LEARNING TO PLAY GAMES IN EXTENSIVE FORM BY VALUATION

PHILIPPE JEHIEL AND DOV SAMET

Abstract. A valuation for a board game is an assignment of numeric values to different states of the board. The valuation reflects the desirability of the states for the player. It can be used by a player to decide on her next move during the play. We assume a myopic player, who chooses a move with the highest valuation. Valuations can also be revised, and hopefully improved, after each play of the game. Here, a very simple valuation revision is considered, in which the states of the board visited in a play are assigned the payoff obtained in the play. We show that by adopting such a learning process a player who has a winning strategy in a win-lose game can almost surely guarantee a win in a repeated game. When a player has more than two payoffs, a more elaborate learning procedure is required. We consider one that associates with each state the average payoff in the rounds in which this node was reached. When all players adopt this learning procedure, with some perturbations, then, with probability 1, strategies that are close to subgame perfect equilibrium are played after some time. A single player who adopts this procedure can guarantee only her individually rational payoff.

1. Introduction

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 fictitious-player-like 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 wrote a checkers-playing learning program as far back as 1955, which marks the beginning of reinforcement learning (see Samuel (1959)). Since then many other sophisticated algorithms, heuristics, and computer programs, have been developed, which are

Date: May, 2000.

based on reinforcement learning. (Sutton and Barto (1998)). Such programs try neither to learn the behavior of a specific opponent, nor to find the distribution of 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 fictitious-player-like models. Reinforcement learning has only attracted the attention of game theorists in the last decade in theoretical works like Gilboa and schmeidler (1995), Camerer and Ho (1997), Sarin and Vahid (1999), and in experimental works like Erev and Roth (1997). In all these studies the basic model is given in a strategic form, and the learning player identifies those of her strategies that perform better. This approach seems inadequate where learning of games in extensive form is concerned. Except for the simplest games in extensive form, the size of the strategy space is so large that learning, by human beings or even machines, cannot involve the set of all strategies. This is certainly true for the game of chess, where the number of strategies exceeds the number of particles in the universe. But even a simple game like tic-tac-toe is not perceived by human players in the full extent of its strategic form.

The process of learning games in extensive form can involve only a relatively small number of simple strategies. But when the strategic form is the basic model, no subset of strategies can be singled out. Thus, for games in extensive form the structure of the game tree should be taken into consideration. Instead of *strategies* being reinforced, as for games in strategic form, it is the *moves* of the game that should be reinforced for games in extensive form.

This, indeed, is the approach of heuristics for playing games which were developed by AI theorists.[1] One of the most common building block of such heuristics 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 of the player;
- A *revision rule*, which specifies how the valuation is revised after playing the game.

[1]Perhaps the concentration of the AI literature on moves rather than strategies is the 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 (1998) and *Reinforcement Learning: An Introduction*, Sutton and Barto (1998).

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.

First, we study stage games in which the learning player has only two payoffs, 1 (win) and 0 (lose). Two-person win-lose games are a special case. But here, there is no restriction on the number of the other players or their payoffs.

For these games we adopt the simple *myopic strategy rule*. By this rule, the player chooses in each of her decision node 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.

As a revision rule we adopt the simple *memoryless revision*: after each round the player revises only the valuation of the moves made in the round. The valuation of such a move becomes the payoff (0 or 1) in that round.

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

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.

> 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.

When the learning player has more than two payoffs, the previous learning process is of no help. In this case we study the *exploratory myopic strategy rule*, by which the player opts for the maximally valued move, but chooses also, with small probability, moves that do not maximize the valuation.

The introduction of such perturbations makes it necessary to strengthen the revision rule. We consider the *averaging revision*. Like the memoryless revision, the player revises only the valuation of moves made in the last round. The valuation of such a move is the average of the payoffs in all previous rounds in which this move was made.

> 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 is close to her individually rational payoff (the maxmin payoff) in the stage game.

The two previous results indicate that reinforcement learning achieves learning of playing the stage game itself, rather than playing against certain opponents. The learning processes described guarantee the player her individually rational payoff (which is the win in the first result). This is exactly the payoff that she can guarantee even when the other players are disregarded.

Our next result concerns the case where all the players learn 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 process yields the players more than just learning the stage game. Indeed, they 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 perfect equilibrium.

Although valuation is defined for all moves, the learning player needs no information concerning the game when she start playing it. Indeed, the initial valuation can be constant. To play the stage game with this valuation, the player needs to know which moves are possible to her, only when it is her turn to play, and then choose one of them at random. 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.

However, the learning processes discussed here are inefficient for games with large number of nodes (or states of the board). In chess, for example, almost any state of the board, except for the few first ones, has been seen in recorded history only once. In order to make these processes more efficient, similar moves (or states of the board) should be grouped together, such that the number of similarity classes is manageable. When the valuation of a move is revised, so are all the moves similar to it. We will deal with such learning processes, as well as with games with incomplete information, in a later paper.

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 $M_i(n) = \{m \mid (n, m) \in A\}$. Denote $M_i = \cup_{n \in N_i} M_i(n)$. For each i the function $f_i \colon Z \to R$ 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 $t \geq 0$, 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 $v \colon M_i \to R$.

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

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 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.*

The memoryless revision rule. *For a history $h = (z)$ of length 1, 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 initial nonnegative 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 Figure 1, where the payoffs are player 1's.

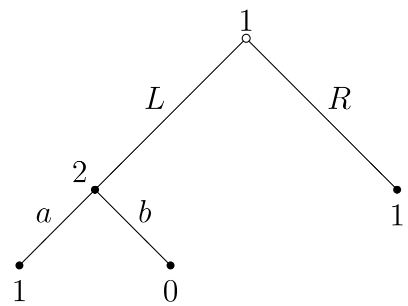

FIGURE 1. Two payoffs

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. (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 going 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.

The next example shows that when the payoff function of a player has more than two values, 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 Figure 2.

In this game player 1 can guarantee a payoff of 10, and therefore we expect a learning process to lead player 1 to this payoff. But, no reasonable restriction on the initial valuation can guarantee that the learning process induced by the myopic strategy and the memoryless revision results in the payoff 10 in the long run. For example, for any constant initial valuation, there is a positive probability that the valuation $(-10,2)$ for (L,R) is obtained, which is absorbing.

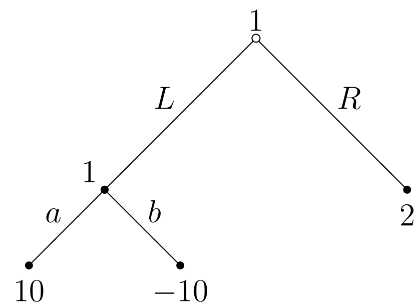

FIGURE 2. More than two payoffs

The remedy for this is straightforward. The player should always allow for exploratory moves that remind her of the high payoffs, so that she is not stuck 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. This guarantees that she will never be stuck in the valuation $(-10, 2)$.

Unfortunately, adding exploratory moves 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 fist 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) form 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.

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

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, the failure to achieve the payoff 10 after introducing the exploratory moves is the result of these moves, and has nothing to do with the number of values of the payoff function.

3.2. **The case of payoff functions with more than two values.** Example 2 leads us to consider new strategy and revision rules for payoff functions with more than two values.

The δ-exploratory myopic strategy rule. *This rule associates with each valuation v the strategy σ^v_δ, where for each node $n \in N_i$, $\sigma^v_\delta(n) = (1-\delta)\sigma^v(n) + \delta\mu(n)$. 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.*

The averaging revision rule. *For a node $m \in M_i$, and a history $h = (z_1, \dots, z_t)$, if the node m was never reached in h, then $v^h(m) = v(m)$. Else, let $t_1, \dots, t_k$ 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 little exploration, and averaging revision, player i can guarantee to be close to his individually rational (maxmin) payoff in G.

Theorem 2. *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 $h = (z_1, z_2, \dots)$,

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

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 3. *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^δ_i 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 $h = (z_1, \dots, z_t, \dots)$, 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. **Stochastic repeated games.** We prove all the theorems by induction on the depth of the game tree. For this we need to be able to deduce properties of Γ from properties of repeated games of stage games G' which are subgames of G. This can be more naturally done when we consider a wider class of repeated games which we call *stochastic repeated games*. Within this class the repeated game of G' can be imbedded in the repeated game of G, thus enabling us to make the required deductions.

Let S be a countable set of states which also includes an *end state* e. We consider 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 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 $t \geq 0$ the set of finite history 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 $t \geq 0$ 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 $H^p \backslash F$ 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

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

for $h \in H^p$.

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 $\Sigma_i(h) = \sigma^{v^{\bar{h}}}$.

4.2. **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 tag 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 these 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 $h \in H^p$, such that node n is never reached in h. Obviously, S' subsumes S, and in particular includes the end state e. Note that the set H'_∞ of infinite history in $\Gamma_n^{S'}$ can be naturally viewed as a subset of H_∞, H' as a subset of H, and H'^p as a subset of H^p. We use this fact to define the transition probability distribution τ' 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, $\tau'(h')(e) = P(E \mid h')$, 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 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*

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

for each $h' \in H'^p$, *where the right-hand side is the restriction of* $\Sigma(h')$ *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 $P(h', s')\Sigma(h', s')(n)\Sigma'(h', s')(z')$. 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. ∎

4.3. **Win-lose games.** 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 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, $P(R^+_{t+1} \subseteq R^+_t \mid R^+_t \neq \emptyset) = 1$.

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 $t \geq T$. 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 $t \geq T$. 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

$t \geq T$ 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 $P(\bar{E^+}) = 0$. ∎

4.4. **The case of payoff functions with more than two values.** We prove Theorem 2 for stochastic repeated games, where the conclusion of the theorem holds for P_δ-almost all *infinite* histories.

Proof of Theorem 2. 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 tagged. 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 $\bar{f}^t_j(h)$ i's average payoff until t *at the times* n_j *was reached*, that is,

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

The history h can be viewed as an infinite history h' in $\Gamma^{S'}_{n_j}$. Moreover, for each l, $\bar{f}^l(h') = \bar{f}^{t_l}_j(h)$. By the definition of $\bar{f}^t_j(h)$, 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 $t > T$, $\bar{f}^t_j(h) > \rho_j - \varepsilon$. 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 $t > T$, $\bar{f}^t_j(h) > \rho_j - \varepsilon$.

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 $\nu^t_j(h)$ be the number of times n_j is reached in h until time t. Then,

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

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

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

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 $j = 1\ldots k$ 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

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

where E_δ is the expectation with respect to P_δ.

The 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 $v^{h'_{t+1}} = \big(\nu v^{h_t}(n_j) + f(z_{t+1})\big)/(\nu+1)$, where ν 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, ν 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 (1963) 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

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

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

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 $\underline{\lim}_{t\to\infty} \frac{1}{t}\sum_{l=1}^{k} E(X_l \mid h_l)) \geq 1-\delta$. Therefore, almost surely $\underline{\lim}_{t\to\infty} \frac{1}{t}\sum_{l=1}^{k} E(X_l \mid X_1, \ldots, X_{l-1})) \geq 1-\delta$. This is so, because the field generated by the the random variables $(X_1, \ldots, X_{l-1})$ 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,

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

By the definition of X_t,

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

where $\underline{\mathrm{M}}$ is the minimal payoff in G. If we choose δ_0 such that $(1-\delta_0)(\rho-\varepsilon/2)+\delta_0\underline{\mathrm{M}} > \rho-\varepsilon$, 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 3 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 3. 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 $0 < \delta < \delta_0$, node n_j, and player 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 $t+1$.

The 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 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 2, 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 2, 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

Camerer, C. and T. Ho (1997). Experience-Weighted Attraction Learning in Games: A Unifying Approach, *Econometrica.*

Erev, I. and A. Roth (1997). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibrium. *American Economic Rev.*

Fudenberg, D. and D. Levine (1998). *The Theory of Learning in Games*, The MIT Press.

Gilboa, I. and D. Schmeidler (1999). Case Base Decision Theory. *Quart. J. Econom*, 110, pp. 605–639.

Loève, M. (1963). *Probability Theory*, D. Van Nostrand, Third ed.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers, *IBM J. Res. and Devel.*, 3, p. 210—229.

Sarin, R. and F. Vahid (1999). Payoff Assessments without Probabilities: A Simple Dynamic Model of Choice. *Games and Economic Behaviour*, 28, pp. 294–309.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction*, The MIT Press.