Reinforcement Learning in Perfect-Information Games*

Maxwell Pak†

Queen's University

July 2006

Key words: reinforcement learning, extensive-form games

JEL Classification: D83

MSC2000 subject classification: primary: 91A18, 91A22; secondary: 68Q32

OR/MS subject classification: primary: games/group decisions

[*] This paper could not have been written without the help and encouragement from my advisor Chris Shannon. Her suggestions have been indispensable for this work from its inception, and I owe her my deepest gratitude. I have also benefited greatly from many discussions with Botond Koszegi. I would like to thank Robert Anderson, Jim Bergin, Steve Goldman, Matthew Rabin, and Robert Powell for their helpful comments as well. Support from the NSF grants SES-9710424 and SES-9818759 is gratefully acknowledged.

[†] Department of Economics, Queen's University, Kingston, Ontario K7L 3N6, Canada. Phone: 613.533.2251. Fax: 613.533.6668. E-mail: pakm@qed.econ.queensu.ca.

Abstract

This paper studies action-based reinforcement learning in finite perfection-information games. Restrictions on the valuation updating rule that guarantee that the play eventually converges to a subgame-perfect Nash equilibrium (SPNE) are identified. These conditions are mild enough to contain interesting and plausible learning behavior. We provide two examples of such updating rule that suggest that the extent of knowledge and rationality assumptions needed to support a SPNE outcome in finite perfect-information games may be minimal.

1 Introduction

Most learning models in game theory feature strategy-based learning in which players attempt to learn their optimal strategies.[1] While this may be reasonable in simultaneous-move games, strategy-based learning models fail to provide a good descriptive model of learning in extensive-form games because they require the players to form complete contingent plans of actions at all of their decision nodes before the game is played. For example, even in a relatively simple game like tic-tac-toe in which each player makes at most four action choices, the game tree, ignoring symmetry, contains roughly 9! nodes, and forming a complete strategy for the game is beyond the ability of human players.

Since much of the real learning that takes place in extensive-form games appears to involve players who do not use complete strategies, this paper considers action-based reinforcement learning model in which players make an action choice at a decision node only when that node is reached during the game play.[2] In particular, we study a model in a setting where players repeatedly play a finite perfect-information game but treat each play myopically so that they are only concerned with the current period payoffs. When a decision node is reached during the game play, the player who moves at that node assigns valuations to her available actions and chooses the action with the highest valuation. The valuations are based on the payoff experience the player has had in the past periods in which that action had been played. We identify conditions on the valuation updating rule that leads to the players eventually playing a subgame-perfect Nash equilibrium (SPNE).

Our approach is similar to Jehiel and Samet [4] and Laslier and Walliser [7].

[1] For example, in fictitious play models players keep track of their opponents' plays and choose a strategy that is a best response to the empirical frequency of their opponents' plays. In replicator dynamic models, players adjust the frequency with which they play a pure strategy in proportion to the current payoff corresponding to that strategy.

[2] Good surveys of reinforcement learning in the machine learning literature are Sutton and Barto [9] and Kaelbling, Littman, and Moore [5]. The papers collected in Sutton [8] provide a more technical introduction to the reinforcement learning literature.

However, while these papers specify one particular valuation updating rule and show that the play converges to a SPNE in some appropriate sense, we provide a more general result in that we identify general restrictions on the valuation updating rules that guarantee convergence to a SPNE. As we show through examples, these restrictions are mild enough to encompass a wide range of learning behaviors.

Moreover, Jehiel and Samet employ ε-greedy learning rule in which players take the action with the highest valuation with probability $1 - \varepsilon$ or explore by taking a random action with probability ε. However, whether the exploration is viewed as an experimentation or a mistake, assuming that exploration probability stays the same no matter how much experience a player has gained seems antithetic in a model of learning. A more plausible model of learning should reflect the fact that the rate at which a player experiments, or makes mistakes, when playing the same game for the millionth time is lower than when she has played it only for the first few times. In addition, the constant exploration probability is also undesirable in a prescriptive model of learning because it is difficult to know the appropriate value for the exploration probability. The desire to converge to a near optimal behavior in the limit requires the probability of exploration to be set at small value. However, exploring with small probability means that it will be difficult for the player to escape suboptimal behavior in the early plays of the game. Thus, good limiting behavior comes at the price of poor finite time performance.

To resolve the tension between the desired limit behavior and the desired finite time behavior, a player should explore often in early periods and less frequently in later periods. Here, we accomplish this by endogenizing the exploration probability. In fact, in our model the players do not explicitly experiment. Rather, they always take an action with the highest valuation but are nevertheless induced to explore because of the imperfection in forming valuations.[3] We show that if the valuation of

[3] We believe this approach to be more natural when the players, as in Jehiel and Samet [4] and Laslier and Walliser [7], are assumed to be myopic and treat each game as an end in itself. In such setting, it is not clear why players would choose to experiment. Since the players are not concerned with future payoffs, there is no reason why players would be willing to sacrifice current payoff and take an action that they believe to be suboptimal.

an action becomes more accurate as, and only as, the number of times that action is taken increases, then the player is induced to explore with ever decreasing frequency but still infinitely often. This, in turn, leads the play to converge to a SPNE.

The remainder of the paper is organized as follows. Section 2 describes the reinforcement learning model considered here. Because endogenizing exploration comes with technical burden, the main results in their full generality are discussed in Appendix A. Instead, Section 3 restricts attention to additively separable valuations that are sum of two terms, an empirical term and an error term, and presents four conditions that together guarantee convergence to a SPNE. The two substantive conditions required for these valuations have intuitive interpretations. The first condition requires that the error term converges to zero as the player's experience with that action grows. The second condition requires that if the fraction of time a player receives some payoff u when an action is chosen converges to one, then the empirical term converges to that payoff.

In Section 3.1, two examples of additively separable valuations, sample averaging model and more primitive, and more interesting, simple recollection model are presented and shown to satisfy the four conditions. Section 3.2 presents the main results for additively separable valuations. The first result shows that if a valuation process satisfies the first substantive condition, then players explore every action infinitely often. The second result shows that if it satisfies both conditions, then the play converges to a SPNE outcome in probability and the fraction of time in which a SPNE outcome is played converges to one almost surely in finite perfect-information games with no relevant ties in the payoffs. The proofs of the results are deferred to Appendix A, and instead an intuition for the results are developed using the simple recollection model.

Finally, a simulation result suggesting that the two examples considered here attain near-SPNE behavior relatively fast is given in Section 3.3. The simulation results also illustrate the tension inherent in the ε-greedy algorithm and show that models considered in this paper can outperform the ε-greedy models in both finite time and the limit. The paper concludes in Section 4.

2 The Model

We consider a set of players who repeatedly play a finite perfect-information game $\mathcal{G}$ but treat each game myopically. Let G be the set of decision nodes of $\mathcal{G}$, z_0 the root node, and G_T the set of terminal nodes. For each node $z \in G \backslash G_T$, $i(z)$ denotes the player who moves at z and A_z denotes the set of actions available at z. In abuse of notation, $i(a)$ is used to denote the player to whom the action a belongs to so that $i(a) = i(z)$, where $a \in A_z$. Let $\mathcal{G}_z$ denote the subgame starting at z and let G_z denote the set of nodes in $\mathcal{G}_z$. For each $z \in G_T$, let $u^i(z)$ denote player i's payoff from z. Let $\underline{C}_i = \min\{u^i(z) : z \in G_T\}$ and $\overline{C}_i = \max\{u^i(z) : z \in G_T\}$. Let A be the set of all actions in $\mathcal{G}$. For later convenience we add the "null action" a_0, interpreted as the action immediately preceding the root node z_0, to A. With this addition, we can define a bijection $\zeta : A \to G$ that maps each action to the node that immediately succeeds it.

The players are assumed to play the game in the following sequential manner. Let $\mathbb{T} = \{1, 2, 3, ...\}$ be the time index. For each $t \in \mathbb{T}$, the game in period t begins by player $i(z_0)$ choosing an action from A_{z_0}. Player $i(z_0)$ is assumed to have some valuation $v_t(a)$ for each action $a \in A_{z_0}$ and choose an action a' with the highest valuation. When there is a tie, the player chooses according to some arbitrary tie-breaking rule. Once action a' has been chosen, the game proceeds to node $z' = \zeta(a')$ and player $i(z')$ moves next. Player $i(z')$ is also assumed to have some valuation $v_t(a)$ for each action $a \in A_{z'}$ and choose an action a'' with the highest valuation. The game proceeds in this manner until a terminal node is reached and each player i receives her payoff, which is denoted u_t^i. The outcome of the t-th play of the game is identified by the path, denoted ξ_t, that was taken during the play.[4] How the game

[4] That is, the path ξ_t is the unique sequence of actions that starts from a_0 and leads to the

play evolves over time is governed by how the valuation are updated, and identifying the restrictions on the valuation updating rule, represented by the valuation process $\{v_t(a) : t \in \mathbb{T}\}$ for each $a \in A$ or collectively by $\{v_t : t \in \mathbb{T}\}$, that leads the game play to evolve towards a SPNE is the main goal of the paper.

3 Additively Separable Valuation Process

For valuations satisfying additive separability, the sufficient conditions that guarantee convergence to a SPNE are particularly intuitive and easily verified. Therefore, this section restricts attention to valuation processes such that for all $a \in A$ and $t \in \mathbb{T}$, $v_t(a) = f_t(a) + g_t(a)$, where $f_t(a)$, interpreted as the empirical term, has support inside $[\underline{C}_{i(a)}, \overline{C}_{i(a)}]$ and $g_t(a)$, interpreted as the error term, has full support.[5] We show that if an additively separable valuation process satisfies the four assumptions listed below, then the play converges to a SPNE.

In the following, let $\mathcal{F}$ denote the σ-field generated by the underlying valuation updating rule, and let sub-σ-field $\mathcal{F}_t$ denote the collection of events up to time t. For any event B, indicator variable $I(B)$ is the random variable taking the value 1 or 0 depending on whether event B has occurred or not. Let $N_t(a) = \sum_{k=1}^{t} I(a \in \xi_t)$ denote the number of times action a has occurred up to time t.

Assumption (A1). *For all $a', a'' \in A_z$ with $a' \neq a''$, $v_{t+1}(a')$ and $v_{t+1}(a'')$ are conditionally independent given $\mathcal{F}_t$. That is, for all Borel $B', B'' \subset \mathbb{R}$,*

$$P\left(v_{t+1}(a') \in B' \text{ and } v_{t+1}(a'') \in B'' \mid \mathcal{F}_t\right) = P\left(v_{t+1}(a') \in B' \mid \mathcal{F}_t\right)\ P\left(v_{t+1}(a'') \in B'' \mid \mathcal{F}_t\right)$$

almost surely.

Assumption (A2). *For all $a \in A$, conditional distribution of $v_{t+1}(a)$ changes only*

terminal node reached at the end of the t-th play of the game.

[5] Full support assumption is made for convenience. For further discussion, see general assumption (GA3) in Appendix A.

after action a has been sampled. That is, for any Borel $B \subset \mathbb{R}$,

$$P\left(v_{t+2}(a) \in B \mid \mathcal{F}_{t+1}\right) = P\left(v_{t+1}(a) \in B \mid \mathcal{F}_t\right) \text{ on } \{a \notin \xi_{t+1}\}.$$

Assumption (A3). *For all $a \in A$, error term converges to zero as the number of times action a is taken increases. That is, for all $\varepsilon > 0$,*

$$P\left(|g_{t+1}(a)| > \varepsilon \mid \mathcal{F}_t\right) \to 0 \text{ on } \{N_t(a) \to \infty\}.$$

Assumption (A4). *For all $a \in A$, whenever*

$$\frac{\sum_{n=1}^{t} I(a \in \xi_n) I\left(u_n^{i(a)} = u\right)}{N_t(a)} \to 1 \text{ as } N_t(a) \to \infty$$

for some constant u, then $f_t(a) \to u$ as $t \to \infty$.[6]

Assumptions (A1) and (A2) formalize how valuation process for action-based learning should behave. Assumption (A1) requires that the valuations of different actions are independent when conditioned on the past history. Assumption (A2) requires that the conditional distribution of the valuation of an action changes only after that action has been taken. Assumption (A3) requires that the error term in the valuation of an action decreases to zero, in appropriate probabilistic sense, as the player's experience with that action grows. Lastly, to interpret assumption (A4), suppose that the fraction of time a player receives some payoff u when action a is taken converges to one. Assumption (A4) then requires that the empirical term corresponding to action a must also converge, again in appropriate sense, to u.

3.1 Examples of Additively Separable Valuation Processes

For an immediate example of an additively separable valuation process satisfying assumptions (A1)-(A4), consider the following model of learning, which we call

[6] Formally, $f_t(a) \to u$ as $t \to \infty$ on $\left\{\frac{\sum_{n=1}^{t} I(a \in \xi_n) I\left(u_n^{i} = u\right)}{N_t(a)} \to 1 \text{ as } N_t(a) \to \infty\right\}$.

sample averaging model. When evaluating an action, players in this model try to use the average of the past payoffs associated with the action. Players are assumed to have imperfect ability to calculate the historical average, so the valuation assigned to an action is a perturbed average of the payoffs received in the periods in which that action had been taken; however, the error associated with evaluating an action is assumed to decrease as the number of times that action has been taken by the player increases.

For all $a \in A$, let $\{\varepsilon_t^a : t \in \mathbb{T}\}$ be independent copies of a random variable ε^a that has full support. The valuation process $\{v_t : t \in \mathbb{T}\}$ for the model is given by the following.

$$v_t(a) = \frac{\sum_{n=1}^{t-1} I(a \in \xi_n)\, u_n^i}{\widetilde{N}_{t-1}(a)} + \frac{\varepsilon_t^a}{\widetilde{N}_{t-1}(a)},$$

where $i = i(a)$ and $\widetilde{N}_{t-1}(a) = \max\{1, N_{t-1}(a)\}$. Proposition 1 shows formally that sample averaging model satisfies assumptions (A1)-(A4).

Proposition 1. *Sample averaging model satisfies assumptions (A1)-(A4).*

Proof. It is easy to see that sample averaging model satisfies (A1) and (A2). Fix any $z' \in G_T$. Letting

$$f_t(a) = \frac{u^i(z')\, I(N_{t-1} = 0)}{\widetilde{N}_{t-1}(a)} + \frac{\sum_{n=1}^{t-1} I(a \in \xi_n)\, u_n^i}{\widetilde{N}_{t-1}(a)},$$

and

$$g_t(a) = \frac{\varepsilon_t^a - u^i(z')\, I(N_{t-1} = 0)}{\widetilde{N}_{t-1}(a)},$$

where $i = i(a)$, it is also easy to see that $g_t(a)$ has full support and that $f_t(a) \in [\underline{C}_i, \overline{C}_i]$ almost surely. For any $\varepsilon > 0$,

$$\begin{aligned} P\left(|g_{t+1}(a)| > \varepsilon \mid \mathcal{F}_t\right) &= P\left(|\varepsilon_{t+1}^a - u^i(z')\, I(N_t = 0)| > \varepsilon \widetilde{N}_t(a) \mid \mathcal{F}_t\right) \\ &= P\left(|\varepsilon^a - u^i(z')\, I(N_t = 0)| > \varepsilon \widetilde{N}_t(a) \mid \mathcal{F}_t\right) \\ &\to 0 \text{ on } \{N_t(a) \to \infty\}. \end{aligned}$$

So, (A3) is satisfied. Lastly, since

$$\frac{u^i(z')\, I(N_{t-1} = 0)}{\widetilde{N}_{t-1}(a)} \to 0$$

on $\{N_t(a) \to \infty\}$, (A4) is readily satisfied.

□

For more interesting example, consider the next model, which we call simple recollection model. This model tries to capture the following primitive learning behavior. When evaluating an action, players try to remember what the payoffs had been in the previous periods in which that action had been chosen and assign one of the past payoffs as the value of that action. Naturally, the more a player receives a particular payoff after playing action a, the more likely it is that the value attached to action a is that particular payoff. Players are assumed to have imperfect memory so that there is always a chance that players make an erroneous recall. However, the chance of making an error when evaluating an action decreases as the number of times that action has been taken by the player increases.

For all $a \in A$, let $\{\eta_t^a : t \in \mathbb{T}\}$ be independent copies of uniform$[0,1]$ random variable, and let $\{\varepsilon_t^a : t \in \mathbb{T}\}$ be independent copies of an independent random variable ε^a that has full support. Let τ_k^a denote the period in which action a was chosen for the k-th time. The valuation process $\{v_t : t \in \mathbb{T}\}$ is given formally by:

$$v_t(a) = \sum_{k=1}^{N_{t-1}(a)} I\left(\eta_t^a \in \left(\frac{k}{1+N_{t-1}(a)},\ \frac{k+1}{1+N_{t-1}(a)}\right]\right) u^i_{\tau_k^a} + I\left(\eta_t^a \leq \frac{1}{1+N_{t-1}(a)}\right) \varepsilon_t^a,$$

where $i = i(a)$.

This process behaves as if there is an urn, or a memory bank, corresponding to each action. Each urn initially contains one card, called the wild card. Suppose node z is reached during the course of the t-th play. Player $i(z)$ assigns a value $v_t(a)$ to each $a \in A_z$ by drawing a card from the urn corresponding to action a. If a card that is drawn is a wild card then the value assigned to the action is the outcome of a draw from an independent random variable ε_t^a. If the card is not a wild card then the value assigned to the action is the pre-recorded value on the card. In either case the card is placed back into the urn once the value has been assigned to a.

If during the course of the t-th play action a was chosen, then at the end of the t-th play player $i(a)$'s payoff in the t-th play is recorded on a new card and placed into the urn for that action. Thus, each time action a is chosen, the number of cards in the corresponding urn increases by one. Therefore, the chance of receiving a random valuation declines as players gain experience but never disappears. On the other hand, the distribution of $v_t(a)$ converges to the empirical distribution of the payoffs received when action a was chosen as the number of the times in which action a is chosen goes to infinity. For example, if player i receives the same payoff u after choosing action a for all but finitely many times, then $v_t(a)$ converges to u with probability one as t goes to infinity.

Proposition 2. *Simple recollection model satisfies assumptions (A1)-(A4).*

Proof. It is easy to see that simple recollection model satisfies (A1) and (A2). Fix any $z' \in G_T$. Letting

$$f_t(a) = I\left(\eta_t^a \leq \frac{1}{1+N_{t-1}(a)}\right) u^i(z') + \sum_{k=1}^{N_{t-1}(a)} I\left(\eta_t^a \in \left(\frac{k}{1+N_{t-1}(a)},\ \frac{k+1}{1+N_{t-1}(a)}\right]\right) u_{\tau_k^a},$$

and

$$g_t(a) = I\left(\eta_t^a \leq \frac{1}{1+N_{t-1}(a)}\right)\left(\varepsilon_t^a - u^i(z')\right),$$

where $i = i(a)$, it is also easy to see that $g_t(a)$ has full support and that $f_t(a) \in [\underline{C}_i, \overline{C}_i]$ almost surely. For any $\varepsilon > 0$,

$$\begin{aligned} P\left(|g_{t+1}(a)| > \varepsilon \mid \mathcal{F}_t\right) &= P\left(\eta_{t+1}^a \leq \frac{1}{1+N_t(a)} \text{ and } |\varepsilon_{t+1}^a - u^i(z')| > \varepsilon \,\middle|\, \mathcal{F}_t\right) \\ &= P\left(\eta^a \leq \frac{1}{1+N_t(a)} \,\middle|\, \mathcal{F}_t\right) P\left(|\varepsilon^a - u^i(z')| > \varepsilon \,\middle|\, \mathcal{F}_t\right) \\ &\to 0 \text{ on } \{N_t(a) \to \infty\}. \end{aligned}$$

So, (A3) is satisfied.

Lastly, on $\left\{\frac{\sum_{n=1}^{t} I(a\in\xi_n)I\left(u_n^i=u\right)}{N_t(a)} \to 1 \text{ as } N_t(a)\to\infty\right\}$,

$$I\left(\eta_t^a \le \frac{1}{1+N_{t-1}(a)}\right) u^i(z') \to 0$$

and

$$\sum_{k=1}^{N_{t-1}(a)} I\left(\eta_t^a \in \left(\frac{k}{1+N_{t-1}(a)},\ \frac{k+1}{1+N_{t-1}(a)}\right]\right) u_{\tau_k^a} \to u$$

as $t\to\infty$. So, (A4) is satisfied. □

3.2 Main Results for Additively Separable Valuation Process

Theorem 1 shows that valuation processes satisfying assumptions (A1)-(A3) generate sufficient exploration by inducing players to play every action in the game infinitely often with probability one. In particular, this means that SPNE paths occur infinitely often. Of course, this is not a strong result since the same theorem also shows that every path occurs infinitely often as well. So, this theorem also serves as a negative result showing that the probability of non-SPNE paths occurring only finitely many times is zero.

Theorem 1. *Let G be a finite perfect-information game. Suppose an additively separable valuation process $\{v_t : t\in\mathbb{T}\}$ satisfies assumptions (A1)-(A3). Then $\forall a\in A,\ N_t(a)\to\infty$ almost surely as $t\to\infty$.*

Given the weak restriction placed on valuation processes by assumption (A3), it should not be surprising that more restriction is needed to obtain positive convergence results. The additional restriction required is provided by assumption (A4). Because non-SPNE paths occur infinitely often almost surely, it is clear that the play cannot converge almost surely to a SPNE path. However, Theorem 2 shows that the play can converge to a SPNE path in probability so that the occurrence of non-SPNE paths becomes increasingly rare.

Since ties in the payoffs can present difficulties, the convergence results in Theorem 2 are limited to the following class of games. Let Γ be the collection of all finite perfect-information games such that $\forall z', z'' \in G_T$, $u^i(z') = u^i(z'')$ if and only if $u^j(z') = u^j(z'')$ for all j. So, Γ includes games with generic payoffs, which have no ties in the payoffs, and games where ties are irrelevant like "win-lose-draw" games. Theorem 2 shows that if $\mathcal{G} \in \Gamma$ and the valuation process satisfies assumptions (A1)-(A4), then the probability of playing a SPNE path converges to one as the number of plays goes to infinity. In addition, the fraction of time in which a SPNE path is played converges to one almost surely.[7]

Theorem 2. *Let $\mathcal{G} \in \Gamma$. Suppose an additively separable valuation process $\{v_t : t \in \mathbb{T}\}$ satisfies assumptions (A1)-(A4). Then the probability of playing a SPNE path converges to one, and the fraction of time a SPNE path of G is played converges to one almost surely as $t \to \infty$.*

Theorem 8 in the appendix show that assumptions (A1)-(A4) are special cases of general assumptions (GA1)-(GA4). Therefore, Theorem 1 and Theorem 2 follow from corresponding theorems for general valuation processes, Theorems 3 and 6 in Appendix A. Rather than discuss the general theorems and proofs here, we defer their discussion to the appendix and instead develop intuition for the results using the simple recollection model.

Consider the 2-player game given in Figure 1. Since there are only two actions at the root node, one of the two actions, say action L_1, must be played infinitely often. Since the valuation $v_t(L_1)$ takes the random wild card value with probability $\frac{1}{1+N_{t-1}(L_1)}$, the probability of $v_t(L_1)$ taking a wild card value goes to zero as the number of plays goes to infinity. So, the probability that $v_t(L_1) = 3$ converges to one.

[7] Moreover, as shown in Theorem 6, of which Theorem 2 is a special case, this is true for all subgames of $\mathcal{G}$ as well. That is, the probability of playing a SPNE path of $\mathcal{G}_z$ when node z is reached converges to one. Also, the ratio of the number of times in which a SPNE path of $\mathcal{G}_z$ is played to the number of times in which node z is reached converges to one almost surely. The players, therefore, learn the optimal action at every node and not just at nodes on a SPNE path.

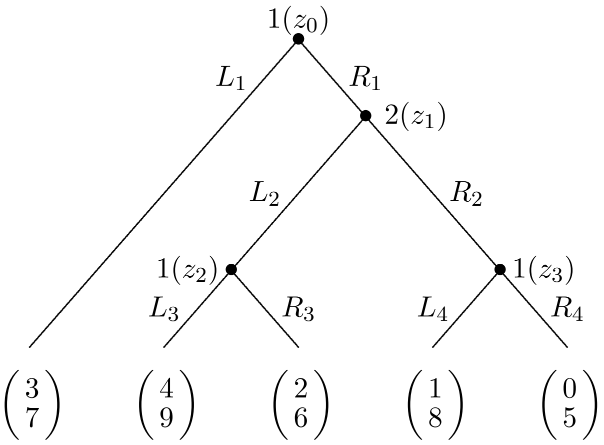

Figure 1: Example

Now, suppose action R_1 occurs only finitely often, say $M-1$ many times, so that time T is the last time action R_1 occurs. Then, for all $t > T$, the probability that $v_t(R_1)$ takes a random wild card value is at least $\frac{1}{M}$. So, for all large $t > T$, the probability that $v_t(R_1)$ is greater than $v_t(L_1)$, which is at least $P(v_t(L_1) = 3) \times P(\text{wild card is drawn at } R_1) \times P(\text{wild card at } R_1 > 3)$, is approximately $\frac{P(\varepsilon^{R_1} > 3)}{M}$. Since this is true for every large $t > T$, it is not possible that the event $\{v_t(R_1) > v_t(L_1)\}$ never occurs after time T, contrary to our assumption. Therefore, action R_1 must occur infinitely often. Similar intuition shows that at each node, every action must occur infinitely often. Of course, this argument is not entirely correct since, among other things, T is random and the events considered here are not independent events. The proof of Theorem 3 provides a formal argument.

Next, the convergence of the play to the SPNE path can be demonstrated by an induction argument. Consider node z_2. Since both actions L_3 and R_3 are played infinitely often, the probability that $v_t(L_3) = 4$ and $v_t(R_3) = 2$ converges to 1 as $t \to \infty$. So, the probability that player 1 plays action L_3 converges to one. Likewise, the probability that player 1 plays action L_4 converges to one as well. So, the probability that player 2 receives payoff 9 after choosing L_2 converges to one, which in turn makes the probability that $v_t(L_2) = 9$ converge to one. Likewise, the probability that $v_t(R_2) = 8$ converges to one as well. Therefore, the probability that player 2 chooses action L_2 converges to one. Through induction, it is not hard to see that the probability that player 1 will choose R_1 converges to one. Thus, the play converges to the SPNE of the game. Since not all the subgames are played at every period, complications arise in formalizing this induction argument. The proof of Theorem 6 solves this problem with the use of stopping times.

3.3 Finite Time Properties: Simulation Results

The theoretical results obtained here are that of asymptotic convergence. However, often it is argued that asymptotic results, while reassuring, have little practical implications.[8] In light of such views, this section concludes with simulation results that roughly assess the finite time properties of the two valuation processes. Specifically, 1000 copies of the simple 3 player game given in Figure 2 were generated, each with payoffs drawn randomly from normal distribution with mean zero and variance 100. The learning behavior of simple recollection model and sample averaging model on each of these games, both with ε_t^a's drawn from normal distribution with mean zero and variance 100, were simulated and compared to the simulated learning behavior generated by the sample averaging ε-greedy models and the sample averaging greedy ($\varepsilon = 0$) model.

For each simulation, the fraction of time the SPNE outcome had occurred up to the t-th play was tracked for each $t \leq 2000$ in each of the learning models. The graph in Figure 3 presents the average of these ratios over 1000 simulations for each learning model. The results suggest that the simple recollection model and the sample averaging model not only reach reasonable optimality relatively quickly but also outperform the ε-greedy models quickly.

Since simple recollection model starts with greater randomness than ε-greedy

[8] See, for example, the remarks on the use of eventual convergence to optimal behavior in measuring learning performance in Kaelbling, Littman, and Moore [5]. As another example, Ellison [2] questions the relevance of the asymptotic convergence results of Kandori, Mailath, and Rob [6] since the time required before the system reaches its limit is beyond the time scale relevant for humans.

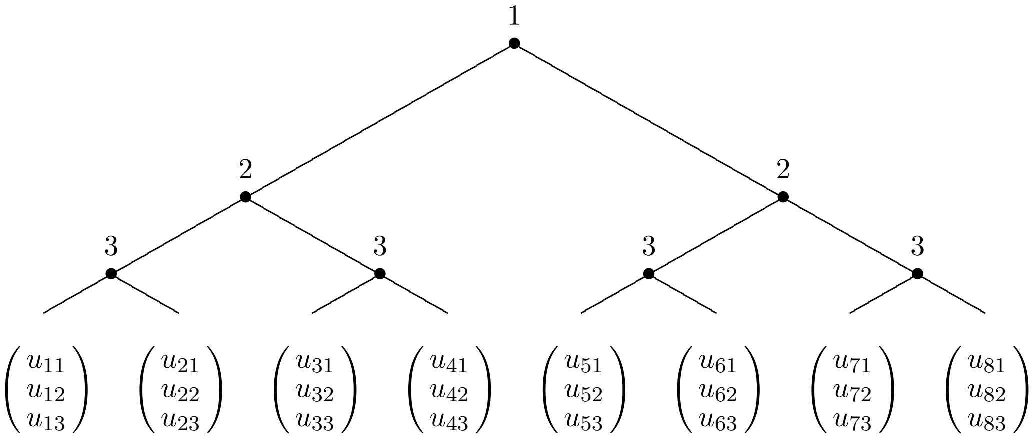
1
2
2
3
3
3
3
u_{11} u_{12} u_{13}
u_{21} u_{22} u_{23}
u_{31} u_{32} u_{33}
u_{41} u_{42} u_{43}
u_{51} u_{52} u_{53}
u_{61} u_{62} u_{63}
u_{71} u_{72} u_{73}
u_{81} u_{82} u_{83}

Figure 2: Test Problem

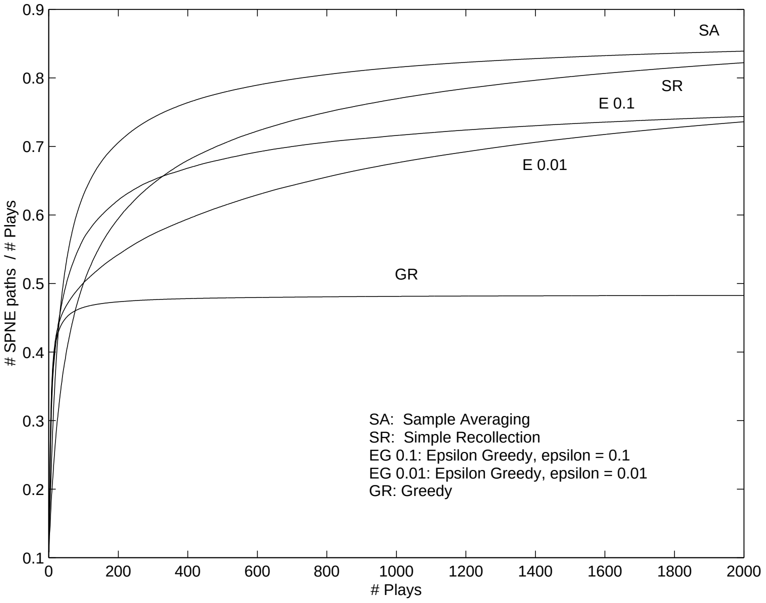

Figure 3: Simulation Results

models, it initially performs worse than ε-greedy models. However, whereas ε-greedy models are forced to explore with the same probability no matter how much experience the players have gained, simple recollection model allows the exploration probabilities to decline with experience, so it outperforms the ε-greedy models eventually. In these simulations, this occurs before the 400th play.

In addition, these simulations also illustrate the inherent tension between finite time and limit behavior present in ε-greedy models. As discussed before, while smaller values of ε leads to better limit behavior, it comes at the price of poorer finite time performance. Indeed, the simulations show that while ε-greedy model with $\varepsilon = 0.01$ will eventually have a better limit behavior than the model with $\varepsilon = 0.1$, it performs worse for the first 2000 periods considered. Sample averaging model with endogenous exploration probability introduced in this paper solves this tension and performs better in both finite time and the limit.

Lastly, to see the importance of maintaining exploration, consider the learning curve corresponding to the greedy method. Because the greedy method starts out with no randomness, the greedy method initially performs as well as or better than the other methods. However, because it never explores, it is trapped in taking suboptimal actions in the later plays.

4 Concluding Remarks

This paper studies sufficient conditions on the valuation updating rule that guarantee that the play converges to a subgame-perfect Nash equilibrium in finite perfect-information games with no relevant ties in the payoffs. The restrictions we identify are mild enough to contain interesting and plausible learning behavior. The two examples of such valuation updating rule we provide correspond to learning behaviors that are primitive but still satisfy the two basic principles of learning outlined in Erev and Roth [3]: Law of Effect (Thorndike [10]) and the Power Law of Practice

(Blackburn [1]), which assert, respectively, that the more often an action leads to a good outcome the more likely it is that this action will be chosen in the future, and that learning curves rise steeply initially but flatten out later. The very naivete of the learning behaviors associated with these models suggests that in this class of games, the extent of rationality that is needed to support a subgame-perfect Nash equilibrium may be minimal.

A General Valuation Process

This section presents the restrictions on general valuation updating rules that guarantee convergence to a SPNE. In particular, additive separability assumption is dropped. Let $(\Omega, \mathcal{F}, P)$ be the probability space on which a valuation process $\{v_t : t \in \mathbb{T}\}$ is defined. Let $\mathcal{F}_0 = \{\emptyset, \Omega)$, and let $\mathcal{F}_t = \sigma(v_1, ..., v_t)$ be the sub-σ-field consisting of events up to time t. For any action $a \in A$, let τ_n^a denote the time of the n-th occurrence of action a. That is, $\tau_0^a = 0$ and $\forall n \in \mathbb{T}$, $\tau_n^a = \inf\{t > \tau_{n-1} : a \in \xi_t\}$. Then, $\{\tau_n^a : n \in \mathbb{Z}_+\}$ is a sequence of stopping times. The following facts about stopping times are used throughout this section. For any stopping time τ, $\mathcal{F}_\tau = \{B \in \mathcal{F} : \forall n \; B \cap \{\tau \leq n\} \in \mathcal{F}_n\}$ is a σ-field. Suppose $\tau_0 < \tau_1 < \tau_2 < ...$ almost surely, then $\{\mathcal{F}_{\tau_n} : n \in \mathbb{Z}_+\}$ is a filtration. Moreover, if $\{Y_t : t \in \mathbb{Z}_+\}$ is adapted to $\{\mathcal{F}_t : t \in \mathbb{Z}_+\}$, then Y_{τ_n} is adapted to $\{\mathcal{F}_{\tau_n} : n \in \mathbb{Z}_+\}$, and if $Y_t \to Y$ almost surely as $t \to \infty$, then $Y_{\tau_n} \to Y$ almost surely as $n \to \infty$.

Each of the four conditions (GA1)-(GA4) below plays the same role as the corresponding assumptions (A1)-(A4) for additively separable valuations. Assumptions (GA1) and (GA2) formalize how valuation process for reinforcement learning should behave: (GA1) requires that valuations of actions are independent when conditioned on the past history, and (GA2) requires that the conditional distribution of valuation of an action changes only after that action has been taken. The substantive conditions are assumptions (GA3) and (GA4): (GA3) guarantees that every action is taken infinitely often, and (GA4) provides the additional condition that guarantees that the play path converges to a SPNE path.

Assumption (GA1). *For all $a \in A$, $a', a'' \in A_{\zeta(a)}$ with $a' \neq a''$, and Borel $B', B'' \subset \mathbb{R}$,*

$$P\left(v_{\tau_{n+1}^a}(a') \in B' \text{ and } v_{\tau_{n+1}^a}(a'') \in B'' \mid \mathcal{F}_{\tau_n^a}\right)$$

$$= P\left(v_{\tau_{n+1}^a}(a') \in B' \mid \mathcal{F}_{\tau_n^a}\right) \; P\left(v_{\tau_{n+1}^a}(a'') \in B'' \mid \mathcal{F}_{\tau_n^a}\right)$$

almost surely.

Assumption (GA2). *For all $a \in A$, $a' \in A_{\zeta(a)}$, and Borel $B \subset \mathbb{R}$,*

$$P\left(v_{\tau^a_{n+2}}(a') \in B \mid \mathcal{F}_{\tau^a_{n+1}}\right) = P\left(v_{\tau^a_{n+1}}(a') \in B \mid \mathcal{F}_{\tau^a_n}\right) \;\; on \;\; \left\{a' \notin \xi_{\tau^a_{n+1}}\right\}.$$

In the following, let $C = 1 + \max\{|u^i(z)| : i \in \mathcal{I},\; z \in G_T\}$, where $\mathcal{I}$ is the set of players of $\mathcal{G}$.[9]

Assumption (GA3). *For all $a \in A$ and $a' \in A_{\zeta(a)}$,*

$$\left\{P(v_{\tau^a_{n+1}}(a') \geq C \mid \mathcal{F}_{\tau^a_n}) \to 0 \;\; as \; n \to \infty\right\} = \left\{N_{\tau^a_n}(a') \to \infty \;\; as \; n \to \infty\right\}.$$

To interpret assumption (GA3), first note that the event $\{v_{\tau^a_n}(a') \geq C\}$ is the event that the valuation of action a' is "overly optimistic" and $\left\{N_{\tau^a_n}(a') \to \infty \text{ as } n \to \infty\right\}$ is the event that the number of times action a' is taken goes to infinity. Therefore, assumption (GA3) essentially requires that the probability of a player having an overly optimistic valuation of an action declines to zero if and only if the number of times a' is sampled goes to infinity.[10]

Assumption (GA4). *For all $a \in A$, $i = i(\zeta(a))$, and $a', a'' \in A_{\zeta(a)}$, if there exist constants $u' > u''$ such that*

$$\frac{\sum_{n=1}^{t} I(a' \in \xi_n) I\left(u^i_n = u'\right)}{N_t(a')} \to 1 \;\; a.s. \;\; as \; N_t(a') \to \infty$$

and

$$\frac{\sum_{n=1}^{t} I(a'' \in \xi_n) I\left(u^i_n = u''\right)}{N_t(a'')} \to 1 \;\; a.s. \;\; as \; N_t(a'') \to \infty,$$

then

$$P\left(v_{\tau^a_{n+1}}(a') > v_{\tau^a_{n+1}}(a'') \;\middle|\; \mathcal{F}_{\tau^a_n}\right) \to 1 \;\; a.s. \;\; as \; n \to \infty.$$

[9] It is not hard to see from the proof of Theorem 3 that C need not be this large. It is enough for C to be the player $i(\zeta(a))$'s highest possible payoff in the subgame $G_{\zeta(a)}$. The particular choice of C made here simplifies notation and the proofs.

[10] A simpler way to state this requirement would have been the expression "$P(v_{\tau^a_n}(a') \geq C) \to 0$ if $N_{\tau_n}(a') \to \infty$." However, because the events $\{v_{\tau^a_n}(a') \geq C\}$, $n \in \mathbb{T}$, are not independent, a conditional version of this idea as stated in assumption (GA3) is needed.

To interpret assumption (GA4), suppose that the fraction of time a player receives some payoff u' when an action a' is taken and the fraction of time that the player receives some payoff u'' when an action a'' is taken both converges to one, where $u' > u''$. Assumption (GA4) then requires that the conditional probability of the player's valuation of action a' being higher than the valuation of action a'' goes to one.

Theorem 3 shows that valuation processes satisfying assumptions (GA1)-(GA3) induce players to play every action in the game infinitely often with probability one.

Theorem 3. *Let G be a finite perfect-information game. Suppose a valuation process $\{v_t : t \in \mathbb{T}\}$ satisfies assumptions (GA1)-(GA3). Then $\forall a \in A,\ N_t(a) \to \infty$ almost surely as $t \to \infty$.*

Proof. Take any $a \in A$, and let $z = \zeta(a)$. The proof proceeds by showing that if $a \in \xi_t$ infinitely often a.s., then $\forall a' \in A_z,\ N_t(a') \to \infty$ a.s. So, assume $a \in \xi_t$ infinitely often a.s. so that $\tau_n^a < \infty$ a.s. for all n. Since $|A_z| < \infty$ there must be at least one action in A_z that occurs infinitely often a.s. Suppose, towards contradiction, that there exists nonempty $\hat{A}_z \subset A_z$ such that $P(B) > 0$, where B is the event $\{\forall a' \in \hat{A}_z,\ a' \in \xi_t \text{ f.o. and } \forall a' \in A_z \backslash \hat{A}_z,\ a' \in \xi_t \text{ i.o.}\}$.

Fix $\hat{a} \in \hat{A}_z$, and let constant C be as in assumption (GA3). Then there exists $Y > 0$ a.s. such that $P\left(v_{\tau_{n+1}^a}(\hat{a}) \geq C \mid \mathcal{F}_{\tau_n^a}\right) \geq Y$ for all n on B. Moreover, $P\left(v_{\tau_{n+1}^a}(a') \geq C \mid \mathcal{F}_{\tau_n^a}\right) \to 0$ for all $a' \in A_z \backslash \hat{A}_z$ on B. Therefore, on B,

$$\begin{aligned}
& \sum_{n=1}^{\infty} P\left(v_{\tau_{n+1}^a}(\hat{a}) > \max_{a' \in A_z \backslash \hat{A}_z} \{v_{\tau_{n+1}^a}(a')\} \,\middle|\, \mathcal{F}_{\tau_n^a}\right) \\
\geq \quad & \sum_{n=1}^{\infty} P\left(v_{\tau_{n+1}^a}(\hat{a}) \geq C \text{ and } \forall a' \in A_z \backslash \hat{A}_z,\ v_{\tau_{n+1}^a}(a') < C \,\middle|\, \mathcal{F}_{\tau_n^a}\right).
\end{aligned}$$

By conditional independence,

$$= \quad \sum_{n=1}^{\infty} \left(P\left(v_{\tau_{n+1}^a}(\hat{a}) \geq C \,\middle|\, \mathcal{F}_{\tau_n^a}\right) \prod_{a' \in A_z \backslash \hat{A}_z} P\left(v_{\tau_{n+1}^a}(a') < C \,\middle|\, \mathcal{F}_{\tau_n^a}\right) \right)$$

$$
\begin{aligned}
&\geq \sum_{n=1}^{\infty} \left(Y \prod_{a' \in A_z \setminus \hat{A}_z} \left(1 - P\left(v_{\tau^a_{n+1}}(a') \geq C \,\middle|\, \mathcal{F}_{\tau^a_n} \right) \right) \right) \\
&= \infty. \\
\Longrightarrow \quad & v_{\tau^a_{n+1}}(\hat{a}) > \max_{a' \in A_z \setminus \hat{A}_z} \{ v_{\tau^a_{n+1}}(a') \} \text{ i.o. by the conditional Borel-Cantelli Lemma.}
\end{aligned}
$$

However, $\max_{a' \in A_z \setminus \hat{A}_z} \{v_{\tau^a_{n+1}}(a')\} > \max_{a' \in \hat{A}_z} \{v_{\tau^a_{n+1}}(a')\}$ for all but finitely many times on B. Since $\hat{a} \in \hat{A}_z$, this is a contradiction. Thus, $\forall a' \in A_z$, $N_t(a') \to \infty$ a.s. as $t \to \infty$.

Finally, since a_0 is the null action, $a_0 \in \xi_t$ infinitely often a.s. So, the theorem follows by induction. □

The following two corollaries follow trivially from Theorem 3, so the proofs are omitted.

Corollary 4. *If the assumptions of Theorem 3 are satisfied, every SPNE path occurs infinitely often with probability one.*

Corollary 5. *If the assumptions of Theorem 3 are satisfied, the probability of SPNE paths occurring for all but finitely many times is zero.*

Recall that Γ is defined as the collection of all finite perfect-information games such that $\forall z', z'' \in G_T$, $u^i(z') = u^i(z'')$ if and only if $u^j(z') = u^j(z'')$ for all j. Theorem 6 shows that if $\mathcal{G} \in \Gamma$ and the valuation process satisfies assumptions (GA1)-(GA4), then the probability of playing a SPNE path converges to one as the number of plays goes to infinity. In addition, the fraction of time in which a SPNE path is played converges to one almost surely as the number of plays goes to infinity.

Moreover, Theorem 6 shows that this is true for all subgames of $\mathcal{G}$ as well. That is, the probability of playing a SPNE path of $\mathcal{G}_z$ when node z is reached converges to one as the number of times z is played goes to infinity. Also, the ratio of the number of times in which a SPNE path of $\mathcal{G}_z$ is played to the number of times in which node z is reached converges to one almost surely as the number of plays goes to infinity. The players, therefore, eventually learn the optimal action at every node and not just at nodes on a SPNE path.

In the following, if $a \in \xi$ let $\xi_a = \xi \cap \{a' \in A_{z'} : z' \in \mathcal{G}_{\zeta(a)}\}$ denote the continuation path of ξ. Let $\Xi_a = \{\text{path } \xi \text{ of } \mathcal{G} : a \in \xi \text{ and } \xi_a \text{ is a SPNE path of } \mathcal{G}_{\zeta(a)}\}$, and let $S_t(a) = \sum_{n=1}^{t} I(\xi_n \in \Xi_a)$.

Theorem 6. *Suppose $\mathcal{G} \in \Gamma$ and the valuation process $\{v_t : t \in \mathbb{T}\}$ satisfies assumptions (GA1)-(GA4). Then for all $a \in A$, $P(\xi_{\tau_n^a} \in \Xi_a) \to 1$ as $n \to \infty$, and $\frac{S_t(a)}{N_t(a)} \to 1$ a.s. as $t \to \infty$.*[11]

Proof. The proof proceeds by induction on subgames of $\mathcal{G}$. So, let $L(\mathcal{G}_z)$ denote the maximum length of a path in $\mathcal{G}_z$. For all $z \in G \setminus G_T$, let $\tilde{A}_z = \{\tilde{a} \in A_z : \text{there exists a SPNE of } \mathcal{G}_z \text{ in which } \tilde{a} \text{ occurs with positive probability}\}$.

Let $a \in A$ be such that $L(\mathcal{G}_{\zeta(a)}) = 1$. Let $z = \zeta(a)$ and $i = i(z)$. By Theorem 3, $\forall a' \in A_z$, $N_t(a') \to \infty$ a.s. Since $L(\mathcal{G}_z) = 1$, $\zeta(a') \in G_T$ for all $a' \in A_z$. So, $I(a' \in \xi_t) I\left(u_t^i = u^i(\zeta(a'))\right) = I(a' \in \xi_t)$ a.s. So, $\frac{\sum_{n=1}^{t} I(a' \in \xi_n) I\left(u_n^i = u^i(\zeta(a'))\right)}{N_t(a')} = \frac{\sum_{n=1}^{t} I(a' \in \xi_n)}{N_t(a')} = 1$ a.s. for all $t \in \mathbb{T}$. Let $\tilde{a} \in \tilde{A}_z$. Then, a.s.,

$$\begin{aligned}
1 &\geq E[I(\xi_{\tau_{n+1}^a} \in \Xi_a) \mid \mathcal{F}_{\tau_n^a}] \\
&= P\left(\xi_{\tau_{n+1}^a} \in \Xi_a) \mid \mathcal{F}_{\tau_n^a}\right) \\
&\geq P\left(v_{\tau_{n+1}^a}(\tilde{a}) > v_{\tau_{n+1}^a}(a') \text{ for all } a' \in A_z \setminus \tilde{A}_z \;\middle|\; \mathcal{F}_{\tau_n^a}\right) \\
&= \prod_{a' \in A_z \setminus \tilde{A}_z} P\left(v_{\tau_{n+1}^a}(\tilde{a}) > v_{\tau_{n+1}^a}(a') \;\middle|\; \mathcal{F}_{\tau_n^a}\right) \text{ by conditional independence.} \\
&\to 1 \text{ as } n \to \infty \text{ by assumption (GA4).}
\end{aligned}$$

Therefore, $E[I(\xi_{\tau_{n+1}^a} \in \Xi_a) \mid \mathcal{F}_{\tau_n^a}] \to 1$ a.s. as $n \to \infty$, and $P(\xi_{\tau_n^a} \in \Xi_a) \to 1$ as

[11] The first version of this theorem showed that $\frac{S_t(a)}{N_t(a)} \to 1$ in expectation. It was not until I saw the use of a stability theorem in Jehiel and Samet [4] that I realized that the proof can be easily strengthened to show almost sure convergence. I would like to thank them for their insight. The version of the stability theorem used is stated and proved as Lemma B.2.

$n \to \infty$ by the dominated convergence theorem. Next,

$$\begin{array}{rl}
& \dfrac{\sum_{k=1}^{n} E\left[I(\xi_{\tau_{k+1}^a} \in \Xi_a) \mid \mathcal{F}_{\tau_k}\right]}{n} \to 1 \text{ a.s. as } n \to \infty \text{ by Lemma B.1,} \\
\text{and} & \dfrac{\sum_{k=1}^{n} \left(I(\xi_{\tau_{k+1}^a} \in \Xi_a) - E\left[I(\xi_{\tau_{k+1}^a} \in \Xi_a) \mid \mathcal{F}_{\tau_k}\right]\right)}{n} \to 0 \text{ a.s. as } n \to \infty \text{ by Lemma B.2.} \\
\implies & \dfrac{\sum_{k=1}^{n} I(\xi_{\tau_k^a} \in \Xi_a)}{n} \to 1 \text{ a.s. as } n \to \infty. \\
\implies & \dfrac{S_t(a)}{N_t(a)} = \dfrac{\sum_{n=1}^{t} I(\xi_n \in \Xi_a)}{N_t(a)} = \dfrac{\sum_{k=1}^{N_t(a)} I(\xi_{\tau_k^a} \in \Xi_a)}{N_t(a)} \to 1 \text{ a.s. as } t \to \infty.
\end{array}$$

Next, suppose for all subgames $\mathcal{G}_{\zeta(a)}$ of $\mathcal{G}$ such that $L(\mathcal{G}_{\zeta(a)}) \leq m$, $E[I(\xi_{\tau_{n+1}^a} \in \Xi_a) \mid \mathcal{F}_{\tau_n^a}] \to 1$ a.s. as $n \to \infty$ and $\frac{S_t(a)}{N_t(a)} \to 1$ a.s. as $t \to \infty$. Let $a \in A$ be such that $L(\mathcal{G}_{\zeta(a)}) =$ m + 1. Let $z = \zeta(a)$ and $i = i(z)$. By Theorem 3, $\forall a' \in A_z$, $N_t(a') \to \infty$ a.s. Then, for any $\tilde{a} \in \tilde{A}_z$, a.s.,

$$\begin{array}{rcl}
1 & \geq & P\left(v_{\tau_{n+1}^a}(\tilde{a}) > v_{\tau_{n+1}^a}(a') \text{ for all } a' \in A_z \setminus \tilde{A}_z \;\middle|\; \mathcal{F}_{\tau_n^a}\right) \\
& = & \displaystyle\prod_{a' \in A_z \setminus \tilde{A}_z} P\left(v_{\tau_{n+1}^a}(\tilde{a}) > v_{\tau_{n+1}^a}(a') \;\middle|\; \mathcal{F}_{\tau_n^a}\right) \text{ by conditional independence.} \\
& & \text{By the induction hypothesis } \frac{S_t(a'')}{N_t(a'')} \to 1 \text{ a.s. } \forall a'' \in A_z. \text{ So,} \\
& \to & 1 \text{ as } n \to \infty \text{ by assumption (GA4).}
\end{array}$$

Therefore, a.s.,

$$\begin{array}{rcl}
1 & \geq & E[I(\xi_{\tau_{n+1}^a} \in \Xi_a) \mid \mathcal{F}_{\tau_n^a}] \\
& = & E\left[\displaystyle\sum_{\tilde{a} \in \tilde{A}_z} I\left(v_{\tau_{n+1}^a}(\tilde{a}) = \max_{a' \in A_z}\{v_{\tau_{n+1}^a}(a')\}\right) I\left(\xi_{\tau_{n+1}^a} \in \Xi_{\tilde{a}}\right) \;\middle|\; \mathcal{F}_{\tau_n^a}\right] \\
& = & E\left[\displaystyle\sum_{\tilde{a} \in \tilde{A}_z} I\left(v_{\tau_{n+1}^a}(\tilde{a}) = \max_{a' \in A_z}\{v_{\tau_{n+1}^a}(a')\}\right) \sum_{\tilde{a} \in \tilde{A}_z} I\left(\xi_{\tau_{n+1}^a} \in \Xi_{\tilde{a}}\right) \;\middle|\; \mathcal{F}_{\tau_n^a}\right] \\
& = & P\left(\displaystyle\max_{\tilde{a} \in \tilde{A}_z}\left\{v_{\tau_{n+1}^a}(\tilde{a})\right\} > \max_{a' \in A_z \setminus \tilde{A}_z}\left\{v_{\tau_{n+1}^a}(a')\right\} \;\middle|\; \mathcal{F}_{\tau_n^a}\right) \sum_{\tilde{a} \in \tilde{A}_z} P(\xi_{\tau_{n+1}^a} \in \Xi_{\tilde{a}} \mid \mathcal{F}_{\tau_n^a}) \\
& \to & 1 \text{ as } n \to \infty \text{ by the induction hypothesis.}
\end{array}$$

So, $E[I(\xi_{\tau^a_{n+1}} \in \Xi_a) \mid \mathcal{F}_{\tau^a_n}] \to 1$ a.s. as $n \to \infty$. Then, $P(\xi_{\tau^a_n} \in \Xi_a) \to 1$ as $n \to \infty$, and $\frac{S_t(a)}{N_t(a)} \to 1$ a.s. as $t \to \infty$ by the same argument as before. □

Corollary 7. *Let $\mathcal{G}$ be a finite perfect-information game with generic payoffs, and let valuation process $\{v_t : t \in \mathbb{T}\}$ satisfy assumptions (GA1)-(GA4). Then the probability of playing the SPNE path converges to one, and the fraction of time the SPNE path of $\mathcal{G}$ is played converges to one almost surely as $t \to \infty$.*

Proof. Since G has a unique SPNE path, the corollary follows trivially from Theorem 6. □

We close with a theorem showing that additive separability and assumptions (A1)-(A4) are special case of general assumptions (GA1)-(GA4).

Theorem 8. *Suppose an additively separable valuation process $\{v_t : t \in \mathbb{T}\}$ satisfies assumptions (A1)-(A3). Then the valuation process satisfies the general assumptions (GA1)-(GA3). If, in addition, the process satisfies assumption (A4), then it also satisfies general assumption (GA4).*

Proof. Let $a = a_0$. Since a is the null action, $\tau^a_n < \infty$ a.s. Then by (A1) and strong Markov property, for all $a', a'' \in G_{\zeta(a)}$ with $a' \neq a''$, condition (GA1) is satisfied. Likewise, by (A2) and strong Markov property, for all $a' \in G_{\zeta(a)}$, condition (GA2) is satisfied. Let $i = i(\zeta(a))$. For all $a' \in A_{\zeta(a)}$,

$$\begin{aligned} P\left(v_{\tau^a_{n+1}}(a') \geq C \mid \mathcal{F}_{\tau^a_n}\right) &\leq P\left(f_{\tau^a_{n+1}}(a') \in [\underline{C}_i, \overline{C}_i] \text{ and } g_{\tau^a_{n+1}}(a') \geq C - \underline{C}_i \mid \mathcal{F}_{\tau^a_n}\right) \\ &= P\left(f_{\tau^a_{n+1}}(a') \in [\underline{C}_i, \overline{C}_i] \mid \mathcal{F}_{\tau^a_n}\right) P\left(g_{\tau^a_{n+1}}(a') \geq C - \underline{C}_i \mid \mathcal{F}_{\tau^a_n}\right) \\ &= P\left(g_{\tau^a_{n+1}}(a') \geq C - \underline{C}_i \mid \mathcal{F}_{\tau^a_n}\right) \\ &\to 0 \text{ as } n \to \infty \text{ on } \{N_{\tau^a_n}(a') \to \infty \text{ as } n \to \infty\} \end{aligned}$$

by (A3) and strong Markov property. Conversely, on $\left\{P\left(v_{\tau^a_{n+1}}(a') \geq C \mid \mathcal{F}_{\tau^a_n}\right) \to 0\right\}$, $N_{\tau^a_n}(a') \to \infty$ since the conditional distribution of $v_{\tau^a_{n+1}}(a')$ differs from that of $v_{\tau^a_n}(a')$ only if $a' \in \xi_{\tau_n}$. Therefore, condition (GA3) is satisfied.

The induction step in the proof of Theorem 3, shows that then $\tau_n^{a'} < \infty$ a.s. for all $a' \in A_{\zeta(a)}$. Therefore, by induction, $\{v_t(a) : t \in \mathbb{T}\}$ satisfies (GA1)-(GA3) for all $a \in A$.

For (GA4), consider any $a \in A$, and let $i = i(\zeta(a))$. Suppose for some a' and a'' in $A_{\zeta(a)}$, there exist constants $u' > u''$ such that

$$\frac{\sum_{n=1}^{t} I(a' \in \xi_n) I\left(u_n^i = u'\right)}{N_t(a')} \to 1 \text{ a.s. as } N_t(a') \to \infty,$$

and

$$\frac{\sum_{n=1}^{t} I(a'' \in \xi_n) I\left(u_n^i = u''\right)}{N_t(a'')} \to 1 \text{ a.s. as } N_t(a'') \to \infty.$$

By (A4), $f_t(a') \to u'$ and $f_t(a'') \to u''$ a.s. as $t \to \infty$. Let $\varepsilon = \frac{u'-u''}{3}$. Then a.s.,

$$\begin{aligned}
& P\left(v_{\tau_n^a}(a') > v_{\tau_n^a}(a'') \;\middle|\; \mathcal{F}_{\tau_{n-1}^a}\right) \\
\geq\ & P\left(|g_{\tau_n^a}(a')| < \varepsilon \text{ and } |g_{\tau_n^a}(a'')| < \varepsilon \;\middle|\; \mathcal{F}_{\tau_{n-1}^a}\right) \\
& \quad \times P\left(f_{\tau_n^a}(a') \in [u'-\varepsilon, u'+\varepsilon] \text{ and } f_{\tau_n^a}(a'') \in [u''-\varepsilon, u''+\varepsilon] \;\middle|\; \mathcal{F}_{\tau_{n-1}^a}\right) \\
\to\ & 1 \text{ a.s. as } n \to \infty.
\end{aligned}$$

□

B Stability Lemmas

This section presents the lemmas that are used in the proof of Theorem 6.[12]

Lemma B.1. *Let $Z_n \to 1$ a.s. as $n \to \infty$, and $0 \leq Z_n \leq 1$ a.s. $\forall n$. Then, $\frac{\sum_{k=1}^{n} Z_k}{n} \to 1$ a.s. as $n \to \infty$.*

Proof. Consider any $\omega \in \Omega$ such that $Z_n(\omega) \to 1$ as $n \to \infty$. Fix any $\varepsilon > 0$. Then,

[12] The stability theorem is a well known fact that often appears in standard probability textbook as an exercise. However, since a proof for the version that is needed could not be located, it is presented as Lemma B.2 and a proof is provided.

$\exists N_\varepsilon > 0$ such that $\forall n > N_\varepsilon$, $Z_n(\omega) \geq 1 - \varepsilon$. Then, $\forall n > N_\varepsilon$,

$$\frac{\sum_{k=1}^{n} Z_k(\omega)}{n} = \frac{\sum_{k=1}^{N_\varepsilon} Z_k(\omega)}{n} + \frac{\sum_{k=N_\varepsilon+1}^{n} Z_k(\omega)}{n} \geq \frac{\sum_{k=1}^{N_\varepsilon} Z_k(\omega)}{n} + \frac{(n - N_\varepsilon)(1-\varepsilon)}{n}$$

So, $\forall \varepsilon > 0$,

$$\liminf \frac{\sum_{k=1}^{n} Z_k(\omega)}{n} \geq \liminf \left(\frac{\sum_{k=1}^{N_\varepsilon} Z_k(\omega)}{n} + \frac{(n - N_\varepsilon)(1-\varepsilon)}{n} \right) = 1 - \varepsilon.$$

Therefore, $\frac{\sum_{k=1}^{n} Z_k(\omega)}{n} \to 1$, and $\frac{\sum_{k=1}^{n} Z_k}{n} \to 1$ a.s. □

Lemma B.2. *Let $\{Z_n : n \in \mathbb{T}\}$ be adapted to filtration $\{\mathcal{F}_n : n \in \mathbb{T}\}$, and let $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Suppose there exists a constant M such that $|Z_n| < M$ a.s. for all n. Then, $\frac{\sum_{k=1}^{n}(Z_k - E[Z_k \mid \mathcal{F}_{k-1}])}{n} \to 0$ a.s. as $n \to \infty$.*

Proof. Let $Y_n = \sum_{k=1}^{n} \frac{Z_k - E[Z_k \mid \mathcal{F}_{k-1}]}{k}$. Then, $EY_n^2 < \infty$. Also, a.s.,

$$\begin{aligned}
& E[Y_{n+1} \mid \mathcal{F}_n] \\
= & \sum_{k=1}^{n+1} \frac{E[Z_k \mid \mathcal{F}_n] - E[E[Z_k \mid \mathcal{F}_{k-1}] \mid \mathcal{F}_n]}{k} \\
= & \frac{E[Z_{n+1} \mid \mathcal{F}_n] - E[Z_{n+1} \mid \mathcal{F}_n]}{n+1} + \sum_{k=1}^{n} \frac{E[Z_k \mid \mathcal{F}_n] - E[E[Z_k \mid \mathcal{F}_{k-1}] \mid \mathcal{F}_n]}{k} \\
= & \sum_{k=1}^{n} \frac{Z_k - E[Z_k \mid \mathcal{F}_{k-1}]}{k} = Y_n \text{ since } \forall k \leq n,\ Z_k \text{ is } \mathcal{F}_n\text{-measurable and } \mathcal{F}_{k-1} \subset \mathcal{F}_n.
\end{aligned}$$

So, $\{Y_n : n \in \mathbb{T}\}$ is a martingale with bounded second moments. By the martingale convergence theorem, there exists Y such that

$$\sum_{k=1}^{n} \frac{Z_k - E[Z_k \mid \mathcal{F}_{k-1}]}{k} = Y_n \to Y \text{ a.s. as } n \to \infty.$$

Then by Kronecker's lemma, $\frac{\sum_{k=1}^{n}(Z_k - E[Z_k \mid \mathcal{F}_{k-1}])}{n} \to 0$ a.s. as $n \to \infty$. □

References

[1] Blackburn, J. (1936). Acquisition of Skill: An Analysis of Learning Curves. *IHRB Report, No. 73.*

[2] Ellison, G. (1993). Learning, Local Interaction, and Coordination. *Econometrica*, 61: 1047-1072.

[3] Erev, I. and A. Roth. (1998). Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria. *American Economic Review*, 88: 848-881.

[4] Jehiel, P. and D. Samet. (2000). Learning to Play Games in Extensive Form by Valuation. Working Paper.

[5] Kaelbling, L., M. Littman, and A. Moore. (1996). Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research*, 4: 237-285.

[6] Kandori, M., G. Mailath, and R. Rob. (1993). Learning, Mutation, and Long Run Equilibria in Games. *Econometrica*, 61: 29-56.

[7] Laslier, J-F. and B. Walliser. (2005) A Reinforcement Learning Process in Extensive Form Games. *International Journal of Game Theory*, 33: 219-227.

[8] Sutton, R. (ed.). (1992). *Reinforcement Learning.* Boston: Kluwer Academic Publishers.

[9] Sutton, R. and A. Barto. (1998). *Reinforcement Learning: An Introduction.* Cambridge: MIT Press.

[10] Thorndike, E. (1911). *Animal Intelligence.* Macmillan: New York.