2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Logit Learning by Valuation in Extensive-Form Games with Simultaneous Moves

Jason Castiglione and Gürdal Arslan

Abstract— **We study the long-term behavior of the logit learning rule in multiplayer repeated extensive-form games. Our model involves the possibility of simultaneous moves by multiple players as well as chance moves by nature in every node of the game tree. The logit learning rule considered in this paper is based on average payoff valuations. In certain class of extensive-form games with simultaneous moves, we show that player strategies converge to a perturbed subgame perfect equilibrium of the stage game when every player uses the logit learning rule in the repeated game. In extensive-form games with perfect information, we also show that the long run average payoff of a player using the logit learning rule is guaranteed to be nearly as high as the player's maxmin payoff in the stage game.**

I. INTRODUCTION

There is currently a substantial body of literature on learning in multiplayer games; see the books [1], [2] and the references therein. By and large, the existing work in this area is concentrated on repeated normal-form games. For example, well-known learning rules, fictitious play [3], [4], reinforcement learning [5], no-regret learning [6], Bayesian learning [7], hypothesis testing [8], and their variants have been well studied in the context of normal-form games. Typical results in this literature include convergence to equilibria under self play when all players employ the same learning rule. In comparison with the normal-form games, the extensive-form games have received much smaller attention in terms of learning in repeated play [9], [10], [11], [12], [13]. The main objective of this paper is to advance the state of the art on learning in repeated extensive-form games, in particular, in the case of imperfect information. We are ultimately interested in developing simple and effective learning rules that can be used in large extensive-form games. A *valuation* for a player is a mapping from all of the player's possible moves to $\mathbb{R}$. As such, we are interested in valuation-based learning rules where players form valuations for possible moves, or sets of possible moves that are considered similar in some (possibly adaptive)[1] way, at each node as opposed to forming valuations for possible game strategies, that is typically a much larger set. Valuation-based learning, for example [12], [13] updates the valuation from the game play history. The player's strategy is calculated directly from the updated valuation.

The reference [9] studies a fictitious playlike process for extensive games in which each player makes moves that maximize his/her immediate expected payoffs under the (incorrect) assumption that the other players make moves according to empirical frequencies of past play, and show that the stable points of the learning process may not be Nash equilibria. The reference [10] identifies general conditions on valuation rules for the convergence of the repeated play to a subgame perfect equilibrium in extensive games with perfect information where no player is indifferent between two outcomes without every other player being also indifferent. In [10], each player always makes a move with maximum valuation, i.e., no explicit exploration; however, implicit exploration can be introduced through imperfections in the valuation process.

The most relevant references for this work are [12] and [11] where learning by valuation rules are studied for extensive-form games with perfect information. At any node visited during the repeated play, a player using "the δ-exploratory myopic strategy with the averaging revision rule" in [12] makes a move, with probability $1-\delta$, that has resulted in the highest average payoff in the previous rounds; with probability δ, the player makes a random move for exploration. When every player adheres to such a learning rule with small $\delta > 0$, the reference [12] shows that player strategies would be close to an equilibrium in the long run. It is furthermore shown in [12] that the long-term average payoff of a player using the δ-exploratory myopic strategy with the averaging revision rule would be nearly as high as the player's maxmin payoff in the stage game, provided that the exploration rate $\delta > 0$ is small. As expected, the role of random exploration in obtaining such results is critical since, without persistent exploration, players might easily get stuck at arbitrary strategies due to the bias in their initial valuations of possible moves.

It is also recognized in the literature that not all forms of random exploration lead to the same outcome; see for example [14]. As in the δ-exploratory myopic strategy with the averaging revision rule in [12], assigning a fixed (and a typically high) probability to moves that resulted in the highest average payoff in the past makes a player quite predictable in games with simultaneous moves, which can be exploited by a sophisticated opponent leading to very poor performance for the player. The root cause of this vulnerability is to assign almost all of the probability mass to a single move even when that move has only infinitesimally

G. Arslan is with the Department of Electrical Engineering, University of Hawaii at Manoa, 440 Holmes Hall, 2540 Dole Street, Honolulu, HI 96822, USA gurdal@hawaii.edu

J. Castiglione is with the Department of Electrical Engineering, University of Hawaii at Manoa, 440 Holmes Hall, 2540 Dole Street, Honolulu, HI 96822, USA jcastig@hawaii.edu

[1]By adaptive is meant the sets of similar moves are recalculated each iteration.

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

higher valuation than the other possible moves. This calls for a form of random exploration in which moves with comparable valuations are assigned comparable probabilities of selection while ensuring that every possible move is always assigned a probability that is uniformly bounded above zero to achieve persistent random exploration. The reference [11] studies the cumulative proportional reinforcement (CPR) learning in extensive games where the selection probabilities for the possible moves are proportional to the corresponding valuations. The reference [11] shows that an action-based (as opposed to a strategy-based) CPR process converges to the (unique) subgame equilibrium in generic extensive games with perfect information and with positive payoffs. The idea of comparable probabilities for comparable valuations can also be achieved by utilizing the well-known logistic response function to map the valuations of possible moves to the selection probabilities of those moves.

Accordingly, in this paper, we explore the behavior of an alternative valuation-based learning rule, and a close variant of it, in extensive-form games with simultaneous moves. The valuations of possible moves at possible nodes are the average payoffs obtained in the past through those nodes and moves; however, the probability of selection for each possible move at each possible node is determined by applying the logistic response function to the corresponding valuations. This leads to what is called in this paper the logit learning rule and the modified logit learning rule. One of the contributions of this paper is to show that, when every player uses the modified logit learning rule in the repeated play, player strategies converge to a perturbed subgame-perfect equilibrium of the stage game, which is an extensive-form game with simultaneous moves. Our other main contribution is to obtain a lower bound on the long-term average payoff of a player using the logit learning rule, regardless of the learning rules used by the other players, in an extensive-form game with perfect information, that is the player's maxmin payoff in the stage game. This result is a counterpart of a similar robustness result in [12] for the δ-exploratory myopic strategy with the averaging revision rule. We believe that this work contributes to the goal of developing simple learning rules that can be deployed, with provable performance, in repeated extensive-form games with imperfect information.

Section II introduces our repeated game model and learning rules. Our main results are presented in Section III followed by simulation results in Section IV. Some concluding remark are given in Section V and the proofs of the main results are provided in Appendix.

Notation: $\mathbb{N}$ and $\mathbb{N}_0$ denote the sets of positive and non-negative integers, respectively, whereas $\mathbb{R}$ denotes the set of real numbers; $|A|$ denotes the number of elements of a finite set A; $\mathcal{P}(A)$ denotes the set of probability distributions over a finite set A; $(x)^+ = \max\{x, 0\}$ for any $x \in \mathbb{R}$.

II. SETUP AND MOTIVATION

We adopt the framework of extensive-form games in which multiple players are allowed make simultaneous moves at each node of the game tree. Such games are referred to as "extensive-form games with perfect information and simultaneous moves" or "extensive-form games with almost perfect information". If only one player is allowed to make moves, then such games are referred to as "extensive-form games with perfect information". Our model includes a non-strategic player, called the nature, who makes chance moves at some nodes of the game tree according to some fixed probability distributions. A precise model is provided below.

A. Stage Game

The stage game, denoted by G, has a finite set $I = \{1, \ldots, |I|\}$ of strategic players. $I^0 = \{0, 1, \ldots, |I|\}$ denotes the set of players including the nature, that is a non-strategic player referred to as player 0. The game tree consists of two finite (disjoint) sets of nodes, called non-terminal nodes N and terminal nodes Z, and a finite set of arcs A. One of the nodes in N is the root node r in which the stage game starts. Each arc in A is an ordered pair of nodes (n, m) where $m \in N \cup Z$ is the immediate successor of $n \in N$. The length of a node n is the length of a longest path starting at n and ending at a terminal node. The length of the root node is denoted by $L \in \mathbb{N}$. $N(k)$ denotes the set of nodes with length $k \in \mathbb{N}_0$, where $N(0) = Z$. We denote the subgame rooted at a node $n \in N$ by $G(n)$, and the set of terminal nodes in $G(n)$ by $Z(n)$.

The set of nodes in which player $i \in I^0$ makes moves is dented by N^i, where $N^i \subset N$ and $\cup_{i \in I^0} N^i = N$. $I^0(n) := \{i \in I^0 : n \in N^i\}$ denotes the set of players who make moves at a node $n \in N$. $I(n) := \{i \in I : n \in N^i\}$ denotes the set of strategic players who make moves at a node $n \in N$. $M^i(n)$ denotes the set of moves of a player $i \in I^0(n)$ at a node $n \in N$. The set of arcs between a node $n \in N$ and its immediate successors is identified with the set of joint moves $M(n) := \times_{i \in I^0(n)} M^i(n)$ at n. As a result, we will use the notation $M(n)$ also to denote the set of nodes that are immediate successors of $n \in N$. Starting from the root node, at each non-terminal node $n \in N$ reached during the play of the stage game, the set of players $I^0(n)$ move simultaneously and their joint move $m \in M(n)$ determines the next node reached. The stage game ends when a terminal node z is reached, at which point each player $i \in I$ receives a payoff denoted by $f^i(z) \in \mathbb{R}$.

Each player $i \in I^0$ makes his/her moves using a strategy σ^i defined on N^i such that $\sigma^i(n) \in \mathcal{P}(M^i(n))$ at each $n \in N^i$. If a node $n \in N^i$ is reached during the play, player $i \in I^0$ makes his/her move m^i at n according to $\sigma^i(n)$ where m^i is an independent draw from $\sigma^i(n)$. A strategy σ^i is a *pure* if $\sigma^i(n)$ assigns probability one to a single move $m^i \in M^i(n)$ at every node $n \in N^i$.

We use the notation $\sigma := (\sigma^j)_{j \in I}$ to denote the joint strategy employed by all strategic players, and $\sigma^{-i} := (\sigma^j)_{j \in I \setminus \{i\}}$ to denote the joint strategy employed by all strategic players except a player $i \in I$. In general, we refer to the set of strategic players other than player $i \in I$ as $-i$, i.e., $-i$ refers to $I \setminus \{i\}$. We sometimes write σ as $\sigma = (\sigma^i, \sigma^{-i})$ for any $i \in I$. Throughout the paper, we suppress the dependency of various quantities on the nature's strategy σ^0,

which is assumed to assign non-negative probability to every move $m^0 \in M^0(n)$ at each node $n \in N^0$.

We let P_σ denote the probability distribution induced by σ over Z, and $P_\sigma[z|n]$ denote the probability distribution induced by σ over Z given that the node $n \in N$ is reached. $f^i_\sigma(n)$ denotes the expected payoff of player $i \in I$ under the joint strategy σ in $G(n)$, $n \in N$, i.e.,

$$f^i_\sigma(n) := E_\sigma[f^i|n] = \sum_{z \in Z} P_\sigma[z|n] f^i(z).$$

Definition 1: A joint strategy σ is called a *subgame perfect equilibrium* of the stage game G if

$$f_\sigma(n) = \max_{\sigma^i} f^i_{(\sigma^i, \sigma^{-i})}(n)$$

for all $i \in I$ and $n \in N$.

Definition 2: The *maxmin* (or *minmax*) payoff of each player $i \in I$ in $G(n)$, $n \in N$, is defined as

$$\rho^i(n) := \max_{\sigma^i} \min_{\sigma^{-i}} f^i_{(\sigma^i, \sigma^{-i})}(n) = \min_{\sigma^{-i}} \max_{\sigma^i} f^i_{(\sigma^i, \sigma^{-i})}(n).$$

We next introduce the notion of a perturbed subgame perfect equilibrium. For this, we first introduce the logistic response function β^λ that maps any $f \in \mathbb{R}^d$, $d \in \mathbb{N}$, to the probability vector

$$\beta^\lambda(f) = \frac{1}{\sum_{\ell=1}^d e^{\frac{1}{\lambda} f_\ell}} (e^{\frac{1}{\lambda} f_1}, \ldots, e^{\frac{1}{\lambda} f_d})$$

where $\lambda > 0$. It is well-known that, given $f \in \mathbb{R}^d$, $\beta^\lambda(f)$ assigns arbitrarily high probability to $\arg\max_{\ell \in \{1,\ldots,d\}} f_\ell$ for sufficiently small $\lambda > 0$. Consequently,

$$\lim_{\lambda \to 0} \sum_{\ell=1}^d f_\ell \beta^\lambda_\ell(f) = \max_{\ell \in \{1,\ldots,d\}} f_\ell$$

where $\beta^\lambda_\ell(f)$ is the ℓ-th component of $\beta^\lambda(f)$.

Definition 3: A joint strategy σ^λ, $\lambda > 0$, is called a *λ-perturbed subgame perfect equilibrium* of the stage game G if

$$\sigma^{i,\lambda}(n) = \beta^\lambda(f^i_{\sigma^{-i,\lambda}}(n))$$

for all $i \in I$ and $n \in N$ where

$$f^i_{\sigma^{-i,\lambda}}(n) := \left(E_{\sigma^\lambda}[f^i(z)|n, m^i]\right)_{m^i \in M^i(n)}$$

and $E_{\sigma^\lambda}[f^i(z)|n, m^i]$ is player i's expected payoff under σ^λ given that the node n is reached and player i makes the move $m^i \in M^i(n)$ at n.

Proposition 1: Let $\{\sigma^{\lambda_k}\}_{k \in \mathbb{N}}$ be such that $\lambda_k \to 0$ and σ^{λ_k} is a λ_k-perturbed subgame perfect equilibrium of G for each $k \in \mathbb{N}$. If $\sigma_k \to \sigma$, then σ is a subgame perfect equilibrium of G.

Proposition 1 implies that, for sufficiently small $\lambda > 0$, a λ-perturbed subgame perfect equilibrium will be near a subgame perfect equilibrium, which follows from Theorem 2 in [15].

B. Repeated Game and Learning Rules

We now introduce a repeated game in which the stage game G is repeated over an infinite number of rounds $t \in \mathbb{N}$. The terminal node reached in round $t \in \mathbb{N}$ is denoted by $z_t \in Z$. At the end of each round $t \in \mathbb{N}$, the history of the play is $h_t = (z_1, \ldots, z_t)$ (h^0 is the null history). The strategy σ^i_t used by player $i \in I$ in round $t \in \mathbb{N}$ can be chosen on the basis of h_{t-1} (the nature uses the same strategy σ^0 in every round $t \in \mathbb{N}$). We use the notation $\bar{f}^i_t(n)$ to denote the average payoff received by player $i \in I$ in rounds prior to round $t \in \mathbb{N}$ in which the node $n \in N^i$ is visited; $\bar{f}^i_1(n) \in \mathbb{R}$ is arbitrary. Similarly, we use the notation $\bar{f}^i_t(n, m^i)$ to denote the average payoff received by player $i \in I$ in rounds prior to round $t \in \mathbb{N}$ in which player has made the move $m \in M^i(n)$ at a node $n \in N^i$; $\bar{f}^i_1(n, m^i) \in \mathbb{R}$ is arbitrary. The average payoffs $\bar{f}^i_t(n)$ and $\bar{f}^i_t(n, m^i)$ are determined by the history h_{t-1}, $t \in \mathbb{N}$.

In the repeated game, each player i uses a learning rule Σ^i to determine his/her/her strategy σ^i_t for each round $t \in \mathbb{N}$ after observing the history h_{t-1}, i.e., $\sigma^i_t = \Sigma_i(h_{t-1})$. The probability distribution induced by a joint learning rule $\Sigma = (\Sigma^i)_{i \in I}$ over the histories of finite length is denoted by P_Σ. We are interested in analyzing the long-term behavior of certain payoff-based learning rules, introduced next. A simple payoff-based learning rule, called the δ-exploratory myopic strategy and averaging revision rule where $\delta > 0$ is a small exploration probability, is introduced in [12]. We will refer to this learning rule as $\Sigma^{i,\delta}$. Upon reaching a node $n \in N^i$ in round t, a player $i \in I$ using $\Sigma^{i,\delta}$ makes a move $m^i \in M^i(n)$, with probability $1 - \delta$, that has achieved the highest average payoff $\max_{m^i \in M^i(n)} \bar{f}^i_t(n, m^i)$ in rounds in which player i has made the move m^i at n prior to t (ties are broken by dividing the probability $1 - \delta$ equally); and player i makes a move m^i, with probability δ, that is chosen randomly from $M^i(n)$ with equal likelihood.

In extensive-form games with perfect information (a single player moves at every node and there is no nature player), the reference [12] shows that, for sufficiently small $\delta > 0$, player strategies almost surely converge to a strategy that is close to the (assumed) unique perfect equilibrium of the stage game under $\Sigma^\delta = (\Sigma^{i,\delta})_{i \in I}$. Again in the perfect information case, the reference [12] also shows that, for sufficiently small $\delta > 0$, the long-run average payoff to a player i using $\Sigma^{i,\delta}$ is no worse than player i's maxmin payoff in the stage game (minus small $\epsilon > 0$) almost surely, regardless of the learning rule used by the other players. Although $\Sigma^{i,\delta}$ is a natural and appealing learning rule, the results in [12] cannot be extended to the imperfect information case, in particular to the extensive-form games with simultaneous moves, in a satisfactory manner.

To see why $\Sigma^{i,\delta}$ would not produce satisfactory results in the case of simultaneous moves, consider the matching pennies game depicted in fig. 1. Each of the two players chooses either Heads (H) or Tails (T) simultaneously with the other player. If player 1's choice matches player 2's choice, player 1 receives a one unit payoff from player

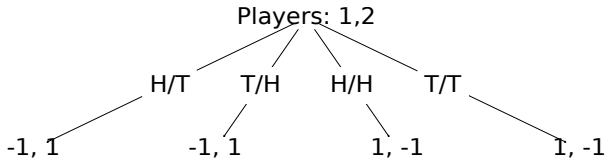

Fig. 1. Matching Pennies (H=Heads / T=Tails)

2; otherwise, player 2 receives a one unit payoff from player 1. This game can be represented by a tree where the only non-terminal node is the root node with four terminal nodes. The edges, $H/T, T/H, H/H, T/T$, correspond to each possible joint move at the root node. If player 1 uses $\Sigma^{1,\delta}$ and player 2 uses a constant strategy σ^2 (with $\sigma^2(r,H), \sigma^2(r,T) > 0$), then player 1's average payoff for H and T converges to $\sigma^2(r,H)$ and $\sigma^2(r,T)$ respectively. In this case, player 1's average payoff would converge to $(1-\delta/2)\max\{\sigma^2(r,H), \sigma^2(r,T)\} + (\delta/2)\min\{\sigma^2(r,H), \sigma^2(r,T)\}$, which would be nearly optimal for player 1 for sufficiently small $\delta > 0$.

However, if player 2 uses a strategy which selects T (resp. H) in each round t when $\bar{f}_t^1(r,H) > \bar{f}_t^1(r,T)$ (resp. $\bar{f}_t^1(r,T) > \bar{f}_t^1(r,H)$), then player 1 would be winning each round only with probability $\delta/2$ unless $\bar{f}_t^1(r,H) = \bar{f}_t^1(r,T)$ (conditioned on the prior history). This would lead to the long-term average payoff $-1+\delta$ that is nearly player 1's minimum payoff for small $\delta > 0$. The root cause of this unsatisfactory performance of $\Sigma^{i,\delta}$ is the fact that, when $\delta > 0$ is small (as it should be), it assigns nearly all of the probability mass to a move corresponding to the highest average payoff in prior rounds even when an alternative move achieves an average payoff in prior rounds that is only slightly below the highest. As a result, player i using $\Sigma^{i,\delta}$ with small $\delta > 0$ becomes quite predictable by a sophisticated opponent. This shows the significance of random exploration as well as the manner in which it is conducted. We refer the interested reader to [14] for a detailed discussion on this point.

The discussion above motivates us to introduce an alternative learning rule which will call the *logit learning rule* and refer to it as $\bar{\Sigma}^{i,\lambda}$. A player i using $\bar{\Sigma}^{i,\lambda}$ makes his/her moves in round t according to the strategy

$$\bar{\sigma}_t^{i,\lambda}(n) := \beta^\lambda(\bar{f}_t^i(n,\cdot)), \qquad \forall n \in N^i$$

where $\bar{f}_t^i(n,\cdot) := (\bar{f}_t^i(n,m^i))_{m^i \in M^i(n)}$ and $\lambda > 0$. Unlike $\Sigma^{i,\delta}$, $\bar{\Sigma}^{i,\lambda}$ assigns probabilities to actions in a way that is commensurate with the corresponding average payoffs in the prior rounds; in particular, actions with nearly equal average payoffs in the prior rounds are assigned nearly equal probabilities for small $\lambda > 0$.

We finally introduce a slight modification of $\bar{\Sigma}^{i,\lambda}$ to simplify our analysis and make use of the existing results in the literature; the analysis of the long term behavior of $\bar{\Sigma}^{i,\lambda}$ in repeated normal-form games does not seem to be readily available in the literature to our knowledge. For any $\lambda > 0$, the modified logit learning rule $\hat{\Sigma}^{i,\lambda}$ is obtained from $\bar{\Sigma}^{i,\lambda}$ by replacing the averages $\{\bar{f}_t^i\}_{t\in\mathbb{N}}$ with $\{\hat{f}_t^i\}_{t\in\mathbb{N}}$ generated by

$$\hat{f}_{t+1}^i(n,m^i) = \hat{f}_t^i(n,m^i) + x_t(n,m^i)\frac{f^i(z_t) - \hat{f}_t^i(n,m^i)}{\nu_{t+1}(n)\beta_{m^i}^\lambda(\hat{f}_t^i(n,\cdot))} \tag{1}$$

starting from some arbitrary $\hat{f}_1^i(n,m^i)$, for all $n \in N^i$ and $m^i \in M^i(n)$, where

$$x_t(n,m^i) := \begin{cases} 1 & \text{if } n \text{ is visited and } m^i \text{ is chosen} \\ & \text{by player } i \text{ in round } t \\ 0 & \text{else} \end{cases}$$

$$\nu_{t+1}(n,m^i) := \sum_{k=1}^{t} x_t(n,m^i)$$

$$\nu_{t+1}(n) := \sum_{m^i \in M^i(n)} \nu_{t+1}(n,m^i)$$

$$\hat{f}_t^i(n,\cdot) := (\hat{f}_t^i(n,m^i))_{m^i \in M^i(n)}$$

and $\beta_{m^i}^\lambda(\hat{f}_t^i(n,\cdot))$ is the probability assigned by $\beta^\lambda(\hat{f}_t^i(n,\cdot))$ to the move m^i. We note that, in (1), $\nu_{t+1}(n)\beta_{m^i}^\lambda(\hat{f}_t^i(n,\cdot))$ is expected to approximate $\nu_{t+1}(n,m^i)$ in the long run. In fact, if $\nu_{t+1}(n)\beta_{m^i}^\lambda(\hat{f}_t^i(n,\cdot))$ is replaced with $\nu_{t+1}(n,m^i)$, the recursion (1) would generate $\{\bar{f}_t^i\}_{t\in\mathbb{N}}$.

When every player i uses $\hat{\Sigma}^{i,\lambda}$ in a repeated normal-form game, the long term behavior of player strategies is analyzed in [16] using the Ordinary Differential Equation (ODE) method of stochastic approximation. The joint learning rule $\hat{\Sigma}^\lambda := (\hat{\Sigma}^{i,\lambda})_{i\in I}$ leads to an autonomous ODE which is related to the ODE associated with the smooth fictitious play algorithm in [16]. The reference [16] shows that player strategies generated by $\hat{\Sigma}^\lambda$ converge to a λ-perturbed equilibrium of the stage game in repeated two-player zero-sum and two-player partnership normal-form games (with countably many λ-perturbed equilibrium) under the following boundedness assumption.

Assumption 1: $\{\hat{f}_t^i\}_{i\in I, t\in\mathbb{N}}$ generated by $\hat{\Sigma}^\lambda$, $\lambda > 0$, is bounded almost surely.

Assumption 1 can be removed by projecting the iterates $\{\hat{f}_t^i\}_{i\in I, t\in\mathbb{N}}$ to large compact sets at the cost of a more complicated analysis [17]. Without further mention, we will also let Assumption 1 hold throughout the paper. A consequence of Assumption 1 is that all node-move pairs for all players are visited infinitely often, i.e., $\lim_{t\to\infty}\nu_t(n,m^i) = \infty$ for all $i \in I$, $n \in N$, $m^i \in M^i(n)$, almost surely under $\hat{\Sigma}^\lambda$, $\lambda > 0$.

In the next section, we present our main results on the long term behavior of player strategies and average cost generated by $\hat{\Sigma}^\lambda$ in extensive-form games with simultaneous moves.

III. MAIN RESULTS

We will present two results whose proofs can be found in Appendix. We will first present convergence results under self-play, i.e., every player i uses the modified logit learning rule $\hat{\Sigma}^{i,\lambda}$. We will then present a robustness result for a player i using $\bar{\Sigma}^{i,\lambda}$ against the other players using arbitrary learning rules.

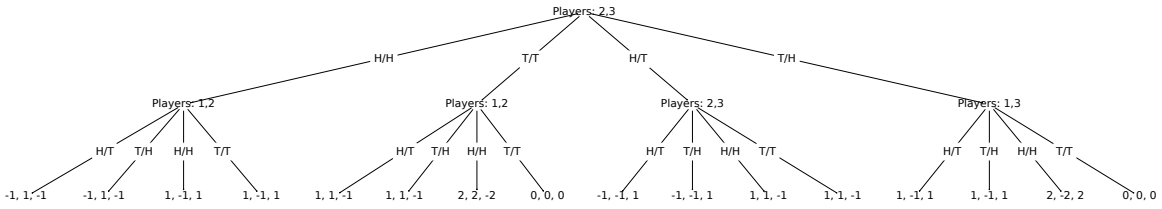

Fig. 2. Three Player Game with Simultaneous Moves (H=Heads / T=Tails)

For each $\lambda > 0$, we consider the set $\mathcal{G}^\lambda$ of extensive-form games G with simultaneous moves satisfying

1) At most two strategic players can make moves at any node, i.e., $\max_{n \in N} |I(n)| \leq 2$.
2) If $|I(n)| = 2$, then either
$$f^i(z) = -f^{-i}(z), \qquad \forall i \in I(n), z \in Z(n)$$
or
$$f^i(z) = f^{-i}(z), \qquad \forall i \in I(n), z \in Z(n)$$
3) Every subgame $G(n)$ has a finite number of λ-perturbed equilibrium.

The second condition above implies that, if player i and $-i$ can make moves at any node $n \in N$, then either player i and $-i$ receive the same payoffs at each terminal node $z \in Z(n)$, or their payoffs sum to zero at each $z \in Z(n)$. In other words, either player i and $-i$ face a zero-sum or a partnership game in $G(n)$ for any fixed strategies of the other players in $I \backslash I(n)$. The third condition is a generic one, i.e., almost all games G satisfies the third condition (in fact, with an odd number of λ-perturbed subgame perfect equilibrium) for almost all $\lambda > 0$; see Theorem 3 in [15]. Furthermore, every zero-sum extensive-form game with simultaneous moves has a unique λ-perturbed subgame perfect equilibrium; see [18] for the case of zero-sum normal-form games.

Theorem 1: Player strategies $\sigma_t = (\sigma_t^i)_{i \in I}$ generated by the modified logit learning rule $\hat{\Sigma}^\lambda$, $\lambda > 0$, in an extensive-form game $G \in \mathcal{G}^\lambda$ with simultaneous moves, converges almost surely to a λ-perturbed subgame perfect equilibrium of G.

Remark 1: A very appealing robustness result for the modified logit learning rule is included in [14]. In Section 6 of [14], it is argued that the modified logit learning rule achieves "ϵ-universal consistency". Loosely speaking, this means that a player i using the modified logit learning rule $\hat{\Sigma}^{i,\lambda}$ with small $\lambda > 0$ would achieve near optimal performance in the long run with respect to the empirical frequency distribution of the other player's moves. In particular, this would imply that a player i using $\hat{\Sigma}^{i,\lambda}$ with small $\lambda > 0$ would be more or less playing optimally in the long run against the other players using constant strategies. We conjecture that this ϵ-universal consistency result can be extended to the extensive-form games with simultaneous moves under $\hat{\Sigma}^{i,\lambda}$.

Remark 2: We expect that the counterparts of the results obtained for the modified logit learning rule can also be obtained for the logit learning rule, that is a somewhat more natural learning rule. This would first require obtaining counterparts of Proposition 4.2 in [16] and the ϵ-universal consistency result in [14] for the logit learning rule in normal-form games. A promising approach to obtain such results would be to employ the asynchronous stochastic approximation techniques [19], [20].

We next present a robustness result for a player $i \in I$ using the logit learning rule $\bar{\Sigma}^{i,\lambda}$ in an extensive-form game with perfect information.

Theorem 2: For every extensive-form game G with perfect information and $\epsilon > 0$, there exists $\bar{\lambda} > 0$ such that, under any joint learning rule Σ where $\Sigma^i = \bar{\Sigma}^{i,\lambda}$ for some player $i \in I$ with $\lambda \in (0, \bar{\lambda}]$, we have

$$\underline{\lim}_{t \to \infty} \frac{1}{t} \sum_{k=1}^{t} f^i(z_k) \geq \rho^i(r) - \epsilon$$

almost surely (P_Σ).

In extensive-form games with perfect information, Theorem 2 provides a guaranteed long-run average payoff to a player using the logit learning rule, that is the player's maxmin payoff minus small $\epsilon > 0$, against all sophisticated opponents. For example, this would imply that, in two-player constant-sum win-lose games, a player using the logit learning rule would eventually almost always win if a guaranteed winning strategy exists for the player. It is straightforward to see that a player using the logit learning rule, or its modified version, would in general achieve higher, in fact nearly optimal, long-run average payoffs against unsophisticated opponents using constant strategies. Extending Theorem 2 to extensive-form games with simultaneous moves and obtaining an ϵ-universal consistency result are interesting research problems, which we plan to address in the fuller version of the paper.

IV. SIMULATION

To illustrate convergence using logit learning, we consider a three player game with simultaneous moves, see fig. (2). The sub-games played after initial plays H/H and H/T are zero sum games, and the other two sub-games are identical interest between the active players at the node. Note for players two and three that from condition 2 of Theorem 1, that their payoffs are opposite at each node. For this game, the expected payoffs at any subgame perfect equilibrium are $0, 0, 0$ for players one, two, and three. The non-terminal nodes are labeled with the players, and the terminal nodes

are labeled with the payoffs. In the simulation all three players use logit learning with $\lambda = 0.1$, as in the context of Theorem 1. The simulation was run for $10,000$ trials where for each trial the players play $10,000$ stage games. As λ becomes smaller the payoffs get closer to the equilibrium, but also take longer to converge.

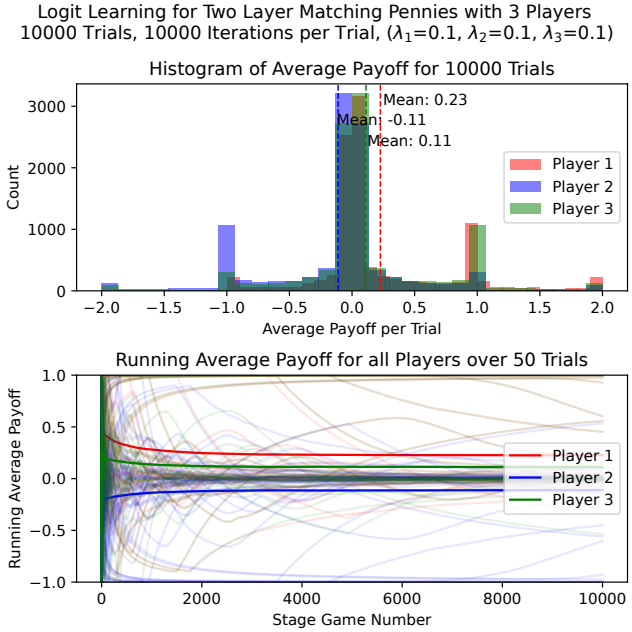

Fig. 3. Simulation of Theorem 1 for Multiplayer Game in fig. (2)

V. CONCLUSION

Much work remains to be done to obtain a fuller picture of the behavior of the logit learning by valuation in extensive-form games with imperfect information, some of which is pointed out in the sequel. We considered the long-term behavior of the logit learning in extensive-form games with simultaneous moves. We showed convergence to perturbed subgame perfect equilibria under self play in a certain class of extensive-form games with simultaneous moves. We also obtained a robustness result for a player using the logit learning rule in the case of perfect information. Extending these results to more general extensive-form games with imperfect information is a future research topic. Finally, obtaining counterparts of such results in large games where the moves of players are partitioned into similarity classes as in [13] would be another interesting topic for future research.

REFERENCES

[1] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.

[2] H. P. Young, *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. New Jersey: Princeton University Press, 1998.

[3] G. W. Brown, "Iterative solutions of games by fictitious play," in *Activity Analysis of Production and Allocation*, T. Koopmans, Ed. New York: Wiley, 1951, pp. 374–376.

[4] ——, "Iterative solution of games by fictitious play," *Act. Anal. Prod Allocation*, vol. 13, no. 1, p. 374, 1951.

[5] C. Camerer and T. Hua Ho, "Experience-weighted attraction learning in normal form games," *Econometrica*, vol. 67, no. 4, pp. 827–874, 1999.

[6] D. P. Foster and R. V. Vohra, "Calibrated learning and correlated equilibrium," *Games and Economic Behavior*, vol. 21, no. 1-2, p. 40, 1997.

[7] E. Kalai and E. Lehrer, "Rational learning leads to nash equilibrium," *Econometrica: Journal of the Econometric Society*, pp. 1019–1045, 1993.

[8] D. P. Foster and H. P. Young, "Learning, hypothesis testing, and nash equilibrium," *Games and Economic Behavior*, vol. 45, no. 1, pp. 73–96, 2003.

[9] D. Fudenberg and D. M. Kreps, "Learning in extensive-form games i. self-confirming equilibria," *Games and Economic Behavior*, vol. 8, no. 1, pp. 20–55, 1995.

[10] M. Pak and B. Xu, "Generalized reinforcement learning in perfect-information games," *International Journal of Game Theory*, vol. 45, pp. 985–1011, 2016.

[11] J.-F. Laslier and B. Walliser, "A reinforcement learning process in extensive form games," *International Journal of Game Theory*, vol. 33, pp. 219–227, 2005.

[12] P. Jehiel and D. Samet, "Learning to play games in extensive form by valuation," *Journal of Economic Theory*, vol. 124, no. 2, pp. 129–148, 2005.

[13] ——, "Valuation equilibrium," *Theoretical Economics*, vol. 2, no. 2, pp. 163–185, 2007.

[14] D. Fudenberg and D. K. Levine, "Consistency and cautious fictitious play," *Journal of Economic Dynamics and Control*, vol. **19**, pp. 1065–1089, 1995.

[15] R. D. McKelvey and T. R. Palfrey, "Quantal response equilibria for normal form games," *Games and Economic Behavior*, vol. **10**, pp. 6–38, 1995.

[16] D. Leslie and E. Collins, "Individual Q-learning in normal form games," *SIAM Journal on Control and Optimization*, vol. 44, pp. 495–514, 2005.

[17] H. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, ser. Stochastic Modelling and Applied Probability. Springer New York, 2003.

[18] J. Hofbauer and E. Hopkins, "Learning in perturbed asymmetric games," *Games and Economic Behavior*, vol. 52, no. 1, pp. 133–152, 2005.

[19] V. S. Borkar, "Asynchronous stochastic approximations," *SIAM Journal on Control and Optimization*, vol. 36, no. 3, pp. 840–851, 1998.

[20] V. S. Borkar and S. P. Meyn, "The ode method for convergence of stochastic approximation and reinforcement learning," *SIAM Journal on Control and Optimization*, vol. 38, no. 2, pp. 447–469, 2000.

[21] D. P. Foster and R. V. Vohra, "Asymptotic calibration," *Biometrika*, vol. 85, no. 2, pp. 379–390, 1998.

[22] S. Hart and A. Mas-Colell, "A simple adaptive procedure leading to correlated equilibrium," *Econometrica*, vol. 68, no. 5, pp. 1127–1150, 2000.

[23] M. Benaïm, "Dynamics of stochastic approximation algorithms," in *Seminaire de probabilites XXXIII*. Springer, 2006, pp. 1–68.