1 = Q^2$, then the choice probability that the agent chooses i depends on his tie break rule and C^i may become a correspondence.

4.4.1 Extensive Form Games with Perfect Information

In this subsection, I investigate the case where each stage game in a finitely repeated game consists of an extensive form game with perfect information. Therefore, each information set is a singleton. I first introduce the updating rule of assessments for this case; I consider the updating rule which is introduced by Sarin and Vahid (1999) (SV, hereafter). If a terminal history h_T is realized, then the assessment of action a is updated as follows; for a history h and $a \in A_h$

$$Q^i_{n+1,h}(a) = Q^i_{n,h}(a) + \lambda_{n+1}\mathbf{1}_{h,a}(\pi^i_{h_T} - Q^i_{n,h}(a))$$

where (1) $\mathbf{1}_{h,a}$ is an indicator function such that $\mathbf{1}_{h,a} = 1$ if history h is realized and a is chosen after history h and 0 otherwise, (2) $\{\lambda_n\}_{n\in N}$ is a sequence of weighting parameters, which is a deterministic sequence and satisfies the following conditions[1];

$$\sum_{n\geq 1}\lambda_n = \infty,\ \sum_{n\geq 1}(\lambda_n)^2 < \infty.$$

Proposition 14. *If in infinitely many periods, players play a finitely repeated game in which each stage game consists of an extensive form game with perfect information and players follow the SV updating rule, then their behavioural strategy profiles converge to the unique AQRE of the game almost surely.*

Proof. See Appendix. □

It is obvious from the proof of Proposition 14 that the behavioural strategies of adaptive players converge to the AQRE when the stage game is repeated only once;

[1]See Chapter 2 for the idea of the conditions

Corollary 5. *When adaptive players with the SV updating rule play an extensive form game with perfect information in infinitely repeated periods, then their behavioural strategies converge to the unique AQRE of the game with probability one.*

4.4.2 Normal Form Games

In this subsection, I consider the case where each stage game of a finitely repeated game consists of a normal form game. Thus the stage game consists of (1) the set of players N, (2) the set of actions A^i available to player i for each $i \in N$, and (3) the payoff function of player i, $\pi^i : \prod_i A^i \to \mathbb{R}$, for each $i \in N$.

The updating rule adopted in this subsection is akin to the one in Q-learning model, especially for the situation where a normal form game is repeated T times. At the end of each period, the assessment of the chosen action at each round is adjusted toward the sum of a payoff received at the round and the highest estimated payoff that the player can receive after the round. In detail, when a terminal history $h_T \in Z$ is realized, the assessment of action a_t at information set I^i, $Q^i_{n+1,I^i_t}(a_t)$, is updated in the following manner; for $t \leq T-1$,

$$\begin{aligned} Q^i_{n+1,I^i_t}(a_t) &= Q^i_{n,I^i_t}(a_t) \\ &+ \lambda_{n+1}\mathbf{1}_{I^i_t,a_t}\left(\pi^i_{t,h_T} + \max_{a_{t+1}\in A^i} Q^i_{n,I^i_{t+1,h_T}}(a_{t+1}) - Q^i_{n,I^i_t}(a_t)\right), \end{aligned}$$

where (1) $\mathbf{1}_{I^i_t,a_t}$ is an indicator function such that $\mathbf{1}_{I^i_t,a_t} = 1$ if I^i_t is reached and a_t is chosen and 0 otherwise; and (2) I^i_{t+1,h_T} is the information set at round $t+1$ which is reached by h_T. For the assessment of action a_T at information set I^i_T, the updating rule

is as follows;

$$Q^i_{n+1,I^i_T}(a_T) = Q^i_{n,I^i_T}(a_T) + \lambda_{n+1}\mathbf{1}_{I^i_T,a_T}(\pi^i_{T,h_T} - Q^i_{n,I^i_T}(a_T)).$$

The sequence of weighting parameters, $\{\lambda_n\}_n$, is again assumed to be a deterministic sequence satisfying the following conditions;

$$\sum_{n\geq 1}\lambda_n = \infty,\ \sum_{n\geq 1}(\lambda_n)^2 < \infty.$$

To show convergence to an AQRE in this case, I need to introduce some notation. Consider the case where a normal form game is played once in each of infinitely many periods. Let $\pi_C = (\pi^i_C)_i : \mathbb{R}^{\sum_i |A^i|} \to \mathbb{R}^{\sum_i |A^i|}$ be a function which gives players' expected payoffs given players' assessments: $\pi^i_C(Q) = (\pi^i_{C(Q)}(a))_{a\in A^i}$ is player i's expected payoffs for his actions, where the expected payoffs are obtained by players' choice probabilities C and assessments $Q = (Q^i_a)_{i,a}$. It is shown by Cominetti et al. (2010) that players' choice probabilities converge to the unique quantal response equilibrium if the stage game is repeated only once and π_C is a $\|\cdot\|_\infty -$ contraction;

Lemma (Cominetti et al. 2010, Theorem 4, p75). *If π_C is a $\|\cdot\|_\infty -$ contraction and a normal form game is repeated only once in each of infinitely repeated periods, then players' behavioural strategy profiles converge to the unique quantal response equilibrium almost surely.*

I now assume that for any $i \in N$ and $a \in A^i$, action a's emotional noise for all histories have an identical distribution. Note that it is still allowed for the distributions of the noise to be different among players and actions. For example, the case where i.i.d noise with

the extreme value distribution is assumed to obtain the logistic choice rule satisfies this assumption. Cominetti et al. (2010) show that under a logistic choice rule, π_C is a $\|\cdot\|_\infty -$ contraction if the noise term is big enough or the difference of player's payoff when other player changes his action is small enough.

Now I show that when the stage game is repeated finitely many times, players' behavioural strategies converge to the agent quantal response equilibrium of the game with probability one.

Proposition 15. *Assume that players play a finitely repeated game in infinitely many periods. Then with probability one, players' behavioural strategy profiles converge to the unique AQRE if (i) each stage game consists of a normal form game, (ii) players follow Q-learning updating rule, and (iii) π_C is a $\|\cdot\|_\infty -$ contraction.*

Proof. See Appendix. □

4.5 Finitely Repeated Games without Emotional Noise

In this section, I assume that in each period, players play a finitely repeated game where each stage game consists of a normal form game. I also assume that players do not experience noise on their assessments. Therefore, in any period, each player chooses an action which has the highest subjective assessment[1]: for any $i \in N$, $I^i \in \mathcal{I}^i$, $h \in I^i$, $n \in \mathbb{N}$,

$$s_n^i(h) = \arg\max_{a \in A^i} Q_{n,I^i}^i(a),$$

[1]I do not assume a specific tie break rule for players here. I may assume that players pick uniformly one of actions which have the highest subjective assessment. However, the results in this section do not depend on the assumption on tie break rule.

where s_n^i is a strategy of player i in period n, which specifies the action at each of his own information sets.

I assume that players follow the Q-learning updating rule with $\{\lambda_n^i\}_{i,n}$ being a random sequence where (i) $\lambda_n^i \in (0,1)$ for all i and n, (ii) $\{\lambda_n^i\}_{i,n}$ is i.i.d. among players and periods and (iii) $\text{Prob}(\lambda_n^i \in J) > 0$ for any interval J for all i and n [1].

I assume that the initial assessment of each action for each player satisfies the following condition; for I_t^i with a_s^i being the action taken by player i at round s,

$$Q^i_{1,I_t^i}(a_t^i) > \min_{a_t^{-i} \in A^{-i}} \pi_t^i(a_t^i, a_t^{-i}) + \max_{(a_s^i)_{s=t+1}^T} \min_{(a_s^{-i})_{s=t+1}^T} \sum_{s=t+1}^{T} \pi_s^i(a_s^i, a_s^{-i})$$

and

$$Q^i_{1,I_T^i}(a_T^i) > \min_{a_T^{-i} \in A^{-i}} \pi_T^i(a_T^i, a_T^{-i}).$$

The conditions say that the initial assessment of an action is strictly greater than the minimum payoff from the action. The conditions also exclude trivial cases where the assessment of each action is lower than the minimum payoff from the action so that players play the strategy profile which is chosen in the first period in the subsequent periods.

I first consider the finitely repeated prisoner's dilemma, which has the following stage game payoff matrix:

	Cooperate	Defect
Cooperate	π^1_{CC},π^2_{CC}	π^1_{CD},π^2_{DC}
Defect	π^1_{DC},π^2_{CD}	π^1_{DD},π^2_{DD}

where $\pi^i_{DC} > \pi^i_{CC} > \pi^i_{DD} > \pi^i_{CD}$. In the following Proposition, I show that if the

[1] See Chapter 3 for the ideas of these assumptions

payoff π^i_{CD} is low enough, then players end up playing a strategy profile in which players show mutual cooperation or defection at each round. To show that, I introduce some technical notations.

Let $(s^i, s^{-i}) : \bigcup_t H_t \setminus Z \to A^i \times A^{-i}$ be the strategy profile of player i and $-i$ where $(s^i, s^{-i})(h) = (s^i(h), s^{-i}(h, s^i(h)))$. Let $S_{a,b}$ denote the set of strategy profiles such that

1. $(s^i, s^{-i})(h_0) = (a_1, a_1)$,
2. $(s^i, s^{-i})((a_1, a_1)) = (a_2, a_2)$, and
3. $(s^i, s^{-i})((a_1, a_1, a_2, a_2, ..., a_{t-1}, a_{t-1})) = (a_t, a_t)$ where $a_t \in A = \{a, b\}$ for any $t \in \mathbf{T} = \{1, 2, ..., T\}$.

This means that players always pick (a, a) or (b, b), coordinate on the same action, at on-path games. Using the notations above, the following result is shown:

Proposition 16. *With* $\mathbf{T} = \{1, 2, ..., T\}$,

$$\Pr(\{\lim_{n\to\infty}(s^i_n, s^{-i}_n) = (s^{i*}, s^{-i*}) \text{ where } (s^{i*}, s^{-i*}) \in S_{C,D}\}) = 1$$

if $\pi^i_{CD} + (T-t)\pi^i_{CC} < (T-t+1)\pi^i_{DD}$, $\forall i \in \{1, 2\}$, $\forall t \in \mathbf{T}$.

Proof. See Appendix. □

It shows that if payoff from π^i_{CD} is low enough, then player i will not sacrifice one-round payoff for the future payoffs and mutual cooperation or mutual defection is achieved at each round.

I next consider the case where each stage game consists of a 2×2 coordination game. For instance, the stag hunt game has the following payoff matrix;

	Stag	Rabbit
Stag	π^1_{SS}, π^2_{SS}	π^1_{SR}, π^2_{RS}
Rabbit	π^1_{RS}, π^2_{SR}	π^1_{RR}, π^2_{RR}

where $\pi^i_{SS} > \pi^i_{RS} \geq \pi^i_{RR} > \pi^i_{SR}$ for $i \in \{1, 2\}$. In general, a 2×2 normal form game

	a	b
a	π^1_{aa}, π^2_{aa}	π^1_{ab}, π^2_{ba}
b	π^1_{ba}, π^2_{ab}	π^1_{bb}, π^2_{bb}

is a coordination game if $\pi^i_{aa} > \pi^i_{ba}$ and $\pi^i_{bb} > \pi^i_{ab}$ for $i \in \{1, 2\}$. I now restrict my attention to specific 2×2 coordination games where (i) all off-diagonal payoffs for both players are the same or (ii) at each off-diagonal action profile, there exists one player who receives the worst payoff. The battle of the sexes game, the stag hunt game, and a pure coordination game are examples of the coordination games, where Chapter 3 shows that players end up playing a pure Nash equilibrium in the long run[1] if these games are repeated only once in each period. It can be also shown from Proposition 16 that in the long run, players succeed to cooperate in each stage game if the stage game is repeated more than once, but finite times, in each period. Let $k^i := \arg\max_{a \in A} \pi^i_{a,a}$ and $l^i \neq k^i$ where $l^i \in A$. Then from the argument of Proposition 16, it is easy to show the condition under which mutual cooperation is achieved in a finitely repeated coordination;

Corollary 6. *Consider the case in which players play a* 2×2 *coordination game* T *times in each period where at least one player receives the worst payoff at each non-Nash equilibrium*

[1]There are some cases that players may not end up playing pure Nash equilibrium in some modification of the battle of the sexes game. Here, I consider the standard version of the battle of the sexes game. For details, see Chapter 3.

profile in the stage game. With $\mathbf{T} = \{1, 2, ..., T\}$,

$$\Pr(\{\lim_{n\to\infty}(s_n^i, s_n^{-i}) = (s^{i*}, s^{-i*}) \ where \ (s^{i*}, s^{-i*}) \in S_{a,b}\}) = 1$$

if

$$\min(\pi_{a,b}^i, \pi_{b,a}^i) + (T-t)\pi_{k^i,k^i}^i < (T-t+1)\pi_{l^i,l^i}^i$$

$\forall i \in \{1, 2\}$ *and* $t \in \mathbf{T}$.

4.6 Conclusion and Discussion

In this chapter, I consider adaptive players who follow Q-learning and SV updating rules. I show that players' behavioural strategy profiles converge to a unique agent quantal response equilibrium when temporal emotional noise affects their assessments. If there is no noise on their assessments, then both players (1) show mutual cooperation or defection at each stage game in the finitely repeated prisoner's dilemma and (2) succeed to coordinate at each stage game in finitely repeated coordination games.

Note that players in this chapter assign subjective assessments on actions which follow any history. If the size of extensive form game becomes bigger, then it may be difficult for players to remember all assessments. Then it may be reasonable for players to group some information sets and assign representative assessments to actions available in the group. The analysis for such a case is left for future research.

4.7 Appendix

4.7.1 Proof of Proposition 14

Proposition 14. *If at each of infinitely repeated periods, players play a finitely repeated game in which each stage game consists of an extensive form game with perfect information and players follow the SV updating rule, then their behavioural strategy profiles converge to the unique AQRE of the game almost surely.*

Noteh that the uniqueness of the AQRE of the game is due to the assumption of the stochastic emotional noise; it is assumed that the density of the stochastic noise for each action's assessment is strictly positive on its domain. For the details, see McKelvey and Palfrey (1998).

I now show that the behavioural strategies of players converge to the AQRE. It can be shown that any subgame can be reached infinitely often with probability one. Then it is shown by backward induction that at each subgame the assessment of each action converges to the AQRE payoff from the action so that the behavioural strategy of each player in the limit corresponds to the AQRE behavioural strategy.

First, I consider a history which requires one more action to be a terminal history; let $h_{K-1} = (a_1, ..., a_{K-1})$ be a history such that $(a_1, ..., a_{K-1}, a) \in Z$ for some action $a \in A_{h_{K-1}}$, where $A_{h_{K-1}}$ is the set of actions available after history h_{K-1}. Then the updating rule of the assessment of the action $a \in A_{h_{K-1}}$ is as follows;

$$Q^i_{n+1,h_{K-1}}(a) = Q^i_{n,h_{K-1}}(a) + \lambda_{n+1}\mathbf{1}_{h_{K-1},a}(\pi^i_{h_{K-1},a} - Q^i_{n,h_{K-1}}(a)),$$

where $h_{K-1}, a := (a_1, ..., a_{K-1}, a)$ is a terminal history. Notice that the payoff $\pi^i_{h_{K-1},a}$ is constant. Therefore, it is obvious that with probability one, $Q^i_{n,h_{K-1}}(a)$ converges to $\pi^i_{h_{K-1},a} =: F^{i*}_{h_{K-1}}(a)$ almost surely. Note that the choice probability of the action of player i after history h_{K-1} in the limit corresponds to the probability which is derived by the quantal response equilibrium strategy of player i after h_{K-1}, $\beta^{i*}(h_{K-1})(a)$.

Next, I consider the assessment of an action $a \in A_{h_{K-2}}$ which follows history h_{K-2} such that the longest terminal history which contains history $h_{K-2}, a := (a_1, ..., a_{K-2}, a)$ as a partial history has the length of K. Then the updating rule of the action is as follows:

$$Q^i_{n+1,h_{K-2}}(a) = Q^i_{n,h_{K-2}}(a) + \lambda_{n+1}\mathbf{1}_{h_{K-2},a}(\pi^i_{h_{K-2},a} - Q^i_{n,h_{K-2}}(a)), \tag{4.1}$$

where $\pi^i_{h_{K-2},a}$ is defined as follows;

$$\pi^i_{h_{K-2},a} = \sum_{a' \in A_{K-2,a}} \mathbf{1}_{a'|h_{K-2},a}\pi^i_{h_{K-2},a,a'},$$

where (1) $\mathbf{1}_{a'|h_{K-2},a} = 1$ if given the event that history h_{K-2}, a is realized, a' is chosen and 0 otherwise, (2)$h_{K-2}, a, a' := (a_1, ..., a_{K-2}, a, a')$. Now I define $F^{i*}_{h_{K-2}}(a)$ in the following manner;

$$\begin{aligned} F^{i*}_{h_{K-2}}(a) &= \sum_{a' \in A_{K-2,a}} C^j(Q^{j*}_{h_{K-2},a})(a')\pi^i_{h_{K-2},a,a'} \\ &= \sum_{a' \in A_{K-2,a}} C^j(Q^{j*}_{h_{K-2},a})(a')F^{i*}_{h_{K-2},a,a'} \end{aligned}$$

where $j = P(h_{K-2}, a)$ and $(Q^{j*}_{h_{K-2},a}(a'))_{a' \in A^j}$ is agent j's limit assessments, which is $(\pi^j_{h_{K-2},a,a'})_{a' \in A_{h_{K-2},a}}$. Therefore, $(C^j(Q^{j*}_{h_{K-2},a})(a'))_{a' \in A_{h_{K-2},a}}$ corresponds to player j's AQRE

behavioural strategy profile. Note that $F^{i*}_{h_{K-2}}(a)$ is constant for each $a \in A_{h_{K-2}}$. I can rewrite the equation (4.1) as follows;

$$\begin{aligned} Q^i_{n+1,h_{K-2}}(a) &= Q^i_{n,h_{K-2}}(a) + \lambda_{n+1}\mathbf{1}_{h_{K-2},a}\Bigg((F^{i*}_{h_{K-2}}(a) - Q^i_{n,h_{K-2}}(a)) \\ &- \Big(\sum_{a' \in A_{K-2,a}} C^j(Q^j_{n,h_{K-2},a})(a')\pi^i_{h_{K-2},a,a'} - \sum_{a' \in A_{K-2,a}} \mathbf{1}_{a'|h_{K-2},a}\pi^i_{h_{K-2},a,a'}\Big) \\ &- (F^{i*}_{h_{K-2}}(a) - \sum_{a' \in A_{K-2,a}} C^j(Q^j_{n,h_{K-2},a})(a')\pi^i_{h_{K-2},a,a'})\Bigg). \end{aligned}$$

Since the term in the second line is martingale difference noise and the term in the last line converges to 0 almost surely if n goes to infinity, by a stochastic approximation method[1], we know that $Q^i_{n,h_{K-2}}(a) \overset{a.s.}{\to} Q^{i*}_{h_{K-2}}(a) = F^{i*}_{h_{K-2}}(a)$ for $a \in A_{h_{K-2}}$. Notice that the player i plays the agent quantal response equilibrium strategy after history h_{K-2}.

Last, I prove for the remaining cases by induction. I first define $F^{i*}_{h_l}(a)$ as follows; for any $l \in \{1, 2, ..., K-2\}$,

$$F^{i*}_{h_l}(a) := \sum_{a' \in A_{h_l,a}} C^j(Q^{j*}_{h_l,a})(a')F^{i*}_{h_l,a}(a'),$$

where $j = P(h_l, a)$ and $F^i_{h_{K-2},a,}(a') = \pi^i_{h_{K-2},a,a'}$. Since $\pi^i_{h_{K-2},a,a'}$ is constant, $F^{i*}_{h_l}(a)$ is also constant for $l \in \{1, ..., K-3, K-2\}$. Assuming that for any $l \in \{K-k, ..., K-3, K-2\}$, h_l and $a \in A_{h_l}$, $Q^i_{n,h_l}(a) \overset{a.s.}{\to} Q^{i*}_{h_l}(a) = F^{i*}_{h_l}(a)$, I show that $Q^i_{n,h_{K-(k+1)}}(a) \overset{a.s.}{\to} Q^{i*}_{h_{K-(k+1)}}(a) = F^{i*}_{h_{K-(k+1)}}(a)$ for $a \in A_{h_{K-(k+1)}}$. Now the updating rule of action $a \in A_{h_{K-(k+1)}}$ can be

[1]For example, the interested reader may refer Chapter 2 in Borkar (2008) for the stochastic approximation method used in this proof.

expressed as follows;

$$\begin{aligned} Q^i_{n+1,h_{K-(k+1)}}(a) &= Q^i_{n,h_{K-(k+1)}}(a) + \lambda_{n+1}\mathbf{1}_{h_{K-(k+1)},a}\Bigg((F^{i*}_{h_{K-(k+1)}}(a) - Q^i_{n,h_{K-(k+1)}}(a)) \\ &- (F^i_{n,h_{K-(k+1)}}(a) - \sum_{a'\in A_{h_k,a}} \mathbf{1}_{a'|h_{K-(k+1)},a}\pi^i_{h_{K-(k+1),a,a'}}) \\ &- (F^{i*}_{h_{K-(k+1)}}(a) - F^i_{n,h_{K-(k+1)}}(a))\Bigg), \end{aligned}$$

where

$$\pi^i_{h_{K-(k+1)},a,a'} := \sum_{a''\in A_{h_{K-(k+1)},a,a'}} \mathbf{1}_{a''|h_{K-(k+1)},a,a'}\pi^i_{h_{K-(k+1)},a,a',a''}$$

if $h_{K-(k+1)}, a, a'$ is not a terminal history and

$$F^i_{n,h_{K-(k+1)}}(a) := \sum_{a'\in A_{h_{K-(k+1)},a}} C^j(Q^j_{n,h_{K-(k+1)},a})(a')F^i_{h_l,a}(a').$$

Notice that the term in the second line is martingale difference noise. Also notice that the term in the third line converges to zero almost surely by an induction argument. Therefore, by a stochastic approximation method, we have $Q^i_{n,h_k}(a) \overset{a.s.}{\to} F^{i*}_{h_k,a}$ for any $a \in A_{h_k}$. □

4.7.2 Proof of Proposition 15

Proposition 15. *Assume that players play a finitely repeated game in each of infinitely many periods. Then with probability one, players' behavioural strategy profiles converge to the unique AQRE if (i) each stage game consists of a normal form game, (ii) players follow Q-learning updating rule, and (iii) π_C is a $\|\cdot\|_\infty -$ contraction.*

I assume the Q-learning updating rule for each player. It can also be shown easily that any subgame is reached infinitely often with probability one. In this proof, again backward induction method is used to show that the assessment of each action converges to the AQRE payoff from the action so that the behavioural strategy of each player in the limit corresponds to the AQRE behavior strategy.

Now I consider the last subgame which follows some history h_{T-1} and let I_T^i be player i's information set at the subgame. Note that the last subgame is a one shot normal form game. Therefore, by the nature of Q-learning updating rule at the last period and Cominetti et al. (2010), it is shown that behavioural strategies of players converge to the unique quantal response equilibrium with probability one. In fact, because of the structure of choice probabilities, the equilibrium behavioural strategies coincide with the agent quantal response equilibrium behavioural strategies of the finitely repeated game[1]. Note that $\max_{b^i \in A^i} Q^i_{n,I_T^i}(b^i)$ converges to $\max_{b^i \in A} \pi^i(b^i, \beta^{-i,*}_{I_T^{-i}})$ almost surely, where $\beta^{-i,*}_{I_t^{-i}}$, $t \in \{1, ..., T\}$, is player $-i$'s quantal response equilibrium behavioural strategy at I_t^{-i}. Note that the distribution of the emotional noise of an action and the stage games at round T do not depend on history. Thus for any information set of each player at the last subgame, I_T^i, let $\beta_T = (\beta_{I_T^i})_i$ be a behavioural strategy profile at round T.

I next consider a stage game in period $T-1$. Then the updating rule of an assessment of player i at round $T-1$ is as follows; letting I^i_{T-1} be player i's information set at the stage game and history h_T being realized,

$$Q^i_{n+1,I^i_{T-1}}(a^i) = Q^i_{n,I^i_{T-1}}(a^i) + \lambda_{n+1}\mathbf{1}_{I^i_{T-1},a^i}(\pi^i_{T-1,h_T} + \max_{b^i \in A^i} Q^i_{n,I^i_{T,h_T}}(b^i) - Q^i_{n,I^i_{T-1}}(a^i)).$$

[1]Note that the choice probability at some subgame depends on the differences among expected payoffs from actions available at the game. Also, notice that at the last subgame, the differences are determined by the payoffs from the game.

Since $\max_{b^i \in A^i} Q^i_{n,I^i_T}(b^i)$ converges to $\max_{b^i \in A^i} \pi^i(b^i, \beta^{-i,*}_T)$ almost surely, it is shown that they are approximated by the following ordinary differential equations;

$$\dot{Q}^i_{\tau,I^i_{T-1}}(a^i) = C_{I^i_{T-1},a^i}(Q_\tau)(\pi^i_{T-1}(a^i, \beta^{-i}_{T-1}) + K^i_T - Q^i_{\tau,h_{T-2}}(a^i)), a^i \in A^i,$$

where (i) $C_{I,a}(Q)$ is a probability which is derived by players' choice probabilities and assessments Q, of the realization of I and a; and (ii) $K^i_T = \max_{b^i \in A^i} \pi^i(b^i, \beta^{-i,*}_T)$. Note that $C_{I^i_{t-1},a}(Q)$ is determined by choice probabilities of players until round t. Now defining $\bar{\pi}^i_{T-1} := \pi^i_{T-1} + K^i_T$, we have

$$\dot{Q}^i_{\tau,I^i_{T-1}}(a^i) = C_{I^i_{T-1},a^i}(Q_\tau)(\bar{\pi}^i_{T-1}(a^i, \beta^{-i}_{T-1}) - Q^i_{\tau,I^i_{T-1}}(a^i)).$$

Notice that the game with the payoff function $\bar{\pi}^i_{T-1}$ is equivalent to the stage game. It is also easy to show that $\bar{\pi}_C$, which is players' expected payoffs for all actions derived by $\bar{\pi}^i_{T-1}$ for each i, is also a $\|\cdot\|_\infty -$ contraction if π_C is a $\|\cdot\|_\infty -$ contraction. Then by Cominetti et al. (2010), player i' behavioural strategies converge to the agent quantal response equilibrium strategy and $Q^i_{n,I^i_{T-1}}(a^i)$ converges to $\pi^i(a^i, \beta^{-i,*}_{T-1}) + K^i_T$ almost surely, where $\beta^{-i,*}_{T-1}$ is player $-i$'s agent quantal response equilibrium strategy at round $T-1$[1] . Notice that letting $K^i_{T-1} := \max_{a^i} \pi^i(a^i, \beta^{-i,*}_{T-1}) + K^i_T$, $\max_{b^i \in A} Q^i_{n,I^i_{T-1}}(b^i)$ converges to K^i_{T-1} almost surely.

Now I prove for the other cases by backward induction. I assume that $Q^i_{n,I^i_{t+1}}(a^i)$ converges to $\pi^i(a^i, \beta^{-i,*}_{t+1}) + K^i_{t+2}$, where $K^i_{t+2} := \max_{a^i} \pi^i(a^i, \beta^{-i,*}_{t+2}) + K^i_{t+3}$ for $0 \leq t \leq T-3$. Then we know that $\max_{b^i \in A^i} Q^i_{n,I^i_{t+1}}(b^i)$ converges to K^i_{t+1} almost surely. Consider the

[1]Since the expected payoffs from the last subgames which follow the game are all equivalent, therefore, again, the differences among expected payoffs of the actions are determined by the payoffs from the game.

updating rule of the assessment of action a^i at a t-th round stage game when history h_T is realized:

$$Q^i_{n+1,I^i_t}(a^i) = Q^i_{n,I^i_t}(a^i) + \lambda_{n+1}\mathbf{1}_{I^i_t,a^i}(\pi^i_{t,h_T} + \max_{b^i \in A^i} Q^i_{n,I^i_{t+1,h_T}}(b^i) - Q^i_{n,I^i_t}(a^i)).$$

Since $\max_{b^i \in A^i} Q^i_{n,I^i_{t+1}}(b^i)$ converges to K^i_{t+1} almost surely and K^i_{t+1} is equal across player i's information sets at round $t+1$, by a stochastic approximation method, we have

$$\dot{Q}^i_{\tau,I^i_t}(a^i) = C_{I^i_t,a^i}(Q_\tau)(\pi^i_t(a^i, \beta^{-i}_t) + K^i_{t+1} - Q^i_{\tau,I^i_t}(a^i)).$$

By setting $\bar{\pi}^i := \pi^i + K^i_{t+1}$, we have

$$\dot{Q}^i_{\tau,I^i_t}(a^i) = C_{I^i_t,a^i}(Q_\tau)(\bar{\pi}^i_t(a^i, \beta^{-i}_t) - Q^i_{\tau,I^i_t}(a^i)).$$

Since the game with the payoff function $\bar{\pi}^i$ is equivalent to the stage game, it can be shown that $Q^i_{n,I^i_t}(a^i)$ converges to $\pi^i(a^i, \beta^{-i*}_t) + K^i_{t+1}$ where β^{-i*}_t is player $-i$'s agent quantal response equilibrium strategy. □

4.7.3 Proof of Proposition 16

Proposition 16. *With* $\mathbf{T} = \{1, 2, ..., T\}$,

$$\Pr(\{\lim_{n\to\infty}(s^i_n, s^{-i}_n) = (s^{i*}, s^{-i*}) \textit{ where } (s^{i*}, s^{-i*}) \in S_{C,D}\}) = 1$$

if $\pi^i_{CD} + (T-t)\pi^i_{CC} < (T-t+1)\pi^i_{DD}$, $\forall i \in \{1,2\}$, $\forall t \in \mathbf{T}$.

First, I consider the simplest case of finitely repeated case where $\mathbf{T} = \{1, 2\}$. It can be shown that with probability one, players play (C, C) or (D, D) at any second stage game, which follows any history if the history is realized infinitely many times[1]. Note that (i) (C, C) is the absorbing state of the game and (ii) at off-diagonal action profiles, one player receives the worst payoff so that players never play the off-diagonal action profiles infinitely many times.

I now show that players cannot play $((C, D), (D, D))$ and $((C, D), (C, C))$ infinitely many times. The intuition of the proof is that the payoff of the player who chooses C at (C, D) is so low that he does not choose C to receive the better payoff in the nest stage.

I first consider the proof for $((C, D), (D, D))$. When $((C, D), (D, D))$ is played, the assessment of action C of player i approaches $\pi^i_{CD} + \pi^i_{DD}$. While the assessment of action D at the first period is at least $2\pi^i_{DD}$. Therefore, with probability one, the assessment of action C of player i becomes lower than $2\pi^i_{DD}$ if $((C, D), (D, D))$ is played infinitely many times and after that, $((C, D), (D, D))$ is not played; which contradicts the hypothesis. Now suppose that $((C, D), (C, C))$ is played infinitely many times. Then the assessment of C of player i at the first period approaches to $\pi^i_{CD} + \pi^i_{CC}$, while the assessment of action D is at least $2\pi^i_{DD}$. Therefore, with probability one, the assessment of action C becomes lower than $2\pi^i_{DD}$, in which case the player never plays C, which contradicts the hypothesis. Note that players end up playing one action profile, since the weighting parameters are i.i.d. random variables among players [2].

Now I show for the general case. It has been shown that at the last round, players end up playing (C, C) or (D, D) with probability one. To prove this by induction, assume that

[1]See the argument in Chapter 3 or the main theorem of Sarin (1999).
[2]See Chapter 3 for detailed argument.

after round t, players play (C, C) or (D, D). If players play (C, D) at round t infinitely often, then the assessment of C becomes lower than $(T-t+1)\pi^i_{DD}$ with positive probability in each of the periods, since

$$\pi^i_{CD} + \sum_{s=t+1}^{T} \pi^i_{t,h'_T} \leq \pi^i_{CD} + (T-t)\pi^i_{CC} < (T-t+1)\pi^i_{DD},$$

where, in the history h'_T, players play (C, C) or (D, D) after period t. If (C, D) is played at round t in infinitely many periods, then the assessment of the C becomes lower than $(T-t+1)\pi^i_{DD}$ with probability one, which contradicts the hypothesis. □

CHAPTER 5

CONCLUSION AND SUGGESTIONS FOR FUTURE RESEARCH

5.1 Conclusion

In this thesis, I investigate the behaviour of adaptive decision makers in the long run in a decision problem, normal form games, and finitely repeated games. They are assumed to have limited information, as in many situations in the real world, and assess each action based on its past payoffs. Given the assessments, each of them chooses the action which he thinks is the best; the action which has the highest assessment. After receiving payoff information, they update their assessments adaptively using the information.

In Chapter 2, I investigate the case in which an adaptive decision maker observes objective payoff information and in addition, obtains foregone payoff information. When the noise on each assessment is big enough, then with probability one, the assessments converge to the weighted average of expected objective and distorted payoffs with weights depending on the limit choice probabilities. It is also shown that he picks the optimal action most frequently in the long run if expected distorted payoff of the action is greater than the ones of the other actions. For example, the decision makers in the EWA learning

model and in the stochastic fictitious play model, which are special cases of this model, pick the optimal action most frequently in the long run. However, there is a case where the decision maker chooses a non-optimal action most frequently; it happens when he distorts the foregone payoff of the action greater than the one of the optimal action.

In Chapter 3, I investigate the case in which adaptive players face a normal form game with strict Nash equilibria in infinitely many periods. It is shown that players end up playing a strict Nash equilibrium if (1) at non-Nash equilibrium action profile, there exists at least one player who can find another action which always gives better payoffs than his current payoff or (2) each player's payoffs at all non-Nash equilibrium profiles are equivalent. For example, the stag hunt game satisfies condition (1) while battle of the sexes games and pure coordination games satisfy condition (2), meaning that players end up playing a pure Nash equilibrium in these games. The convergence result is also shown for the first order statistic game.

In Chapter 4, I investigate the case in which adaptive players play a finitely repeated game in each of infinitely repeated periods. I consider two updating rules; Q-learning updating rule and SV updating rule. When each stage game consists of an extensive form game with perfect information, then players' behavioural strategies converge to the agent quantal response equilibrium introduced by McKelvey and Palfrey (1998). I also give a condition, which is from Cominetti et al. (2010), for the convergence when each stage game consists of a normal form game. While the results are based on the model with noise on each assessment, I also consider the model without the noise. I show that when the finitely repeated prisoner's dilemma is played, players may end up cooperating in each stage game, while when they play a finitely repeated coordination game, they end

up cooperating on one of pure Nash equilibria in each stage game.

In each of these chapters, I also show the conditions under which they may fail to learn the optimal action in the decision problem, Nash equilibrium in the normal form and subgame perfect equilibrium in extensive form games. The intuition behind the failure of learning the optimal action or Nash equilibrium is that (1) the lack of exploration prevents players from learning the values of actions, where the exploration is caused by emotional stochastic shocks on players' assessments and (2) distortion of foregone payoff information makes a non-optimal action more attractive. However, note that even the lack of exploration, convergence to Nash equilibrium is shown in Chapter 3.

5.2 Extensions

In Chapter 2, I restrict my attention to a decision problem but the model can be extended to normal form and extensive form games. It may be interesting to see how the distortion affects the long run outcomes of learning in games. In Chapter 3, I may consider the situation where players experience emotional shocks on assessments. In Chapter 4, it is natural to extend the analysis to more games and learning rules. Also, I may focus on different aspects of players' cognitive limitations in adaptive learning. For instance, in Chapter 4, players assign assessments on actions which follow any history. This seems adequate if the size of the game is not big, but it may be a problem otherwise, because of players' memory limitations. One natural thought is that they may categorize information sets so that they need to remember only representative assessments of actions in each of categories of sets. These extensions are left for future research.

LIST OF REFERENCES

[1] A. W. Beggs, On the convergence of reinforcement learning, *Journal of Economic Theory* 122 (2005) 1-36.

[2] M. Benaïm, Dynamics of stochastic approximation algorithms, *Le Séminaire de Probabilité*, XXXIII, Lecture Notes in Mathematics, vol.1709, Springer, Berlin, 1999, pp. 1–68.

[3] V. Borkar, Stochastic approximation: a dynamical systems viewpoint, Cambridge University Press, 2008.

[4] C. Camerer, T. H. Ho, Experience-weighted attraction learning in normal form games, *Econometrica* 67 (1999) 827-874.

[5] C. Camerer, T. H. Ho, X. Wang, Individual differences in ewa learning with partial payoff information, *Economic Journal* 118 (2008) 37-59.

[6] Y.Chen, Y. Khoroshilov, Learning under limited information, *Games and Economic Behavior* 44 (2003) 1-25.

[7] R. Cominetti, E. Melo, S. Sorin, A payoff-based learning and its application to traffic games, *Games and Economic Behavior* 70 (2010) 71-83.

[8] T. G. Conley, C. R. Udry, Learning about a new technology: pineapple in Ghana, *American Economic Review* 100 (2010) 35-69.

[9] R. W. Cooper, D. V. DeJong, R. Forsythe and T. W. Ross, Selection criteria in coordination games: some experimental results, *American Economic Review* 80 (1990) 218-233.

[10] G. Coricelli, A Rustichini, Counterfactual thinking and emotions: regret and envy learning, *Philosophical Transactions of the Royal Society B* 365 (2010) 241-247.

[11] J. Duffy, N. Feltovich, Does observation of others affect learning in strategic environments? An experimental study, *International Journal of Game Theory* 28 (1999) 131-152.

[12] I. Erev, A. Roth, Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria, *American Economic Review* 88 (1998) 848-881.

[13] A. W. Foster, M. R. Rosenzweig, Learning by doing and learning from others: human capital and technical change in agriculture, *Journal of Political Economy* 103 (1995) 1176-1209.

[14] D. Fudenberg, D. M. Kreps, Learning mixed equilibria, *Games and Economic Behavior* 5 (1993) 320-367.

[15] D. Fudenberg, D. M. Kreps, Learning in extensive-form games I. self-confirming equilibria, *Games and Economic Behavior* 8(1995) 20-55.

[16] D. Fudenberg, D. K. Levine, The theory of learning in games, MIT Press, Cambridge MA, 1998.

[17] E. Groes, H. J. Jacobsen, B. Sloth, Adaptive learning in extensive form games and sequential equilibrium, *Economic Theory* 13 (1999) 125- 142.

[18] B. Grosskopf, I. Erev, E. Yechiam, Foregone with the wind: indirect payoff information and its implications for choice, *International Journal of Game Theory* 34 (2006) 285-302.

[19] J. Grygolec, G Coricelli, A Rustichini, Positive interaction of social comparison and personal responsibility for outcomes, *Frontiers in Psychology* 3:25 (2012).

[20] P. Hall, C. C. Heyde, Martingale limit theory and its application, Academic Press, New York, 1980

[21] D. Heller, R. Sarin, Adaptive learning with indirect payoff information, Mimeo, 2001.

[22] E. Hendon, H. J. Jacobsen, B. Sloth, Fictitious play in extensive form games, *Games and Economic Behavior* 15 (1996) 177 - 202.

[23] J. Hofbauer, W.H. Sandholm, On the global convergence of stochastic fictitious play, *Econometrica* 70 (2002) 2265–2294.

[24] P. Jehiel, D. Samet, Learning to play games in extensive form by valuation, *Journal of Economic Theory* 124 (2005) 129-148.

[25] J. F. Laslier, R. Topol, B. Walliser, A behavioral learning process in games, *Games and Economic Behavior* 37 (2001) 340 - 366.

[26] J. F. Laslier, B. Walliser, A reinforcement learning process in extensive form games, *International Journal of Game Theory* 33 (2005) 219-227.

[27] D. S. Leslie, E. J. Collins, Individual q-learning in normal form games, *SIAM Journal on Control and Optimization* 44 (2005) 495-514.

[28] D. S. Leslie, E. J. Collins, Generalized weakened fictitious play, *Games and Economic Behavior* 56 (2006) 285-298.

[29] R. McKelvey, T. R. Palfrey, An experimental study of the centipede game, *Econometrica* 60 (1992) 803-836.

[30] R. McKelvey, T. R. Palfrey, Quantal response equilibria for normal form games, *Games and Economic Behavior* 10 (1995) 6-38.

[31] R. McKelvey, T. R. Palfrey, Quantal response equilibria for extensive form games, *Experimental Economics* 1 (1998) 9-41.

[32] D. Monderer, L.S. Shapley, Fictitious play for games with identical interests, *Journal of Economic Theory* 68 (1996) 258-265.

[33] M. J. Osborne, A. Rubinstein, A course in game theory, MIT Press, 1994

[34] I. Palacious-Huerta, O. Volij, Field centipedes, *American Economic Review* 99 (2009) 1619-1635.

[35] R. Sarin, Simple play in the prisoner's dilemma, *Journal of Economic Behavior and Organization* 40 (1999) 105-113.

[36] R. Sarin, F. Vahid, Payoff assessments without probabilities: a simple dynamic model of choice, *Games and Economic Behavior* 28 (1999) 294-309.

[37] R. Sarin, F. Vahid, Predicting how people play games: a simple dynamic model of choice, *Games and Economic Behavior* 34 (2001) 104-122.

[38] R. Selten, R. Stoecker, End behavior in sequences of finite prisoner's dilemma supergames, *Journal of Economic Behavior and Organization* 7 (1986) 47-70.

[39] J. Van Huyck, R. C. Battalio, and R. O. Beil, Tacit coordination games, strategic uncertainty, and coordination failure, *American Economic Review* 80 (1990) 234-248.

[40] C. J. C. H. Watkins, P. Dayan, Q-learning, *Machine Learning* 8 (1992) 279-292.

[41] E. Yechiam, J. R. Busemeyer, Evaluating generalizability and parameter consistency in learning models, *Games and Economic Behavior* 63 (2008) 370-394.