Reinforcement Learning with Foregone Payoff Information in Normal Form Games

Naoki Funai

Department of Economics, Ryutsu Keizai University, Japan

February 27, 2019

Abstract

This paper studies the reinforcement learning of Erev and Roth with foregone payoff information in normal form games: players observe not only the realised payoffs but also foregone payoffs, the ones which they could have obtained if they had chosen the other actions. We provide conditions under which the reinforcement learning process almost surely converges to a mixed action profile at which each action is chosen with a probability proportional to its expected payoff. In particular, in symmetric 2×2 partnership games without a Pareto-dominant Nash equilibrium and matching-pennies games, the action profile corresponds to the mixed Nash equilibrium. However, in general, the action profile does not correspond to a Nash equilibrium: it corresponds to a Nash equilibrium if and only if for each player, all the actions are chosen with equal probability at the equilibrium. Instead, we show that the action profile corresponds to a perturbed equilibrium, regular quantal response equilibrium (Goeree et al., 2005), when there is no distortion on the foregone payoffs. Therefore, for any normal form games, if the conditions hold, the process almost surely converges to the equilibrium.

Keywords: Reinforcement learning; foregone payoff information; regular quantal response equilibrium; normal form games; asynchronous stochastic approximation

1 Introduction

In this paper, we investigate the long-run outcomes of the reinforcement learning of Roth and Erev (1995) and Erev and Roth (1998) with foregone payoff information in normal form games. That is, players observe payoffs from not only chosen actions but also unchosen actions. This paper provides conditions under which the reinforcement learning process converges to a mixed action profile at which each action is chosen with a probability proportional to the expected payoff of the action. In particular, in symmetric 2×2 partnership games without a Pareto-dominant Nash equilibrium and matching-pennies games, the action profile corresponds to the mixed Nash equilibrium. However, the action profile does not correspond to a Nash equilibrium unless for each player, all the actions are chosen with equal probability at the equilibrium. Instead, we show that the mixed action profile corresponds to a regular quantal response equilibrium (Goeree et al., 2005) when there is no distortion on the foregone payoff information. Therefore, under these conditions, the reinforcement learning process almost surely converges to the equilibrium in normal form games.

We consider the following situation. In each period, players face a fixed normal form game and choose actions based on their past experience. Specifically, each player assigns some weights to his actions and chooses an action with a probability proportional to its weight. After choosing an action, each player observes the realised payoff, which he/she actually obtains from the play, and the foregone payoffs, which he/she could have obtained if he/she had chosen the other actions. Using the payoff information, he/she updates the weight of each action by adding the corresponding payoff to the weight. Therefore, the weight of each action represents the accumulated payoff of the action, and the probability of choosing the action is proportional to the accumulated payoff. We also consider the case in which each player may not treat the foregone payoffs as the realised payoffs and may distort the foregone payoff information.

By taking into account the foregone payoff information in the learning model of Erev and Roth, we investigate the case in which players learn not only from their own experience but also from others' experience in a social setting. For instance, we can consider the situation in which each player in a large group of players who are involved in a similar environment makes a decision and observes what has happened to the other players:[1] e.g. when having a cup of coffee alone at a coffee shop in the morning, we may observe colleagues enjoying a conversation at another coffee shop on the opposite side of the street and think we could have enjoyed ourselves more if we had chosen the other coffee shop. Also, we can consider the situation in which each player cannot directly observe others' experience but can ask about their experience to determine what could have happened if he/she had chosen the

[1] Rustichini (1999) calls the environment "full information". In this interpretation of observing foregone payoffs, it is not required for players to know the payoff structure. In another interpretation, players actually know the payoff structure and "imagine" what they could have obtained if they had chosen the other actions (e.g. Camerer and Ho, 1999).

other actions: in the coordination problem of choosing a coffee shop in the morning, we can ask our colleagues about their experience at the coffee shops which we have not chosen.

When obtaining the payoff information from others, players may not process the information in the same manner as the payoff information from their own experience. For instance, when asking neighbours about what has happened to them, players may believe that the neighbours exaggerate what they have experienced, and take the discounted payoff information into account. Also, information gained from their own experience may have a stronger impact on their behaviour than information gained from observing the others (Simonsohn et al., 2008): this idea can be expressed by discounting the effect of foregone payoffs.[2] In this paper, we express the way each player distorts the foregone payoff information by δ_i. In particular, we focus on the case in which for each player i, $\delta_i \in (0, 1]$ and is stationary over periods. The assumptions mean that players (i) discount the effect of foregone payoffs and (ii) never experience and learn what the other players have actually experienced and do not change their behaviour toward the foregone payoff information over periods.

In this paper, by utilising the asynchronous stochastic approximation method of Tsitsiklis (1994), we show that the reinforcement learning process almost surely converges to a mixed action profile when a function which maps from a mixed action profile of players to a profile each of whose components is associated with some player's action and consists of the fraction of a distorted expected payoff of the action over the sum of the payoffs of the player's available actions is a contraction mapping. In particular, we show that in several games, the function becomes a contraction mapping when δ_i is close to 1 for each player. We also show that the action profile corresponds to the mixed Nash equilibrium in symmetric 2×2 partnership games without a Pareto-dominant Nash equilibrium and matching-pennies games. However, in general, it does not correspond to a Nash equilibrium: we show that the action profile corresponds to a Nash equilibrium if and only if for each player, all the actions are chosen with equal probability at the equilibrium. Instead, we can view the action profile as a perturbed equilibrium. We first introduce the concept of regular quantal response equilibrium introduced by Goeree et al. (2005). Then we show that the action profile corresponds to the equilibrium when there is no distortion on the foregone payoffs. Therefore, under these conditions, the reinforcement learning process of this paper almost surely converges to the equilibrium in normal form games. Lastly, we show that given some payoff normalisation and $\delta_i = 1$ for each player, the function becomes a contraction mapping in any 2×2 game and thus the process converges to a regular quantal response equilibrium in the game.

The original reinforcement learning model which this paper is based on is introduced by Roth and Erev (1995) and Erev and Roth (1998) and theoretically investigated by Beggs (2005), Hopkins and Posch (2005), Ianni (2014) and Laslier et al. (2001). How-

[2] Camerer and Ho (1999) also mention this idea in the context of learning by imitation: the actions that the other players have chosen are also reinforced, so that the probability of imitating them increases, but are not as fully reinforced as the actions which have actually been chosen.

ever, the original model does not take into account the payoff information from unchosen actions. By taking into account this aspect, Rustichini (1999) extends the model to single-person decision problems. Camerer and Ho (1999) also consider this aspect and introduce the experience-weighted attraction learning model, which nests the reinforcement learning models of Erev and Roth and this paper and the stochastic fictitious play learning model of Fudenberg and Kreps (1993). Using some experimental data, they estimate the parameters of their model and show that in constant-sum games, median-action games and beauty contests, foregone payoffs are discounted. In particular, in constant-sum games and median-action games, they show that their model corresponds to the reinforcement learning model with foregone payoff information. However, they investigate empirically and do not provide any theoretical investigation.[3] In this paper, we complement the analysis of Rustichini (1999) and Camerer and Ho (1999) with the theoretical investigation of the model with foregone payoff information in normal form games.

In particular, we extend the result of Rustichini (1999) to normal form games. He shows that in a stationary single-person decision problem, the probability of choosing each action converges to the fraction of the expected payoff of the action over the sum of the expected payoffs of the available actions.[4] We show that a similar result is obtained even when players face a non-stationary decision problem: we investigate the situation in which reinforcement learners interact with each other and play a fixed normal form game repeatedly.

It is worth noting that, under reinforcement learning without foregone payoff information, Hopkins and Posch (2005) show that the reinforcement learning process converges to a pure Nash equilibrium with probability one in partnership games. This paper shows a different result: there exists a range of the distortion on the foregone payoffs such that the reinforcement learning process with foregone payoff information converges to the mixed Nash equilibrium in partnership games without a Pareto-dominant Nash equilibrium almost surely. Therefore, the observation of foregone payoff information affects their behaviour in the long run, shifting from the pure Nash equilibrium to the mixed one.

The result of convergence to the mixed Nash equilibrium with each action being chosen equally also coincides with that of Freedman (1965): he shows that in the Pólya urn model, which is often used to describe and analyse the original reinforcement learning model, with some modification in that not only balls of the same colour as the drawn ball but also balls of the opposite colour are added to the urn, the fraction of each ball being chosen converges to $\frac{1}{2}$. Beggs (2005) also mentions the result in a single-person decision problem context and provides conjectures such that when unchosen actions are also reinforced, (i)

[3] Experience-weighted attraction learning, mainly the belief-based model, is also theoretically investigated by Funai (2018).

[4] Rustichini (1999) also considers the case in which the choice probability of each player is expressed as the logit choice function and obtains different results in single-person decision problems. The convergence analysis of the process when players follow the choice rule and face a normal form game is left for future research.

the reinforcement learning process in 2×2 games does not converge to a mixed Nash equilibrium unless the probability of one action chosen is $\frac{1}{2}$ at the equilibrium, and (ii) the process may rather converge to some approximate equilibrium close by. In this paper, we verify his conjectures by considering a more general case, in that there exist more than two players with more than two actions available, and specifying the concept of the equilibrium that the reinforcement learning process converges.

Lastly, even though Beggs (2005), as well as Erev and Roth (1998) and Hopkins (2002), considers the case in which unchosen actions are also reinforced, the reinforcement on each of the unchosen actions does not depend on the corresponding foregone payoff: in their models, even though one unchosen action could provide a much better payoff than another unchosen action, those two actions are equally reinforced. This is due to the fact that they investigate the way in which experimentation or exploration on unchosen actions affects the learning process; in this paper, we investigate the way in which extra payoff information about unchosen actions affects the learning process. Hopkins (2002) shows that the reinforcement learning process with experimentation globally converges to a perturbed equilibrium in 2×2 games with a unique mixed Nash equilibrium, and the perturbed equilibrium corresponds to the mixed Nash equilibrium only when each action is chosen with probability $\frac{1}{2}$ at the equilibrium. Even though we obtain a similar result, the difference between the models also brings about differences in the convergence results. For instance, for partnership games without a Pareto-dominant Nash equilibrium, the reinforcement learning process of this paper converges uniquely to the mixed Nash equilibrium, while their reinforcement learning processes may not. Note also that Hopkins (2002) provides mostly local convergence results, while in this paper, we focus on global convergence results.

2 Model

In each period, $n \in \mathbb{N} \cup \{0\}$, players play a fixed normal form game $(\mathcal{N}, S, \pi)$ which consists of (i) the set of players $\mathcal{N} := \{1, ..., N\}$; (ii) the finite set of actions $S := \times_{i \in \mathcal{N}} S_i$, where S_i denotes the set of player i's actions and $s = (s_1, ..., s_N)$ denotes an element of S; and (iii) the payoff function $\pi : S \to \mathbb{R}^N$, where $\pi_i : S \to \mathbb{R}$ denotes the i-th component and corresponds to player i's payoff function: for any s, $\pi(s) = (\pi_1(s), ..., \pi_N(s))$. We assume that the payoffs are strictly positive: for any i and s, $\pi_i(s) > 0$.

We extend the payoff function to the set of mixed actions. Let $\Delta_i := \{x_i = (x_{i,s_i})_{s_i \in S_i} \in [0,1]^{|S_i|} : \sum_{s_i \in S_i} x_{i,s_i} = 1\}$ be the set of player i's mixed actions and $\Delta := \times_{i \in \mathcal{N}} \Delta_i$ be the set of mixed action profiles. Then for each i and $x \in \Delta$, the payoff of player i given mixed action profile x is expressed as $\pi_i(x) = \sum_{s \in S} \pi_i(s) \prod_{j \in \mathcal{N}} x_{j,s_j}$. Also, we express the payoff of player i when choosing s_i with probability one as $\pi_i(s_i, x_{-i}) = \sum_{s_{-i} \in S_{-i}} \pi_i(s_i, s_{-i}) \prod_{j \neq i} x_{j,s_j}$, where $-i := (1, ..., i-1, i+1, ..., N)$ denotes all players except i, $S_{-i} := \times_{j \neq i} S_j$ denotes the set of those players' actions, $s_{-i} := (s_1, ..., s_{i-1}, s_{i+1}, ..., s_N)$ denotes an element of S_{-i}, $\Delta_{-i} := \times_{j \neq i} \Delta_j$ denotes the set of mixed actions of all players except i and x_{-i} denotes an element of Δ_{-i}.

Next, we describe the behaviour rule of the players. In each period, the choice behaviour of each player is associated with some weights on his actions. For each n, i and s_i, let w_{n,i,s_i} denote the weight of action $s_i \in S_i$ in period n. Then each player chooses an action with a probability proportional to the weight of the action. In detail, for each n, i and s_i, player i chooses s_i in period n with probability p_{n,i,s_i}, which is defined as follows:

$$p_{n,i,s_i} = \frac{w_{n,i,s_i}}{\sum_{t_i} w_{n,i,t_i}}.$$

Here, we call the sequence $\{p_n = (p_{n,i,s_i})_{i,s_i} : n \in \mathbb{N} \cup \{0\}\}$ a reinforcement learning process. We assume that the initial weight of each action is positive: $w_{0,i,s_i} > 0$ for each i and s_i. Also, we assume that, given their weights, players choose their actions independently.

After choosing actions, players obtain payoff information and revise their choice behaviour by updating their weights. In particular, the weight of each action is updated as follows: for each n, i and s_i,

$$w_{n+1,i,s_i} = w_{n,i,s_i} + \pi_{n,i,s_i}$$

where π_{n,i,s_i} describes the payoff that player i observes for action s_i and is added to the weight in the next period. In this paper, we consider the situation in which each player also observes the foregone payoff information, and thus π_{n,i,s_i} is defined in the following manner: if s_{-i} is chosen in period n,

$$\pi_{n,i,s_i} = \begin{cases} \pi_i(s_i, s_{-i}) & \text{if } s_i \text{ is chosen in period } n, \\ \delta_i \pi_i(s_i, s_{-i}) & \text{otherwise} \end{cases}$$

where $\delta_i \in (0, 1]$ describes player i's distortion factor on the foregone payoffs. Therefore, players update their weights of not only chosen actions but also unchosen actions, whose payoff information is subject to some distortion. We assume that the way of distorting the foregone payoff information is consistent over periods. Note that π_{n,i,s_i} can be expressed as follows:

$$\pi_{n,i,s_i} = \sum_{s_{-i} \in S_{-i}} (\mathbb{1}_{n,s_i} + \delta_i(1 - \mathbb{1}_{n,s_i}))\pi_i(s_i, s_{-i})\mathbb{1}_{n,s_{-i}}$$

where $\mathbb{1}_{n,s_i}$ and $\mathbb{1}_{n,s_{-i}}$ represent the indicator functions for the events that s_i and s_{-i}, respectively, are chosen in period n.

For the purpose of formal analysis, we introduce the following notation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the probability space on which all the random variables that appear in this paper are defined. Let $\{\mathcal{F}_n\}$ be a sequence of increasing σ-fields which are subsets of $\mathcal{F}$ such that for each n, $\mathcal{F}_n$ is generated by $\{(\mathbb{1}_{m,s_i})_{i,s_i} : 0 \leq m \leq n-1\}$ and $(w_{0,i,s_i})_{i,s_i}$. Here, $\mathcal{F}_n$ is considered to be the information about players' choices up to period n but not including their choices in that period. Note that for each i and s_i, w_{n,i,s_i} and p_{n,i,s_i} are $\mathcal{F}_n$-measurable and $\mathbb{1}_{n,s_i}$ and π_{n,i,s_i} are $\mathcal{F}_{n+1}$-measurable. Then we express the choice behaviour of players as follows: for each i and s_i,

$$\mathbb{P}(s_i \text{ is chosen in period } n \mid \mathcal{F}_n) = p_{n,i,s_i}$$

and

$$\mathbb{P}(s = (s_i)_i \text{ is chosen in period } n \mid \mathcal{F}_n) = \prod_i p_{n,i,s_i}.$$

Therefore, given a choice history of players in period n, the players' decisions are characterised by p_n and are conditionally independent. Lastly, let $\overline{\pi}_{n,i,s_i}$ denote the conditional expected payoff of action s_i in period n:

$$\begin{aligned}\overline{\pi}_{n,i,s_i} :=& \mathbb{E}[\pi_{n,i,s_i} \mid \mathcal{F}_n] \\ =& \sum_{s_{-i}\in S_{-i}} (p_{n,i,s_i} + \delta_i(1 - p_{n,i,s_i}))\pi_i(s_i, s_{-i}) \prod_{j\neq i} p_{n,j,s_j} \\ =& (p_{n,i,s_i} + \delta_i(1 - p_{n,i,s_i}))\pi_i(s_i, p_{n,-i}).\end{aligned}$$

3 Convergence Results

3.1 Main results

In this section, we provide conditions under which we obtain the convergence of the reinforcement learning process. To state the main result, we introduce the following functions. Let $F = (F_{i,s_i})_{i,s_i} : \Delta \to \mathbb{R}^{|S|}$ be such that for any $x \in \Delta$,

$$\begin{aligned}F_{i,s_i}(x) =& \sum_{s_{-i}\in S_{-i}} (x_{i,s_i} + \delta_i(1 - x_{i,s_i}))\pi_i(s_i, s_{-i}) \prod_{j\neq i} x_{j,s_j} \\ =& (x_{i,s_i} + \delta_i(1 - x_{i,s_i}))\pi_i(s_i, x_{-i})\end{aligned}$$

for each i and s_i, and let $G = (G_{i,s_i})_{i,s_i} : \Delta \to \Delta$ be such that

$$G_{i,s_i}(x) = \frac{F_{i,s_i}(x)}{\sum_{t_i\in S_i} F_{i,t_i}(x)}$$

for each i and s_i.

Theorem 1. *If there exists $p^* = (p^*_{i,s_i})_{i,s_i} \in \Delta$ and $\beta \in [0, 1)$ such that*

$$||G(p) - p^*||_\infty \leq \beta||p - p^*||_\infty \tag{1}$$

for any $p \in \Delta$, then $p_n \to p^*$ almost surely, where $||\cdot||_\infty$ denotes the maximum norm.

Proof. See Appendix A. □

Figure 1: 2 × 2 game

	s_2	t_2
s_1	a^1_{11}, a^2_{11}	a^1_{12}, a^2_{21}
t_1	a^1_{21}, a^2_{12}	a^1_{22}, a^2_{22}

To understand condition (1), it is worth noting that the condition holds when G is a contraction mapping with the unique fixed point $p^* = G(p^*)$. In this case, the reinforcement learning process converges to the mixed action profile in which each action is chosen with a probability proportional to its expected payoff.

Corollary 1. *If G is a contraction mapping, then p_n converges to the fixed point p^* almost surely, where*

$$p^*_{i,s_i} = \frac{\pi^*_{i,s_i}}{\sum_{t_i} \pi^*_{i,t_i}} \tag{2}$$

and

$$\begin{aligned}\pi^*_{i,s_i} :=& \sum_{s_{-i}} (p^*_{i,s_i} + \delta_i(1 - p^*_{i,s_i}))\pi_i(s_i, s_{-i}) \prod_{j \neq i} p^*_{j,s_j} \\ =& (p^*_{i,s_i} + \delta_i(1 - p^*_{i,s_i}))\pi_i(s_i, p^*_{-i})\end{aligned}$$

for each i and s_i.

Proof. Note that if G is a contraction mapping, there exists $\beta \in [0, 1)$ such that for any p, p',

$$||G(p) - G(p')||_\infty \leq \beta ||p - p'||_\infty$$

with the unique fixed point $p^* = G(p^*)$. Then condition (1) is satisfied and thus p_n converges to the fixed point p^*, which is characterised by condition (2). □

In the following argument, we first focus on 2 × 2 games to investigate the condition under which G becomes a contraction mapping. In particular, we show that when each player's distortion factor on the foregone payoff information is close to one, G becomes a contraction mapping for partnership games without a Pareto-dominant Nash equilibrium and matching-pennies games. Moreover, under this condition, we show that the reinforcement leaning process converges to a Nash equilibrium in the games. However, in general, the fixed point of G does not correspond to a Nash equilibrium, which we further investigate by focusing on some specific games: partnership games with a Pareto-dominant Nash equilibrium and games with a dominant strategy.

Figure 2: 2 × 2 partnership game without a Pareto-dominant equilibrium

	s_2	t_2
s_1	a, a	b, b
t_1	b, b	a, a

Figure 3: matching-pennies game

	s_2	t_2
s_1	a, b	b, a
t_1	b, a	a, b

3.2 Contraction mapping in 2×2 games

In this section, we focus on 2×2 games, which are expressed by the payoff matrix in Figure 1, and investigate the condition under which G is a contraction mapping. Let

$$\beta := \max_{i \in \{1,2\}} \frac{|a^i_{11}a^i_{22} - a^i_{12}a^i_{21}| + (1-\delta_i^2)\max_{j,k\in\{1,2\}} a^i_{1j}a^i_{2k}}{\left(\min_{k\in\{1,2\},x\in\{0,1-\delta_i\}}\left((\delta_i + x)a^i_{1k} + (1-x)a^i_{2k}\right)\right)^2}.$$

Proposition 1. *If $\beta < 1$, then G is a contraction mapping and thus p_n converges to the fixed point of G almost surely.*

Proof. See Appendix C. □

To enhance our understanding, it is helpful to focus on some specific examples. First, we consider the partnership game of Figure 2 and the matching-pennies game of Figure 3 with $a > b$. Let $\delta := \min_i \delta_i$, which denotes the minimum discount factor.

Corollary 2. *In the partnership game of Figure 2 and the matching-pennies game of Figure 3, if $\frac{(a^2-b^2)+(1-\delta^2)a^2}{(a\delta+b)^2} < 1$, then p_n converges to the mixed Nash equilibrium almost surely.*

Proof. Note that in both games,

$$\beta = \frac{(a^2-b^2)+(1-\delta^2)a^2}{(a\delta+b)^2}.$$

Therefore, if $\frac{(a^2-b^2)+(1-\delta^2)a^2}{(a\delta+b)^2} < 1$, then G is a contraction mapping. Also, notice that condition (2) holds at p^* such that $p^*_{i,s_i} = \frac{1}{2}$ for each i and s_i, which corresponds to the mixed Nash equilibrium of the games. Since the fixed point is unique when G is a contraction mapping, the learning process converges to the mixed Nash equilibrium almost surely. □

Remark. If $a = 2$ and $b = 1$, the condition of Corollary 2 holds when $\delta > \frac{\sqrt{13}-1}{4} \approx 0.65$.

Figure 4: partnership game with a Pareto-dominant equilibrium

	s_2	t_2
s_1	$a+k, a+k$	b, b
t_1	b, b	a, a

Figure 5: dominant strategy game ($k > 1$)

	s_2	t_2
s_1	a, a	b, ka
t_1	ka, b	kb, kb

In general, for any $a, b > 0$, the condition holds if δ is close to one, in that the reinforcement learning process converges to the mixed Nash equilibrium almost surely in partnership games without a Pareto-dominant Nash equilibrium and matching-pennies games.[5]

It is worth noting that in general, the learning process does not converge to a Nash equilibrium. To investigate the convergence properties further, we first focus on partnership games with a Pareto-dominant Nash equilibrium and then focus on games with a dominant strategy. In the following argument, let $\sigma^* = (\sigma^*_{i,s_i})_{i,s_i} \in \Delta$ denote a mixed Nash equilibrium.

First, consider the partnership game of Figure 4 with $k > 0$ and $a > b$. In the game,

$$\beta = \frac{(2-\delta^2)(a+k)a - b^2}{(a\delta + b)^2}.$$

Therefore, for small enough k, if δ is close to 1, we have $\beta < 1$ and thus G becomes a contraction mapping and the learning process converges to the fixed point.[6] If we assume that $\delta = 1$, then at the fixed point, we have

$$p^*_{i,s_i} = \frac{k - 2b + \sqrt{4b^2 + k^2}}{2k}.$$

In addition, when $a = 2$, $b = 1$ and $k = 1$,

$$p^*_{i,s_i} = \frac{-1+\sqrt{5}}{2} \approx 0.62.$$

Note that at the mixed Nash equilibrium, $\sigma^*_{i,s_i} = \frac{1}{3}$.

[5] Note that $\beta < 1$ when $\delta = 1$. In addition, since $\beta > 1$ at $\delta = 0$ and $\frac{\partial\beta}{\partial\delta} < 0$, there exists $\overline{\delta} \in (0,1)$ such that $\beta < 1$ for $\delta \in (\overline{\delta}, 1]$.

[6] $\beta = \frac{a(a+k)-b^2}{(a+b)^2}$ when $\delta = 1$ and becomes less than 1 if $k < 2b + 2\frac{b^2}{a}$. Since $\beta > 1$ at $\delta = 0$ and $\frac{\partial\beta}{\partial\delta} < 0$, for small enough k, $\beta < 1$ if δ is close to 1. When $a = 2$, $b = 1$ and $k = 1$, $\beta < 1$ if $\delta > \frac{-1+\sqrt{26}}{5} \approx 0.82$.

Next, consider the game of Figure 5 with $k > 1$ and $a > b$. If $a > kb$, then the game corresponds to the prisoner's dilemma game. Note that

$$\beta = \frac{(1-\delta^2)ka^2}{(1+k\delta)^2 b^2}.$$

Therefore, if δ is close to 1, $\beta < 1$ and thus the learning process converges to the fixed point.[7] If we assume that $\delta = 1$, then at the fixed point, we have

$$p^*_{i,s_i} = \frac{1}{1+k}.$$

In addition, if $k = 2$, $p^*_{i,s_i} = \frac{1}{3}$. Note that at the Nash equilibrium, $\sigma^*_{i,s_i} = 0$.

3.3 Convergence to a Nash equilibrium

In the previous section, we show that in several 2×2 games, the learning process converges when δ is close to one, that is, when each player does not distort the foregone payoff information so much. In addition, we show that the process converges to the mixed Nash equilibrium in partnership games without a Pareto-dominant Nash equilibrium and matching-pennies games. However, we also show that the process does not converge to a Nash equilibrium in partnership games with a Pareto-dominant Nash equilibrium or games with a dominant action. In this section, we provide conditions under which the process converges to a Nash equilibrium. In particular, we show that the fixed point of G, which is the convergence target, corresponds to a Nash equilibrium if and only if for each player, all the actions are chosen with equal probability at the equilibrium.

Proposition 2. *When G is a contraction mapping, the fixed point corresponds to a Nash equilibrium if and only if for each player, all the actions are chosen with equal probability at the equilibrium.*

Proof. We first assume that the fixed point corresponds to a Nash equilibrium. Since we assume that payoffs are positive, for any i, s_i and x, $(x_{i,s_i} + \delta_i(1 - x_{i,s_i}))\pi_i(s_i, x_{-i}) > 0$. Then, from condition (2), we have $p^*_{i,s_i} > 0$ for each i and s_i. As the expected payoffs of the actions which are chosen with positive probability at a mixed Nash equilibrium are equivalent, $\pi_i(s_i, p^*_{-i}) = \pi_i(t_i, p^*_{-i})$ for any s_i and t_i such that $p^*_{i,s_i} > 0$ and $p^*_{i,t_i} > 0$, and condition (2) is expressed as follows: for each i and s_i,

$$p^*_{i,s_i} = \frac{(p^*_{i,s_i} + \delta_i(1 - p^*_{i,s_i}))}{\sum_{t_i}(p^*_{i,t_i} + \delta_i(1 - p^*_{i,t_i}))} = \frac{p^*_{i,s_i} + \delta_i(1 - p^*_{i,s_i})}{1 + \delta_i|S_i| - \delta_i}.$$

[7] Since $\beta < 1$ if $\delta = 1$, $\beta > 1$ at $\delta = 0$ and $\frac{\partial \beta}{\partial \delta} < 0$, there exists $\overline{\delta}$ such that $\beta < 1$ if $\delta \in (\overline{\delta}, 1]$. When $a = 3$, $b = 1$ and $k = 2$, $\beta < 1$ if $\delta > \frac{-2+3\sqrt{42}}{22} \approx 0.79$.

By solving the equation, we have $p^*_{i,s_i} = \frac{1}{|S_i|}$ for each i and s_i. Conversely, we assume that for each player, all the actions are chosen with equal probability at a Nash equilibrium: $\sigma^*_{i,s_i} = \frac{1}{|S_i|}$ for any i and $s_i \in S_i$. Since at any mixed Nash equilibrium, the expected payoffs of the actions which are chosen with positive probability are equivalent, we have

$$\frac{(\sigma^*_{i,s_i} + \delta_i(1 - \sigma^*_{i,s_i}))\pi_i(s_i, \sigma^*_{-i})}{\sum_{t_i}(\sigma^*_{i,t_i} + \delta_i(1 - \sigma^*_{i,t_i}))\pi_i(t_i, \sigma^*_{-i})} = \frac{1}{|S_i|} = \sigma^*_{i,s_i}$$

for any i and s_i, which shows that the mixed Nash equilibrium is a fixed point of G. Since G is a contraction mapping, the fixed point uniquely exists and thus the fixed point corresponds to the mixed Nash equilibrium. □

3.4 Convergence to a regular quantal response equilibrium

In section 3.2, in partnership games without a Pareto-dominant Nash equilibrium and matching-pennies games, we show that the reinforcement learning process converges to a Nash equilibrium when each player does not distort the foregone payoff information so much. However, in the previous section, we show that the process does not converge to a Nash equilibrium in general. Instead, in this section, we introduce an alternative equilibrium concept, regular quantal response equilibrium, to which the learning process converges. First, we provide the definition of the equilibrium according to Goeree et al. (2005).

Definition 1. *A regular quantal response equilibrium of the normal form game* $(\mathcal{N}, S, \pi)$ *is a mixed action profile* $x^* = (x^*_{i,s_i})_{i,s_i}$ *such that for each* i *and* s_i*,*

$$x^*_{i,s_i} = f_{i,s_i}((\pi_i(t_i, x^*_{-i}))_{t_i \in S_i}),$$

where for each i*,* $f_i : \mathbb{R}^{|S_i|} \to \Delta_i$ *with* f_{i,s_i} *being the* s_i*-th component is a regular response function satisfying the following four conditions.*

1. *Interiority:* $f_{i,s_i}(y_i) > 0$ *for any* s_i *and* y_i*.*
2. *Continuity:* f_{i,s_i} *is continuous and differentiable.*
3. *Responsiveness:* $\frac{\partial f_{i,s_i}(y_i)}{\partial y_{i,s_i}} > 0$ *for any* s_i *and* y_i*.*
4. *Monotonicity:* $f_{i,s_i}(y_i) > f_{i,t_i}(y_i)$ *if* $y_{i,s_i} > y_{i,t_i}$ *for any* s_i *and* t_i*.*

In the following argument, we show that the fixed point of G corresponds to a regular quantal response equilibrium when $\delta = 1$. Therefore, in any normal form game, if G is a contraction mapping and there is no distortion on the foregone payoffs, the reinforcement learning process converges to the equilibrium almost surely.

Figure 6: Normalised 2 × 2 game

	s_2	t_2
s_1	$\frac{a^1_{11}}{a^1_{11}+a^1_{21}}, \frac{a^2_{11}}{a^2_{11}+a^2_{21}}$	$\frac{a^1_{12}}{a^1_{12}+a^1_{22}}, \frac{a^2_{21}}{a^2_{11}+a^2_{21}}$
t_1	$\frac{a^1_{21}}{a^1_{11}+a^1_{21}}, \frac{a^2_{12}}{a^2_{12}+a^2_{22}}$	$\frac{a^1_{22}}{a^1_{12}+a^1_{22}}, \frac{a^2_{22}}{a^2_{12}+a^2_{22}}$

Proposition 3. *For any normal form game, if $\delta = 1$, the fixed point of G, p^*, corresponds to a regular quantal response equilibrium.*

Proof. For each i, let $f_i : \mathbb{R}^{|S_i|} \to \Delta_i$ be a mapping such that for each s_i and $y_i \in \mathbb{R}^{|S_i|}$,

$$f_{i,s_i}(y_i) = \frac{g(y_{i,s_i})}{\sum_{t_i} g(y_{i,t_i})}$$

where letting $\pi_{\max} := \max_{i,s} \pi_i(s)$ and $\pi_{\min} := \min_{i,s} \pi_i(s)$, $g : \mathbb{R} \to \mathbb{R}$ is defined such that

$$g(x) = \begin{cases} x & \text{if } \pi_{\min} \leq x \leq \pi_{\max}, \\ \pi_{\max} e^{\frac{x-\pi_{\max}}{\pi_{\max}}} & \text{if } x > \pi_{\max}, \\ \pi_{\min} e^{\frac{x-\pi_{\min}}{\pi_{\min}}} & \text{if } x < \pi_{\min}. \end{cases}$$

Note that for any i, s_i and $p \in \Delta$, $\pi_i(s_i, p_{-i}) \in [\pi_{\min}, \pi_{\max}]$, $g(\pi_i(s_i, p_{-i})) = \pi_i(s_i, p_{-i})$ and $f_{i,s_i}((\pi_i(t_i, p_{-i}))_{t_i \in S_i}) = \frac{\pi_i(s_i,p_{-i})}{\sum_{t_i} \pi_i(t_i,p_{-i})}$. Therefore, condition (2) can be expressed as $p^*_{i,s_i} = f_{i,s_i}((\pi_i(t_i, p^*_{-i}))_{t_i \in S_i})$. To show that p^* is a regular quantal response equilibrium, we have to show that f_i is a regular quantal response function for each i. First, since $g(x) > 0$ for any x, the interiority condition holds. Second, since g is continuously differentiable and $\frac{\partial f_{i,s_i}}{\partial y_{i,t_i}}$ is continuous for each $t_i \in S_i$, the continuity condition holds.[8] Third, since $\frac{\partial f_{i,s_i}(y_i)}{\partial y_{i,s_i}} > 0$ for any s_i and y_i, the responsiveness condition holds. Lastly, note that if $y_{i,s_i} > y_{i,t_i}$, then we have $g(y_{i,s_i}) > g(y_{i,t_i})$, and thus $f_{i,s_i}(y_i) > f_{i,t_i}(y_i)$. Therefore, the monotonicity condition holds. □

Remark. In section 3.2, we have shown the fixed points of several 2 × 2 games when $\delta = 1$. From the proposition above, we now know that the fixed points correspond to regular quantal response equilibria of the games.

[8] Note that for each y_i,

$$\frac{\partial f_{i,s_i}(y_i)}{\partial y_{i,t_i}} = \begin{cases} \frac{g'(y_{i,s_i})(\sum_{u_i \neq s_i} g(y_{i,u_i}))}{(\sum_{u_i} g(y_{i,u_i}))^2} & \text{if } t_i = s_i, \\ \frac{-g(y_{i,s_i})g'(y_{i,t_i})}{(\sum_{u_i} g(y_{i,u_i}))^2} & \text{otherwise.} \end{cases}$$

Lastly, we focus on the 2 × 2 game of Figure 6, which is obtained from the game of Figure 1 after some payoff normalisation. In detail, the payoffs are normalised so that given the opponent player's action, the sum of each player's payoffs becomes 1. Note that the number of Nash equilibria and the dominance relation among actions are equivalent between the original and normalised games. We then show that given the normalisation and $\delta = 1$, the reinforcement learning process almost surely converges to a regular quantal response equilibrium in the game.

Corollary 3. *In the* 2×2 *game of Figure 6, if* $\delta = 1$*, the reinforcement learning process* p_n *almost surely converges to a regular quantal equilibrium of the game.*

Proof. Note that when $\delta = 1$, we have

$$\beta = \max_{i \in \{1,2\}} |\frac{a^i_{11}}{a^i_{11} + a^i_{21}} \frac{a^i_{22}}{a^i_{21} + a^i_{22}} - \frac{a^i_{12}}{a^i_{12} + a^i_{22}} \frac{a^i_{21}}{a^i_{11} + a^i_{21}}| < 1.$$

Therefore, from the results above, the reinforcement learning process converges to a regular quantal response equilibrium almost surely. □

4 Conclusion

In this paper, we consider the reinforcement learning model of Erev and Roth with foregone payoff information. Players observe not only the payoffs from chosen actions but also foregone payoffs, the ones from unchosen actions. They may not treat the payoff information from unchosen actions the same as information from chosen actions. We provide conditions under which the process converges to a mixed action profile at which each action is chosen with a probability proportional to its expected payoff. In particular, in symmetric 2 × 2 partnership games without a Pareto-dominant Nash equilibrium and matching-pennies games, when each player's discount factor on the foregone payoffs is close to one, the reinforcement learning process converges to the mixed Nash equilibrium almost surely. However, in general, the mixed action profile to which the process converges does not correspond to a Nash equilibrium: it corresponds to a Nash equilibrium if and only if for each player, all the actions are played with equal probability at the equilibrium. Instead, we show that the action profile corresponds to a perturbed equilibrium, the regular quantal response equilibrium of Goeree et al. (2005), when each player does not distort the foregone payoff at all. Therefore, if the conditions are satisfied, the reinforcement learning process almost surely converges to a regular quantal response equilibrium in normal form games. In particular, given some payoff normalisation and the fact that each player does not distort the foregone payoff information, we obtain the convergence in any 2 × 2 game.

Appendix A The proof of Theorem 1.

We can rewrite the updating rule of the choice probability as follows:

$$\begin{aligned}
p_{n+1,i,s_i} &= \frac{w_{n+1,i,s_i}}{\sum_{t_i} w_{n+1,i,t_i}} \\
&= \frac{w_{n,i,s_i} + \pi_{n,i,s_i}}{\sum_{t_i} w_{n,i,{}_i} + \sum_{t_i} \pi_{n,i,t_i}} \\
&= p_{n,i,s_i} + \frac{1}{\sum_{t_i} w_{n+1,i,t_i}} \big(\pi_{n,i,s_i} - p_{n,i,s_i} \sum_{t_i} \pi_{n,i,t_i}\big) \\
&= p_{n,i,s_i} + \frac{1}{\sum_{t_i} w_{n+1,i,t_i}} \big(\overline{\pi}_{n,i,s_i} - p_{n,i,s_i} \sum_{t_i} \overline{\pi}_{n,i,t_i} + d_{n,i,s_i}\big) \\
&= p_{n,i,s_i} + \frac{\sum_{t_i} \overline{\pi}_{n,i,t_i}}{\sum_{t_i} w_{n+1,i,t_i}} \big(\frac{\overline{\pi}_{n,i,s_i}}{\sum_{t_i} \overline{\pi}_{n,i,t_i}} - p_{n,i,s_i} + D_{n,i,s_i}\big) \\
&= p_{n,i,s_i} + \frac{\sum_{t_i} \overline{\pi}_{n,i,t_i}}{\sum_{t_i} w_{n,i,t_i}} \big(\frac{\overline{\pi}_{n,i,s_i}}{\sum_{t_i} \overline{\pi}_{n,i,t_i}} - p_{n,i,s_i} + D_{n,i,s_i} + \eta_{n,i,s_i}\big)
\end{aligned}$$

where for each n, i and s_i,

$$\begin{aligned}
d_{n,i,s_i} &:= \big(\pi_{n,i,s_i} - p_{n,i,s_i} \sum_{t_i} \pi_{n,i,t_i}\big) - \big(\overline{\pi}_{n,i,s_i} - p_{n,i,s_i} \sum_{t_i} \overline{\pi}_{n,i,t_i}\big) \\
D_{n,i,s_i} &:= \frac{d_{n,i,s_i}}{\sum_{t_i} \overline{\pi}_{n,i,t_i}} \\
\eta_{n,i,s_i} &:= -\frac{\sum_{t_i} \pi_{n,i,t_i}}{\sum_{t_i} w_{n,i,t_i} + \sum_{t_i} \pi_{n,i,t_i}} \big(\frac{\overline{\pi}_{n,i,s_i}}{\sum_{t_i} \overline{\pi}_{n,i,t_i}} - p_{n,i,s_i} + \frac{d_{n,i,s_i}}{\sum_{t_i} \overline{\pi}_{n,i,t_i}}\big).
\end{aligned}$$

Lemma 1. *For each i and s_i:*

(i) $\{(D_{n,i,s_i})\}$ is a martingale difference sequence and there exists K such that

$$\mathbb{E}[(D_{n,i,s_i})^2 \mid \mathcal{F}_n] < K \tag{3}$$

for each n;

(ii) η_{n,i,s_i} converges to zero almost surely;

(iii) $\{\lambda_{n,i,s_i}\}$, where $\lambda_{n,i,s_i} := \frac{\sum_{t_i} \overline{\pi}_{n,i,t_i}}{\sum_{t_i} w_{n,i,t_i}}$ for each n, satisfies the following conditions:

(a) λ_{n,i,s_i} is $\mathcal{F}_n$-measurable for each n;

(b) $\sum_n \lambda_{n,i,s_i} = \infty$ and $\sum_n (\lambda_{n,i,s_i})^2 < \infty$ almost surely.

Proof. (i) It is easy to check that for each n, i and s_i, d_{n,i,s_i} is $\mathcal{F}_{n+1}$-measurable, $\mathbb{E}[d_{n,i,s_i} \mid \mathcal{F}_n] = 0$ and there exists C such that $\mathbb{E}[(d_{n,i,s_i})^2 \mid \mathcal{F}_n] < C$, as the set of actions is finite. Therefore, $\{D_{n,i,s_i}\}$ is a martingale difference sequence and satisfies condition (3). (ii) For each n, i, s_i, since the payoffs, martingale difference sequence and choice probabilities are bounded and the sum of the weights ($\sum_{t_i} w_{n,i,t_i}$) goes to infinity as n grows, η_{n,i,s_i} converges to zero almost surely.
(iii) For each n, i and s_i, since $\overline{\pi}_{n,i,s_i}$ and w_{n,i,s_i} are $\mathcal{F}_n$-measurable, λ_{n,i,s_i} is also $\mathcal{F}_n$-measurable. Also, note that for each n, i and s_i,

$$\frac{\min_i \delta_i \sum_{t_i} \min_{s_{-i}} \pi_i(t_i, s_{-i})}{n \sum_{t_i} \max_{s_{-i}} \pi_i(t_i, s_{-i})} \leq \frac{\sum_{t_i} \overline{\pi}_{n,i,t_i}}{\sum_{t_i} w_{n,i,t_i}} \leq \frac{\sum_{t_i} \max_{s_{-i}} \pi_i(t_i, s_{-i})}{n \min_i \delta_i \sum_{t_i} \min_{s_{-i}} \pi_i(t_i, s_{-i})}$$

almost surely and thus condition (b) holds. □

Therefore, by utilising the asynchronous stochastic approximation method of Tsitsiklis (1994), we can show that p_n converges to p^* if condition (1) holds. Note that comparing with the asynchronous stochastic process of Tsitsiklis (1994), we have an extra term $\eta_n = (\eta_{n,i,s_i})_{i,s_i}$ which converges to zero almost surely. It is easy to show that the result still holds with the noise. In the following section, we briefly introduce the asynchronous stochastic approximation method with the modification.

Appendix B The asynchronous stochastic approximation method of Tsitsiklis (1994)

In this section, we introduce a modified result of Tsitsiklis [16]: with an additional noise term, which disappears with probability one, we provide conditions under which the asynchronous stochastic approximation process converges uniquely with probability one. To introduce the method, we follow the argument of Tsitsiklis (1994).

Consider the updating rule of a vector $Q \in \mathbb{R}^M$ which consists of M components: $Q = (Q_1, ..., Q_M)$. Let $F : \mathbb{R}^M \to \mathbb{R}^M$ be a mapping with M components: $F = (F_1, ..., F_M)$ where $F_m : \mathbb{R}^M \to \mathbb{R}$ is the m-th component of F. Let $Q_{n,m}$ be the m-th component of Q_n, which is the value of Q in period n. Then the updating rule of the value is defined as follows:

$$Q_{n+1,m} = Q_{n,m} + \alpha_{n,m}\Big(F_m(Q_{n,m}) - Q_{n,m} + \eta_{n,m} + \omega_{n,m}\Big),$$

where (i) $\alpha_{n,m} \in [0,1]$ is a weighting parameter of the m-th component in period n; and (ii) $\eta_{n,m}$ and $\omega_{n,m}$ are noise terms. Now we provide additional assumptions for the convergence result:

(a) Q_0 is $\mathcal{F}_0$-measurable;

(b) there exists some D_0 such that $\|Q_n\|_\infty \leq D_0$ for all n;

(c) $\lim_{n\to\infty} \eta_{n,m} = 0$ with probability one for all m;

(d) $\omega_{n,m}$ is $\mathcal{F}_{n+1}$-measurable for all n and m;

(e) $E[\omega_{n,m} \mid \mathcal{F}_n] = 0$ for all m;

(f) there exist constants A and B such that

$$E[\omega_{n,m}^2 \mid \mathcal{F}_n] \leq A + B \max_{l\leq n} ||Q_l||_\infty^2 \quad \forall n, m;$$

(g) $\alpha_{n,m}$ is $\mathcal{F}_n$-measurable for all m;

(h) for all m,

$$\sum_{n=0}^{\infty} \alpha_{n,m} = \infty \quad \text{w.p.1;}$$

(i) there exists some constant C such that for all m,

$$\sum_{n=0}^{\infty} \alpha_{n,m}^2 \leq C \quad \text{w.p.1;}$$

(j) there exists a scalar $\beta \in [0, 1)$ and a vector $Q^* \in \mathbb{R}^*$ such that

$$\|F(Q) - Q^*\|_\infty \leq \beta\|Q - Q^*\|_\infty, \forall Q \in \mathbb{R}^M.$$

Then we have the following result.

Proposition 4. *If assumptions (a) to (j) hold, then $Q_n \to Q^*$ a.s..*

Proof. In the following argument, we extend the argument in Section 6 of Tsitsiklis (1994) with additional noise term η_n.

To prove the claim, we first introduce Lemma 1 of Tsitsiklis (1994):

Lemma (Lemma 1, Tsitsiklis, 1994). Let $\{\mathcal{F}_n\}$ be an increasing sequence of σ-fields. For each n, let α_n, ω_{n-1} and B_n be $\mathcal{F}_n$-measurable scalar random variables. Let C be a deterministic constant. Suppose that the following hold with probability 1:

1. $E[\omega_n \mid \mathcal{F}_n] = 0$;
2. $E[\omega_n^2 \mid \mathcal{F}_n] \leq B_n$;
3. $\alpha_n \in [0, 1]$;

4. $\sum_{n=0}^{\infty} \alpha_n = \infty$;
5. $\sum_{n=0}^{\infty} \alpha_n^2 \leq C$.

Suppose that the sequence $\{B_n\}$ is bounded with probability 1. Let W_n satisfy the recursion

$$W_{n+1} = (1 - \alpha_n) W_n + \alpha_n \omega_n.$$

Then $\lim_{n\to\infty} W_n = 0$ with probability 1.

Without loss of generality, we assume that $Q^* = 0$. Fix some $\varepsilon > 0$ and $\eta > 0$ such that $\beta(1 + 2\varepsilon + 2\eta) < 1$. We now define

$$D_{k+1} = \beta(1 + 2\varepsilon + 2\eta) D_k, \quad k \geq 0.$$

It is obvious that D_k converges to zero.

Suppose that there exists some period n_k such that $\|Q_n\|_\infty \leq D_k$ for all $n \geq n_k$. We will show that this implies that there exists some period n_{k+1} such that $\|Q_n\|_\infty \leq D_{k+1}$ for all $n \geq n_{k+1}$. This will complete the proof of convergence of Q_n to zero.

Let $W_m(0) = 0$ and

$$W_m(n+1) = (1 - \alpha_{n,m}) W_m(n) + \alpha_{n,m} \omega_{n,m}.$$

By Lemma 1 of Tsitsiklis (1994), we have that $\lim_{n\to\infty} W_m(n) = 0$. For any period n_0, we also define $W_m(n_0; n_0) = 0$ and

$$W_m(n+1; n_0) = (1 - \alpha_{n,m}) W_m(n; n_0) + \alpha_{n,m} \omega_{n,m}, \quad n \geq n_0.$$

Following the same argument as in the proof of Lemma 2 of Tsitsiklis (1994), we have that for any $\delta > 0$, there exists some N such that $|W_m(n; n_0)| < \delta$ for all $n_0 \geq N$ and $n \geq n_0$.[9]

Let $\nu_k \geq n_k$ be such that $|W_m(n; \nu_k)| \leq \beta\varepsilon D_k$ and $\|Q_{n,m}\|_\infty \leq D_k$ for all $n \geq \nu_k$ and all m. Since $\lim_{n\to\infty} \eta_{n,m} = 0$ for all $m \in \{1, ..., M\}$, we can find ν_k' such that $|\eta_{n,m}| \leq \beta\eta D_k$ for all $n \geq \nu_k'$ and m. Let $\mu_k = \max\{\nu_k, \nu_k'\}$.

We now define $Y_m(\mu_k) = D_k$ and

$$Y_m(n+1) = (1 - \alpha_{n,m}) Y_m(n) + \alpha_{n,m} \beta(1 + \eta) D_k, \quad n \geq \mu_k.$$

Lemma 2.

$$-Y_m(n) + W_m(n; \mu_k) \leq Q_{n,m} \leq Y_m(n) + W_m(n; \mu_k) \quad \forall n \geq \mu_k.$$

[9] Since $W_m(n; 0) = \prod_{\tau=n_0}^{n-1} (1 - \alpha_{\tau,m}) W_m(n_0, 0) + W_m(n; n_0)$, we have $|W_m(n; n_0)| \leq |W_m(n; 0)| + |W_m(n_0; 0)|$. Therefore, we can pick N large such that $|W_m(n, n_0)| < \delta$ for all n_0. Also, see Lemma 2 of Tsitsiklis (1994).

Proof: We use induction on n. Since $Y_m(\mu_k) = D_k$ and $W_m(\mu_k; \mu_k) = 0$, the result is true for $n = \mu_k$. Suppose that the equation holds for some $n \geq \mu_k$. We then have

$$\begin{aligned} Q_{n+1,m} &\leq (1 - \alpha_{n,m})(Y_m(n) + W_m(n; \mu_k)) + \alpha_{n,m}\beta D_k + \alpha_{n,m}\omega_{n,m} + \alpha_{n,m}\beta\eta D_k \\ &= Y_m(n+1) + W_m(n+1; \mu_k). \end{aligned}$$

A symmetric argument yields $Q_{n+1,m} \geq -Y_m(n+1) + W_m(n+1; \mu_k)$ and the inductive proof is complete. □

It is obvious from the recursive definition of $Y_m(n)$ and assumptions that $Y_m(n)$ converges to $\beta(1+\eta)D_k$ as $n \to \infty$. Therefore, from this and the result of Lemma 2, we have that

$$\limsup_{n\to\infty} |Q_{n,m}| \leq \beta(1 + \varepsilon + \eta)D_k < D_{k+1}.$$

□

Appendix C The proof of Proposition 1.

Since there exist only two actions available for both players, we make the notation simpler by letting $p_i := p_{i,s_i}$ and $p_{-i} := p_{-i,s_{-i}}$. First, G is expressed as follows: for $i \in \{1, 2\}$ and $s_i \in S_i$,

$$G_{i,s_i}(p) = \frac{\delta_{s_i,p}X_p}{\delta_{s_ip}X_p + \delta_{t_ip}Y_p}$$

where:

- $\delta_{s_i,p} := p_i + \delta_i(1 - p_i)$, $\delta_{t_i,p} := (1 - p_i) + \delta_i p_i$,
- $X_p := (a^i_{11}p_{-i} + a^i_{12}(1 - p_{-i}))$,
- $Y_p := (a^i_{21}p_{-i} + a^i_{22}(1 - p_{-i}))$.

Next, we consider the absolute difference of G_{i,s_i} for p and p'. Then

$$|G_{i,s_i}(p) - G_{i,s_i}(p')| \leq \frac{|(\delta_{s_i,p}\delta_{t_i,p'} - \delta_{s_i,p'}\delta_{t_i,p})X_pY_{p'}| + |\delta_{s_i,p'}\delta_{t_i,p}(X_pY_{p'} - X_{p'}Y_p)|}{K_{s_i,p}K_{s_i,p'}}$$

where $K_{s_i,p} := \delta_{s_i,p}X_p + \delta_{t_i,p}Y_p$. Note that

$$|(\delta_{s_i,p}\delta_{t_i,p'} - \delta_{s_i,p'}\delta_{t_i,p})X_pY_{p'}| \leq (1 - \delta_i^2)(\max_{j,k\in\{1,2\}} a^i_{1j}a^i_{2k})|p_i - p'_i|,$$

$$|X_pY_{p'} - X_{p'}Y_p| = |a^i_{11}a^i_{22} - a^i_{12}a^i_{21}||p_{-i} - p'_{-i}|$$

and

$$K_{s_i,p}K_{s_i,p'} \geq \Big(\min_{k\in\{1,2\},x\in\{0,1-\delta_i\}} \big((\delta_i + x)a^i_{1k} + (1-x)a^i_{2k}\big)\Big)^2.$$

Therefore, letting

$$\beta := \max_{i\in\{1,2\}} \frac{|a^i_{11}a^i_{22} - a^i_{12}a^i_{21}| + (1-\delta_i^2)\max_{j,k\in\{1,2\}} a^i_{1j}a^i_{2k}}{\Big(\min_{k\in\{1,2\},x\in\{0,1-\delta_i\}} \big((\delta_i + x)a^i_{1k} + (1-x)a^i_{2k}\big)\Big)^2}$$

we have

$$|G_{i,s_i}(p) - G_{i,s_i}(p')| \leq \beta||p-p'||_\infty.$$

Since $G_{i,t_i}(p) = (1 - G_{i,s_i}(p))$, G is a contraction mapping if $\beta < 1$.

Acknowledgements

I am grateful to Andrea Collevecchio, Marco LiCalzi and Massimo Warglien for their support at the Department of Management, Ca' Foscari University of Venice. I also thank seminar audiences at Monash University for suggestions and helpful comments. Financial support from the MatheMACS project (FP7-ICT, #318723), the COPE project (supported by a Sapere Aude grant from the Danish Research Council for Independent Research) and Ca' Foscari University of Venice is gratefully acknowledged. All remaining errors are mine.

References

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

[2] Camerer, C., Ho, T. H., 1999. Experience-weighted attraction learning in normal form games. Econometrica 67, 827–874.

[3] Erev, I., Roth, A. E., 1998. Predicting how people play games: reinforcement learning in experimental games with unique mixed strategy equilibria. American Economic Review 88, 848–881.

[4] Freedman, D. A., 1965. Bernard Friedman's urn. The Annals of Mathematical Statistics 36, 956–970.

[5] Fudenberg, D., Kreps, D. M., 1993. Learning mixed equilibria. Games and Economic Behavior 5, 320–367.

[6] Funai, N., 2018. Convergence results on stochastic adaptive learning. Economic Theory, https://doi.org/10.1007/s00199-018-1150-8.

[7] Goeree, J. K., Holt, C. A., Palfrey, T. R., 2005. Regular quantal response equilibrium. Experimental Economics 8, 347–367.

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

[9] Hopkins, E., 2002. Two competing models of how people learn in games. Econometrica 70, 2141–2166.

[10] Hopkins, E., Posch, M., 2005. Attainability of boundary points under reinforcement learning. Games and Economic Behavior 53, 110–125.

[11] Ianni, A., 2014. Learning strict Nash equilibria through reinforcement. Journal of Mathematical Economics 50, 148–155.

[12] Laslier, J. F., Topol, R., Walliser, B., 2001. A behavioural learning process in games. Games and Economic Behavior 37, 340–366.

[13] Roth, A. E., Erev, I., 1995. Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. Games and Economic Behavior 8, 164–212.

[14] Rustichini, A., 1999. Optimal properties of stimulus-response learning models. Games and Economic Behavior 29, 244–273.

[15] Simonsohn, U., Karlsson, N., Loewenstein, G., Ariely, D., 2008. The tree of experience in the forest of information: overweighing experienced relative to observed information. Games and Economic Behavior 62, 263–286.

[16] Tsitsiklis, J. N., 1994. Asynchronous stochastic approximation and Q-learning. Machine Learning 16, 185–202.