2011 American Control Conference
on O'Farrell Street, San Francisco, CA, USA
June 29 - July 01, 2011

Distributed Strategic Learning with Application to Network Security

Quanyan Zhu, Hamidou Tembine and Tamer Başar

Abstract— **We consider in this paper a class of two-player nonzero-sum stochastic games with incomplete information. We develop fully distributed reinforcement learning algorithms, which require for each player a minimal amount of information regarding the other player. At each time, each player can be in an active mode or in a sleep mode. If a player is in an active mode, she updates her strategy and estimates of unknown quantities using a specific pure or hybrid learning pattern. We use stochastic approximation techniques to show that, under appropriate conditions, the pure or hybrid learning schemes with random updates can be studied using their deterministic ordinary differential equation (ODE) counterparts. Convergence to state-independent equilibria is analyzed under specific payoff functions. Results are applied to a class of security games in which the attacker and the defender adopt different learning schemes and update their strategies at random times.**

I. Introduction

In recent years, game-theoretic methods have been applied to study resource allocation problems in communication networks, security mechanisms for network security and privacy [1], and economic pricing in power networks [7]. Most frameworks have assumed the rationality of the agents or the decision-makers as well as complete information about their payoffs and strategies. However, in practice, due to noise and uncertainties in the environment, agents often have information limitations in their knowledge not only of other players' payoffs and strategies, but also of their own. For this reason, we must consider the learning aspects of the decision-makers and address their estimation and assessment of their payoffs and strategies based on the information available to them.

Learning in games has been investigated in many recent papers. In [9], [16], a fictitious-play algorithm is used to find Nash equilibrium in a nonzero-sum game. Players observe opponents' actions and update their strategies in reaction to others' actions in a best-response fashion. The authors in [14] propose a modified version of the fictitious play called joint fictitious play with inertia for potential games, in which players alternate their updates at different time slots. In standard fictitious play (Brown 1951, Robinson 1951), players have to monitor the actions of every other player and need to know their own payoff so as to find their optimal actions. In this paper, we are interested in fully distributed learning procedures, where players do not need any information about the actions or payoffs of the other players, and, moreover, they do not need to have complete information of their own payoff structure. The focus of this paper is on finite games, where the existence of mixed Nash equilibrium is ensured by [15]. Some recent work has been done for infinite (or continuous-kernel) games under incomplete information, where extremum-seeking methods have been used for (local) convergence to pure-strategy Nash equilibria (see [8] and several of the references therein).

A similar setting was adopted in [18], which however dealt with zero-sum games. Here we extend these results in a non-trivial way to general-sum two-person games and introduce the new paradigm of *hybrid learning*, where players can choose different pure learning schemes at different times based on their rationality and preferences. The heterogenous learning in [18] can be seen as a special case of the generalized hybrid learning of this paper. In order to render the learning more practical in the context of network security, we introduce additional features of the game: (F1) In addition to exogenous environment uncertainties, we introduce inherent mode uncertainties in players. Each player can be in an *active* mode or a *sleeping* mode. Players learn their strategies and average payoffs only when they are in an *active* mode. (F2) We allow the interaction between the players to occur at random times unknown by the players. We use stochastic approximation techniques to show that the hybrid learning schemes with random updates can be studied using their deterministic ordinary differential equation (ODE) counterparts. The ODE obtained for hybrid learning is a linear combination of ODEs from pure learning schemes. We show the convergence properties of the learning algorithms for special classes of games, namely, games with two actions, as well as potential games, and demonstrate their applicability in a network security environment.

The paper is structured as follows. In Section II, we formulate the two-player nonzero-sum stochastic game with incomplete information and introduce the solution concept of state-independent Nash equilibrium. In Section III, we present a number of distinct learning schemes and discuss their properties. In Section IV, we present main results on learning for general-sum games. In Section V, we apply the learning algorithms to a network security application. Section VI concludes the paper.

II. Two-Person Game

In this section, we consider a finite two-person nonzero-sum game (NZSG) in which each player has stochastic payoffs and the interactions between the players are random. Let $\Xi := \langle \mathcal{N}, \{\mathcal{S}_i\}_{i\in\mathcal{N}}, \{\Omega_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \{U_i(s, B^2, .)\}_{s\in\mathcal{S}, b\in\mathcal{B}, i\in\mathcal{N}}\rangle$ be the stochastic NZSG, where $\mathcal{N} = \{1, 2\}$ is the set of players P1 and P2 who maximize their payoffs, and $\mathcal{A}_1, \mathcal{A}_2$

This work was supported in part by grants from AFOSR and DOE.
Q. Zhu and T. Başar are with Dept. ECE and CSL, University of Illinois, 1308 West Main, Urbana, IL, 61801, USA., University of Illinois at Urbana-Champaign. {zhu31, basar1}@illinois.edu. H. Tembine is with Telecommunication Department, Supelec, France. E-mail: tembine@ieee.org

978-1-4577-0079-8/11/$26.00 ©2011 AACC

are the finite sets of actions available to players P1 and P2, respectively. The set $\mathcal{S}_i := [s_{i,1}, s_{i,2}, \cdots, s_{i,N_S^i}]$ comprises all possible N_S^i external states of Pi, which describes the environment where Pi resides. We assume that the state space $\mathcal{S} := \prod_{i \in \mathcal{N}} \mathcal{S}_i$ and the probability transition on the states are both unknown to the players. A state s_i is randomly and independently chosen at each time from the set $\mathcal{S}_i$. We assume that the action spaces are the same in each state.

In addition, players do not interact at all times. A player can be in one of the two modes: *active mode* or *sleep mode*, denoted by mode $B_i = 1$ and $B_i = 0$, respectively. Let $B_i, i \in \mathcal{N}$, be an i.i.d. random variable on $\Omega_i := \{0, 1\}$ whose probability mass function is given by $\rho_B^i = \begin{cases} p_i, & B^i = 1, \\ 1 - p_i, & B^i = 0 \end{cases}, i \in \mathcal{N}$. The player modes can be viewed as internal states that are governed by the inherent randomness of the player. The system mode $B^2 \in \Omega := \Omega_1 \times \Omega_2$ is a set of independent modes of the players and we denote by $\mathcal{B}^2 \subseteq \mathcal{N}$ as the corresponding set of active players to B^2.

The NZSG is characterized by utility functions $U_i : \mathcal{S} \times \Omega_i \times \mathcal{A}_1 \times \mathcal{A}_2 \to \mathbb{R}$. P$i$ collects a payoff $U_i(s, B^2, a_1, a_2)$ when P1 chooses $a_1 \in \mathcal{A}_1$ and P2 uses $a_2 \in \mathcal{A}_2$ at state $s \in \mathcal{S}$ and mode B^2.

Remark 1: The preceding game model can be viewed as a special class of stochastic games in which the state transitions are independent of the player actions as well as the current state and we assume that the state processes and the activities of the players are i.i.d. random variables.

We have slotted time, $t \in \{0, 1, \ldots\}$, when players pick their mixed strategies as functions of what has transpired in the past, to the extent the information available to them allows. Toward this end, we let $x_{i,t}(a_i)$ denote the probabilities of Pi choosing $a_i \in \mathcal{A}_i$ at time t, and let $\mathbf{x}_{i,t} = [x_{i,t}(a_i)]_{a_i \in \mathcal{A}_i}$ be the mixed strategies of Pi at time t, where more precisely, $\mathbf{x}_{i,t} \in \mathcal{X}_i := \left\{ \mathbf{x}_i \in \mathbb{R}^{|\mathcal{A}_i|} : \ x_i(a_i) \in [0,1], \sum_{a_i \in \mathcal{A}_i} x_i(a_i) = 1 \right\}$. In particular, we define $e_{a_i} \in \mathbb{R}^{|\mathcal{A}_i|}$, with $a_i \in \mathcal{A}_i$, as unit vectors of sizes $|\mathcal{A}_i|$, whose entry that corresponds to a_i is 1 while others are zeros. We assume that the mixed strategies of the players are independent of the current state s and the player mode B_i. For any given pair of mixed strategies, $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X}_1 \times \mathcal{X}_2$, and for a fixed $s_i \in S_i, B^2 \in \Omega$, we define the expected utility (as expected payoff to Pi) as $\mathbb{U}_i(s, B^2, \mathbf{x}_1, \mathbf{x}_2) := \mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2} U_i(s, B^2, a_1, a_2)$, where $\mathbb{E}_{\mathbf{x}_1,\mathbf{x}_2} U_i$ denotes expectation of U_i over the action sets of the players under the given mixed strategies. A further expectation of this quantity over the states s and B^2, denoted by $\mathbb{E}_{s,B^2}$, yields the performance index of the *expected game*. We now define the equilibrium concept of interest for this game, that is the equilibrium of the expected game:

Definition 1 (State-independent equilibrium): A strategy profile $(\mathbf{x}_1^*, \mathbf{x}_2^*) \in \mathcal{X}_1 \times \mathcal{X}_2$ is a state-independent equilibrium of the game Ξ if it is equilibrium of the expected game, i.e., $\forall \mathbf{x}_1 \in \mathcal{X}_1, \mathbf{x}_2 \in \mathcal{X}_2$, $\mathbb{E}_{s,B^2} \mathbb{U}_1(s, B^2, \mathbf{x}_1^*, \mathbf{x}_2^*) \geq \mathbb{E}_{s,B^2} \mathbb{U}_1(s, B^2, \mathbf{x}_1, \mathbf{x}_2^*)$, and $\mathbb{E}_{s,B^2} \mathbb{U}_2(s, B^2, \mathbf{x}_1^*, \mathbf{x}_2^*) \geq \mathbb{E}_{s,B^2} \mathbb{U}_2(s, B^2, \mathbf{x}_1^*, \mathbf{x}_2)$.

Since the expected game is a two-player game with finite actions for each player, the existence of an equilibrium follows from Nash's existence theorem [15], and hence we have the following lemma.

Lemma 1: The stochastic NSZG Ξ with unknown states and changing modes admits a state-independent equilibrium.

III. Learning in NZSGs

A. Learning Procedures

In many practical applications, players in two-person NZSGs do not have complete knowledge of each other's utility functions and the state of their environment. Moreover, they do not know whether they interact with the other player or not. Hence, the equilibrium strategy has to be learned online by observing the realized payoffs during each time slot. A general learning procedure is outlined as follows. At each time slot $t \in \mathbb{Z}_+$, each player generates an internal mode B_i to determine whether to participate in the game or not. If both players are active, they interact and receive a payoff after the play. If only one of the players is active, then the active player receives his payoff as an outcome of his action at t only without interaction with the other player. If players do not have the knowledge of their active mode probability p_i, then each player keeps a count of its interaction with others by updating its vectors $\theta_{ij,t} \in \mathbb{R}^2, i, j \in \{1, 2\}$, as follows: $\theta_{ij,t} = \theta_{ij,t-1} + \mathbb{1}_{\{B_j = 1\}}$, where $\theta_{ij,t}$ is Pi's count of Pj's number of activity since $t \geq 0$ and the initial condition is given by $\theta_{ij} = 0, \forall i, j \in \{1, 2\}$. The active players choose an action $a_{i,t} \in \mathcal{A}_i$ at time t and observe or measure an output $u_{j,t} \in \mathbb{R}$ as an outcome of their actions. Players estimate their payoffs by updating the entry of the estimated payoff vector $\hat{\mathbf{u}}_{i,t+1} \in \mathbb{R}^{|\mathcal{A}_i|}$ that corresponds to the chosen action $a_{i,t}$. In a similar way, players update their strategy vectors $\mathbf{x}_{i,t+1}$ based on a specific learning scheme (to be introduced later). The update of the strategy vectors can exploit the payoff information $\hat{u}_{i,t}$ from the previous time step. In this case, we say the learning is *combined*.

The general combined learning updates on the strategy and utility vectors take the following form:

$$\left\{ \begin{array}{lcl} \mathbf{x}_{i,t+1} & = & \mathbf{x}_{i,t} + \Pi_{i,t}(\lambda_{i,t}, a_{i,t}, u_{i,t}, \hat{\mathbf{u}}_{i,t}, \mathbf{x}_{i,t}), \\ \hat{\mathbf{u}}_{i,t+1} & = & \hat{\mathbf{u}}_{i,t} + \Sigma_{i,t}(\nu_{i,t}, a_{i,t}, u_{i,t}, \mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t}), \end{array} \right. \tag{1}$$

where $\Pi_{i,t}, \Sigma_{i,t}, i \in \mathcal{N}$, are properly chosen functions for strategy and utility updates, respectively. The parameters $\lambda_{i,t}, \nu_{i,t}$ are learning rates indicating players' capabilities of information retrieval and update. The vectors $\mathbf{x}_{i,t} \in \mathcal{X}_i$ are mixed strategies of the players at time t. $\hat{\mathbf{u}}_{i,t}, i \in \mathcal{N}$, are estimated average payoffs updated at each iteration t, and $u_{i,t}, i \in \mathcal{N}$, are the observed payoffs received by players at time t. The learning rates $\lambda_{i,t}, \nu_{i,t} \in \mathbb{R}_+$ need to satisfy the conditions (C1) $\sum_{t \geq 0} |\lambda_{i,t}|^2 < \infty$, $\sum_{t \geq 0} |\nu_{i,t}|^2 < \infty$; (C2) $\sum_{t \geq 0} |\lambda_{i,t}| = +\infty$, $\sum_{t \geq 0} |\nu_{i,t}| = +\infty$.

The learning rates of Pi are relative to their frequency of activity. In general, they are functions of $\theta_{ii}, i \in \mathcal{N}$, and can be written as $\lambda_{i,\theta_{ii}(t)}, \nu_{i,\theta_{ii}(t)}$. We need to adopt a time reference for the game using maximum learning rates among the active players, i.e., $\lambda_t^* := \max_{i\in\mathcal{B}^2(t)} \lambda_{i,\theta_{ii}(t)}$, $\nu_t^* := \max_{i\in\mathcal{B}^2(t)} \nu_{i,\theta_{ii}(t)}$. It can be verified that the reference learning rates λ_t^*, ν_t^* satisfy (C1) and (C2) if $\lambda_{i,t}, \nu_{i,t}$ satisfy the conditions for every $i \in \mathcal{N}$. The learning rates λ_t^*, ν_t^*, as we will see later, affect the ODE approximation.

We call the learning in (1) a COmbined DIstributed PAyoff and Strategy Reinforcement Learning (CODIPAS-RL) [18]. The players can have different learning rates for their utility and strategy updates. The payoff learning rate is on a faster time scale than strategy learning rate if $\lambda_{i,t}/\nu_{i,t} \to 0$ as $t \to \infty$; if it is the other way around, $\nu_{i,t}/\lambda_{i,t} \to 0$ as $t \to \infty$. In the former case, the payoff learning can be seen as quasi-static compared to the strategy learning, and *vice versa* for the latter.

B. Learning Schemes

We introduce different learning schemes in the form of (1) for the stochastic NZSG. Let $\mathcal{L} = \{\mathcal{L}_k, k \in \{1, 2, \cdots, 5\}\}$ be the set of five pure learning schemes. A player Pi chooses a learning schemes $\mathcal{P}_i$ from the set $\mathcal{L}$. We call the learning *homogeneous* if both players use the same pure learning schemes and *heterogeneous* if players use different learning schemes, i.e., $\mathcal{P}_1 \neq \mathcal{P}_2$.

1) Bush-Mosteller-based CODIPAS-RL $\mathcal{L}_1$: Let $\Gamma_i \in \mathbb{R}$ be a reference level of Pi and $\widetilde{\Gamma}_{i,t} := \frac{u_{i,t}-\Gamma_i}{\sup_{s,B^2,\mathbf{a}} |U_i(s,B^2,\mathbf{a})-\Gamma_i|}$. The learning pattern $\mathcal{L}_1$ is given by

$$\begin{cases} x_{i,t+1}(a_i) &= x_{i,t}(a_i) + \lambda_{i,t}\mathbb{1}_{\{i\in\mathcal{B}^2(t)\}}\times \\ & \quad \widetilde{\Gamma}_{i,t}\left(\mathbb{1}_{\{a_{i,t}=a_i\}} - x_{i,t}(a_i)\right), \\ \hat{u}_{i,t+1}(a_i) &= \hat{u}_{i,t}(a_i) + \nu_{i,t}\mathbb{1}_{\{a_{i,t}=a_i, i\in\mathcal{B}^2(t)\}}\times \\ & \quad (u_{i,t} - \hat{u}_{i,t}(a_i)). \end{cases}$$

The updates on the strategy and the estimated payoff are decoupled but they are implicitly dependent. The strategy update does not exploit the knowledge of estimated payoff but only relies on the observed payoffs during each time slot. The strategy update of $\mathcal{L}_1$ is widely studied in machine learning and has been initially proposed by Bush and Mosteller in [6]. Combined with the payoff update, we obtain a COPIDAS-RL based on Bush-Mosteller learning. When $\Gamma_i = 0$, we obtain the learning schemes in [2], [4].

2) Boltzmann-Gibbs-based CODIPAS-RL $\mathcal{L}_2$: Let $\tilde{\beta}_{i,\epsilon} : \mathbb{R}^{|\mathcal{A}_i|} \to \mathbb{R}^{|\mathcal{A}_i|}$ be the Boltzmann-Gibbs (B-G) strategy mapping given by $\tilde{\beta}_{i,\epsilon}(\hat{\mathbf{u}}_{i,t})(a_i) := \frac{e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i)}}{\sum_{a_i'\in\mathcal{A}_i} e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i')}}$, $a_i \in \mathcal{A}_i$. It is also known as the soft-max function. When $\epsilon \to 0$, the B-G strategy yields a (pure) strategy that picks the maximum entry of the payoff vector $\hat{\mathbf{u}}_{i,t}$. The learning pattern $\mathcal{L}_2$ is given by

$$\begin{cases} x_{i,t+1}(a_i) &= x_{i,t}(a_i) + \lambda_{i,t}\mathbb{1}_{\{i\in\mathcal{B}^2(t)\}}\times \\ & \quad \left(\tilde{\beta}_{i,\epsilon}(\hat{\mathbf{u}}_{i,t})(a_{i,t}) - x_{i,t}(a_i)\right), \\ \hat{u}_{i,t+1}(a_i) &= \hat{u}_{i,t}(a_i) + \nu_{i,t}\mathbb{1}_{\{a_{i,t}=a_i, i\in\mathcal{B}^2(t)\}}\times \\ & \quad (u_{i,t} - \hat{u}_{i,t}(a_i)). \end{cases}$$

The strategy and the estimated payoff are updated in a coupled fashion. The numerical value of experiment is used in the estimation, and the estimated payoffs are used to built the strategy (here the estimations are crucial since a player does not know the numerical value of the payoff corresponding to his other actions that he did not use). The strategy update is a B-G based reinforcement learning. Combined together one gets the B-G based CODIPAS-RL. The rest point $\mathcal{L}_2$ can be seen as the equilibrium for a modified game with the perturbed payoff $\mathbb{E}_{s,B^2}\mathbb{U}_i + \epsilon_i H_i$, where H_i is the extra entropy term as discussed in [16].

3) Imitative B-G CODIPAS-RL $\mathcal{L}_3$: Let $\beta_{i,\epsilon,t}^I : \mathcal{X}_i \times \mathbb{R}^{|\mathcal{A}_i|} \to \mathbb{R}^{|\mathcal{A}_i|}$ be the imitative B-G strategy mapping given by $\tilde{\beta}_{i,\epsilon,t}^I(\mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t})(a_i) = \frac{x_{i,t}(a_i)e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i)}}{\sum_{a_i'\in\mathcal{A}_i} x_{i,t}(a_i')e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i')}}$, $a_i \in \mathcal{A}_i$. The learning pattern $\mathcal{L}_3$ is given by

$$\begin{cases} x_{i,t+1}(a_i) &= x_{i,t}(a_i) + \lambda_{i,t}\mathbb{1}_{\{i\in\mathcal{B}^2(t)\}}\times \\ & \quad \left(\tilde{\beta}_{i,\epsilon,t}^I(\hat{\mathbf{u}}_{i,t})(a_i) - x_{i,t}(a_i)\right), \\ \hat{u}_{i,t+1}(a_i) &= \hat{u}_{i,t}(a_i) + \nu_{i,t}\mathbb{1}_{\{a_{i,t}=a_i, i\in\mathcal{B}^2(t)\}}\times \\ & \quad (u_{i,t} - \hat{u}_{i,t}(a_i)). \end{cases}$$

The imitative B-G learning weights the B-G strategy with the current strategy vector $\mathbf{x}_{i,t}$ and the strategy mapping $\tilde{\beta}_{i,\epsilon,t}^I$ is time-dependent. It allows the learning strategies to be attained at the boundary of the simplex $\mathcal{X}_i$.

4) Weighted Imitative B-G CODIPAS-RL $\mathcal{L}_4$: Let $\tilde{\beta}_{i,t}^W : \mathcal{X}_i \times \mathbb{R} \times \mathbb{R}^{|\mathcal{A}_i|} \to \mathbb{R}^{|\mathcal{A}_i|}$ be the imitative weighted B-G strategy mapping given by $\tilde{\beta}_{i,t}^W(\mathbf{x}_{i,t}, \lambda_{i,t}, \hat{\mathbf{u}}_{i,t})(a_i) := \frac{x_{i,t}(a_i)(1+\lambda_{i,t})^{\hat{u}_{i,t}(a_i)}}{\sum_{a_i'\in\mathcal{A}_i} x_{i,t}(a_i')(1+\lambda_{i,t})^{\hat{u}_{i,t}(a_i')}}$, for every $a_i \in \mathcal{A}_i$. The learning pattern $\mathcal{L}_4$ is given by

$$\begin{cases} x_{i,t+1}(a_i) &= x_{i,t}(a_i) + \mathbb{1}_{\{i\in\mathcal{B}^2(t)\}}\times \\ & \quad \left(\tilde{\beta}_{i,t}^W(\mathbf{x}_{i,t}, \lambda_{i,t}, \hat{\mathbf{u}}_{i,t})(a_i) - x_{i,t}(a_i)\right), \\ \hat{u}_{i,t+1}(a_i) &= \hat{u}_{i,t}(a_i) + \nu_{i,t}\mathbb{1}_{\{a_{i,t}=a_i, i\in\mathcal{B}^2(t)\}}\times \\ & \quad (u_{i,t} - \hat{u}_{i,t}(a_i)). \end{cases}$$

Note that the exploitation function learning $\tilde{\beta}_{i,t}^W$ is time dependent in $\mathcal{L}_4$ and is independent of parameter ϵ. If the learning yields an interior point as the equilibrium, then it is the exact equilibrium of the expected game, while the equilibrium in $\mathcal{L}_2$ is an approximated one for the $\epsilon-$perturbed game.

5) Weakened Fictitious-Play $\mathcal{L}_5$: Let $\tilde{\beta}_{i,t}^F : \mathbb{R}^{|\mathcal{A}_i|} \to 2^{\mathbb{R}^{|\mathcal{A}_i|}}$ be a point-to-set mapping (correspondence) $\tilde{\beta}_{i,t}^F(\hat{\mathbf{u}}_{i,t}) := (1-\epsilon)\delta_{\beta_i(\hat{\mathbf{u}}_{i,t})} + \frac{\epsilon}{|\mathcal{A}_i|}\mathbf{1}$, where $\mathbf{1} \in \mathcal{R}^{|\mathcal{A}_i|}$ is a vector with all its entries being 1; $\beta_i : \mathbb{R}^{|\mathcal{A}_i|} \to 2^{\mathcal{A}_i}$ is the best-response correspondence: $\beta_i(\hat{\mathbf{u}}_{i,t}) \in \arg\max_{a_i'\in\mathcal{A}_i} \hat{u}_{i,t}(a_i')$and $\delta_{\mathcal{Z}}, \mathcal{Z} \subseteq \mathcal{A}_i$, denotes a set of unit vectors $\{e_{a_i}, a_i \in \mathcal{Z}\}$. The learning pattern $\mathcal{L}_5$ is given by

$$\begin{cases} x_{i,t+1}(a_i) &= x_{i,t}(a_i) \in \mathbb{1}_{\{i\in\mathcal{B}^2(t)\}}\times \\ & \quad \left(\tilde{\beta}_{i,t}^F(\hat{\mathbf{u}}_{i,t}) - x_{i,t}(a_i)\right), \\ \hat{u}_{i,t+1}(a_i) &= \hat{u}_{i,t}(a_i) + \nu_{i,t}\mathbb{1}_{\{a_{i,t}=a_i, i\in\mathcal{B}^2(t)\}}\times \\ & \quad (u_{i,t} - \hat{u}_{i,t}(a_i)). \end{cases}$$

The weakened fictitious play $\mathcal{L}_5$ has been discussed in [12], [14]. Different from the classical fictitious play, a player does not observe the action played by the other player at the previous step and the utility function is unknown. Each player estimates its payoff by updating $\hat{\mathbf{u}}_{i,t}$ using perceived payoffs. The strategy update equation is composed of two parts. A player chooses one of his optimal actions with probability $(1-\epsilon)$ by optimizing the up-to-date payoff estimation $\hat{u}_{i,t}$, and plays an arbitrary action with equal probability ϵ.

IV. Main Results

A. Stochastic approximation of the pure learning schemes

The pure learning schemes introduced in Section III share the same learning structure for average utility but differ in their strategy learning. Denote by $\Pi_{i,t}^{(l)}$ the strategy learning function for $l \in \mathcal{L}$ in the general form (1). Following the multiple time-scale stochastic approximation framework developed in [3], [5], [11], [13], one can write the pure learning schemes into the form

$$\left\{ \begin{array}{lcl} \mathbf{x}_{i,t+1} - \mathbf{x}_{i,t} & \in & q_{i,t}\left(f_i^{(l)}(\mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t}) + M_{i,t+1}^{(l)}\right) \\ \hat{\mathbf{u}}_{i,t+1} - \hat{\mathbf{u}}_{i,t} & \in & \bar{q}_{i,t}\left(\mathbb{E}_{s,\mathbf{x}_{-i,t},\mathcal{B}^2} U_i - \hat{\mathbf{u}}_{i,t} + \bar{M}_{i,t+1}\right) \end{array} \right. ,$$

where $f_i^l = \mathbb{E}[\Pi_{i,t+1}^{(l)} | \mathcal{F}_t]$, $l \in \mathcal{L}$, is a learning pattern in the form of stochastic approximation. $q_{i,t}$ is a time-scaling factor which is a function of the learning rates $\lambda_{i,t}$ and the probability of Pi in active mode at time t, denoted by $\mathbb{P}(i \in \mathcal{B}^2(t))$; $\bar{q}_{i,t}$ is the time-scaling factor for $\hat{u}_{i,t}$. To use ODE approximation, we check first the assumptions given in the Appendix. The term $M_{i,t+1}^{(l)}$ is a bounded martingale difference because the strategies are in the product of simplices which are convex and compact, and the conditional expectation of $M_{i,t+1}$ given the sigma-algebra generated by the random variables $s_{t'}, \mathbf{x}_{t'}, u_{t'}, \hat{\mathbf{u}}_{t'},\ t' \leq t$, is zero. Similar properties hold for $\bar{M}_{t+1}$. The function f is a regular function, and hence Lipschitz over a compact set, which implies linear growth. Note that the case of constant learning rates can be analyzed under the same setting but the convergence result is weaker Thus, the asymptotic pseudo-trajectories for the non-vanishing time-scale ratio, i.e., $\lambda_{i,t}/\nu_{i,t} \rightarrow \gamma_i$ for some $\gamma_i \in \mathbb{R}_{++}$ are

$$\left\{ \begin{array}{lcl} \frac{d}{dt}\mathbf{x}_{i,t} & \in & g_{i,t}\left(f_i^{(l)}(\mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t})\right) \\ \frac{d}{dt}\hat{\mathbf{u}}_{i,t} & = & \bar{g}_{i,t}\left(\mathbb{E}_{s,\mathbf{x}_{-i,t},\mathcal{B}^2} U_i - \hat{\mathbf{u}}_{i,t}\right) \end{array} \right. ,$$

where $g_{i,t}$ (resp. $\bar{g}_{i,t}$) are the asymptotic functions of $q_{i,t}, \lambda_t^*, p_i$ (resp. $\bar{q}_{i,t}, , \nu_t^*, p_i$).

If the learning rates have the vanishing ratio, i.e., $\frac{\lambda_t}{\mu_t} \rightarrow 0$, the asymptotic pseudo-trajectories are

$$\left\{ \begin{array}{lcl} \frac{d}{dt}\mathbf{x}_{i,t} & \in & g_{i,t}\left(f_i^{(l)}(\mathbf{x}_{i,t}, \mathbb{E}_{s,\mathbf{x}_{-i,t}} U_i)\right) \\ \hat{\mathbf{u}}_{i,t} & \longrightarrow & \mathbb{E}_{s,\mathbf{x}_{-i},\mathcal{B}^2} U_i. \end{array} \right.$$

B. Stochastic approximation of the hybrid learning scheme

Players can choose different patterns during different time slots. Consider the hybrid and switching learning

$$\left\{ \begin{array}{l} \mathbf{x}_{i,t+1} - \mathbf{x}_{i,t} \in q_{i,t}(\sum_{l \in \mathcal{L}} \mathbb{1}_{\{l_{i,t}=l\}} f_i^{(l)}(\mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t}) + M_{i,t+1}^{(l)}) \\ \hat{\mathbf{u}}_{i,t+1} - \hat{\mathbf{u}}_{i,t} \in \bar{q}_{i,t}\left(\mathbb{E}_{s,\mathbf{x}_{-i,t}} U_i - \hat{\mathbf{u}}_{i,t} + \bar{M}_{i,t+1}\right) \end{array} \right.$$

TABLE I

Asymptotic pseudo-trajectories of pure learning

Learning patterns	Class of ODE
$\mathcal{L}_1$	Adjusted replicator dynamics
$\mathcal{L}_2$	Smooth best response dynamics
$\mathcal{L}_3$	Imitative BG dynamics
$\mathcal{L}_4$	Time-scaled replicator dynamics
$\mathcal{L}_5$	Perturbed best response dynamics

where $l_{i,t} \in \mathcal{L}$ is the learning pattern chosen by Pi at time t.

Theorem 1: Assume that each player Pi, $i \in \mathcal{N}$, adopts one of the CODIPAS-RLs in $\mathcal{L}$ with probability $\omega_i = [\omega_{i,l'}]_{l' \in \mathcal{L}} \in \Delta(\mathcal{L})$ and the learning rates satisfy conditions (C1) and (C2). Then, the asymptotic pseudo-trajectories of the hybrid and switching learning can be written into the form

$$\left\{ \begin{array}{lcl} \frac{d}{dt}\mathbf{x}_{i,t} & \in & g_{i,t}\left(\sum_{l \in \mathcal{L}} \omega_{i,l} f_i^{(l)}(\mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t})\right) \\ \frac{d}{dt}\hat{\mathbf{u}}_{i,t} & = & \bar{g}_{i,t}\left(\mathbb{E}_{s,\mathbf{x}_{-i,t}} U_i - \hat{\mathbf{u}}_{i,t}\right) \end{array} \right.$$

for the non-vanishing time-scale learning ratio $\lambda_{i,t}/\nu_{i,t}$; and,

$$\left\{ \begin{array}{lcl} \frac{d}{dt}\mathbf{x}_{i,t} & \in & g_{i,t}\left(\sum_{l \in \mathcal{L}} \omega_{i,l} f_i^{(l)}(\mathbf{x}_{i,t}, \mathbb{E}_{s,\mathbf{x}_{-i,t},\mathcal{B}^2} U_i)\right) \\ \hat{\mathbf{u}}_{i,t} & \longrightarrow & \mathbb{E}_{s,\mathbf{x}_{-i},\mathcal{B}^2} U_i \end{array} \right.$$

for the vanishing learning ratio $\lambda_{i,t}/\nu_{i,t}$.

In Table 2, we give the asymptotic pseudo-trajectory of the pure learning when the rate of payoff learning is faster than the strategy learning. Let $\overline{\mathbb{U}}_j(\mathbf{x}) := \mathbb{E}_{s,B^2} \mathbb{U}_j(s, B^2, \mathbf{x}), j \in \mathcal{N}$. In Table 2, the replicator dynamics are given by $\dot{x}_j(a_j) = q_j x_j(a_j) \left[\overline{\mathbb{U}}_j(e_{a_j}, \mathbf{x}_{-j}) - \sum_{a'_j \in \mathcal{A}_j} \overline{\mathbb{U}}_j(e_{a'_j}, \mathbf{x}_{-j}) x_j(a'_j)\right]$.The smooth best response dynamics are given by $\dot{x}_j(a_j) = q_j(\frac{e^{\frac{1}{\epsilon}\overline{\mathbb{U}}_j(e_{a_j}, \mathbf{x}_{-j})}}{\sum_{a'_j} e^{\frac{1}{\epsilon}\overline{\mathbb{U}}_j(e_{a'_j}, \mathbf{x}_{-j})}} - x_j(a_j))$. The imitative Boltzman-Gibbs dynamics are given by $\dot{x}_j(a_j) = q_j(\frac{x_j(a_j) e^{\frac{1}{\epsilon}\overline{\mathbb{U}}_j(e_{a_j}, \mathbf{x}_{-j})}}{\sum_{a'_j} x_j(a'_j) e^{\frac{1}{\epsilon}\overline{\mathbb{U}}_j(e_{a'_j}, \mathbf{x}_{-j})}} - x_j(a_j))$. The best response dynamics are given by $\dot{\mathbf{x}}_j \in q_j(\beta_j(\mathbf{x}_{-j}) - \mathbf{x}_j)$, and the payoff dynamics are $\frac{d}{dt}\hat{u}_j(a_j) = \bar{q}_j x_j(a_j)(\overline{\mathbb{U}}_j(e_{a_j}, \mathbf{x}_{-j}) - \hat{u}_j(a_j))$.

C. Connection with equilibria of the expected game

We study the convergence properties of the dynamics and their connection with the state-independent Nash equilibrium for three special classes of games.

1) Games with two actions: For two-player games with two actions, i.e, $\mathcal{A}_1 = \{a_1^1, a_1^2\}, \mathcal{A}_2 = \{a_2^1, a_2^2\}$, one can transform the system of ODEs of the strategy-learning into a planar system under the form

$$\dot{\alpha}_1 = Q_1(\alpha_1, \alpha_2),\ \dot{\alpha}_2 = Q_2(\alpha_1, \alpha_2), \tag{2}$$

where we let $\alpha_i = x_i(a_i^1)$. The dynamics for Pi can be expressed in terms of α_1, α_2 only as $x_1(a_1^2) = 1 - x_1(a_1^2)$, and $x_2(a_2^2) = 1 - x_2(a_2^2)$.

We use the Poincaré-Bendixson theorem and the Dulac criterion [10] to establish a convergence result for (2).

Theorem 2 ([10]): Consider an autonomous planar vector field as in (2).Let $\gamma(.)$ be a scalar function defined on the unit square $[0,1]^2$. If $\frac{\partial[\gamma(\alpha))\dot{\alpha}_1]}{\partial\alpha_1}+\frac{\partial[\gamma(\alpha)\dot{\alpha}_2]}{\partial\alpha_2}$ is not identically zero and does not change sign in $[0,1]^2$, then there are no cycles lying entirely in $[0,1]^2$.

Corollary 1: Consider a two-player two-action game. Assume that each player adopts the Boltzmann-Gibbs CODIPAS-RL with $\frac{\lambda_{i,t}}{\nu_{i,t}}=\frac{\lambda_t}{\nu_t}\longrightarrow 0$. Then, the asymptotic pseudo-trajectory reduces to a planar system in the form $\dot{\alpha}_1=\beta_{1,\epsilon}(u_1(e_{a_1},\alpha_2))-\alpha_1;\dot{\alpha}_2=\beta_{2,\epsilon}(u_2(\alpha_1,e_{a_2}))-\alpha_2$. Moreover, the system satisfies the conditions of Theorem 2 (known as the Dulac's criterion).

Note that for the replicator dynamics, the Dulac criterion reduces to $(1-2\alpha_1)(\overline{\mathbb{U}}_1(e_{a_1^1},\alpha_2)-\overline{\mathbb{U}}_1(e_{a_1^2},\alpha_2))+(1-2\alpha_2)(\overline{\mathbb{U}}_2(\alpha_1,e_{a_2^1})-\overline{\mathbb{U}}_2(\alpha_1,e_{a_2^2}))$ which vanishes for $(\alpha_1,\alpha_2)=(1/2,1/2)$. It is possible to have limit cycles in replicator dynamics and hence the Dulac criterion does not apply. However, the stability of the replicator dynamics can be directly studied in the two-action case by identifying the game as belonging to one of the types: coordination, anti-coordination, prisoner's dilemma, hawk-and-dove.

The following corollary now follows from Theorem 2.

Corollary 2:

(CR1) *Heterogeneous learning:* If P1 is with Boltzmann-Gibbs CODIPAS-RL and P2's learning leads to replicator dynamics, then the convergence condition reduces to $(1-2\alpha_2)(u_2(\alpha_1,e_{a_2^1})-u_2(\alpha_1,e_{a_2^2}))<1$ for all (α_1,α_2).

(CR2) *Hybrid learning:* If the players use an hybrid learning obtained by combining Boltzmann-Gibbs CODIPAS-RL with weight $\omega_{i,1}$ and the replicator dynamics with weight $1-\omega_{i,1}$, then the Dulac criterion reduces to $\omega_{1,2}[(1-2\alpha_1)(u_1(e_{a_1^1},\alpha_2)-u_1(e_{a_1^2},\alpha_2))]+\omega_{2,2}[(1-2\alpha_2)(u_2(\alpha_1,e_{a_2^1})-u_2(\alpha_1,e_{a_2^2}))]<w_{1,1}+w_{2,2}$ for all (α_1,α_2).

Remark 2 (Symmetric games with three actions): If the expected game is a symmetric game with three actions per player, then, the symmetric game dynamics reduce to the two-dimensional dynamical system. This allows us to apply the Dulac criterion.

2) Lyapunov games: We say that the game Ξ is a *Lyapunov game* under the hybrid dynamics if the resulting dynamics has a Lyapunov function.

Theorem 3: Consider a Lyapunov game under the learning schemes $\mathcal{L}_1,\mathcal{L}_4$. Then, the learning procedure has global convergence to the set of equilibria of the expected robust game for all interior initial conditions.

Note that this result holds also for $n-$player stochastic games with random updates.

We say that the stochastic game Ξ is an *expected robust potential game* if the expected payoff derives from a potential function.

Potential games constitute a special class of games where the payoff functions of the players are governed by a potential function $\Phi:\mathbb{R}^{\sum_{i\in\mathcal{N}}|\mathcal{A}_i|}\to\mathbb{R}$, i.e., $\mathbb{U}_i(e_{a_i},x_{-i})=\frac{\partial\Phi(\mathbf{x})}{\partial x_i(a_i)},i\in\mathcal{N},a_i\in\mathcal{A}_i$. We use a Lyapunov approach to show the global convergence of hybrid learning for potential games.

Lemma 2: Assume that the stochastic NZSG Φ has a potential function Φ. Then, there exists a Lyapunov function $V^R(\mathbf{x}_1,\mathbf{x}_2):\mathbb{R}^{|\mathcal{A}_1|+|\mathcal{A}_2|}\to\mathbb{R}$ for learning schemes $\mathcal{L}_1,\mathcal{L}_4$-associated replicator dynamics and it is given by its potential $V^R=\Phi$. Hence, the replicator dynamics converge to a rest point. In addition, if starting from an interior point of the simplex, the dynamics converge to the Nash equilibrium of the game Ξ.

Lemma 3: Let $V^B(\mathbf{x}_1,\mathbf{x}_2):\mathbb{R}^{|\mathcal{A}_1|+|\mathcal{A}_2|}\to\mathbb{R}$ be a Lyapunov function for learning pattern $\mathcal{L}_l$-associated replicator dynamics $f^l,l=2$, such that $V^B(\mathbf{x}_1,\mathbf{x}_2)=\Phi(\mathbf{x}_1,\mathbf{x}_2)+\epsilon_1H_1(\mathbf{x}_1)+\epsilon_2H_2(\mathbf{x}_2)$,where $H_i:\mathbb{R}^{|\mathcal{A}_i|}\to\mathbb{R}$ are strictly concave perturbation functions which can take different forms depending on the pure learning scheme l. The ODEs corresponding to the learning schemes converge to a set of perturbed equilibria of the game Ξ.

Theorem 4: Assume that the stochastic NZSG Ξ has a potential function Φ. The hybrid learning with $\mathcal{L}_1$ and $\mathcal{L}_2$ converges locally to a perturbed state-independent Nash equilibrium $\mathbf{x}_1^*,\mathbf{x}_2^*$ of the potential game Ξ for sufficiently small ϵ_i.

The proof of Lemmas 2, 3, and Theorem 4 can be found in the internal technical report [17].

V. Security Application

In this section, we use the learning algorithm to study a two-person security game in a network between an intruder and an administrator. An administrator P1 can use different levels of protection. The intruder P2 can launch an attack that can be of high or low intensity. Let the action sets for P1 and P2 be $\mathcal{A}_1:=\{H,L\}$ and $\mathcal{A}_2:=\{S,W\}$, respectively. The network administrator is assumed to be always on alert while the intruder attacks with a probability p. Hence, the set $\mathcal{B}^2(t)$ can be of two types, i.e., (C1) {P1, P2} or (C2) {P1}. The former case (C1) corresponds to the scenario where the intruder and the administrator attack and defend, respectively, whereas the latter (C2) suggests that the administrator faces no threat. We represent the payoff under these two scenarios by $\mathbf{M}_1$ and $\mathbf{M}_2$, respectively: $\mathbf{M}_1:=\left[\begin{array}{ccc} & S & W\\ H & 1,-1 & 1,0\\ L & -2,1 & 2,0\end{array}\right]$, $\mathbf{M}_2:=\left[\begin{array}{cc} H & 1\\ L & 2\end{array}\right]$. In (C1), a successful defense against attack yields a payoff of 2 for P1 while a failure results in a payoff of -2. A successful attack yields P2 a payoff of 1 while a failed attack yields a zero payoff. The employment of strong defense (H) or strong attack (S) costs an extra unit of effort as compared to the low defense (L) and the weak attack (W) for P1 and P2, respectively. In (C2), P1 stays secure without the threat from the intruder hence yields a payoff of 2. However, the high security level costs an extra unit of energy from the player.

The payoffs in $\mathbf{M}_1$ and $\mathbf{M}_2$ are subject to exogenous noise which depends on the environmental state s. The state-independent equilibrium of the game is found to be at $\mathbf{x}_1^*=[\frac{1}{2},\frac{1}{2}]^T,\mathbf{x}_2^*=[\frac{1}{3},\frac{2}{3}]^T$ and the optimal average payoffs are $\hat{\mathbf{u}}_1^*=[\frac{2}{3},\frac{2}{3}]^T,\hat{\mathbf{u}}_2^*=[0,0]^T$. In Figures 1 and 2, we show the payoffs and mixed strategies of both players when both players use the learning pattern $\mathcal{L}_1$. We can

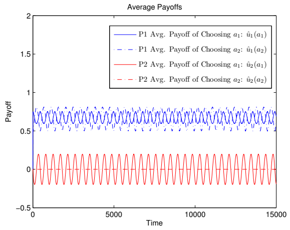

Fig. 1. The payoffs to the players with both players using $\mathcal{L}_1$.

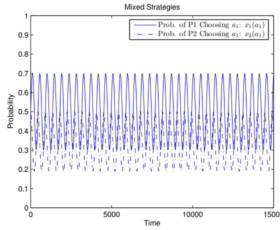

Fig. 2. The mixed strategies of the players with both players using $\mathcal{L}_1$.

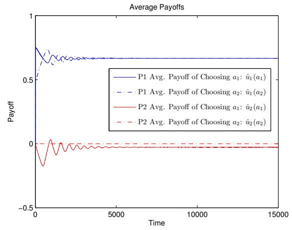

Fig. 3. The payoffs to the players with both players using $\mathcal{L}_2$.

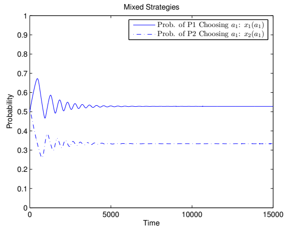

Fig. 4. The mixed strategies of the players with both players using $\mathcal{L}_2$.

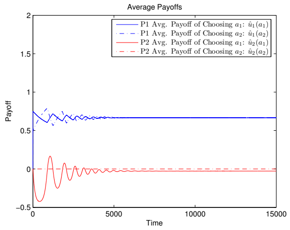

Fig. 5. The payoffs to the heterogeneous players with P1 using $\mathcal{L}_1$ and P2 using $\mathcal{L}_2$.

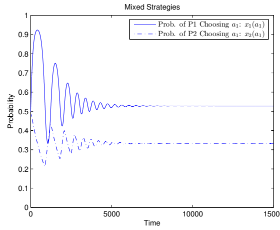

Fig. 6. The mixed strategies of the heterogeneous players with P1 using $\mathcal{L}_1$ and P2 using $\mathcal{L}_2$.

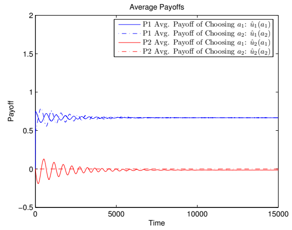

Fig. 7. The payoffs to the players with both players using hybrid learning scheme with equal weights on $\mathcal{L}_1$ and $\mathcal{L}_2$.

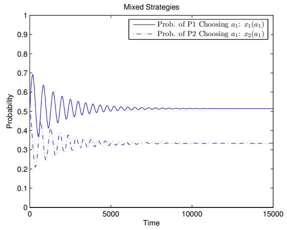

Fig. 8. The mixed strategies of the players with both players using hybrid learning scheme with equal weights on $\mathcal{L}_1$ and $\mathcal{L}_2$.

see that the replicator dynamics from $\mathcal{L}_1$ do not converge. However the time average strategies $\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbf{x}_{i,t}dt$ converges to $\mathbf{x}_1^*, \mathbf{x}_2^*$, respectively, and, the time average payoffs $\lim_{T\to\infty}\frac{1}{T}\int_0^T \hat{\mathbf{u}}_{i,t}dt$ converges to $\hat{\mathbf{u}}_1^*, \hat{\mathbf{u}}_2^*$, respectively. In Figures 3 and 4, we show the payoffs and mixed strategies of the players when they both adopt the learning pattern $\mathcal{L}_2$. We choose $\epsilon = 1/50$ and observe that the mixed strategies converge to $\bar{\mathbf{x}}_1 = [0.5277, 0.4723]^T, \bar{\mathbf{x}}_1 = [0.3333, 0.6667]^T$ and the payoffs converge to $\bar{\hat{\mathbf{u}}}_1 = [0.6667, 0.6667]^T, \bar{\hat{\mathbf{u}}}_2 = [-0.027, 0]^T$, which are in the close neighborhood of $\hat{\mathbf{u}}_1^*, \hat{\mathbf{u}}_2^*$. In Figures 5 and 6, we show the convergence of the heterogeneous learning scheme where P1 uses $\mathcal{L}_1$ and P2 uses $\mathcal{L}_2$. With $\epsilon = 1/50$, we find the converging strategies at $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_1$ and the payoffs at $\bar{\hat{\mathbf{u}}}_1, \bar{\hat{\mathbf{u}}}_2$. In Figures 7 and 8, we show the convergence of the hybrid learning scheme where P1 and P2 adopt $\mathcal{L}_1$ and $\mathcal{L}_2$ with equal weights. The strategies converge to $[0.5145, 0.4855]^T, [0.3334, 0.6666]^T$ for P1 and P2, respectively, whereas the payoffs converge to $[0.6666, 0.6666]^T, [-0.01459, 0]^T$ for P1 and P2, respectively.

VI. Conclusions and Future Works

We have presented distributed strategic learning algorithms for two-person nonzero-sum stochastic games along with their general convergence or non-convergence properties. Interesting work that we leave for the future is to extend this learning framework to an arbitrary number of players, each of them adopting hybrid learning with a diffusion term leading to *stochastic differential equations*. Another extension will be to more general stochastic games where the state evolution depends on the actions used the players and their states. This situation is more complicated because the noises are correlated and depend on states and actions and the convergence issue in that case is a very challenging open problem.

References

[1] T. Alpcan and T. Başar. *Network Security: A Decision and Game Theoretic Approach*. Cambridge University Press, 2011.

[2] W. B. Arthur. On designing economic agents that behave like human agents. *J. Evolutionary Econ. 3*, pages 1–22, 1993.

[3] M. Benaïm and M. Faure. Stochastic approximations, cooperative dynamics and supermodular games. *Preprint available at http://members.unine.ch/michel.benaim/perso/papers1.html*, 2010.

[4] T. Borgers and R. Sarin. Learning through reinforcement and replicator dynamics. *Mimeo, University College London.*, 1993.

[5] V. S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*. 2008.

[6] R. Bush and F. Mosteller. *Stochastic Models of Learning*. 1955.

[7] R.W. Ferrero, S.M. Shahidehpour, and V.C. Ramesh. Transaction analysis in deregulated power systems using game theory. *Power Systems, IEEE Transactions on*, 12(3):1340 –1347, Aug. 1997.

[8] P. Frihauf, M. Krstic, and T. Başar. Nash equilibrium seeking with infinitely-many players. *in Proceedings of American Control Conference (ACC)*, 2011.

[9] D. Fudenberg and D. Levine. *Learning in Games*. 1998.

[10] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. 1983.

[11] Kushner H. Stochastic approximation: a survey. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:87–96, 2010.

[12] D. Leslie and E. Collins. Individual q-learning in normal form games. *SIAM J. Control Optim.*, 44:495–514, 2005.

[13] D. S. Leslie and E. J. Collins. Convergent multiple timescales reinforcement learning algorithms in normal form games. *The Annals of Applied Probability*, 13(4):1231–1251, 2003.

[14] J. R. Marden, G. Arslan, and J. S. Shamma. Joint strategy fictitious play with inertia for potential games. *in Proc. 44th IEEE Conf. Decision Control*, pages 6692–6697, Dec. 2005.

[15] J. Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

[16] J. S. Shamma and G. Arslan. Dynamic fictitious play, dynamic gradient play, and distributed convergence to Nash equilibria. *IEEE Trans Automatic Control*, 50(3):312–327, March 2005.

[17] Q. Zhu, T. Hamidou, and T. Başar. Distributed strategic learning with application to network security. *Internal Technical Report, CSL, UIUC*, 2011.

[18] Q. Zhu, H. Tembine, and T. Başar. Heterogeneous learning in zero-sum stochastic games with incomplete information. *in 49th IEEE Conf. on Decision and Control, Atlanta, GA, USA*, 2010.