Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Economic Theory 124 (2005) 129–148
JOURNAL OF Economic Theory
www.elsevier.com/locate/jet

Learning to play games in extensive form by valuation

Philippe Jehiel*, Dov Samet

CERAS-ENPC, CNRS (URA 2036), 48 Bd. Jourdan, 75014 Paris, France

Received 3 September 2003; final version received 1 September 2004
Available online 29 December 2004

Abstract

Game theoretic models of learning which are based on the strategic form of the game cannot explain learning in games with large extensive form. We study learning in such games by using valuation of moves. A valuation for a player is a numeric assessment of her moves that purports to reflect their desirability. We consider a myopic player, who chooses moves with the highest valuation. Each time the game is played, the player revises her valuation by assigning the payoff obtained in the play to each of the moves she has made. We show for a repeated win–lose game that if the player has a winning strategy in the stage game, there is almost surely a time after which she always wins. When a player has more than two payoffs, a more elaborate learning procedure is required. We consider one that associates with each move the average payoff in the rounds in which this move was made. When all players adopt this learning procedure, with some perturbations, then, with probability 1 there is a time after which strategies that are close to subgame perfect equilibrium are played. A single player who adopts this procedure can guarantee only her individually rational payoff.

Keywords: Reinforcement learning; Valuation; Perfect equilibrium

1. Introduction

1.1. Moves vs. strategies

Game theory has developed scores of models which describe how players learn to play games. But invariably, these models describe learning in terms of the strategic form of the

* Corresponding author.
E-mail address: jehiel@enpc.fr (P. Jehiel).

0022-0531/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jet.2004.09.004

game. [1] Implementing these learning models, say by computer programs, requires that the strategic form of the game is used as an input. This is, of course, practically impossible for games in extensive form, the strategic form of which is too big to be effectively described. Thus, game theory has not yet provided an explanation of learning in such games.

This explains why game theory has ignored the developing of learning programs in artificial intelligence, starting with the first such program by Samuel [18]—the checkers-playing learning program, and ending with the chess-playing program "deep-blue". We do not know a single game theoretic study that proposes a rigorous theoretic explanation of the success of these programs or indicates the way to such theory. Here, we are still far from being able to provide such an explanation, but we hope that we are providing a first step in the right direction.

In contrast to the existing learning models in game theory, we base our model not on the strategic form, but rather on the *moves* in the games. As a result the models employed here can be *effectively* implemented for games of any size. [2]

1.2. Reinforcement vs. response

The other way in which this paper differs from most of the learning models in game theory is the data used by the player for learning.

Models of learning in games fall roughly into two categories. In the first, the learning player forms beliefs about the future behavior of other players and nature, and directs her behavior according to these beliefs. We refer to these as response models. In the second, the player is attuned only to her own performance in the game, and uses it to improve future performance. These are called models of reinforcement learning.

Reinforcement learning has been used extensively in artificial intelligence (AI). Samuel's [18] checkers-playing learning program marks the beginning of reinforcement learning algorithms. Since then many other sophisticated algorithms, heuristics, and computer programs, have been developed, based on reinforcement learning (see [20]). Such playing programs try neither to learn the behavior of a specific opponent, nor to find the distribution of the opponents' behavior in the population. Instead, they learn how to improve their play from the achievements of past behavior.

Until recently, game theorists studied mostly response models. Reinforcement learning has only attracted the attention of game theorists in the last decade in theoretical works like Gilboa and Schmeidler [10], Börgers and Sarin [1], Sarin and Vahid [19] or Karandikar et al. [15] and Cho and Matsui [3], and in experimental works like Erev and Roth [5] and

[1] This is true even for the few studies of learning in games that are given in extensive form. See Fudenberg and Levine [9] for a survey of these studies. In the context of evolutionary models, Hart [11] may be viewed as exception, as he provides an analysis of extensive form games based on the agent-normal form (one different player per node), and thus uses moves rather than strategies as the basic building block. See Cressman [4] for a recent account of evolutionary approaches in game theory.

[2] The concentration of the AI literature on moves rather than strategies is the main reason why there seems to be almost no overlap between two major books on learning, each in its field: *The Theory of Learning in Games*, Fudenberg and Levine [6] and *Reinforcement Learning: An Introduction*, Sutton and Barto [20].

Camerer and Ho [2]. [3] In all these studies the basic model is given in a strategic form, and the learning player reinforces those of her strategies that perform better. This approach, as we argued before, is inadequate where learning of games in extensive form is concerned. Here, as opposed to all the game theoretic models of reinforcement it is the moves of the game that are reinforced and not the strategies.

Reinforcement learning, and concentrating on moves rather than strategies is typical not only of the AI learning models. Consider the very different context of a 2 year old toddler learning how to operate a DVD player, with his efforts being frustrated by two highly rational and strategic players, mom and dad, and perhaps also by nature in the form of the family cat. Our toddler is oblivious of the strategic aspects of the situation. She concentrates mostly on the possible moves available to her, exhibiting reinforcement learning by remembering the button pushes that terminated in a successful operation of the device, and learning how to use them in the right sequence in order to reach the desired goal: watching "A Beautiful Mind".

1.3. Valuation

One of the most common building blocks of AI heuristics for reinforcement learning is the *valuation*, which is a real valued function on the possible moves of the learning player. The valuation of a move reflects, very roughly, the desirability of the move. Given a valuation, a learning process can be defined by specifying two rules:

- A *strategy rule*, which specifies how the game is played for any given valuation function of the player.
- A *revision rule*, which specifies how the valuation is revised after playing the game.

Our purpose here is to study learning-by-valuation processes, based on simple strategy and revision rules. In particular, we want to demonstrate the convergence properties of these processes in repeated games, where the stage game is given in an extensive form with perfect information and any number of players. Converging results of the type we prove here are very common in the literature of game theory. But as noted before, convergence of reinforcement is limited in this literature to strategies rather than moves. Since there is no obvious way to define a valuation of a strategy from a system of move valuations, a simple translation of our learning model in terms of strategies is not straightforward. [4]

---

[3] While Gilboa Schmeidler [10] study an axiomatization motivated by reinforcement learning, Börgers and Sarin [1] establish some connections between certain stochastic versions of reinforcement learning and the replicator dynamics. Karandikar et al. [15] study a learning model based on evolving aspirations (see also Cho and Matsui [3]: As in all reinforcement learning models, the learning player bases her strategy solely on her past performance, but in addition she keeps playing the same strategy (up to perturbations) as long as the strategy gives more than the current level aspiration level (assumed to evolve according to some averaging of past payoffs).

[4] To illustrate the difficulties, consider first the case in which a player must move at several nodes, and consider a path that crosses a move m of this player. In our setting, after this path has been played the valuation of move m is revised. Thus, the assessments of all the strategies that specify the move m are affected. In contrast, when strategies are reinforced, only the valuation of the strategy chosen is revised. Consider next the case where there is a different player at every node. In our setting, when a node is not reached the valuations of the corresponding moves are not revised. In the strategic form approach, all strategies are revised according to their performance no matter what the outcome is.

1.4. The main results

The strategy rule we adopt here is the *exploratory myopic strategy rule*. By this rule, the learning player chooses in each of her decision nodes, with high probability, a move which has the highest valuation among the moves available to her at this node. In case there are several moves with the highest valuation, she chooses one of them at random. But the player chooses also, with small probability, all other moves. [5]

As a revision rule we adopt the *averaging revision*. After each round the player revises only the valuation of the moves made in the round. The valuation of such a move is the average of the payoffs in all previous rounds in which this move was made.

Equipped with these rules, and an initial valuation, the player can play a repeated game. In each round she plays according to the exploratory myopic strategy, defined by the current valuation. At the end of the round she revises her valuation according to the averaging revision.

When one player learns: This learning process, together with the strategies of the other players in the repeated game, induce a probability distribution over the infinite histories of the repeated game. We show the following, with respect to this probability.

> If the learning player obeys the exploratory myopic strategy and the averaging revision rules, then starting with any valuation, there exists, with probability 1, a time after which the player's payoff exceeds her individually rational payoff (the minmax payoff) in the stage game, minus ε.

Thus, the learning process described yields the player approximately the payoff that she can guarantee even when the other players are disregarded. This result indicates that reinforcement learning achieves learning of playing the stage game itself, rather than playing against certain opponents. [6]

When all players learn: Our next result concerns the case where all the players learn how to play the stage game. By the previous result we know that each can guarantee his individually rational payoff. But, it turns out that the synergy of the learning processes yields a stronger convergence result. Indeed, players learn in this case each other's behavior and act rationally on this information.

> Suppose the stage game has a unique perfect equilibrium. If all the players employ the exploratory myopic strategy and the averaging revision rules, then starting with any valuation, with probability 1, there is a time after which their strategy in the stage game is close to the subgame perfect Nash equilibrium (SPNE). [7]

---

[5] The importance of trembles for learning in extensive form games was first noted by Fudenberg and Kreps [6] and Fudenberg and Levine [7]. Without trembles learning convergences to self-confirming equilibria rather than subgame perfect Nash Equilibria.

[6] The idea of deriving results for the behavior of a player irrespective of other players' strategies is in the spirit of universal consistency as defined in Fudenberg and Levine [8].

[7] It should be noted that convergence to the SPNE would also hold if we were to place in each node an agent of the player. This is so, because the stochastic process of valuations would be the same in both cases.

Learning and evolutionary models have had a mixed success in providing support for the SPNE. One main difficulty is that starting from the SPNE strategy profile, a strategy that differs only off the equilibrium path performs as well as the SPNE strategy. Thus, such strategies tend to increase in size through the mutation force, up to a point where the SPNE strategy profile gets unstable (see Noldeke and Samuelson [17] for an illustration). A recent paper by Hart [11] gives support to the SPNE for the case of large populations. As he shows in the large population case the evolutionary pressure dominates the mutation force and the SPNE obtains. Our learning model is different in nature from the evolutionary model both in that it does not require populations of agents [8] (representing each player) and in that the state of the learning system is unaffected at those nodes which are unreached in a given round (that is, there is no analog of the mutation force in our context).

1.5. Win–lose games

The class of win–lose game is of special interest because much effort has been invested in studying learning algorithm for such games. Also, learning to perform simple tasks, like operating a DVD player discussed in Section 1.2, can be modelled as win–lose game.

We study a somewhat larger class of stage games in which the learning player has only two payoffs, 1 (win) and 0 (lose). But no assumption is made on the number of the other players or their payoff functions.

By our main result we know that using the rules described above, the learning player can guarantee approximately her individually rational payoff. Obviously, this result has a bite only when this payoff is 1, that is, when the learning player can guarantee a win.

It turns out, though, that to achieve this result much simpler rules suffice. For a strategy rule we adopt the simple *myopic strategy rule*. This rule differs from the exploratory myopic strategy rule in that moves that do not have the highest valuation among the moves available to her at this node, are played with probability zero.

For a revision rule we use here the simple *memoryless revision*. Like in the averaging revision, after each round the player revises only the valuation of the moves made in the round. But here no averaging is done, and only the last round matters. The valuation of a move made in the last round becomes the player's payoff (0 or 1) in that round, regardless of previous valuations of the move:

> Suppose that the learning player can guarantee a win in the stage game. If she plays according to the myopic strategy and the memoryless revision rules, then starting with any nonnegative valuation, there exists, with probability 1, a time after which the player always wins.

Note, that no assumption is made on how the players, other than the learning player, play the game. In particular, the stochastic process generated in the repeated game is not necessarily a Markov process, and simple techniques of such processes cannot be used.

[8] While Hendon et al. [13] consider a fictitious play model leading to the SPNE, their model as acknowledged by the authors cannot be viewed as a learning model, since players keep updating the strategy of their opponent at nodes which are not reached in a given round. The authors provide a mental process interpretation of their model.

A simpler learning method we might consider for a win–lose game is one in which the learning player deletes her last move in each round when her payoff is 0. This method is not equivalent to the revision method we adopt here: when valuation is used, moves with valuation 0 my have valuation 1 in later rounds, while deleted moves do not reappear. Thus, assigning 0 valuation to a move is not the same as deleting it. But unlike the valuation method, the method of deleting moves does not lend itself to generalizations and seem to be a dead end. Obviously, it cannot be extended to games in which the learning player has more than two payoffs. Second, it cannot be extended to efficient learning models, even in games with 0–1 payoffs. In contrast, valuation can be used in many ways to form strategy and revision rules.

1.6. Information requirements

Although valuation is defined for all moves, the learning player needs no information concerning the game when she starts playing it. Indeed, the initial valuation can be constant, which does not require knowledge of the game. Starting with this valuation, the player needs to be informed of the moves that are possible to her only when it is her turn to play. During the repeated game, the player should be able to record the moves she made and their valuations. Still, the learning procedure does not require that the player knows how many players there are, let alone the moves they can make and their payoffs.

1.7. Efficiency

Unlike strategy-based learning models, the model studied here, which is move-based, can be *effectively* implemented by a computer program. Although the number of moves can be very large, there is no need to record them in advance. Instead, each can be recorded after being first encountered. However, this learning model will not be efficient for large games, because the time required to see a given move again is too long for practical purposes. In chess, for example, almost any state of the board, except for the first few, has been seen in recorded history only once.

In order to make the model more efficient, similarity of moves should be introduced. Thus, moves (or states of the board) should be considered similar if they share certain properties. In chess these can be the number of pieces on the board, for example, or more subtle features of the array. Now, when the valuation of a move is revised, so are the valuations of all the moves similar to it. Similarity of moves can be given exogenously, or preferably, change endogenously during the learning process. The strategic implication of similarity grouping as well as the properties of this similarity that can guarantee convergence of the learning process to a reasonable outcome should be the subject of further research. In a companion paper, Jehiel and Samet [14], we make a first step toward this.

2. Preliminaries

2.1. Games and super games

Consider a finite game G with complete information and a finite set of players I. The game is described by a tree (Z, N, r, A), where Z and N are the sets of terminal and non-terminal

nodes, correspondingly, the root of the tree is r, and the set of arcs is A. Elements of A are ordered pairs (n, m), where m is the immediate successor of n.

The set N_i, for $i \in I$, is the set of nodes in which it is i's turn to play. The sets N_i form a partition of N. The *moves* of player i at node $n \in N_i$ are the nodes in $M_i(n) = \{m|(n, m) \in A\}$. Denote $M_i = \cup_{n\in N_i} M_i(n)$. For each i the function $f_i: Z \to R$ is i's payoff function. The depth of the game is the length of the longest path in the tree. A game with depth 0 is one in which $\{r\} = Z$ and $N = \emptyset$.

A behavioral strategy, (strategy for short) for player i is a function σ_i defined on N_i such that for each $n \in N_i$, $\sigma_i(n)$ is a probability distribution on $M_i(n)$.

The super game Γ is the infinitely repeated game, with stage game G. An infinite history in Γ is an element of Z^ω. A finite history of t rounds, for $t \geqslant 0$, is an element of Z^t. A *super strategy* for player i in Γ is a function Σ_i on finite histories, such that for $h \in Z^t$, $\Sigma_i(h)$ is a strategy of i in G, played in round $t + 1$. The super strategy $\Sigma = (\Sigma_i)_{i\in I}$ induces a probability distribution on histories in the usual way.

2.2. Valuations

We fix one player i (the learning player) and omit subscripts of this player when the context allows it. We first introduce the basic notions of playing by valuation. A *valuation* for player i is a function $v: M_i \to R$.

Playing the repeated game Γ by valuation requires two rules that describe how the stage game G is played for a given valuation, and how a valuation is revised after playing G.

- A *strategy rule* is a function $v \to \sigma^v$. When player i's valuation is v, i's strategy in G is σ^v.
- A *revision rule* is a function $(v, h) \to v^h$, such that for the empty history Λ, $v^\Lambda = v$. When player i's initial valuation is v, then after a history of plays h, i's valuation is v^h.

Definition 1. The *valuation super strategy* for player i, induced by a strategy rule $v \to \sigma^v$, a revision rule $(v, h) \to v^h$, and an initial valuation v, is the super strategy Σ_i^v, which is defined by $\Sigma_i^v(h) = \sigma^{v^h}$ for each finite history h.

3. Main results

Our main results concern a learning procedure based on the δ-exploratory myopic strategy rule and the averaging revision rule to be defined in Section 3.2 below. Theorem 3 claims that a learning player can guarantee approximately her individually rational payoff. Theorem 4 claims that when all players learn using this procedure, then they play approximately a perfect equilibrium strategy.

For the special case of a win–lose game, Theorem 3 means that if the learning player can guarantee a win in the stage game, then she learns how to win with high probability. But for such games a much simpler learning procedure can guarantee that such a player learns to win for sure, namely, the procedure which involves the myopic strategy rule and the memoryless revision rule.

Because of the simplicity of the rules required for win–lose games, we present first the case of these games, and explain why they cannot be used for general payoff-structure games.

3.1. Win–lose games

We consider first the case where player *i* has two possible payoffs in *G*, which are, without loss of generality, 1 (win) and 0 (lose). A two-person win–lose game is a special case, but here we place no restrictions on the number of players or their payoffs.

We assume that learning by valuation is induced by a strategy rule and a revision rule of a simple form.

The myopic strategy rule. *This rule associates with each valuation v the strategy σ^v, where for each node $n \in N_i$, $\sigma^v(n)$ is the uniform distribution over the maximizers of v on $M_i(n)$. That is, in each node of player i, the player selects at random one of the moves with the highest valuation.*[9]

The memoryless revision rule. *For a history of length 1, $h = (z)$, the valuation v is revised to v^z which is defined for each node $m \in M_i(n)$ by*

$$v^z(m) = \begin{cases} f_i(z) & m \text{ is on the path leading from } r \text{ to } z, \\ v(m) & \text{otherwise}. \end{cases}$$

For a history $h = (z_1, \ldots, z_t)$, the current valuation is revised in each round according to the terminal node observed in this round. Thus, $v^h = \left(v_i^{(z_1,\ldots,z_{t-1})}\right)^{z_t}$.

The temporal horizons, future and past, required for these two rules are very narrow. Playing the game *G*, the player takes into consideration just her next move. The revision of the valuation after playing *G* depends only on the current valuation, and the result of this play, and not on the history of past valuations and plays. In addition, the revision is confined only to those moves that were made in the last round.

Theorem 1. *Let G be a game in which player i either wins or loses. Assume that player i has a strategy in G that guarantees him a win. Then for any nonnegative initial valuation v of i, and super strategies Σ in Γ, if Σ_i is the valuation super strategy induced by the myopic strategy and the memoryless revision rules, then with probability 1, there is a time after which i is winning forever.*

The following example demonstrates learning by valuation.

Example 1. Consider the game in Fig. 1, where the payoffs are player 1's.

[9] The requirement that σ uniformly selects one of the moves at n is not essential for our results. It is enough that σ assigns positive weight to every move available at n.

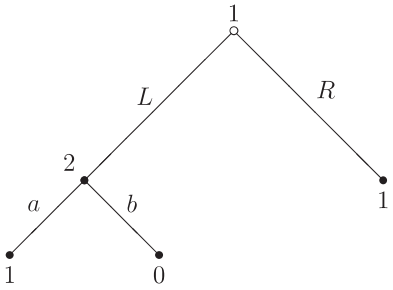

Fig. 1. Two payoffs.

Suppose that 1's initial valuation of each of the moves L and R is 0. The valuations that will follow can be one of $(0, 0)$, $(1, 0)$, and $(0, 1)$, where the first number in each pair is the valuation of L and the second of R. (As we shall see, the valuation $(1, 1)$ cannot be reached from any of these valuations.)

We can think of these possible valuations as states in a stochastic process. The state $(0, 1)$ is absorbing. Once it is reached, player 1 is choosing R and being paid 1 forever. When the valuation is $(1, 0)$, player 1 goes L. She will keep playing L, and winning 1, as long as player 2 is choosing a. Once player 2 chooses b, the valuation goes back to $(0, 0)$. Thus, the only way player 1 can fail to be paid 1 from a certain time on is when $(0, 0)$ recurs infinitely many times. But the probability of this is 0, as the probability of reaching the absorbing state $(0, 1)$ from state $(0, 0)$ is 1/2.

Note that the theorem does not state that with probability 1 there is a time after which player 1's strategy is the one that guarantees him payoff 1. Indeed, in this example, if player 2's strategy is always a, then there is a probability 1/2 that player 1 will play L for ever, which is not the strategy that guarantees player 1 the payoff 1.

3.2. The case of payoff functions with more than two values

We now turn to the case in which payoff functions take more than two values. The next example shows that in this case the myopic strategy and the memoryless revision rules may lead the player astray.

Example 2. Player 1 is the only player in the game in Fig. 2.

The player can guarantee a payoff of 10, and therefore we expect a learning process to yield eventually this payoff. But, in order to guarantee that the learning process induced by the myopic strategy and the memoryless revision results in the payoff 10 in the long run, the initial valuation should reflect the structure of the payoff. [10] If the initial valuation does not reflect it, for example, if it is constant, then there is a positive probability that the valuation $(-10, 2)$ for (L, R) is obtained, which is absorbing.

[10] The valuation of L should be greater than that of R, and the valuation of a should be greater than that of b.

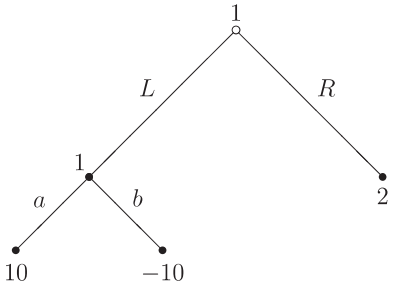

Fig. 2. More than two payoffs.

We cannot state for general payoff functions any theorem analogous to Theorem 1 or even a weaker version of this theorem. But something meaningful can be stated when *all* players play the repeated game according to the myopic strategy and the memoryless revision rules.

We say that game G is *generic* if for every player i and for every pair of distinct terminal nodes z and z', we have $f_i(z) \neq f_i(z')$.

Theorem 2. *Let G be a generic game. Assume that each player i plays Γ according to the myopic strategy rule and uses the memoryless revision rule. Then for any initial valuation profile, with probability 1, there is a time after which the same terminal node is reached in each round.*

The limit plays guaranteed by this theorem depend on the initial valuations and have no special structure in general. [11] Moreover, it is obvious that for any terminal node there are initial valuations that guarantee that this terminal node is reached in all rounds.

We return, now, to the case where only one player learns by reinforcement. In order to prevent a player from being paid an inferior payoff forever, like in Example 2, we change the strategy rule. We allow for exploratory moves that remind her of all possible payoffs in the game, so that she is not stuck permanently in a bad valuation. Assume, then, that having a certain valuation, the player opts for the highest valued nodes, but still allows for other nodes with a small probability δ. Such a rule guarantees that the player in Example 2 will never be stuck permanently in the valuation $(-10, 2)$. We introduce formally this new rule.

The δ-exploratory myopic strategy rule. *This rule associates with each valuation v the strategy σ_δ^v, where for each node $n \in N_i$, $\sigma_\delta^v(n) = (1-\delta)\sigma^v(n) + \delta\mu(n)$. Here, σ^v is the strategy associated with v by the myopic strategy rule, and μ is the strategy that uniformly selects one of the moves at n.* [12]

[11] The emergence of any possible pure outcome is reminiscent of Proposition 1 in Karandikar et al. [15] which was obtained in an evolving aspiration learning model (applied to the prisoner's dilemma). Observe though that unlike the evolving aspiration model in [15] our revision rule has nothing to do with inertia.

[12] Like in the definition of the myopic strategy rule, it is enough to require that μ assigns positive weight to every move available at n.

Unfortunately, adding exploratory moves alone does not help the player to achieve 10 in the long run, as we show now. Assume that the initial valuation of a and b is 10 and -10 correspondingly, and the valuation of the first two moves is also favorable: $(10, 2)$. We assume now that in each of the two nodes player 1 chooses the higher valued node with probability $1-\delta$ and the other with probability δ. The valuation of a and b cannot change over time. The valuation of (L, R) forms an ergodic Markov chain with the two states $\{(10, 2), (-10, 2)\}$. Thus, for example, the probability of transition from $(10, 2)$ to itself occurs when the player chooses either L and a, with probability $(1-\delta)^2$, or R with probability δ, which sum to $1-\delta+\delta^2$.

The following is the transition matrix of this Markov chain.

$$\begin{array}{c} \\ (10,2) \\ (-10,2) \end{array} \begin{array}{c} \begin{array}{cc} (10,2) & (-10,2) \end{array} \\ \begin{pmatrix} 1-\delta+\delta^2 & \delta-\delta^2 \\ \delta-\delta^2 & 1-\delta+\delta^2 \end{pmatrix} \end{array}$$

The two states $(10, 2)$ and $(-10, 2)$ are symmetric and therefore the stationary probability of each is 1/2. Thus, the player is paid 10 and 2, half of the time each.

Note that the exploratory moves are required because the payoff function has more than two values. However, we have shown that a learning player who adopts such a rule fails to achieve the payoff 10. Indeed, even in a win–lose game, a player who has a winning strategy may fail to guarantee a win in the long run by playing according to the rules of δ-exploratory myopic strategy and memoryless revision. To fix this problem we consider the following revision rule:

The averaging revision rule. *For a node $m \in M_i$, and a history $h = (z_1, \ldots, z_t)$, if the node m was never reached in h, then $v^h(m) = v(m)$. Else, let $t_1, \ldots, t_k$ be the times at which m was reached in h, then*

$$v^h(m) = \frac{1}{k} \sum_{l=1}^{k} f(z_{t_l}).$$

We state, now, that by using a little exploration and averaging revision, player i can guarantee a payoff which is above his individually rational (minmax) payoff in G minus ε.

Theorem 3. *Let Σ be a super strategy such that Σ_i is the valuation super strategy induced by the δ-exploratory myopic strategy and the averaging revision rules. Denote by P_δ the distribution over histories in Γ induced by Σ.*

Let ρ be i's individually rational payoff in G. Then for every $\varepsilon > 0$ there exists $\delta_0 > 0$ such that for every $0 < \delta < \delta_0$, for P_δ-almost all infinite histories $h = (z_1, z_2, \ldots)$,

$$\underline{\lim}_{t\to\infty} \frac{1}{t} \sum_{l=1}^{t} f(z_l) > \rho - \varepsilon.$$

We consider now the case where all players learn to play G, using the δ-exploratory myopic strategy and the averaging revision rules. We show that in such a case, in the long

run, the players' strategy in the stage game is close to a perfect equilibrium. We assume for simplicity that the game G has a unique perfect equilibrium (which is true generically).

Theorem 4. *Assume that G has a unique perfect equilibrium $\beta = (\beta_i)_{i\in I}$. Let Σ^δ be the super strategy such that for each i, Σ_i^δ is the valuation super strategy induced by the δ-exploratory myopic strategy, and the averaging revision rules.*

Let P_δ be the distribution over histories induced by Σ^δ. Then there exists δ_0, such that for all $0 < \delta < \delta_0$, for P_δ-almost all infinite histories $h = (z_1, \ldots, z_t, \ldots)$, there exists T, such that for all $t > T$, $\sigma_i^{v^{(z_1,\ldots,z_t)}}(m) = (1-\delta)\beta_i(m) + \delta\mu(m)$, for each player i and node $m \in M_i$.

4. Proofs

4.1. A sketch of the proof of Theorem 1

All the theorems are proved by induction on the depth of the game tree. We first sketch the main idea in the proof of Theorem 1.

Suppose that player i can guarantee a win in the game G, and she has the first move in this game (the other case is simpler). Then, at least one of her moves at the root of G guarantees her a win. Denote by G' a subgame that follows such a move. By the induction hypothesis, the theorem holds for the infinitely repeated game of G'.

Consider the vector of valuations of i's moves at the root. Assume that it is not the 0 vector. At each round in Γ, i chooses one of the moves that has positive valuation. If she wins, it remains positive (indeed, it is 1). If she loses the valuation of the move is reduced to zero. Thus, the set of moves with positive valuation can only shrink. Suppose that in a given history there is a time after which the vector does not become the zero vector. Then, at some later time the set of moves with positive valuation must be fixed, and from that time on i always wins.

Now suppose that in a given history this vector of valuations is 0 infinitely many times. At these times a move is chosen at random, and therefore with probability 1, G' is reached infinitely many times. We now apply the induction hypothesis.

There is a small flaw in the proof just described. The induction hypothesis is about the infinitely repeated game of G' and we need to apply it to histories in Γ. In these histories there are "gaps" between the consecutive times in which G' is played.

To overcome this problem we prove our theorems for a larger family of super games which we call stochastic repeated games. In such a super game, before each round of playing the stage game all the players observe some random signal. This solves the problem mentioned before, because the "gaps" between playing G' can be considered as signals rather than a play.

4.2. Stochastic repeated games

Let S be a countable set of states which also includes an *end state* e. A *stochastic repeated game* is a game Γ^S in which the game G is played repeatedly. Before each round a state

from S is selected according to a probability distribution which depends on the history of the previous terminal nodes and states. When the state e is realized the game ends. The selected state is known to all the players. The strategy played in each round depends on the history of the terminal nodes and states. We now describe Γ^S formally.

Histories. The set of infinite histories in Γ^S, is $H_\infty = (S \times Z)^\omega$. For $t \geqslant 0$ the set of finite histories of t rounds, is $H_t = (S \times Z)^t$, and the set of preplay histories of t rounds is $H_t^p = (S \times Z)^t \times S$. Denote $H = \cup_{t=0}^\infty H_t$ and $H^p = \cup_{t=0}^\infty H_t \times S$. The subset of H^p of histories that terminate with e is denoted by F. For $h \in H_\infty$ and $t \geqslant 0$ we denote by h_t the history in H_t which consists of the first t rounds in h. For finite and infinite histories h we denote by $\bar{h}$ the sequence of terminal nodes in h.

Transition probabilities. For each $h \in H$, $\tau(h)$ is a probability distribution on S. For $s \in S$, $\tau(h)(s)$ is the probability of transition to state s after history h. The probability that the game ends after h is $\tau(h)(e)$.

Super strategies. After t rounds the player observes the history of t pairs of a state and a terminal node, and the state that follows them, and then plays G. Thus, a super strategy for player i is a function Σ_i from $H^p \setminus F$ to i's strategies in G. We denote by $\Sigma(h)(z)$ the probability of reaching terminal node z when $\Sigma(h)$ is played.

The super play distribution. The super strategy Σ induces the *super play distribution* which is a probability distribution P over $H_\infty \cup F$. It is the unique extension of the distribution over finite histories which satisfies

$$P(h, s) = P(h)\tau(h)(s) \tag{1}$$

for $h \in H$, and

$$P(h^p, z) = P(h^p)\Sigma(h^p)(z) \tag{2}$$

for $h^p \in H^p$.

The valuation super strategy. Player i's valuation super strategy in Γ^S, starting with valuation v, is the super strategy Σ_i which satisfies $\Sigma_i(h) = \sigma^{v^{\bar{h}}}$.

4.3. Subgames

We show now how a stochastic repeated game of a subgame of G can be imbedded in Γ^S.

For a node n in G, denote by G_n the subgame starting at n. Fix a super strategy profile Σ in Γ^S and the induced super play distribution P on H_∞. In what follows we describe a stochastic super game $\Gamma_n^{S'}$, in which the stage game is G_n. For this we need to define the state space S'. We denote with primes histories and states in the game $\Gamma_n^{S'}$, as well as terminal nodes in G_n. Our purpose in this construction is to imbed H'_∞ in H_∞. The idea is to regard the rounds in a history h in H_∞ in which node n is not reached as states in S'.

Let S' be defined as the set of all $h^p \in H^p$, such that node n is never reached in h^p. Obviously, S' subsumes S, and in particular includes the end state e. Note that the infinite histories in $\Gamma_n^{S'}$, that is, the elements of H'_∞, can be naturally viewed as the histories in H_∞ in which the node n is repeated infinitely many times. Similarly, the finite histories in H' and H'^p can be identified with those in H an H^p correspondingly. We use this fact to define the transition probability distribution $\tau'(h)$ in $\Gamma_n^{S'}$ as follows.

For any $s' \neq e$ in S' and $h' \in H'$ with $P(h') > 0$,

$$\tau'(h')(s') = P(h', s' \mid h')\Sigma(h', s')(n), \tag{3}$$

where $\Sigma(h', s')(n)$ is the probability that node n is reached under the strategy profile $\Sigma(h', s')$. For e, $\tau'(h')(e) = P(E|h')$, where E consists of all histories $h \in H_\infty \cup F$ with initial segment h' such that n is never reached after this initial segment.

Note that $\tau'(h')(s')$ is the probability of all histories in $H_\infty \cup F$ that start with (h', s') and are followed by a terminal node of the game G_n. These events and the event E described above, form a partition of $H_\infty \cup F$, and therefore τ' is a probability distribution.

Claim 1. *Define a super strategy profile Σ' in $\Gamma_n^{S'}$, by*

$$\Sigma'(h'^p) = \Sigma_n(h'^p) \tag{4}$$

for each $h'^p \in H'^p$, where the right-hand side is the restriction of $\Sigma(h'^p)$ to G_n. Then, the restriction of P to H'_∞ coincides with the super play probability distribution P', induced by Σ'.

Proof. It is enough to show that P and P' coincide on H'. The proof is by induction on the length of $h' \in H'$. Suppose $P'(h') = P(h') > 0$ and consider the history (h, s', z'). Then, by the definition of the super play distribution (1) and (2),

$$P'(h', s', z') = P'(h')\tau'(h')(s')\Sigma'(h', s')(z').$$

By the induction hypothesis and the definitions of τ' in (3), the right-hand side is $P(h', s' \mid h)\Sigma(h', s')(n)\Sigma'(h', s')(z')$. By the definition of Σ' in (4), this is just $P(h', s')\Sigma(h', s')(n)$ $\Sigma_n(h', s')(z')$. The right-hand side, in turn, is just $P(h', s')\Sigma(h', s')(z') = P(h', s', z')$. □

Next, we note that playing by valuation is inherited by subgames.

Claim 2. *Suppose that i's strategy in Γ^S, Σ_i, is the valuation super strategy starting with v, and using either the myopic strategy and the memoryless revision rules, or the δ-exploratory myopic strategy and the averaging revision rules. Then the induced strategy in $\Gamma_n^{S'}$, Σ'_i, is the valuation super strategy starting with v_n—the restriction of v to the subgame G_n—and following the corresponding rules.*

Proof. The valuation super strategy in $\Gamma_n^{S'}$, starting with v_n, requires that after history $h' \in H'$, strategy $\sigma^{v_n^{\bar{h}'}}$ is played. Here, $\bar{h}'$ is the sequence of all terminal nodes in h', which consists of terminal nodes in G_n. These are also all the terminal nodes of G_n, in h', when the latter is viewed as a history in H.

When h' is considered as a history in H, then the strategy $\Sigma_i(h')$ is $\sigma^{v^{\bar{h}'}}$, where $\bar{h}'$ is the sequence of all terminal nodes in h'. $\Sigma'_i(h')$ is the restriction of $\sigma^{v^{\bar{h}'}}$ to G_n. But along the history h', the valuation of nodes in the game G_n does not change in rounds in which terminal nodes which are *not* in G_n are reached. Therefore, $\Sigma'_i(h')$ and $\sigma^{v^{\bar{h}'}}$ are the same. □

4.4. Theorems 1 and 2

The game Γ is in particular a stochastic repeated game, where there is only one state, besides e, and transition to e (that is, termination of the game) has null probability. We prove all three theorems for the wider class of stochastic repeated games. The theorems can be stated verbatim for this wider class of games, with one obvious change: any claim about almost all histories should be replaced by a corresponding claim for almost all *infinite* histories.

All the theorems are proved by induction on the depth of the game G. The proofs for games of depth 0 (that is, games in which payoffs are determined in the root, with no moves) are straightforward and are omitted. In all the proofs, $R = \{n_1, \ldots, n_k\}$ is the set of all the immediate successors of the root r.

Proof of Theorem 1. Assume that the claim of the theorem holds for all the subgames of G. We examine first the case that the first player is not i. By the stipulation of the theorem, player i can guarantee payoff 1 in each of the games G_{n_j} for $j = 1, \ldots, k$.

Consider now the game $\Gamma^{S'}_{n_j}$, the super strategy profile Σ', and the induced super play distribution P'. By the induction hypothesis, and Claim 2, for each j, for P'-almost all infinite histories there is a time after which player i is paid 1. In view of Claim 1, for P-almost all histories in Γ^S in which n_j is reached infinitely many times, there exist a time after which player i is paid 1, whenever n_j is reached. Consider now a nonempty subset Q of R. Let E_Q be the set of infinite histories in Γ^S in which node n_j is reached infinitely many times iff $n_j \in Q$. Then, for P-almost all histories in E_Q there is a time after which player i is paid 1. The events E_Q when Q ranges over all nonempty subsets of R, form a partition of the set of all infinite histories, which completes the proof in this case.

Consider now the case that i is the first player in the game. In this case there is at least one subgame G_{n_j} in which i can guarantee the payoff 1. Assume without loss of generality that this holds for $j = 1$.

For a history h denote by R^+_t the random variable that takes as values the subset of the nodes in R that have a positive valuation after t rounds. When R^+_t is not empty, then i chooses at r, with probability 1, one of the nodes in R^+_t. As a result the valuation of this node after the next round is 0 or 1, while the valuation of all other nodes does not change. Therefore we conclude that R^+_t is weakly decreasing when $R^+_t \neq \emptyset$. That is, $P(R^+_{t+1} \subseteq R^+_t | R^+_t \neq \emptyset) = 1$.

Let E^+ be the event that $R^+_t = \emptyset$ for only finitely many t's. Then, for P-almost all histories in E^+ there exists time T such that R^+_t is decreasing for $t \geqslant T$. Hence, for P-almost all histories in E^+ there is a nonempty subset R' of R, and time T, such that $R^+_t = R'$ for $t \geqslant T$. But in order for the set of nodes in R with positive valuation not to change

after T, player i must be paid 1 in each round after T. Thus we only need to show that $P(\bar{E^+}) = 0$.

Consider the event E^1 that n_1 is reached in infinitely many rounds. As proved before by the induction hypothesis, for P-almost all histories in E^1, there exists T, such that the valuation of n_1 is 1, for each round $t \geqslant T$ in which n_1 is reached. The valuation of this node does not change in rounds in which it is not reached. Thus, $E^1 \subseteq E^+$ P-almost surely.

We conclude that for P-almost all histories in $\bar{E^+}$ there is a time T, such that n_1 is not reached after time T. But P-almost surely for such histories there are infinitely many t's in which the valuation of all nodes in R is 0. In each such history, the probability that n_1 is not reached is $1 - 1/k$, which establishes $P(\bar{E^+}) = 0$. □

Proof of Theorem 2. Let i be the player at the root of G. By the induction hypothesis and Claim 1, for each of the supergames $\Gamma^{S'}_{n_j}$, $j = 1, \ldots, k$, for P'-almost infinite histories in this super game, there is a time after which the same terminal node is reached. By Claim 2, for P-almost all histories of Γ in which n_j recurs infinitely many times there is a time after which i's valuation of this node is constantly the payoff of the same terminal node of G_{n_j}.

It is enough that we show that for P-almost all infinite histories in Γ^S, there is a time after which the same node from R is selected with probability 1 at the root. Suppose that this is not the case. Then there must be a set of histories E with $P(E) > 0$, two nodes n_j and n_l, and two terminal nodes z_j and z_l in G_{n_j} and G_{n_l} correspondingly, that recur infinitely many times in this set. Therefore, for P-almost all histories in E, i's valuation of n_j and n_l is $f_i(z_j)$ and $f_i(z_l)$. Since G is generic, we may assume that $f_i(z_j) > f_i(z_l)$. Thus, for P-almost all histories in E, there is a time after which the conditional probability of n_l given the history is 0. Which is a contradiction. □

4.5. Theorems 3 and 4

We prove Theorem 3 for stochastic repeated games, where the conclusion of the theorem holds for P_δ-almost all *infinite* histories. We first consider a node n_j that follows the root, and histories in which this node recurs infinitely many times. Let ρ_j be i's individually rational payoff at n_j. We prove that for such histories, in the long run, i's average payoff at the times in which n_j was reached, denoted $\bar{f}^t_j$, is not lower than $\rho_j - \varepsilon$. Now, if i is not the player at the root, then i's individually rational payoff ρ is the minimum of the ρ_j's. Since i's average payoff $\bar{f}^t$ is an average over j of the averages $\bar{f}^t_j$, it follows that in the long run $\bar{f}^t$ is not lower than $\rho_j - \varepsilon$. If i is the player at the root, then ρ is the maximum over j of ρ_j. We show that conditional on any history the probability that player i expected payoff in that round is not lower than $\rho - \varepsilon$ is high. To conclude that this holds unconditionally on histories, we use a version of the strong law of large numbers for dependent variables.

Proof of Theorem 3. Assume that the claim holds for all the subgames of G. We denote by ρ_j, i's individually rational (maxmin) payoff in G_{n_j}.

We denote by $\bar{f}^t(h)$, i's average payoff at time t in history h. Fix a subgame G_{n_j}. Histories in the game $\Gamma^{S'}_{n_j}$ are denoted with primes. Thus, $\bar{f}^t(h')$ is i's average payoff at time t in history h' in $\Gamma^{S'}_{n_j}$.

Let h be a history in Γ in which n_j recurs infinitely many times at $t_1, t_2, \ldots$. Let $\bar{h} = (z_1, z_2, \ldots)$. Denote by $\bar{f}_j^t(h)$ i's average payoff until t *at the times* n_j *was reached*, that is,

$$\bar{f}_j^t(h) = \frac{1}{|\{l : t_l < t\}|} \sum_{l:t_l<t} f(z_{t_l}).$$

The history h can be viewed as an infinite history h' in $\Gamma_{n_j}^{S'}$. Moreover, for each l, $\bar{f}^l(h') = \bar{f}_j^{t_l}(h)$. By the definition of $\bar{f}_j^t(h)$, it follows that if there exists L such that for each $l > L$, $\bar{f}^l(h') > \rho_j - \varepsilon$, then there exits T such that for each $t > T$, $\bar{f}_j^t(h) > \rho_j - \varepsilon$. By the induction hypothesis there is δ_0, such that for all $0 < \delta < \delta_0$, for P'_δ-almost all histories h' there exists such an L. Thus, by Claims 1 and 2, there exists δ_0, such that for all j and $0 < \delta < \delta_0$, for P_δ-almost all histories h in Γ^S in which n_j recurs infinitely many times, there exists a time T such that for each $t > T$, $\bar{f}_j^t(h) > \rho_j - \varepsilon$.

We examine first the case that the first player is not i. Obviously, in this case, $\rho = \min_j \rho_j$.

Let Q be a nonempty subset of R, and let E_Q be the set of all infinite histories in which the set of nodes that recurs infinitely many times is Q. Consider a history h in E_Q, with $\bar{h} = (z_1, z_2, \ldots)$. Let $v_j^t(h)$ be the number of times n_j is reached in h until time t. Then,

$$\bar{f}^t(h) = \frac{1}{t} \sum_{j=1}^{k} v_j^t(h) \bar{f}_j^t(h) \geqslant \min_{j:n_j \in Q} \bar{f}_j^t(h),$$

where the inequality holds, because $\sum_j v_j^t(h) = t$, and for $j \notin Q$, $v_j^t(h) = 0$. Thus for P_δ-almost all histories h in E_Q,

$$\begin{aligned}
\varliminf_{t\to\infty} \bar{f}^t(h) &\geqslant \varliminf_{t\to\infty} \min_{j:n_j\in Q} \bar{f}_j^t(h) \\
&\geqslant \min_{j:n_j\in Q} \varliminf_{t\to\infty} \bar{f}_j^t(h) \\
&> \min_{j:n_j\in Q} \rho_j - \varepsilon \\
&\geqslant \rho - \varepsilon.
\end{aligned}$$

Since this is true for all Q, the conclusion of the theorem follows for all infinite histories.

Next, we examine the case that i is the first player. Note that in this case, for each node n_j, $\bar{f}_j^t(h) = v^{h_t}(n_j)$. Observe, also, that for P_δ-almost all infinite histories h in Γ^S, each of the subgames G_{n_j} recurs infinitely many times in h. Indeed, after each finite history, each of the games G_{n_j} is selected by i with probability δ at least. Thus, the event that one of these games is played only finitely many times has probability 0.

Let X_t be a binary random variable over histories such that $X_t(h) = 1$ for histories h in which the node n_{j_0} selected by player i at time t satisfies,

$$v^{h_t}(n_{j_0}) > \rho - \varepsilon/2, \tag{5}$$

and $X_t = 0$ otherwise.

Claim 3. *There exists δ_0 such that for all $j = 1 \dots k$ and any $0 < \delta < \delta_0$, for P_δ-almost all infinite histories h in Γ^S there is time T such that for all $t > T$,*

$$v^{h_t}(n_j) > \rho_j - \varepsilon/4, \tag{6}$$

$$|v^{h_t}(n_j) - v^{h'_{t+1}}(n_j)| < \varepsilon/4, \tag{7}$$

for each history h' such that $h'_t = h_t$, and

$$E_\delta(X_{t+1}|h_t) \geqslant 1 - \delta, \tag{8}$$

where E_δ is the expectation with respect to P_δ.

Inequality (6) follows from the induction hypothesis. For (7), note that if n_j is not reached in round $t+1$ then the difference in (7) is 0. If n_j is reached then $v^{h'_{t+1}} = \big(vv^{h_t}(n_j) + f(z_{t+1})\big)/(v+1)$, where v is the number of times n_j was reached in h_t and $f(z_{t+1})$ is the payoff in round $t+1$. But, v goes to infinity with t, and thus (7) holds for large enough t.

For (8), observe that (6) implies $\max_j v^{h_t}(n_j) > \rho - \varepsilon/4$, as $\rho = max_j \rho_j$. Then, by (7), $\max_j v^{h'_{t+1}}(n_j) > \rho - \varepsilon/2$ for each history h' such that $h'_t = h_t$. Therefore, after h_t, player i chooses, with probability at least δ, a node n_{j_0} that satisfies (5), which shows (8).

The information about the conditional expectations in (8) has a simple implication for the averages of X_t. To see it we use the following convergence theorem from Loève [16] p. 387.

Stability Theorem. *Let X_t be a sequence of random variables with variance σ_t^2. If*

$$\sum_{t=1}^{\infty} \sigma_t^2/t^2 < \infty, \tag{9}$$

then

$$\bar{X}_t - \frac{1}{t}\sum_{l=1}^{t} E(X_l|X_1, \dots, X_{l-1})) \to 0, \tag{10}$$

almost surely, where $\bar{X}_t = (1/t)\sum_{l=1}^{t} X_l$. [13]

Consider now the restriction of the random variables X_t to the set of infinite histories with P_δ conditioned on this space. From (8) it follows that on this space, almost surely $\underline{\lim}_{t\to\infty} \frac{1}{t}\sum_{l=1}^{k} E(X_l|h_l)) \geqslant 1 - \delta$. Therefore, almost surely $\underline{\lim}_{t\to\infty} \frac{1}{t}\sum_{l=1}^{k} E(X_l|X_1, \dots, X_{l-1})) \geqslant 1 - \delta$. This is so, because the field generated by the random variables $(X_1, \dots, X_{l-1})$ is coarser than the field generated by histories h_t. Since condition (9) holds for X_t, it follows by the Stability Theorem that for P_δ-almost all infinite histories h,

$$\underline{\lim}_{t\to\infty} \bar{X}_t \geqslant 1 - \delta. \tag{11}$$

[13] The name stability theorem was give by Loève. Hart and Mas-Colell [12], who also use this theorem in the context of a learning model, refer to it as the strong law of large numbers for dependent random variables.

By the definition of X_t,

$$\bar{f}^t(h) = \frac{1}{t}\sum_{j=1}^{k} v_j^t(h)v^{h_t}(n_j) \geqslant \bar{X}_t(h)(\rho - \varepsilon/2) + (1 - \bar{X}_t(h))\underline{M},$$

where $\underline{M}$ is the minimal payoff in G. If we choose δ_0 such that $(1-\delta_0)(\rho-\varepsilon/2)+\delta_0\underline{M} > \rho - \varepsilon$, then by (11), for each $\delta < \delta_0$, $\underline{\lim}_{t\to\infty} f^t(h) > \rho - \varepsilon$ for P_δ-almost all infinite histories. □

The proof of Theorem 4 is also extended to stochastic repeated games. We show that the conclusion of the theorem holds for P_δ-almost all infinite histories.

Proof of Theorem 4. Assume that the claim of the theorem holds for all the subgames of G. We denote by v_j the restriction of the valuation v to G_{n_j}, and by $\beta_{i,j}$, i's perfect equilibrium strategy there, which is also the restriction of β_i to this game.

Claim 4. *Let i_0 be the player at the root, π_j* be *i_0's payoff in the perfect equilibrium of G_{n_j}, and $\varepsilon > 0$.*

Then there exists $\delta_0 > 0$ such that for all $0 < \delta < \delta_0$, nodes n_j, and players i, for P'_δ almost all infinite histories h' of $\Gamma^{S'}_{n_j}$ there exists T such that for all $t > T$,

$$\sigma_i^{v_j^{h'_t}}(m) = (1-\delta)\beta_{i,j}(m) + \delta\mu(m) \tag{12}$$

for each node $m \in M_i$ in G_{n_j}, and

$$|E_\delta(f_j^{t+1}|h'_t) - \pi_j| < \varepsilon \tag{13}$$

where E_δ is the expectation with respect to P'_δ, and f_j^{t+1} is i's payoff in round $t+1$.

Equality (12) is the induction hypothesis. Consider a history h'_t for which (12) holds. In the round that follows h'_t, the perfect equilibrium path in G_{n_j} is played with probability $(1-\delta)^{d-1}$ at least, where d is the depth of G. Player i_0's payoff in this path is π_j. Thus for small enough δ_0, (13) holds.

By Claims 1 and 2 it follows from (12) that for $0 < \delta < \delta_0$, for P_δ-almost all histories h in Γ, there exists T such that for all $t > T$ the strategies played in each of the games $\Gamma^{S'}_{n_j}$ is the perfect equilibrium of G_{n_j}. Thus, to complete the proof it is enough to show that in addition, at the root, i_0 chooses in these rounds, with probability $1-\delta$, the node n_{j_0} for which $\beta_{i_0}(r) = n_{j_0}$. For this we need to show that i_0's valuation of n_{j_0} is higher than the valuation of all other nodes n_j.

To show it, let 3ε be the difference between π_{j_0} and the second highest payoffs π_j. By the assumption of the uniqueness of the perfect equilibrium, $\varepsilon > 0$. Note that as all players' strategies are fixed for $t > T$, $\lim_{t\to\infty}\frac{1}{t}\sum_{l=1}^{t} E_\delta(f_{i_0}^{t+1}|h'_t)$ exists. Using the stability Theorem, as in Theorem 3, we conclude that $\lim_{t\to\infty}\bar{f}_j^t(h')$ exists, and by (13) the inequality $|\lim_{t\to\infty}\bar{f}_j^t(h') - \pi_j| < \varepsilon$ holds, where $\bar{f}_j^t(h')$ is i_0's average payoff until round t of history h', in the game $\Gamma^{S'}_{n_j}$.

As in the proof of Theorem 3, it follows that for P_δ-almost all infinite histories h in Γ, $|\lim_{t\to\infty} v^{h_t}(n_j) - \pi_j| < \varepsilon$. But then, for P_δ-almost all infinite histories h there exists T such that for all $t > T$, $v^{h_t}(n_{j_0})$ is the highest valuation of all the nodes n_j. □

References

[1] T. Börgers, R. Sarin, Learning through reinforcement and replicator dynamics, J. Econ. Theory 77 (1997) 1–14.
[2] C. Camerer, T. Ho, Experience-weighted attraction learning in games: a unifying approach, Econometrica 67 (1997) 827–874.
[3] I. Cho, A. Matsui, Learning aspiration in repeated games, J. Econ. Theory 124 (2005) 171–201.
[4] R. Cressman, Evolutionary Dynamics and Extensive Form Games, The MIT Press, Cambridge, 2003.
[5] I. Erev, A. Roth, Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibrium, Am. Econ. Rev. 88 (1997) 848–881.
[6] D. Fudenberg, D. Kreps, Learning, experimentation and equilibrium in games, Mimeo, Stanford University, 1988.
[7] D. Fudenberg, D. Levine, Self-confirming equilibrium, Econometrica 61 (1993) 523–545.
[8] D. Fudenberg, D. Levine, Consistency and cautious fictitious play, J. Econ. Dynam. Control 19 (1995) 1065–1090.
[9] D. Fudenberg, D. Levine, The Theory of Learning in Games, The MIT Press, Cambridge, 1998.
[10] I. Gilboa, D. Schmeidler, Case-based decision theory, Quart. J. Econ. 110 (1995) 605–639.
[11] S. Hart, Evolutionary dynamics and backward induction, Games Econ. Behav. 41 (2002) 227–264.
[12] S. Hart, A. Mas-Colell, A simple adaptive procedure leading to correlated equilibrium, Econometrica 68 (2000) 1127–1150.
[13] E. Hendon, H.J. Jacobsen, B. Sloth, Fictitious play in extensive form games, Games Econ. Behav. 15 (1996) 177–202.
[14] P. Jehiel, D. Samet, Valuation equilbria, Mimeo, 2004.
[15] R. Karandikar, D. Mookherjee, D. Ray, F. Vega-Redondo, Evolving aspirations and cooperation, J. Econ. Theory 80 (1998) 292–331
[16] M. Loève, Probability Theory, Van Nostrand, third ed., Princeton, NJ, 1963.
[17] G. Noldeke, L. Samuelson, An evolutionary analysis of backward and forward induction, Games Econ. Behav. 5 (1993) 425–554.
[18] A.L. Samuel, Some studies in machine learning using the game of checkers, IBM J. Res. Devel. 3 (1959) 210–229.
[19] R. Sarin, F. Vahid, Payoff assessments without probabilities: a simple dynamic model of choice, Games Econ. Behav. 28 (1999) 294–309.
[20] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, The MIT Press, Cambridge, 1998.