Generalized Reinforcement Learning in Perfect-Information Games[*]

Maxwell Pak[†] and Bing Xu[‡]

January 2014

Abstract

This paper studies reinforcement learning in which players base their action choice on valuations they have for the actions. We identify two general conditions on valuation updating rules that together guarantee that the probability of playing a subgame perfect Nash equilibrium (SPNE) converges to one in games where no player is indifferent between two outcomes without every other player being also indifferent. The same conditions guarantee that the fraction of times a SPNE is played converges to one almost surely. We also show that for additively separable valuations, in which valuations are the sum of empirical and error terms, the conditions guaranteeing convergence can be made more intuitive. In addition, we give four examples of valuations that satisfy our conditions. These examples represent different degrees of sophistication in learning behavior and include well-known examples of reinforcement learning.

Key words: reinforcement learning, extensive-form games
JEL Classification: D83

[*] The earlier version of this paper benefited greatly from helpful discussions with Chris Shannon. Pak gratefully acknowledges support from NSF grants SES-9710424 and SES-9818759. Xu gratefully acknowledges support from China National Natural Science Foundation grant 71403217.

[†] Research Institute of Economics and Management, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074 P.R. CHINA. E-mail: mpak@swufe.edu.cn

[‡] Research Institute of Economics and Management, Southwestern University of Finance and Economics, Chengdu, Sichuan 610074 P.R. CHINA. E-mail: xubing@swufe.edu.cn

1 Introduction

The notion of subgame perfect Nash equilibrium (SPNE) in an extensive-form game requires the players to be best responding in all the subgames of the game, including those that are not actually played. However, since the ability to form a complete contingent plan of action seems beyond the ability of human players in all but the simplest extensive-form games, it is not clear how a SPNE would arise.[1] Therefore, this paper studies whether action-based reinforcement learning, in which players make an action choice at a decision node only when that node is reached during the game play, can lead to a SPNE.

In particular, we consider players who repeatedly play a finite perfect-information game in which no player is indifferent between two outcomes without every other player being also indifferent. We assume that players treat each play myopically, so 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 based on her past payoff experience, and chooses the action with the highest valuation. We identify conditions on the valuation updating rule that guarantee that the probability of playing a SPNE and the fraction of times a SPNE is played converge to one as the number of times the game is played goes to infinity.

Our framework is similar to Jehiel and Samet [6], who posit that the valuation attached to an action is the average of the payoffs received in the periods in which that action had been taken. When called upon to make a choice, their players follow an ε-greedy rule: with probability $1-\varepsilon$, choosing the action that has the highest valuation and, with total probability ε, experimenting by randomly choosing one of the available actions with equal probability. They show that the players eventually end up playing the SPNE with probability $1-\varepsilon$ in finite perfect-information games with unique SPNE.

Although the framework is similar, our approach differs from Jehiel and Samet in that, rather than studying the long-run property of a single valuation updating rule, we study general conditions on valuation updating rules that will lead to a SPNE. Moreover, players always experimenting with constant probability ε as in Jehiel and Samet's model means that, aside from being somewhat descriptively unnatural, the play can only converge to a SPNE with probability $1-\varepsilon$.[2] This may not seem consequential since ε can be set

[1] For example, even in a relatively simple game like tic-tac-toe, where each player has at most four action choices, the game tree contains 255,168 play paths (terminal nodes). If rotational and reflectional symmetries are considered, the number is reduced to 26,830. Either way, forming a complete strategy for the game, or even solving the game through backward induction, appears to be beyond the ability of human players.

[2] Whether experimentation is viewed as a conscious choice to explore or simply as a mistake, assuming that experimentation probability stays the same no matter how much expe-

arbitrarily small; however, such view belies an important practical consideration. It is not clear how ε should be set if we want the model to be a prescriptive rule for learning. On the one hand, setting ε too small would mean that players do not experiment enough, so the play can get trapped at a "suboptimal" behavior for a very long time.[3] On the other hand, setting ε large to induce players to explore more could mean that players will spend too much time experimenting and not enough time exploiting the valuations that they have learned.

While this tension between exploration and exploitation can be resolved, and the probability of playing a SPNE made to converge to one, by decreasing the experimentation probability over time, the solution is not as trivial as it may first appear. As the following example shows, simply reducing the experimentation probability at some deterministic rate like ε/t, where t is the number of times the game has been played, does not work in general.

Example 1. Consider the following two player game in which player 1 chooses between L and R, and player 2 chooses between l and r. The unique SPNE of the game is (R, r).

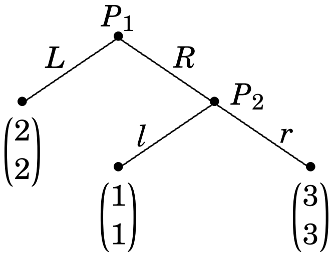

Figure 1: Example 1.

Suppose the players' valuation for an action is the average historical payoff associated with that action, as in Jehiel and Samet's model, but we let the players follow a modified ε-greedy rule in which the probability of experimenting at period t is ε/t. Let B_t be the event that player 1 chooses R through experimentation in period t and player 2 experiments in period t. The decision to experiment is independent of the valuations, so

$$\sum_{t=1}^{\infty} P(B_t) = \sum_{t=1}^{\infty} \left(\frac{\varepsilon}{2t}\right)\left(\frac{\varepsilon}{t}\right) < \infty.$$

rience a player has gained seems unnatural 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 would be lower than when she has played it only few times.

[3] For a similar critique of finite state space Markov learning models, see Ellison[4].

Borel-Cantelli lemma then implies that the probability of B_t occurring infinitely many times is zero, which means that there is a constant $T < \infty$ such that the probability of B_t never occurring after period T is strictly positive, say η.[4]

Now, suppose the players initially have very pessimistic valuations for their "right" actions. In particular, suppose player 1's initial valuations are $v_1(L) = 2$ and $v_1(R) = -3(T+1)$, while player 2's initial valuations are $v_1(l) = 1$ and $v_1(r) = -3(T+1)$. Consider any sample path on which B_t does not occur after period T.[5] Player 1's valuations at period $T+1$ satisfy $v_{T+1}(L) = 2 > v_{T+1}(R)$, and player 2's valuations satisfy $v_{T+1}(l) = 1 > v_{T+1}(r)$. This means that if player 1 chooses R in period $T+1$, it is only as the result of an experimentation. But then player 2 cannot experiment since B_{T+1} does not occur. Thus, (R,r) cannot be the outcome in period $T+1$, so we have $v_{T+2}(L) = 2 > v_{T+2}(R)$ and $v_{T+2}(l) = 1 > v_{T+1}(r) = v_{T+2}(r)$ in period $T+2$. By induction, it is easy to see that (R,r) will not occur at any period after T. Therefore, given these initial valuations, the probability that the SPNE is played in period t is less than or equal to $1-\eta$ for any $t > T$. In particular, it does not converge to one as $t \to \infty$, even though the experimentation probability goes to zero. □

In the above example, the probability of playing the SPNE does not converge to one because the experimentation probability decreases too fast and makes the players stop jointly experimenting before they had the chance to sample sufficient number of (R,r). This suggests that if the players use experimentation schedule that decreases to zero slowly enough so that they jointly experiment infinitely many times, then they will eventually realize that (R,r) is optimal no matter what their initial valuations were. Indeed, rate $\frac{\varepsilon}{\sqrt{t}}$ is one such schedule for this example. However, it is not too hard to see that $\frac{\varepsilon}{\sqrt{t}}$ will be too fast for other games and that, in general, appropriateness of a rate depends on the length of the game tree, which may not be known. More importantly, deterministic schedule such as $\frac{\varepsilon}{\sqrt{t}}$ embodies arbitrariness that is difficult to justify because it forces the experimentation probability to decrease independently of how the game has evolved over time.[6] In contrast, the conditions identified in this paper enable a more natural way of generating infinite exploration by allowing experimentation probability to depend on player's experience.

[4] See, for example, Durrett [3], Theorem 2.3.1 and Theorem 2.3.6 for Borel-Cantelli lemmas.

[5] That is, suppose we are on $\bigcap_{T+1}^{\infty} B_t^c$, which occurs with probability η.

[6] For example, even if a decision node is encountered for the first time in the thousandth time the game has been played, the player will still experiment with probability $\frac{\varepsilon}{\sqrt{1000}}$ at this node, despite the fact that she has learned nothing about the actions at this particular node.

In a related work, Laslier and Walliser [7] resolve the tension between exploration and exploitation by using a more classical reinforcement learning approach, which does not use valuations and instead attaches to each action a variable that represents the "propensity" to choose that action. The probability with which an action is chosen at a decision node is proportional to the propensity of each action at that node.[7] They show that if players follow this behavioral rule, then the probability of playing the SPNE converges to one in finite perfect-information games where no two payoffs are equal. However, their model is restricted to games with strictly positive payoffs because propensity is defined as the sum of payoffs associated with an action.

The remainder of the paper is organized as follows. Section 2 describes the reinforcement learning model considered here. We present our main results in Section 3. We first provide two general conditions that together guarantee that the probability of playing a SPNE and the fraction of time a SPNE is played converge to one. The first condition generates sufficient exploration by ensuring that every action is sampled infinitely many times while the second condition requires that players increasingly exploit the knowledge gained though exploration. We then introduce a class of valuation processes, called additively separable valuations, for which the conditions guaranteeing convergence can be made more intuitive. Section 4 gives four examples of valuations that satisfy our conditions. These examples represent different degrees of sophistication in learning behavior and include a suitably modified version of Jehiel and Samet's model as well as a recast of Laslier and Walliser's model. The paper concludes in Section 5.

2 The Model

Given a finite perfect-information game $\mathscr{G}$, let G be its set of nodes, z_0 the root node, and $\mathscr{I}$ the set of players. For each terminal node $z \in G$, $u^i(z)$ denotes player i's payoff from z. For each decision node $z \in G$, $i(z)$ denotes the player who moves at z, A_z the set of actions available at z, $\mathscr{G}_z$ the subgame starting at z, and G_z the set of nodes in $\mathscr{G}_z$. Let A be the set of all actions in $\mathscr{G}$. For each $a \in A$, $i(a)$ denotes the player to whom action a belongs to, so $i(a) = i(z)$ if $a \in A_z$. We use $\zeta(a)$ to denote the node immediately succeeding a. The sets G^i and A^i denote decision nodes and actions that belong to player i, respectively.

[7] Models similar to this have been studied widely in normal-form games. For example, Sarin and Vahid [14] provide convergence results for a reinforcement learning model in single-player decision problems. Borgers and Sarin [2] connect reinforcement learning with replicator dynamics, and Hopkins [5] explores the connection between reinforcement learning and stochastic fictitious play. Beggs [1] and Laslier, Topol, and Walliser [8] show conditions under which reinforcement learning rule converges to a Nash equilibrium in normal-form games.

Let Γ be the collection of finite perfect-information games satisfying the "no indifference condition," which requires that no player is indifferent between two outcomes without having all the other players be indifferent also. That is, whenever $u^i(z) = u^i(z')$, $u^j(z) = u^j(z')$ for all $j \in \mathscr{I}$.[8] Games satisfying this condition include games that are generic in the sense that no two payoffs of a player is the same, which implies that the SPNE is unique. Non-generic games in Γ, such as "win-lose-or-draw" games, can have multiple SPNE. However, all the SPNE are essentially the same in that every player is indifferent among the SPNE of the game.[9] In particular, this means that if $\tilde{a}_z$ followed by a SPNE of $\mathscr{G}_{\zeta(\tilde{a}_z)}$ is a SPNE of G_z, then so is $\tilde{a}_z$ followed by another SPNE of $\mathscr{G}_{\zeta(\tilde{a}_z)}$.

We assume that players repeatedly play $\mathscr{G} \in \Gamma$ but treat each play myopically as an end in itself. 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 with the highest valuation. If a' is 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 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 game in period t is identified by the path ξ_t that was followed during the play. We use $z \in \xi_t$ to mean that node z was reached during period t and $a \in \xi_t$ to mean that action a was played during period t.

As seen above, players in our model do not explicitly experiment. Rather, they always take an action with the highest valuation, so any "exploration" must occur through imperfections in forming valuations.[10] In addition, aside from requiring players to know when it is their turn to make a choice and the actions available to them, all the rationality and knowledge assumptions are embodied in the valuation updating rule. Thus, how the game play evolves over time is governed by how valuations are updated, and identifying the restrictions on the updating rule that lead the play to evolve towards a SPNE is the main goal of the paper.

In the following, we use $\{v_t^i : t \in \mathbb{Z}_{++}\}$, where $v_t^i = (v_t(a) : a \in A^i)$ and $\mathbb{Z}_{++} =$

[8] Similar condition, called "transfer of decision maker indifference," has been used as a sufficient condition for order independence of removal of weakly dominated strategies in strategic-form games (Marx and Swinkels [10], Østerdal [12]).

[9] See, for example, Osborne and Rubinstein [11], pp. 100-101.

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

$\{1,2,3,\ldots\}$, to denote the valuation process generated by player i's updating rule. We use $\{v_t : t \in \mathbb{Z}_{++}\}$, where $v_t = (v_t^i : i \in \mathscr{I})$, to denote the valuations of all the players collectively as a single process; however, we do not require that players follow the same updating rule. For an example of a valuation process, consider the following model of primitive learning behavior.

Example 2 (Simple Recollection Rule). When evaluating an action, a player using this rule tries to remember what the payoffs had been in the previous periods in which the action had been chosen and assigns one of the past payoffs as the value of the action. The more often the player receives a particular payoff after playing an action, the more likely it is that the value attached to the action is that payoff. The player is also assumed to have imperfect memory so that there is always a chance that she makes an erroneous recall. However, the chance of making an error decreases as the number of times that action had been taken by the player increases.

For a formal description of the rule, let i be the player using this rule, and, for all $a \in A^i$ and $t \in \mathbb{Z}_{++}$, let η_t^a be an independent uniform random variable on $[0,1]$ and ε_t^a be an independent copy of a random variable ε^a that has support $\mathbb{R}$. Let τ_k^a denote the period in which action a was chosen for the k-th time, and let $N_{t-1}(a)$ denote the number of times action a has been chosen up to and including period $t-1$. Letting $\mathbb{1}(\cdot)$ be the indicator function, the valuation for $a \in A^i$ is given by

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

This process behaves as if there is an urn, or a memory bank, corresponding to each action. Each urn initially contains one ball, called the "wild card." Suppose node $z \in G^i$ is reached during period t. Player i assigns a value $v_t(a)$ to each $a \in A_z$ by drawing a ball from the urn corresponding to action a. If the ball that is drawn is the wild card, then the value assigned to the action is the outcome of a draw from a random variable ε_t^a. If the ball is not the wild card, then the value assigned to the action is the pre-recorded value on the ball. In either case, the ball is placed back into the urn after the value has been assigned. For each action a that was chosen during period t, player i's payoff in period t is recorded on a new ball and placed into the urn corresponding to a at the end of the period. □

If all the players use the simple recollection rule, the probability of playing the SPNE of the game in Example 1 converges to one as the number of times the game is played goes to infinity. This is a consequence of Theorem 5 in Section 4. Below, we give an intuition for this result.

Because the number of balls in an urn increases only when the corresponding action is chosen, the chance of drawing the wild card at an urn decreases only when the corresponding action is chosen. This is enough to ensure that every action is sampled infinitely many times. To see this, suppose action L is played only finitely often, say M many times. Let period T be the last time L occurs, and put aside for the moment the complications arising from the fact that M and T are random. Then the probability of drawing the wild card at action L is $\frac{1}{1+M}$ in every period after T.

Consider any $t > T+1$. Since player 1 must be choosing R in every period after T, there must be at least two balls in the urn corresponding to R in period t. Thus,

$$\begin{aligned} P\left(v_t(L) > v_t(R)\right) &\geq P\left(v_t(L) > 3 \text{ and } v_t(R) \leq 3\right) \\ &\geq P\left(\text{wild card is drawn at } L\right) \times P\left(\text{value of wild card at } L > 3\right) \\ &\qquad 1 \times P\left(\text{wild card is not drawn at } R\right) \\ &\geq \left(\frac{1}{1+M}\right) \times P\left(\varepsilon^L > 3\right) \times \left(\frac{1}{2}\right) \; > \; 0. \end{aligned}$$

Since this probability is bounded away from zero and does not depend on t, it means that $v_t(L) > v_t(R)$ at least one more time after $T+1$, contradicting the assumption that L does not occur after period T.[11] Similar argument holds if we assume that R is played only finitely often. Therefore, both L and R must be played infinitely many times.

Since R is played infinitely many times, player 2 also gets to make a choice infinitely often. Therefore, by restricting attention to only the periods in which player 2 gets to make a choice, we can make a similar argument as above to show that both l and r must be played infinitely many times. Thus, every action in the game is played infinitely many times.

Although every action being played infinitely many times means that sufficient exploration is generated, it also means that play cannot converge to the SPNE with probability one since every path, including non-SPNE ones, are played infinitely many times. However, the probability of playing the SPNE does converge to one.[12] This is driven by the fact that the distribution of $v_t(a)$ converges to the empirical distribution of the payoffs received because there is only one wild card in the urn. In particular, if the fraction of times player i receives payoff u after choosing action a goes to one, then the probability of $v_t(a)$ equaling u converges to one as well.

To see this, first restrict attention to periods in which player 1 has chosen R. Since both l and r are played infinitely many times, $P(v_t(l) = 1)$ and

[11] This argument is based on Borel-Cantelli lemmas, but it glosses over the fact that M and T are random and that the events being considered here are not independent. The proofs given in the paper provide a formal argument.

[12] We also show that the fraction of times the SPNE is played converges to one with probability one.

$P(v_t(r)=3)$ both converge to one, so the probability of player 2 choosing r converges to one as well. Moving up a level in the game tree, this means that the fraction of balls with value 3 goes to one in the urn associated with action R, so $P(v_t(R)=3)$ goes to one. Since L is played infinitely often, $P(v_t(L)=2)$ goes to one, which means $P(v_t(R) > v_t(L))$ also converges to one as well. Therefore, the probability of player 1 playing R and player 2 playing r, which is the probability of playing the SPNE, converges to one.

This example suggests that much of the work needed in showing that the play converges to a SPNE is performed by the induction procedure. If, at each decision node, a valuation process can generate enough exploration while making the valuation converge to the most frequently received payoff, then the induction takes care of the rest and leads the play to the SPNE. However, as seen in the next section, the induction arguments are made complicated by the fact that decision nodes are reached in random periods.

3 Main Results

Our main results consist of two parts. We first provide two conditions on general valuation processes that together guarantee that the play converges to a SPNE. Because these conditions are abstract, we then provide more intuitive conditions that guarantee convergence to a SPNE in a smaller class of valuation processes, which we call additively separable valuations.

3.1 General Valuations

The two general conditions form the inductive steps in proofs by induction, which must proceed along the game tree at random times since the process by which a successor node is chosen has random component. Therefore, the conditions need to be stated in terms of random times in which they are required to hold. To that end, we define the following. Let $(\Omega, \mathscr{F}, P)$ be the probability space on which a valuation process $\{v_t : t \in \mathbb{Z}_{++}\}$ is defined. Let $\mathscr{F}_0 = \{\emptyset, \Omega\}$, and let $\mathscr{F}_t = \sigma(v_1, ..., v_t)$ be the sub-σ-field consisting of events up to time t. For each node $z \in G$, let $\tau_0^z = 0$ and for all $n \in \mathbb{Z}_{++}$, let $\tau_n^z = \inf\{t > \tau_{n-1}^z : z \in \xi_t\}$ be the n-th time the node z has been reached.[13]

[13] Random variable τ_n^z is a stopping time. The following facts about stopping times are used throughout the paper. For any stopping time τ, $\mathscr{F}_\tau = \{B \in \mathscr{F} : \forall n \; B \cap \{\tau \le n\} \in \mathscr{F}_n\}$ is a σ-field consisting of events up to (random) time τ. If $\tau_0 < \tau_1 < \tau_2 < \cdots$ almost surely, then $\{\mathscr{F}_{\tau_n} : n \in \mathbb{Z}_+\}$, where $\mathbb{Z}_+ = \{0, 1, 2, ...\}$, is a filtration. Moreover, if $\{Y_t : t \in \mathbb{Z}_+\}$ is adapted to $\{\mathscr{F}_t : t \in \mathbb{Z}_+\}$, then Y_{τ_n} is adapted to $\{\mathscr{F}_{\tau_n} : n \in \mathbb{Z}_+\}$, and if $Y_t \to Y$ almost surely as $t \to \infty$, then $Y_{\tau_n} \to Y$ almost surely as $n \to \infty$.

Our first condition places a limit on how fast the probability of taking any given action is allowed to go to zero by requiring that the sum of these probabilities over the periods in which the choice is being considered is unbounded. Lemma 1 below shows that every action is played infinitely often with probability one if and only if this condition is satisfied for all players.

Assumption 1. For each decision node $z \in G^i$ and $a \in A_z$, the following holds.[14]

$$\text{On } \left\{\tau_n^z < \infty \text{ for all } n\right\}, \text{ we have } \sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau_n^z}(a') \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) = \infty.$$

Lemma 1. *Let $\mathscr{G}$ be a finite perfect-information game. Then $N_t(a) \to \infty$ for all $a \in A$ with probability one if and only if the valuation process $\{v_t^i : t \in \mathbb{Z}_{++}\}$ satisfies Assumption 1 for all i.*

Proof. ($\Leftarrow$) Suppose Assumption 1 is satisfied for all players. We use induction on $\mathscr{G}$ to show that $N_t(a) \to \infty$ almost surely (a.s.) for all $a \in A$. As the basis for the induction, we note that $\tau_n^{z_0} < \infty$ for all n since $\tau_n^{z_0} = n$. Next, as the induction hypothesis, assume that $\tau_n^z < \infty$ for all n a.s. For any $a \in A_z$, we have

$$\begin{aligned}\left\{N_t(a) \to \infty\right\} &= \left\{v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau_n^z}(a') \text{ infinitely often}\right\} \\ &= \left\{\sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau_n^z}(a') \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) = \infty\right\}\end{aligned}$$

by the conditional version of Borel-Cantelli lemma.[15] Assumption 1 thus implies that $N_t(a) \to \infty$ a.s., which in turn means that $\tau_n^{\zeta(a)} < \infty$ for all n a.s. Therefore, by induction, we have $N_t(a) \to \infty$ for all $a \in A$.

($\Rightarrow$) Take any decision node $z \in G$ and $a \in A_z$. Since $N_t(a) \to \infty$ a.s. by assumption, we must have $\tau_n^z < \infty$ a.s. Then the conditional Borel-Cantelli lemma implies

$$\sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau_n^z}(a') \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) = \infty \text{ a.s.,}$$

so Assumption 1 is trivially satisfied for all players. □

Our second condition requires the following. Suppose, for every node z' that succeeds z, the fraction of times a SPNE of $\mathscr{G}_{z'}$ is played when node z' is played converges to one. Then we require that the probability of taking

[14] We use phrase "on B, C" (or equivalently, "C on B") to mean that for every $\omega \in B$, property C holds.

[15] See, for example, Durrett [3], Theorem 5.3.2.

a SPNE action at z must converge to one as well.[16] To state the condition formally, recall that, for terminal node z, $u^i(z)$ is defined as player i's payoff from z. We now extend the definition to non-terminal nodes by letting $u^i(z)$ be player i's (unique) SPNE payoff in subgame $\mathscr{G}_z$. We also use $u^i(a)$ to denote $u^i(\zeta(a))$. Then $\tilde{A}_z = \arg\max_{a \in A_z} u^{i(z)}(a)$ is the set of actions specified at node z by some SPNE of $\mathscr{G}_z$. We let $N_t(z)$ denote the number of times z has been played and $S_t(z)$ denote the number of times some SPNE of $\mathscr{G}_z$, not necessarily the same one every time, has been played up to and including period t.

Assumption 2. For each decision node $z \in G^i$ and $a \in A_z \setminus \tilde{A}_z$, the following holds.

$$\text{On } \left\{ N_t(z') \to \infty \text{ and } \frac{S_t(z')}{N_t(z')} \to 1 \text{ for all } z' \in G_z \setminus \{z\} \right\}, \text{ we have}$$
$$P\left(\max_{\tilde{a}_z \in \tilde{A}_z} \left\{ v_{\tau_n^z}(\tilde{a}_z) \right\} > v_{\tau_n^z}(a) \;\middle|\; \mathscr{F}_{\tau_{n-1}^z} \right) \to 1 \text{ as } n \to \infty.$$

Let $\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset$ mean that an action specified by some SPNE of $\mathscr{G}_z$ was chosen at time τ_n^z. Theorem 2 shows that, for valuations satisfying Assumption 1, the fraction of times some SPNE of $\mathscr{G}_z$ is played, relative to the number of times z is played, converges to one almost surely if and only if Assumption 2 is satisfied for all players. Corollary 3 shows that the probability of playing a SPNE action at node z, conditioned on reaching z, also converges to one. Since these two results apply to the root node as well, the probability of playing a SPNE of $\mathscr{G}$ and the fraction of times some SPNE of the game is played converges to one if Assumptions 1 and 2 are satisfied.

Theorem 2. *Let $\mathscr{G} \in \Gamma$. Suppose valuation process $\{v_t^i : t \in \mathbb{Z}_{++}\}$ satisfies Assumption 1 for all i. Then $P\left(\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \to 1$ with probability one as $n \to \infty$, and $S_t(z)/N_t(z) \to 1$ with probability one as $t \to \infty$ for every decision node $z \in G$ if and only if the valuation process satisfies Assumption 2 for all i.*

Proof. We first note that $\tau_n^z < \infty$ a.s. for all $n \in \mathbb{Z}_+$ and $z \in G$ by Lemma 1.

($\Leftarrow$) Suppose Assumption 2 holds for all i. Let $L(\mathscr{G}_z)$ denote the number of nodes in a longest path from z to a terminal node of $\mathscr{G}_z$. As the basis for the induction, we note that if $L(\mathscr{G}_z) = 1$, then z is a terminal node, which means $S_t(z) = N_t(z)$. Thus, $S_t(z)/N_t(z) \to 1$ a.s. trivially. As the induction hypothesis, assume that for all subgame $\mathscr{G}_{z'}$ such that $L(\mathscr{G}_{z'}) \leq m$, we have $S_t(z')/N_t(z') \to 1$ a.s. as $t \to \infty$.

[16] This assumption may appear strong at first glance. However, the assumption is in “if...then...” form, and it is the hypothesis part of the condition that is strong, which makes the assumption as a whole weak.

Let $z \in G$ be such that $L(\mathscr{G}_z) = m+1$, and let $i = i(z)$. Then by Assumption 2, for all $a \in A_z \setminus \tilde{A}_z$,

$$P\left(\max_{\tilde{a}_z \in \tilde{A}_z} \{v_{\tau_n^z}(\tilde{a}_z)\} > v_{\tau_n^z}(a) \mid \mathscr{F}_{\tau_{n-1}^z}\right) \;\to\; 1 \;\text{ a.s. as } n \to \infty.$$

Therefore,

$$\begin{aligned}
P\left(\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset \mid \mathscr{F}_{\tau_{n-1}^z}\right) &= P\left(\max_{\tilde{a}_z \in \tilde{A}_z} \{v_{\tau_n^z}(\tilde{a}_z)\} > v_{\tau_n^z}(a) \;\text{ for all } a \in A_z \setminus \tilde{A}_z \mid \mathscr{F}_{\tau_{n-1}^z}\right) \\
&= P\left(\bigcap_{a \in A_z \setminus \tilde{A}_z} \left\{\max_{\tilde{a}_z \in \tilde{A}_z} \{v_{\tau_n^z}(\tilde{a}_z)\} > v_{\tau_n^z}(a)\right\} \mid \mathscr{F}_{\tau_{n-1}^z}\right) \\
&= 1 - P\left(\bigcup_{a \in A_z \setminus \tilde{A}_z} \left\{\max_{\tilde{a}_z \in \tilde{A}_z} \{v_{\tau_n^z}(\tilde{a}_z)\} \leq v_{\tau_n^z}(a)\right\} \mid \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\geq 1 - \sum_{a \in A_z \setminus \tilde{A}_z} P\left(\max_{\tilde{a}_z \in \tilde{A}_z} \{v_{\tau_n^z}(\tilde{a}_z)\} \leq v_{\tau_n^z}(a) \mid \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\to 1 \;\text{ a.s. as } n \to \infty.
\end{aligned}$$

Furthermore, this implies that, as $n \to \infty$,

$$\frac{\sum_{k=1}^n E\left[\mathbb{1}\left(\tilde{A}_z \cap \xi_{\tau_k^z} \neq \emptyset\right) \mid \mathscr{F}_{\tau_{k-1}^z}\right]}{n} = \frac{\sum_{k=1}^n P\left(\tilde{A}_z \cap \xi_{\tau_k^z} \neq \emptyset \mid \mathscr{F}_{\tau_{k-1}^z}\right)}{n} \;\to\; 1 \;\text{ a.s.}$$

By the stability theorem for dependent variables,[17]

$$\frac{\sum_{k=1}^n \left(\mathbb{1}\left(\tilde{A}_z \cap \xi_{\tau_k^z} \neq \emptyset\right) - E\left[\mathbb{1}\left(\tilde{A}_z \cap \xi_{\tau_k^z} \neq \emptyset\right) \mid \mathscr{F}_{\tau_{k-1}^z}\right]\right)}{n} \;\to\; 0 \;\text{ a.s. as } n \to \infty.$$

This yields,

$$\frac{\sum_{k=1}^n \mathbb{1}\left(\tilde{A}_z \cap \xi_{\tau_k^z} \neq \emptyset\right)}{n} \;\to\; 1 \;\text{ a.s. as } n \to \infty.$$

Therefore,

$$\begin{aligned}
\sum_{\tilde{a}_z \in \tilde{A}_z} \frac{N_t(\tilde{a}_z)}{N_t(z)} &= \frac{\sum_{k=1}^t \sum_{\tilde{a}_z \in \tilde{A}_z} \mathbb{1}(\tilde{a}_z \in \xi_k)}{N_t(z)} = \frac{\sum_{k=1}^t \mathbb{1}\left(\tilde{A}_z \cap \xi_k \neq \emptyset\right)}{N_t(z)} \\
&= \frac{\sum_{k=1}^{N_t(z)} \mathbb{1}\left(\tilde{A}_z \cap \xi_{\tau_k^z} \neq \emptyset\right)}{N_t(z)} \;\to\; 1 \;\text{ a.s. as } t \to \infty.
\end{aligned}$$

By the induction hypothesis, $S_t(\zeta(\tilde{a}_z))/N_t(\zeta(\tilde{a}_z)) \to 1$ a.s. for all $\tilde{a}_z \in \tilde{A}_z$. Using the fact that $(x_t - x_t y_t) \to 0$ if $y_t \to 1$, we obtain

$$\sum_{\tilde{a}_z \in A_z} \left[\frac{N_t(\tilde{a}_z)}{N_t(z)} - \left(\frac{N_t(\tilde{a}_z)}{N_t(z)}\right)\left(\frac{S_t(\zeta(\tilde{a}_z))}{N_t(\zeta(\tilde{a}_z))}\right)\right] \;\to\; 0 \text{ a.s. as } t \to \infty.$$

[17] See, for example, Loeve [9], p. 53.

Thus,

$$\sum_{\tilde{a}_z \in \tilde{A}_z} \left(\frac{N_t(\tilde{a}_z)}{N_t(z)}\right)\left(\frac{S_t(\zeta(\tilde{a}_z))}{N_t(\zeta(\tilde{a}_z))}\right) \to 1 \text{ a.s. as } t \to \infty.$$

Since SPNE of $\mathscr{G}_z$ are essentially the same, any $\tilde{a}_z \in \tilde{A}_z$ followed by any SPNE of $\mathscr{G}_{\zeta(\tilde{a}_z)}$ is a SPNE of $\mathscr{G}_z$. Thus, $S_t(z) = \sum_{\tilde{a}_z \in \tilde{A}_z} S_t(\zeta(\tilde{a}_z))$. Then, using the fact that $N_t(\tilde{a}_z) = N_t(\zeta(\tilde{a}_z))$, we obtain

$$\frac{S_t(z)}{N_t(z)} = \frac{\sum_{\tilde{a}_z \in \tilde{A}_z} S_t(\zeta(\tilde{a}_z))}{N_t(z)} = \sum_{\tilde{a}_z \in \tilde{A}_z} \left(\frac{N_t(\tilde{a}_z)}{N_t(z)}\right)\left(\frac{S_t(\zeta(\tilde{a}_z))}{N_t(\zeta(\tilde{a}_z))}\right) \to 1 \text{ a.s. as } t \to \infty,$$

as desired.

($\Rightarrow$) Suppose, for any decision node $z \in G$, $P\left(\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset \mid \mathscr{F}_{\tau_{n-1}^z}\right) \to 1$ a.s. as $n \to \infty$. Then

$$P\left(\max_{\tilde{a}_z \in \tilde{A}_z} \left\{v_{\tau_n^z}(\tilde{a}_z)\right\} > v_{\tau_n^z}(a) \mid \mathscr{F}_{\tau_{n-1}^z}\right) \geq P\left(\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset \mid \mathscr{F}_{\tau_{n-1}^z}\right) \to 1 \text{ a.s.}$$

Therefore, Assumption 2 is satisfied trivially. □

Corollary 3. *Suppose valuation process $\{v_t^i : t \in \mathbb{Z}_{++}\}$ satisfies Assumptions 1 and 2 for all i. Then, for every decision node $z \in G$, $P(\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset) \to 1$ as $n \to \infty$.*

Proof. By Theorem 2,

$$E\left[\mathbb{1}\left(\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset\right) \mid \mathscr{F}_{\tau_{n-1}^z}\right] = P\left(\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset \mid \mathscr{F}_{\tau_{n-1}^z}\right) \to 1 \text{ a.s. as } n \to \infty.$$

Taking expectations yield $P(\tilde{A}_z \cap \xi_{\tau_n^z} \neq \emptyset) \to 1$ as $n \to \infty$ by the dominated convergence theorem. □

3.2 Additively Separable Valuations

Let $\bar{u}_z$ and $\underline{u}_z$ be the highest and the lowest possible payoffs for player $i(z)$ in $\mathscr{G}_z$. Similarly, let $\bar{u}_a$ and $\underline{u}_a$ be the highest and the lowest possible payoffs for player $i(a)$ in $\mathscr{G}_{\zeta(a)}$. We say that a valuation process of player i is *additively separable* if, for every decision node $z \in G^i$ and $a \in A_z$, $v_t(a) = f_t(a) + e_t(a)$, where $f_t(a)$, interpreted as the empirical term, has a support in $[\underline{u}_a, \bar{u}_a]$, and $e_t(a)$, interpreted as the error term, has a support containing $[0, (\bar{u}_z - \underline{u}_z) + \bar{c}]$ for some $\bar{c} > 0$. We further assume that for any a and a' in A_z, $e_t(a)$ and $e_t(a')$ are independent when conditioned on $\mathscr{F}_{t-1}$.

The following theorem shows that an additively separable valuation process satisfies Assumptions 1 and 2 if three conditions are satisfied. The conditions can be roughly interpreted in the following way. The first two conditions require that the error term converges to zero “in probability” if and only if the number of times that action had been taken goes to infinity. The third condition requires that the empirical term for an action converges to u “in probability” if the fraction of times payoff u is received after taking that action converges to one.

Theorem 4. *Suppose an additively separable valuation process $\{v_t^i : t \in \mathbb{Z}_{++}\}$ is such that for each decision node $z \in G^i$ and $a \in A_z$,*

(i) On $\{N_t(a) \to \infty \text{ as } t \to \infty\}$, $P\left(|e_{\tau_n^z}(a)| > c \mid \mathscr{F}_{\tau_{n-1}^z}\right) \to 0$ as $n \to \infty$ for all $c > 0$, and

(ii) On $\{N_t(a) \not\to \infty \text{ as } t \to \infty\}$, $P\left(e_{\tau_n^z}(a) > c \mid \mathscr{F}_{\tau_{n-1}^z}\right) \not\to 0$ as $n \to \infty$ for all $c \in (0, \bar{u}_z - \underline{u}_z + \bar{c})$.

Then Assumption 1 is satisfied. Suppose in addition $\{v_t^i : t \in \mathbb{Z}_{++}\}$ satisfies

$$\textit{(iii)}\ \text{On}\ \left\{N_t(a) \to \infty\ \text{ and }\ \frac{\sum_{n=1}^t \mathbb{1}\left(a \in \xi_n \text{ and } u_n^i = u\right)}{N_t(a)} \to 1 \text{ as } t \to \infty\right\},$$

$$P\left(|f_{\tau_n^z}(a) - u| > c \mid \mathscr{F}_{\tau_{n-1}^z}\right) \to 0 \text{ as } n \to \infty \text{ for all } c > 0.$$

Then Assumption 2 is satisfied.

Proof. Suppose $\{v_t^i : t \in \mathbb{Z}_{++}\}$ satisfies conditions (i) and (ii). Assume towards contradiction that Assumption 1 is not satisfied for some $\hat{a} \in A_z$, where $z \in G^i$. For each $a \in A_z$, let

$$\Omega_a = \left\{\tau_n^z < \infty \text{ for all } n \text{ and } \sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau_n^z}(a') \mid \mathscr{F}_{\tau_{n-1}^z}\right) < \infty\right\}.$$

Enumerate the elements of A_z as $a_1, \ldots, a_{n_z}$, where $a_1 = \hat{a}$ and $n_z = |A_z|$, and construct $B_z \subset A_z$ in the following iterative way. Let $B_1 = \{a_1\}$. For $m = 2, 3, \ldots, n_z$, let $B_m = B_{m-1} \cup \{a_m\}$ if

$$\left(\bigcap_{a \in B_{m-1}} \Omega_a\right) \bigcap \Omega_{a_m} \neq \emptyset.$$

Otherwise, set $B_m = B_{m-1}$. Let $B_z = B_{n_z}$ and $\Omega_z = \bigcap_{a \in B_z} \Omega_a$. By construction, Ω_z is non-empty.

In the following, assume that we are on Ω_z. Then

$$\sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau_n^z}(a') \mid \mathscr{F}_{\tau_{n-1}^z}\right) < \infty \text{ for all } a \in B_z$$

$$\sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau_n^z}(a') \mid \mathscr{F}_{\tau_{n-1}^z}\right) = \infty \text{ for all } a \in A_z \setminus B_z.$$

By the conditional Borel-Cantelli lemma, B_z consists of all the actions in A_z that occur only finitely often, and $A_z \setminus B_z$ consists of those that occur infinitely often. Fix any $c \in (0, \bar{c})$. Then, for any $a \in B_z$ and $a' \in A_z \setminus B_z$, we have

$$\begin{aligned}
P\left(v_{\tau_n^z}(a) > \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) &\geq P\left(f_{\tau_n^z}(a) \in [\underline{u}_a, \bar{u}_a] \text{ and } e_{\tau_n^z}(a) > (\bar{u}_z - \underline{u}_a) + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&= P\left(e_{\tau_n^z}(a) > (\bar{u}_z - \underline{u}_a) + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\nrightarrow 0 \;\text{ by condition } (ii)\text{, and} \\
P\left(v_{\tau_n^z}(a') \leq \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) &\geq P\left(f_{\tau_n^z}(a') \in [\underline{u}_{a'}, \bar{u}_{a'}] \text{ and } e_{\tau_n^z}(a') \leq c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&= P\left(e_{\tau_n^z}(a') \leq c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\geq P\left(|e_{\tau_n^z}(a')| \leq c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\to 1 \;\text{ by condition } (i).
\end{aligned}$$

Thus, appealing to the conditional independence of the error terms, we obtain

$$\begin{aligned}
&\sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus B_z} v_{\tau_n^z}(a') \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
\geq\; &\sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \bar{u}_z + c \text{ and } \forall a' \in A_z \setminus B_z,\; v_{\tau_n^z}(a') \leq \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
\geq\; &\sum_{n=1}^{\infty} P\left(e_{\tau_n^z}(a) > \bar{u}_z - \underline{u}_a + c \text{ and } \forall a' \in A_z \setminus B_z,\; e_{\tau_n^z}(a') \leq c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
\geq\; &\sum_{n=1}^{\infty} \left[P\left(e_{\tau_n^z}(a) > \bar{u}_z - \underline{u}_a + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \prod_{a' \in A_z \setminus B_z} P\left(e_{\tau_n^z}(a') \leq c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right)\right] \\
=\; &\infty \;\text{ since } x_n \nrightarrow 0 \text{ and } y_n \to 1 \text{ implies that } \sum_n x_n y_n = \infty.
\end{aligned}$$

Thus, $v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus B_z} v_{\tau_n^z}(a')$ infinitely many times. Since B_z consists of actions that are chosen only finitely often, this means that a must be chosen infinitely often. However, this contradicts the assumption that $a \in B_z$. Therefore Assumption 1 must be satisfied.

To see that Assumption 2 is satisfied, consider any decision node $z \in G^i$ and $\tilde{a}_z \in \tilde{A}_z$, and set

$$c \;=\; \frac{u^i(\tilde{a}_z) - \max\left\{u^i(a) : a \in A_z \setminus \tilde{A}_z\right\}}{5}.$$

For the remainder of the proof, assume that we are on

$$\left\{N_t(z') \to \infty \;\text{ and }\; \frac{S_t(z')}{N_t(z')} \to 1 \text{ for all } z' \in G_z \setminus \{z\}\right\}.$$

Then, for all $a \in A_z$,

$$\frac{\sum_{n=1}^{t} \mathbb{1}\left(a \in \xi_n \text{ and } u_n^i = u^i(a)\right)}{N_t(a)} \;\geq\; \frac{S_t(\zeta(a))}{N_t(\zeta(a))} \;\to\; 1 \quad \text{as } t \to \infty$$

since every SPNE payoff of $\zeta(a)$ is unique. By conditions (*i*) and (*iii*), this implies

$$P\left(|f_{\tau^z_n}(a) - u^i(a)| > c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \;\to\; 0 \quad \text{and} \quad P\left(|e_{\tau^z_n}(a)| > c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \;\to\; 0.$$

Therefore, for all $a \in A_z \setminus \tilde{A}_z$, we have

$$\begin{aligned}
&P\left(v_{\tau^z_n}(\tilde{a}_z) > v_{\tau^z_n}(a) \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right)\\
&\geq P\left(|f_{\tau^z_n}(\tilde{a}_z) - u^i(\tilde{a}_z)| \leq c,\; |e_{\tau^z_n}(\tilde{a}_z)| \leq c,\; |f_{\tau^z_n}(a) - u^i(a)| \leq c, \text{ and } |e_{\tau^z_n}(a)| \leq c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right)\\
&\geq 1 \;-\; P\left(|f_{\tau^z_n}(\tilde{a}_z) - u^i(\tilde{a}_z)| > c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \;-\; P\left(|e_{\tau^z_n}(\tilde{a}_z)| > c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right)\\
&\qquad\qquad - P\left(|f_{\tau^z_n}(a) - u^i(a)| > c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \;-\; P\left(|e_{\tau^z_n}(a)| > c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right)\\
&\to 1.
\end{aligned}$$

Therefore, Assumption 2 is satisfied. □

4 Examples

We give four examples of valuation processes satisfying our conditions. The first two examples are additively separable valuations that satisfy the conditions of Theorem 4. The remaining two are examples of non-additively separable valuations that satisfy Assumptions 1 and 2.

4.1 Simple Recollection

We begin by showing that the simple recollection rule is additively separable. Recall that the valuation process is given by

$$v_t(a) \;=\; \mathbb{1}\left(\eta^a_t \leq \frac{1}{1 + N_{t-1}(a)}\right)\varepsilon^a_t \;+\; \sum_{k=1}^{N_{t-1}(a)} \mathbb{1}\left(\eta^a_t \in \left(\frac{k}{1 + N_{t-1}(a)}, \frac{k+1}{1 + N_{t-1}(a)}\right]\right) u^i_{\tau^a_k},$$

To see that this is additively separable, choose any $u_a \in [\underline{u}_a, \bar{u}_a]$ and let

$$f_t(a) \;=\; \mathbb{1}\left(\eta^a_t \leq \frac{1}{1 + N_{t-1}(a)}\right)u_a \;+\; \sum_{k=1}^{N_{t-1}(a)} \mathbb{1}\left(\eta^a_t \in \left(\frac{k}{1 + N_{t-1}(a)}, \frac{k+1}{1 + N_{t-1}(a)}\right]\right) u^i_{\tau^a_k}$$

$$\text{and} \quad e_t(a) \;=\; \mathbb{1}\left(\eta^a_t \leq \frac{1}{1 + N_{t-1}(a)}\right)\left(\varepsilon^a_t - u_a\right).$$

Choose any $\bar{c} > 0$. Then the support of $f_t(a)$ is in $[\underline{u}_a, \bar{u}_a]$, the support of $e_t(a)$ contains $[0, (\bar{u}_z - \underline{u}_z) + \bar{c}]$, and the error terms are conditionally independent. We now show that it satisfies the conditions of Theorem 4.

Theorem 5. *The simple recollection rule satisfies conditions (i)-(iii) of Theorem 4.*

Proof. Fix any decision node $z \in G^i$ and $a \in A_z$. On $\{N_t(a) \to \infty \text{ as } t \to \infty\}$, we have the following for all $c > 0$:

$$\begin{aligned}
P\left(|e_{\tau^z_n}(a)| > c \mid \mathscr{F}_{\tau^z_{n-1}}\right) &= P\left(\left|\mathbb{1}\left(\eta^a_{\tau^z_n} \leq \frac{1}{1+N_{\tau^z_{n-1}}(a)}\right)\left(\varepsilon^a_{\tau^z_n} - u_a\right)\right| > c \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&= P\left(\eta^a_{\tau^z_n} \leq \frac{1}{1+N_{\tau^z_{n-1}}(a)} \text{ and } |\varepsilon^a_{\tau^z_n} - u_a| > c \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&= P\left(\eta^a \leq \frac{1}{1+N_{\tau^z_{n-1}}(a)} \mid \mathscr{F}_{\tau^z_{n-1}}\right) P\left(|\varepsilon^a - u_a| > c \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&= \frac{P\left(|\varepsilon^a - u_a| > c\right)}{1+N_{\tau^z_{n-1}}(a)} \to 0 \ \text{ since } N_{\tau^z_{n-1}}(a) \to \infty.
\end{aligned}$$

Thus, condition (i) is satisfied.

Next, on $\{N_t(a) \not\to \infty \text{ as } t \to \infty\}$, we have the following for all $c \in (0, \bar{u}_z - \underline{u}_z + \bar{c})$.

$$\begin{aligned}
P\left(e_{\tau^z_n}(a) > c \mid \mathscr{F}_{\tau^z_{n-1}}\right) &= P\left(\eta^a_{\tau^z_n} \leq \frac{1}{1+N_{\tau^z_{n-1}}(a)} \text{ and } \varepsilon^a_{\tau^z_n} - u_a > c \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&= P\left(\eta^a \leq \frac{1}{1+N_{\tau^z_{n-1}}(a)} \mid \mathscr{F}_{\tau^z_{n-1}}\right) P\left(\varepsilon^a - u_a > c \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&= \frac{P\left(\varepsilon^a > u_a + c\right)}{1+N_{\tau^z_{n-1}}(a)} \not\to 0 \ \text{ since } N_{\tau^z_{n-1}}(a) \not\to \infty.
\end{aligned}$$

Therefore, condition (ii) is satisfied.

Lastly, on

$$\left\{ N_t(a) \to \infty \ \text{ and } \ \frac{\sum_{n=1}^{t} \mathbb{1}\left(a \in \xi_n \text{ and } u^i_n = u\right)}{N_t(a)} \to 1 \text{ as } t \to \infty \right\},$$

we have the following.

$$\begin{aligned}
P\left(|f_{\tau^z_n}(a) - u| > c \mid \mathscr{F}_{\tau^z_{n-1}}\right) &\leq 1 - P\left(f_{\tau^z_n}(a) = u \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&\leq 1 - \frac{\sum_{k=1}^{\tau^z_{n-1}} \mathbb{1}\left(a \in \xi_k \text{ and } u^i_k = u\right)}{1+N_{\tau^z_{n-1}}(a)} \to 0 \ \text{ as } n \to \infty.
\end{aligned}$$

Thus, condition (iii) is satisfied as well. □

4.2 Sample Averaging

For another example of an additively separable valuation process satisfying the conditions of Theorem 4, consider the following modification of Jehiel and Samet's model, which we call the sample averaging rule. When evaluating an action, a player using this rule tries to use the average of the past payoffs associated with the action. The player is assumed to have an imperfect ability to calculate historical averages, so the valuation assigned to an action is a perturbed average of the past payoffs; however, the error associated with evaluating an action decreases as the number of times that action had been taken increases.

To state the model formally, for all $a \in A^i$ and $t \in \mathbb{Z}_{++}$, let ε_t^a be an independent copy of random variable ε^a that has support $\mathbb{R}$. The valuation of action a is given by

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

where $\widetilde{N}_{t-1}(a) = \max\{1,\ N_{t-1}(a)\}$.

To see that the sample averaging model is additively separable, choose any $u_a \in [\underline{u}_a, \bar{u}_a]$ and let

$$f_t(a) \;=\; \frac{u_a \mathbb{1}(N_{t-1}(a) = 0)}{\widetilde{N}_{t-1}(a)} \;+\; \frac{\sum_{n=1}^{t-1} u_n^i \mathbb{1}(a \in \xi_n)}{\widetilde{N}_{t-1}(a)}, \quad \text{and}$$

$$e_t(a) \;=\; \frac{\varepsilon_t^a - u_a \mathbb{1}(N_{t-1}(a) = 0)}{\widetilde{N}_{t-1}(a)}.$$

Choose any $\bar{c} > 0$. Then the support of $f_t(a)$ is in $[\underline{u}_a, \bar{u}_a]$, the support of $e_t(a)$ contains $[0, (\bar{u}_z - \underline{u}_z) + \bar{c}]$, and the error terms are conditionally independent. The following theorem shows that it satisfies the conditions of Theorem 4.

Theorem 6. *The sample averaging rule satisfies conditions (i)-(iii) of Theorem 4.*

Proof. Fix any decision node $z \in G^i$ and $a \in A_z$. On $\{N_t(a) \to \infty \text{ as } t \to \infty\}$, we have the following for all $c > 0$:

$$\begin{aligned}
P\left(|e_{\tau_n^z}(a)| > c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \;&=\; P\left(\left|\frac{\varepsilon_{\tau_n^z}^a - u_a \mathbb{1}\left(N_{\tau_{n-1}^z}(a) = 0\right)}{\widetilde{N}_{\tau_{n-1}^z}(a)}\right| > c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\leq\; P\left(\frac{\left|\varepsilon_{\tau_n^z}^a\right| + \left|u_a \mathbb{1}\left(N_{\tau_{n-1}^z}(a) = 0\right)\right|}{\widetilde{N}_{\tau_{n-1}^z}(a)} > c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\leq\; P\left(\left|\varepsilon^a\right| > c\widetilde{N}_{\tau_{n-1}^z}(a) - |u_a| \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\to\; 0 \;\text{ as } n \to \infty \text{ since } N_{\tau_{n-1}^z}(a) \to \infty.
\end{aligned}$$

Thus, condition (*i*) is satisfied.

Next, on $\{N_t(a) \not\to \infty \text{ as } t \to \infty\}$, we have the following for all $c \in (0, \bar{u}_z - \underline{u}_z + \bar{c})$.

$$
\begin{aligned}
P\left(e_{\tau_n^z}(a) > c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) &= P\left(\frac{\varepsilon^a_{\tau_n^z} - u_a \mathbb{1}\left(N_{\tau_{n-1}^z}(a) = 0\right)}{\widetilde{N}_{\tau_{n-1}^z}(a)} > c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&= P\left(\varepsilon^a > c\widetilde{N}_{\tau_{n-1}^z}(a) + u_a \mathbb{1}\left(N_{\tau_{n-1}^z}(a) = 0\right) \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\geq P\left(\varepsilon^a > c\widetilde{N}_{\tau_{n-1}^z}(a) + |u_a| \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\not\to 0 \;\; \text{since the support of } \varepsilon^a \text{ is } \mathbb{R} \text{ and } N_{\tau_{n-1}^z}(a) \not\to \infty.
\end{aligned}
$$

Therefore, condition (*ii*) is satisfied.

Lastly, on

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

we have the following.

$$
f_t(a) = \frac{u_a \mathbb{1}(N_{t-1}(a) = 0)}{\widetilde{N}_{t-1}(a)} + \frac{\sum_{n=1}^{t-1} u_n^i \mathbb{1}(a \in \xi_n)}{\widetilde{N}_{t-1}(a)} \;\to\; u \;\; \text{since } N_t(a) \to \infty.
$$

Thus, condition (*iii*) is satisfied as well. □

4.3 Two-Moves Foresight

In this model, we add one additional layer of sophistication to the sample averaging rule by allowing players to "see two moves." That is, a player using this rule, say player i, is aware of the average historical payoffs associated with actions that are available at the current node and the immediate successor nodes. In particular, we assume that if $a \in A^i$ leads to a terminal node, then player i's valuation of a is just the perturbed average of her historical payoffs associated with a. If a leads to another decision node z' that belongs to player i, then player i looks at the perturbed average of her historical payoffs associated with each action in $A_{z'}$ and assigns the highest one as the valuation of a. If a leads to a decision node z' that belongs to player j, then player i thinks that player j will choose the action in $A_{z'}$ that has the highest valuation for j in player i's estimation. Thus, player i's valuation of a is the perturbed sample average of her payoffs that are associated with what she perceives will be player j's choice.

To formalize the model, first define for all $j \in \mathscr{I}$ and $a \in A$,

$$
v_t^j(a) = \frac{\sum_{n=1}^{t-1} u_n^j \mathbb{1}(a \in \xi_n)}{\widetilde{N}_{t-1}(a)},
$$

where $\widetilde{N}_{t-1}(a) = \max\{1,\ N_{t-1}(a)\}$, so that $v^j_t(a)$ is the sample average of player j's payoffs that are associated with action a. For all $a \in A$ and $t \in \mathbb{Z}_{++}$, let $\varepsilon^{i,a}_t$ be an independent copy of random variable $\varepsilon^{i,a}$ that has support $\mathbb{R}_+$. Letting $z' = \zeta(a)$, the valuation for $a \in A^i$ is given by

$$v_t(a) \;=\; \begin{cases} v^i_t(a) \;+\; \dfrac{\varepsilon^{i,a}_t}{\widetilde{N}_{t-1}(a)} & \text{if } z' \text{ is a terminal node} \\ \max\limits_{a' \in A_{z'}} \left\{ v^i_t(a') \;+\; \dfrac{\varepsilon^{i,a'}_t}{\widetilde{N}_{t-1}(a')} \right\} & \text{if } z' \text{ is a decision node and } i(z') = i \\ v^i_t(\hat{a}) \;+\; \dfrac{\varepsilon^{i,\hat{a}}_t}{\widetilde{N}_{t-1}(\hat{a})} & \text{if } z' \text{ is a decision node and } i(z') \neq i, \end{cases}$$

where

$$\hat{a} \;=\; \operatorname*{arg\,max}_{a' \in A_{z'}} \left\{ v^{i(z')}_t(a') \;+\; \frac{\varepsilon^{i,a'}_t}{\widetilde{N}_{t-1}(a')} \right\}.$$

In simple recollection and sample averaging models, players only respond to their own payoff experiences and do not think about their opponents. In fact, a player using these rules need not even recognize that she is playing a game against other players. In contrast, a player using two-moves foresight rule consciously considers how her opponents may react to her action choice. Thus, this rule models a qualitatively more sophisticated behavior than the first two learning rules. However, because two-moves foresight requires the player to keep track of her opponents' payoff experiences, it is only implementable in situations where opponents' payoffs are observable or where they can be deduced, as in "win-lose-or-draw" games.

Theorem 7. *The two-moves foresight rule satisfies Assumptions 1 and 2.*

Proof. Assume towards contradiction that Assumption 1 is not satisfied for some $z \in G^i$. Then, as shown in the proof of Theorem 4, there are non-empty sets $\Omega_z \subset \Omega$ and $B_z \subset A_z$ such that B_z consists of all the actions in A_z that occur only finitely often on Ω_z, and $A_z \setminus B_z$ consists of all the actions that occur infinitely often. Fix any $c > 0$, and suppose we are on Ω_z. Then for any $a \in B_z$ and $a'' \in A_{\zeta(a)}$, we have $N_t(a'') \leq N_t(a) \not\to \infty$. Therefore,

$$\begin{aligned} & P\left(v_{\tau^z_n}(a) > \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \\ \geq\; & P\left(\min_{a'' \in A_{\zeta(a)}} v^i_{\tau^z_n}(a'') \;+\; \frac{\min_{a'' \in A_{\zeta(a)}} \varepsilon^{i,a''}_{\tau^z_n}}{\max_{a'' \in A_{\zeta(a)}} \widetilde{N}_{\tau^z_{n-1}}(a'')} > \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \\ \geq\; & P\left(\underline{u}_z + \frac{\min_{a'' \in A_{\zeta(a)}} \varepsilon^{i,a''}_{\tau^z_n}}{\widetilde{N}_{\tau^z_{n-1}}(a)} > \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \\ \geq\; & P\left(\min_{a'' \in A_{\zeta(a)}} \varepsilon^{i,a''} > \widetilde{N}_{\tau^z_{n-1}}(a)\big(\bar{u}_z - \underline{u}_z + c\big) \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \not\to\; 0 \;\; \text{since } N_t(a) \not\to \infty. \end{aligned}$$

For all $a' \in A_z \setminus B_z$,

$$\begin{aligned}
&P\left(v_{\tau_n^z}(a') \le \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\ge P\left(\max_{a'' \in A_{\zeta(a')}} v^i_{\tau_n^z}(a'') \;+\; \frac{\max_{a'' \in A_{\zeta(a')}} \varepsilon^{i,a''}_{\tau_n^z}}{\min_{a'' \in A_{\zeta(a')}} \widetilde{N}_{\tau_{n-1}^z}(a'')} \le \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\ge P\left(\bar{u}_z + \max_{a'' \in A_{\zeta(a')}} \varepsilon^{i,a''}_{\tau_n^z} \le \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\ge P\left(\max_{a'' \in A_{\zeta(a')}} \varepsilon^{i,a''} \le c\right) \;=\; c_{a'} \;>\; 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus B_z} v_{\tau_n^z}(a') \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \;&\ge\; \sum_{n=1}^{\infty} \Bigg(P\left(v_{\tau_n^z}(a) > \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&\qquad \times \prod_{a' \in A_z \setminus B_z} P\left(v_{\tau_n^z}(a') \le \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right)\Bigg) \\
&\ge\; \left(\prod_{a' \in A_z \setminus B_z} c_{a'}\right) \sum_{n=1}^{\infty} P\left(v_{\tau_n^z}(a') \le \bar{u}_z + c \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \\
&=\; \infty.
\end{aligned}$$

This implies that a occurs infinitely often, which is a contradiction. Therefore, Assumption 1 is satisfied.

To see that Assumption 2 is satisfied, fix any decision node $z \in G^i$ and for the remainder of the proof assume that we are on

$$\left\{ N_t(z') \to \infty \;\text{ and }\; \frac{S_t(z')}{N_t(z')} \to 1 \text{ for all } z' \in G_z \setminus \{z\} \right\}.$$

Recalling that $u^i(z) = u^i(\tilde{a}_z) > u^i(a)$ for all $\tilde{a}_z \in \tilde{A}_z$ and $a \in A_z \setminus \tilde{A}_z$, let

$$k \;=\; \frac{u^i(z) - \max_{a \in A_z \setminus \tilde{A}_z} u^i(a)}{5}.$$

In the following, we show that for all $a \in A_z$,

$$P\left(v_{\tau_n^z}(a) \in \left(u^i(a) - 2k, u^i(a) + 2k\right) \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) \;\to\; 1 \;\text{ as } n \to \infty.$$

Suppose $a \in A_z$ is such that $\zeta(a)$ is a terminal node. Then $u^i_t = u^i(a)$ if

$a \in \xi_t$. Thus,

$$
\begin{aligned}
&P\left(v_{\tau^z_n}(a) \in \left(u^i(a)-2k,\, u^i(a)+2k\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&= P\left(\left(\frac{\sum_{t=1}^{\tau^z_{n-1}} u^i_t \mathbb{1}(a \in \xi_t)}{\widetilde{N}_{\tau^z_{n-1}}(a)} + \frac{\varepsilon^{i,a}_{\tau^z_n}}{\widetilde{N}_{\tau^z_{n-1}}(a)}\right) \in \left(u^i(a)-2k,\, u^i(a)+2k\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&\geq P\left(u^i(a)\left(\frac{N_{\tau^z_{n-1}}(a)}{\widetilde{N}_{\tau^z_{n-1}}(a)}\right) \in \left(u^i(a)-k,\, u^i(a)+k\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) P\left(\frac{\varepsilon^{i,a}}{\widetilde{N}_{\tau^z_{n-1}}(a)} \in (-k,k) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&\to 1 \;\text{ as } n \to \infty \text{ since } N_{\tau^z_n}(a) \to \infty.
\end{aligned}
$$

Next, suppose $a \in A_z$ is such that $z' = \zeta(a)$ is a decision node and $i(z') = i$. Let

$$
k' \;=\; \min\left\{k,\; \frac{u^i(a) - \max_{a' \in A_{z'} \setminus \tilde{A}_{z'}}\; u^i(a')}{5}\right\}.
$$

Then $k' > 0$ since $u^i(a)$ is the SPNE payoff of $\mathscr{G}_{z'}$. In addition, for any $a' \in A_{z'}$, $u^i(a')$ is the SPNE payoff of $\mathscr{G}_{\zeta(a')}$. Since $S_t(\zeta(a'))/N_t(\zeta(a')) \to 1$ by assumption, $\nu^i_t(a') \to u^i(a')$. Thus,

$$
\begin{aligned}
&P\left(v_{\tau^z_n}(a) \in \left(u^i(a)-2k,\, u^i(a)+2k\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&\geq P\left(\forall a' \in A_{z'},\; \left(\nu^i_{\tau^z_n}(a') \;+\; \frac{\varepsilon^{i,a'}_{\tau^z_n}}{\widetilde{N}_{\tau^z_{n-1}}(a')}\right) \in \left(u^i(a')-2k',\, u^i(a')+2k'\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&= \prod_{a' \in A_{z'}} P\left(\nu^i_{\tau^z_n}(a') + \frac{\varepsilon^{i,a'}_{\tau^z_n}}{\widetilde{N}_{\tau^z_{n-1}}(a')} \in \left(u^i(a')-2k',\, u^i(a')+2k'\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&\geq \prod_{a' \in A_{z'}} P\left(\nu^i_{\tau^z_n}(a') \in \left(u^i(a')-k',\, u^i(a')+k'\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&\qquad\qquad \times \prod_{a' \in A_{z'}} P\left(\frac{\varepsilon^{i,a'}}{\widetilde{N}_{\tau^z_{n-1}}(a')} \in \left(-k',\, k'\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
&\to 1 \;\text{ as } n \to \infty \text{ since } N_t(a') \to \infty \text{ and } \nu^i_t(a') \to u^i(a') \text{ for all } a' \in A_{z'}.
\end{aligned}
$$

Lastly, suppose $a \in A_z$ is such that $z' = \zeta(a)$ is a decision node and $i(z') = j \neq i$. Let

$$
k' \;=\; \min\left\{k,\; \frac{u^j(a) - \max_{a' \in A_{\zeta(a)} \setminus \tilde{A}_{z'}}\; u^j(a')}{5}\right\}.
$$

Letting,

$$
\hat{a} \;=\; \arg\max_{a' \in A_{z'}} \left\{\nu^j_t(a') \;+\; \frac{\varepsilon^{i,a'}_t}{\widetilde{N}_{t-1}(a')}\right\},
$$

we have

$$
\begin{aligned}
& P\left(\hat{a} \in \tilde{A}_{z'} \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
\geq\ & P\left(\forall a' \in A_{z'},\ \left(\nu^j_{\tau^z_n}(a') \ +\ \frac{\varepsilon^{i,a'}_{\tau^z_n}}{\tilde{N}_{\tau^z_{n-1}}(a')}\right) \in \left(u^j(a') - 2k',\, u^j(a') + 2k'\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
\to\ & 1 \ \text{ as } n \to \infty \ \text{ for similar reason as the } i(z') = i \text{ case.}
\end{aligned}
$$

Since $u^i(a) = u^i(z') = u^i(\tilde{a}_{z'})$ for any $\tilde{a}_{z'} \in \tilde{A}_{z'}$, we have

$$
\begin{aligned}
& P\left(v_{\tau^z_n}(a) \in \left(u^i(a) - 2k,\, u^i(a) + 2k\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
\geq\ & P\Bigg(\forall \tilde{a}_{z'} \in \tilde{A}_{z'},\ \left(\nu^i_{\tau^z_n}(\tilde{a}_{z'}) + \frac{\varepsilon^{i,\tilde{a}_{z'}}_{\tau^z_n}}{\tilde{N}_{\tau^z_{n-1}}(\tilde{a}_{z'})}\right) \in \left(u^i(a) - 2k,\, u^i(a) + 2k\right) \\
& \qquad\qquad\qquad \text{and } \hat{a} \in \tilde{A}_{z'} \mid \mathscr{F}_{\tau^z_{n-1}}\Bigg) \\
\geq\ & 1 \ -\sum_{\forall \tilde{a}_{z'} \in \tilde{A}_{z'}} P\Bigg(\Bigg(\left(\nu^i_{\tau^z_n}(\tilde{a}_{z'}) + \frac{\varepsilon^{i,\tilde{a}_{z'}}_{\tau^z_n}}{\tilde{N}_{\tau^z_{n-1}}(\tilde{a}_{z'})}\right) \notin \left(u^i(a) - 2k,\, u^i(a) + 2k\right) \mid \mathscr{F}_{\tau^z_{n-1}}\Bigg) \\
& \qquad\qquad -\ P\left(\hat{a} \notin \tilde{A}_{z'} \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
\to\ & 1 \ \text{ as } n \to \infty \ \text{ for similar reason as the } i(z') = i \text{ case.}
\end{aligned}
$$

Thus, for any $a \in A_z \setminus \tilde{A}_z$,

$$
\begin{aligned}
& P\left(\max_{\tilde{a}_z \in \tilde{A}_z} \left\{v_{\tau^z_n}(\tilde{a}_z)\right\} > v_{\tau^z_n}(a) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
\geq\ & P\left(\forall a \in A_z,\ v_{\tau^z_n}(a) \in \left(u^i(a) - 2k,\, u^i(a) + 2k\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
=\ & \prod_{a \in A_z} \left(v_{\tau^z_n}(a) \in \left(u^i(a) - 2k,\, u^i(a) + 2k\right) \mid \mathscr{F}_{\tau^z_{n-1}}\right) \\
\to\ & 1 \ \text{ as } n \to \infty \text{ by above.}
\end{aligned}
$$

Therefore, Assumption 2 is satisfied. □

4.4 Cumulative Proportional Reinforcement

A classical model of reinforcement learning attaches to each action a variable that represents the "propensity" to choose that action. The probability with which an action is chosen is then assumed to be an increasing function of the propensity. Laslier, Topol, and Walliser [8] proposed a version of this classical formulation, which they called the cumulative proportional reinforcement

(CPR) learning rule. Under the CPR rule, the propensity to choose action a at period $t \geq 1$ is

$$X_t(a) = X_1(a) + \sum_{k=1}^{t-1} u_k^{i(a)} \mathbb{1}(a \in \xi_k),$$

where initial propensity $X_1(a)$ is an arbitrary positive constant. The probability of choosing $a \in A_z$, conditioned on reaching node z, is $X_t(a) / \sum_{a' \in A_z} X_t(a')$. Laslier and Walliser [7] showed that if all the players follow the CPR rule, then the probability of playing the SPNE converges to one in finite perfect-information games where all the payoffs are strictly positive and there are no ties in the payoffs.

As seen above, classical reinforcement rules directly specify the probability of choosing an action, instead of assuming that players play by forming valuations. However, they can easily be made to conform to our valuation-based approach. In the following, we demonstrate this by constructing a valuation process corresponding to the CPR rule and showing that it satisfies Assumptions 1 and 2 in the class of games considered by Laslier and Walliser. To start, let i be the player using the CPR rule and, for each decision node $z \in G^i$, fix an ordering of actions: $A_z = \{a_z^1, \ldots, a_z^{n_z}\}$ such that $a_z^1 = \tilde{a}_z$.[18] Let

$$\begin{aligned}
I_t(a_z^1) &= \left[0, \frac{X_t(a_z^1)}{\sum_{a' \in A_z} X_t(a')}\right) \\
I_t(a_z^2) &= \left[\frac{X_t(a_z^1)}{\sum_{a' \in A_z} X_t(a')}, \frac{X_t(a_z^1) + X_t(a_z^2)}{\sum_{a' \in A_z} X_t(a')}\right) \\
&\vdots \\
I_t(a_z^{n_z}) &= \left[\frac{\sum_{k=1}^{n_z - 1} X_t(a_z^k)}{\sum_{a' \in A_z} X_t(a')}, 1\right]
\end{aligned}$$

so that $\{I_t(a) : a \in A_z\}$ partitions interval $[0,1]$ into subintervals of length $X_t(a) / \sum_{a' \in A_z} X_t(a')$ each. For all $t \in \mathbb{Z}_{++}$, let ε_t^z be independently and uniformly distributed on $[0,1]$, and let

$$v_t(a) = \mathbb{1}\left(\varepsilon_t^z \in I_t(a)\right).$$

Then the probability of choosing a at period τ_n^z is

$$P\left(v_{\tau_n^z}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau_n^z}(a') \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) = P\left(\varepsilon_{\tau_n^z}^z \in I_{\tau_n^z}(a) \;\middle|\; \mathscr{F}_{\tau_{n-1}^z}\right) = \frac{X_{\tau_n^z}(a)}{\sum_{a' \in A_z} X_{\tau_n^z}(a')}.$$

This model can be visualized by the following generalized urn process. To each decision node z, associate an urn, labeled urn z. To each action $a \in A_z$ associate a unique color, labeled color a, and place $X_1(a)$ balls of color a into

[18] Since there are no ties in the payoffs, SPNE of $\mathscr{G}_z$ is unique.

urn z.[19] If node z is reached during period t, pick a ball at random from urn z. Note the color of the ball, say a, place the ball back into the urn, and take the corresponding action. At the end of the period, place additional u^i_t balls of color a into urn z.

Theorem 8. *The CPR rule satisfies Assumptions 1 and 2 in games with strictly positive payoffs and no ties in the payoffs.*

Proof. Fix a decision node $z \in G^i$ and assume that we are on $\{\tau^z_n < \infty \text{ for all } n\}$. Since payoffs are positive, we have

$$\begin{aligned}\sum_{n=1}^{\infty} P\left(v_{\tau^z_n}(a) > \max_{a' \in A_z \setminus \{a\}} v_{\tau^z_n}(a') \mid \mathscr{F}_{\tau^z_{n-1}}\right) &= \sum_{n=1}^{\infty}\left(\frac{X_{\tau^z_n}(a)}{\sum_{a' \in A_z} X_{\tau^z_n}(a')}\right) \\ &\geq \sum_{n=1}^{\infty}\left(\frac{X_1(a)}{(n-1)\bar{u}_z + \sum_{a' \in A_z} X_1(a')}\right) = \infty.\end{aligned}$$

Therefore, Assumption 1 is satisfied.

Since there are no ties in the payoffs, there is a unique SPNE action at z, $\tilde{a}_z$. Let $\hat{a} \in A_z$ be such that $u^i(\hat{a}) = \max_{a \in A_z \setminus \{\tilde{a}_z\}} u^i(a)$ and choose $\delta > 0$ such that $u^i(\tilde{a}_z) - \delta > u^i(\hat{a}) + \delta$. To show that Assumption 2 is satisfied, we construct a new two-color urn process $Y^m = \{Y^m_k(a) : a \in \{\tilde{a}_z, \hat{a}\},\ k \in \mathbb{Z}_+\}$ for each $m \in \mathbb{Z}_{++}$. Let

$$\begin{aligned}Y^m_0(\tilde{a}_z) &= \left(1 + N_{\tau^z_{m-1}}(\tilde{a}_z)\right)\left(u^i(\tilde{a}_z) - \delta\right), \text{ and} \\ Y^m_0(\hat{a}) &= \sum_{a' \in A_z \setminus \{\tilde{a}_z\}} \left(1 + N_{\tau^z_{m-1}}(a')\right)\left(u^i(\hat{a}) + \delta\right),\end{aligned}$$

where $N_t(a)$ still denotes the number of times a has been chosen for X process. For all $k \geq 0$, define $Y^m_{k+1}(a)$ in the following way. Let

$$I^m_k(\tilde{a}_z) = \left[0, \frac{Y^m_k(\tilde{a}_z)}{Y^m_k(\tilde{a}_z) + Y^m_k(\hat{a})}\right) \quad \text{and} \quad I^m_k(\hat{a}) = \left[\frac{Y^m_k(\tilde{a}_z)}{Y^m_k(\tilde{a}_z) + Y^m_k(\hat{a})}, 1\right].$$

Suppose $\varepsilon^z_{\tau^z_{m+k}} \in I^m_k(\tilde{a}_z)$ so that $\varepsilon^z_{\tau^z_{m+k}}$ represents a drawing of a ball of color $\tilde{a}_z$ for process Y^m. Place the ball back into the urn and add $u^i(\tilde{a}_z) - \delta$ additional balls of color $\tilde{a}_z$. That is, $Y^m_{k+1}(\tilde{a}_z) = Y^m_k(\tilde{a}_z) + u^i(\tilde{a}_z) - \delta$. Otherwise, if $\varepsilon^z_{\tau^z_{m+k}} \in I^m_k(\hat{a})$, put the ball back and add $u^i(\hat{a}) + \delta$ balls of color $\hat{a}$ so that $Y^m_{k+1}(\hat{a}) = Y^m_k(\hat{a}) + u^i(\hat{a}) + \delta$.

Since ε^z_t's are being used for both X and Y^m processes, X and Y^m live on the same probability space. However, aside from Y^m being only a two-colored

[19] The urn analogy is more natural if initial propensities and payoffs are assumed to be integers. Otherwise, one imagines an abstract urn process where balls are perfectly divisible.

process, they differ in that the number of balls that are added for Y^m process depends only on the color of the ball drawn at z, where as the number for X can depend on actions chosen in the subsequent nodes. Moreover, the number of balls that are added when $\tilde{a}_z$ is drawn is always greater than the number added for $\hat{a}$. A result on generalized Pólya process implies that, for all m,[20]

$$\frac{Y^m_k(\tilde{a}_z)}{Y^m_k(\tilde{a}_z)+Y^m_k(\hat{a})} \to 1 \text{ a.s. as } k\to\infty.$$

For the remainder of the proof assume that we are on

$$\left\{N_t(z')\to\infty \text{ and } \frac{S_t(z')}{N_t(z')}\to 1 \text{ for all } z'\in G_z\setminus\{z\}\right\}.$$

Then, for all $a\in A_z$, the fraction of times the SPNE payoff of $\mathcal{G}_{\zeta(a)}$ is received when a is played converges to one. Thus,

$$\frac{X_{\tau^z_n}(a)-X_1(a)}{N_{\tau^z_{n-1}}(a)} = \frac{\sum_{k=1}^{n-1} u^i_{\tau^z_k}\mathbb{1}\left(a\in\xi_{\tau^z_k}\right)}{N_{\tau^z_{n-1}}(a)} \to u^i(a) \text{ as } n\to\infty.$$

This yields $X_{\tau^z_n}(a)\Big/\left(1+N_{\tau^z_{n-1}}(a)\right) \to u^i(a)$, which means there exists $M<\infty$ such that, for all $n\geq M$,

$$\frac{X_{\tau^z_n}(a)}{1+N_{\tau^z_{n-1}}(a)} \in \left(u^i(a)-\delta,\ u^i(a)+\delta\right).$$

Thus,

$$\begin{aligned} X_{\tau^z_M}(\tilde{a}_z) &> \left(1+N_{\tau^z_{M-1}}(\tilde{a}_z)\right)\left(u^i(\tilde{a}_z)-\delta\right) = Y^M_0(\tilde{a}_z), \text{ and} \\ \sum_{a'\in A_z\setminus\{\tilde{a}_z\}} X_{\tau^z_M}(a') &< \sum_{a'\in A_z\setminus\{\tilde{a}_z\}} \left(1+N_{\tau^z_{M-1}}(a')\right)\left(u^i(a')+\delta\right) \leq Y^M_0(\hat{a}). \end{aligned}$$

This implies that

$$\begin{aligned} \frac{X_{\tau^z_M}(\tilde{a}_z)}{X_{\tau^z_M}(\tilde{a}_z)+\sum_{a'\in A_z\setminus\{\tilde{a}\}}X_{\tau^z_M}(a')} &> \frac{Y^M_0(\tilde{a}_z)}{Y^M_0(\tilde{a}_z)+\sum_{a'\in A_z\setminus\{\tilde{a}_z\}}X_{\tau^z_M}(a')} \\ &> \frac{Y^M_0(\tilde{a}_z)}{Y^M_0(\tilde{a}_z)+Y^M_0(\hat{a})}. \end{aligned}$$

Therefore, if $\varepsilon^z_{\tau^z_M}\in I^M_0(\tilde{a}_z)$, then we must also have $\varepsilon^z_{\tau^z_M}\in I_M(\tilde{a}_z)$. That is, if $\tilde{a}_z$ is chosen for Y^M process at time τ^z_M, it must be chosen for X process as well. Moreover, in this case, the number of $\tilde{a}_z$ colored balls that are placed in

[20] See, for example, Pemantle [13], Section 3 and Theorem 3.3 in particular.

the urn is larger for X process than for Y^M process. Similarly, whenever $\tilde{a}_z$ is not chosen for the X process, then it is not chosen for Y process as well, and the number of non-$\tilde{a}_z$ colored ball that is placed in X urn is is smaller than the number placed for Y^M. This implies

$$X_{\tau^z_{M+1}}(\tilde{a}_z) > Y^M_1(\tilde{a}_z) \quad \text{and} \quad \sum_{a' \in A_z \setminus \{\tilde{a}_z\}} X_{\tau^z_{M+1}}(a') < Y^M_1(\hat{a}).$$

By using induction, we obtain then obtain for all k,

$$\frac{X_{\tau^z_{M+k}}(\tilde{a}_z)}{X_{\tau^z_{M+k}}(\tilde{a}_z) + \sum_{a' \in A_z \setminus \{\tilde{a}\}} X_{\tau^z_{M+k}}(a')} > \frac{Y^M_k(\tilde{a}_z)}{Y^M_k(\tilde{a}_z) + Y^M_k(\hat{a})}.$$

Since the right hand side of the above inequality converges to one as $k \to \infty$, for all $a \in A_z \setminus \{\tilde{a}_z\}$, we have

$$P\left(v_{\tau^z_n}(\tilde{a}_z) > v_{\tau^z_n}(a) \;\middle|\; \mathscr{F}_{\tau^z_{n-1}}\right) \geq \frac{X_{\tau^z_n}(\tilde{a}_z)}{\sum_{a' \in A_z} X_{\tau^z_n}(a')} \to 1 \text{ as } n \to \infty.$$

□

5 Concluding Remarks

This paper identified two conditions on general valuations and three, more intuitive, conditions on additively separable valuations that together lead the play to converge to a subgame perfect Nash equilibrium in finite perfect-information games satisfying the "no indifference condition." Our examples show that these conditions are mild enough to encompass a wide range of learning behaviors, including primitive ones like simple recollection and more sophisticated ones like two-moves foresight. Moreover, the convergence results hold even if players adopt different learning rules, as long as each player's rule satisfies the conditions given.

Of the two conditions given for general valuations, the first one induces every action to be taken infinitely often. This is clearly not a necessary condition for SPNE play. For example, if the initial valuation of every action is the same as the SPNE payoff of the subgame following that action, and the players never experiment, the play will always result in a SPNE. However, in the absence of such perfect foresight, the “correct” value of each action needs to be learned from payoff experience. We are not aware of any learning rule in such setting that will allow the probability of playing a SPNE to converge to one without inducing players to experiment infinitely often. Given that a learning rule generates infinitely many experiments, the two conditions provided are necessary and sufficient.

References

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

[2] Borgers, T. and R. Sarin (1997). Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory*, 77: 1-14.

[3] Durrett, R. (2010). *Probability: Theory and Examples.* (4th ed.) New York: Duxbury Press.

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

[5] Hopkins, E. (2002). Two Competing Models of How People Learn in Games. *Econometrica*, 70: 2141-2166.

[6] Jehiel, P. and D. Samet. (2000). Learning to Play Games in Extensive Form by Valuation. *Journal of Economic Theory*, 124:129-148.

[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] Laslier, J-F., R. Topol, and B. Walliser (2001). A Behavioral Learning Process in Games. *Games and Economic Behavior*, 37: 340-366.

[9] Loève, M. (1978). *Probability Theory, Vol. II*. (4th ed.) Berlin: Springer-Verlag.

[10] Marx, L. and J. Swinkels (1997). Order Independence for Iterated Weak Dominance. *Games and Economic Behavior*, 18: 219-245.

[11] Osborne, M. and A. Rubinstein (1994). *A Course in Game Theory*. Cambridge, MA: MIT Press.

[12] Østerdal, L. (2005). Iterated Weak Dominance and Subgame Dominance. *Journal of Mathematical Economics*, 41: 637-645.

[13] Pemantle, R. (2007). A survey of random processes with reinforcement. *Probability Surveys*, 4:1-79.

[14] Sarin, R. and F. Vahid. (1999). Payoff Assessments without Probabilities: A Simple Dynamic Model of Choice. *Games and Economic Behavior*, 28: 294-309.