CHAPTER 14

HYBRID LEARNING IN STOCHASTIC GAMES AND ITS APPLICATION IN NETWORK SECURITY†

Quanyan Zhu[1], Hamidou Tembine[2], and Tamer Başar[1]

[1] Coordinated Science Laboratory,
University of Illinois at Urbana-Champaign, Urbana, Illinois, USA

[2] SUPELEC, Gif sur Yvette, France

ABSTRACT

We consider in this chapter a class of two-player nonzero-sum stochastic games with incomplete information, which is inspired by recent applications of game theory in network security. 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

†This material is based upon work supported in part by the U.S. Air Force Office of Scientific Research (AFOSR) under grant number AFOSR MURI FA9550-09-1-0249.

a player is in an active mode, she updates her strategy and estimates of unknown quantities using a specific pure or hybrid learning pattern. The players' intelligence and rationality are captured by the weighted linear combination of different learning patterns. 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 for special classes of games, namely, games with two actions, and potential games. Results are applied to network security games between an intruder and an administrator, where the noncooperative behaviors are well characterized by the features of distributed hybrid learning.

14.1 INTRODUCTION

In recent years, game-theoretic methods have been applied to study resource allocation problems in communication networks [2], security mechanisms for network security and privacy [1, 17], and economic pricing in power networks [9]. Most frameworks have assumed the rationality of the agents or the decision-makers as well as the complete information about their payoffs and strategies. However, in practice, due to the noise and the 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 the payoff and strategy based on the information accessible to them.

In this chapter, we consider a class of two-player nonzero-sum stochastic games with incomplete information. We develop fully distributed payoff and strategy reinforcement learning (CODIPAS-RL) 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. In contrast to the standard reinforcement learning algorithms which focus only on either strategy or payoff reinforcement for the equilibrium learning,

the algorithm that couples the payoff reinforcement learning together with strategy-reinforcement learning allows an immediate prediction and updates the strategies by updated estimations based on recent experiences. The payoff reinforcement learning in our proposed algorithms bears a connection with the Q-learning algorithms in [23, 26], which have been commonly applied to learn the Q-functions of Markov decision processes (MDPs).

We specifically discuss five pure CODIPAS-RL algorithms and 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 such as those in games with two actions, and Lyapunov games.

We apply the learning algorithms to a class of security games where an attacker and an intrusion detection system (IDS) strategically choose their actions to optimize their payoffs. Many forms of security games have been formulated to provide quantitative security and dependability analysis of networked systems [1, 17, 32]. However, technical difficulties in quantifying appropriate security metrics or payoff functions render it difficult to specify the utility functions for the attacker and the defender. In addition, the inevitable false positive and false negative errors in the detection often lead to incomplete information in a dynamic network environment. Our hybrid learning framework for the two-person game with incomplete information provides an appropriate theoretical basis for the on-line implementation of game-theoretic algorithms.

14.1.1 Related Work

Learning in games has been investigated in several papers in the recent literature. In [10, 22], 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 [18] 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 all these

learning schemes, 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 chapter, 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.

Young proposes in [29] such a completely uncoupled learning rule, called interactive trial and error learning. Players occasionally try out new actions and accept them if they lead to higher payoffs. If a player experiences a decrease in payoff due to strategy changes by some other players, he initiates a random search for a new strategy and settles on one with a probability that increases monotonically with its realized payoff. When used by all players, the learning scheme yields pure-strategy Nash equilibrium behavior under an interdependency condition. However, in games without pure-strategy Nash equilibrium, it fails to yield Nash equilibrium strategies.

In [25, 28], strategy reinforcement learning in finite games has been studied. The ordinary differential equation (ODE) approximation of the learning algorithm is shown to be equivalent to an adjusted replicator dynamics [24]. In [15], a multiple-time scale model-free algorithm is introduced and it is shown to be asymptotically equivalent to the smooth fictitious play algorithm. In [31,32], we introduce a class of combined distributed payoff and strategy reinforcement learning schemes (CODIPAS-RL), and propose a heterogeneous learning algorithm for two-person zero-sum stochastic games with incomplete information, where different players can adopt different learning schemes and learning rates. In [30], we propose a Q-learning algorithm for zero-sum stochastic games and apply it to dynamic configuration problems of intrusion detection systems.

14.1.2 Contribution

In this chapter, we consider a class of general-sum two-person games and introduce the new paradigm of *hybrid learning* under the frameworks of combined distributed payoff and strategy reinforcement learning (CODIPAS-RL), where in order to render the learning algorithm practical to implement in the context of network security, we introduce the following 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 to the players.

We use stochastic approximation techniques to show that the hybrid learning schemes with random updates can be studied using their deterministic 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, and potential games, and demonstrate their applications in a network security environment.

14.1.3 Organization of the Chapter

The chapter is structured as follows. In Section 14.2, we formulate the two-player nonzero-sum stochastic game with incomplete information and introduce the solution concept of state-independent Nash equilibrium. In Section 14.3, we present a number of distinct learning schemes and discuss their properties. In Section 14.4, we present main results on learning for general-sum games. In Section 14.5, we apply the learning algorithms to a network security application. Section 14.6 concludes the chapter. In Table 14.1, we summarize the notation used in the chapter for reader's convenience.

14.2 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 $\mathcal{A}_1, \mathcal{A}_2$ 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

Table 14.1 Table of Notations

Symbol	Meaning
$a_{i,t} \in \mathcal{A}_i$	Action of player i (Pi) at time t
$\mathbf{x}_{i,t} \in \mathcal{X}_i$	Mixed strategy of Pi at t
$B_i \in \{0,1\}$	Active or sleep mode of Pi
$s_i \in \mathcal{S}_i$	External state of Pi
$u_{i,t} \in \mathbb{R}$	Observed payoff by Pi at t
$\hat{\mathbf{u}}_{i,t} \in \mathbb{R}^{\lvert\mathcal{A}_i\rvert}$	Estimated payoff vector of Pi at t
$\mathbb{U}_i \in \mathbb{R}$	Mixed extension of the payoff U_i.
$\beta_i(\hat{\mathbf{u}}_{i,t}) \subseteq \mathcal{A}_i$	Best response
$\tilde{\beta}_{i,\epsilon}(\hat{\mathbf{u}}_{i,t}) \in \mathbb{R}^{\lvert\mathcal{A}_i\rvert}$	Boltzmann-Gibbs (B-G) strategy
$\tilde{\beta}_i^I(\mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t}) \in \mathbb{R}^{\lvert\mathcal{A}_i\rvert}$	Imitative B-G strategy
$\tilde{\beta}_i^W(\mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t}) \in \mathbb{R}^{\lvert\mathcal{A}_i\rvert}$	Weighted imitative B-G strategy
$\tilde{\beta}_i^F(\hat{\mathbf{u}}_{i,t}) \in \mathbb{R}^{\lvert\mathcal{A}_i\rvert}$	Weakened fictitious-play strategy
$\nu_{i,t} \in \mathbb{R}_+$	Payoff learning rates of Pi at t
$\lambda_{i,t} \in \mathbb{R}_+$	Strategy learning rates of Pi at t
$e_{a_i} \in \mathbb{R}^{\lvert\mathcal{A}_i\rvert}$	The unit vector with 1 at the position of a_i and 0 otherwise

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

$$\rho_B^i = \begin{cases} p_i, & B^i = 1, \\ 1-p_i, & B^i = 0 \end{cases}, i \in \mathcal{N}. \tag{14.1}$$

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 set of active players corresponding 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.

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.

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 $\mathbb{E}_{s, B^2}$, 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 14.1 (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.,* $\forall \mathbf{x}_1 \in \mathcal{X}_1, \mathbf{x}_2 \in \mathcal{X}_2$,

$$\begin{aligned}\mathbb{E}_{s,B^2}\mathbb{U}_1(s, B^2, \mathbf{x}_1^*, \mathbf{x}_2^*) &\geq \mathbb{E}_{s,B^2}\mathbb{U}_1(s, B^2, \mathbf{x}_1, \mathbf{x}_2^*),\\ \mathbb{E}_{s,B^2}\mathbb{U}_2(s, B^2, \mathbf{x}_1^*, \mathbf{x}_2^*) &\geq \mathbb{E}_{s,B^2}\mathbb{U}_2(s, B^2, \mathbf{x}_1^*, \mathbf{x}_2).\end{aligned}$$

Since the expected game is a two-player game with finite actions for each player, we can show the existence of the equilibrium using Nash's existence theorem [20] and state the following lemma.

Lemma 14.1 (Existence) *The stochastic NSZG* Ξ *with unknown states and changing modes admits a state-independent equilibrium.*

14.3 LEARNING IN NZSGS

In this section, we introduce different learning schemes for the stochastic NZSGs introduced in Section 14.2 and discuss the stochastic approximation of the learning schemes with ODEs.

14.3.1 Learning Procedures

In many practical applications, the players in the two-person NZSGs do not have the complete knowledge of their payoff 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 payoffs at 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 the interaction with the other player. If players do not have the knowledge of their active mode probability p_i, then each player keeps 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 activities 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 coupled; otherwise, we say that it is uncoupled.

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

$$\begin{cases} \mathbf{x}_{i,t+1} &= \mathbf{x}_{i,t} + \Pi_{i,t}(\lambda_{i,t}, a_{i,t}, u_{i,t}, \hat{\mathbf{u}}_{i,t}, \mathbf{x}_{i,t}), \\ \hat{\mathbf{u}}_{i,t+1} &= \hat{\mathbf{u}}_{i,t} + \Sigma_{i,t}(\nu_{i,t}, a_{i,t}, u_{i,t}, \hat{\mathbf{u}}_{i,t}, \mathbf{x}_{i,t}), \end{cases} \tag{14.2}$$

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 rate of Pi is relative to its frequency of activity. In general, the learning rates 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 (14.2) a COmbined DIstributed PAyoff and Strategy Reinforcement Learning (CODIPAS-RL) [31]. 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 $t \to \infty$; it is on a slower time scale if $\nu_{i,t}/\lambda_{i,t} \to 0$ 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.

14.3.2 Learning Schemes

We introduce different learning schemes in the form of (14.2) for the stochastic NZSG. Let $\mathcal{L} = \{\mathcal{L}_k, k = 1, 2, \cdots, 5\}$ be a 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., $\mathcal{P}_1 \neq \mathcal{P}_2$.

14.3.2.1 Bush-Mosteller-based CODIPAS-RL $\boldsymbol{\mathcal{L}_1}$ Let $\Gamma_i \in \mathbb{R}$ be a reference level of Pi and

$$\widetilde{\Gamma}_{i,t} := \frac{u_{i,t} - \Gamma_i}{\sup_{s,B^2,\mathbf{a}} |U_i(s, B^2, \mathbf{a}) - \Gamma_i|}. \tag{14.3}$$

The learning pattern $\mathcal{L}_1$ is given by

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

The updates on the strategy and the estimated payoff are decoupled. The strategy update does not exploit the knowledge of estimated payoff but only relies on the observed payoffs at 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 [8]. 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 [3, 6].

14.3.2.2 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

$$\tilde{\beta}_{i,\epsilon}(\hat{\mathbf{u}}_{i,t})(a_i) := \frac{e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i)}}{\sum_{a_i' \in \mathcal{A}_i} e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i')}},\ a_i \in \mathcal{A}_i. \tag{14.4}$$

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

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

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 the 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 [22].

14.3.2.3 Imitative B-G CODIPAS-RL $\mathcal{L}_3$ Let $\beta^I_{i,\epsilon,t} : \mathcal{X}_i \times \mathbb{R}^{|\mathcal{A}_i|} \to \mathbb{R}^{|\mathcal{A}_i|}$ be the imitative B-G strategy mapping given by

$$\tilde{\beta}^I_{i,\epsilon,t}(\mathbf{x}_{i,t}, \hat{\mathbf{u}}_{i,t})(a_i) = \frac{x_{i,t}(a_i) e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i)}}{\sum_{a_i' \in \mathcal{A}_i} x_{i,t}(a_i') e^{\frac{1}{\epsilon}\hat{u}_{i,t}(a_i')}},\ a_i \in \mathcal{A}_i. \tag{14.5}$$

The learning pattern $\mathcal{L}_3$ is given by

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

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

14.3.2.4 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

$$\tilde{\beta}_{i,t}^W(\mathbf{x}_{i,t}, \lambda_{i,t}, \hat{\mathbf{u}}_{i,t})(a_i) := \frac{x_{i,t}(a_i)(1+\lambda_{i,t})^{\hat{u}_{i,t}(a_i)}}{\sum_{a_i' \in \mathcal{A}_i} x_{i,t}(a_i')(1+\lambda_{i,t})^{\hat{u}_{i,t}(a_i')}}, \tag{14.6}$$

for every $a_i \in \mathcal{A}_i$. The learning pattern $\mathcal{L}_4$ is given by

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

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.

14.3.2.5 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)

$$\tilde{\beta}_{i,t}^F(\hat{\mathbf{u}}_{i,t}) := (1-\epsilon)\delta_{\beta_i(\hat{\mathbf{u}}_{i,t})} + \frac{\epsilon}{|\mathcal{A}_i|}\mathbf{1}, \tag{14.7}$$

where $\mathbf{1} \in \mathbb{R}^{|\mathcal{A}_i|}$ 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:

$$\beta_i(\hat{\mathbf{u}}_{i,t}) \in \arg\max_{a_i' \in \mathcal{A}_i} \hat{u}_{i,t}(a_i') \tag{14.8}$$

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

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

The weakened fictitious play $\mathcal{L}_5$ has been discussed in [15, 18]. Different from the classical fictitious play, a player does not observe the action played by the other player at the previous step and the payoff 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 estimate $\hat{u}_{i,t}$, and plays an arbitrary action with equal probability ϵ.

Remark 14.1 *We note that the average payoff-learning in the five pure learning schemes can be seen as the reinforcement learning of Q-functions in MDPs, which have been introduced in [23, 26]. Since we have considered a stochastic game with state transitions that are independent of the actions of the players, the Q-function in MDPs is reduced to the average-payoff function* $\hat{\mathbf{u}}_{i,t}$.

14.4 MAIN RESULTS

In this section, we introduce the new paradigm of hybrid learning, present the main results on learning in two-person general-sum games, and discuss their convergence properties for some special classes of games.

14.4.1 Stochastic approximation of the pure learning schemes

The pure learning schemes introduced in Section 14.3 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 (14.2). Following the multiple time-scale stochastic approximation framework developed in [5, 7, 14, 16], one can write the pure learning schemes into the form

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

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 conditions 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., $\lambda_{i,t}/\nu_{i,t} \to \gamma_i$ 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., $\frac{\lambda_t}{\mu_t} \to 0$, the asymptotic pseudo-trajectories are

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

14.4.2 Stochastic approximation of the hybrid learning scheme

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

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

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

Theorem 14.1 *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*

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

Table 14.2 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

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

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

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

Proof: We first examine the strategy learning given by

$$\mathbf{x}_{i,t+1}-\mathbf{x}_{i,t}\in\mathbb{1}_{\{i\in\mathcal{B}^2(t)\}}\lambda_{i,t}\left(\sum_{l\in\mathcal{L}}\mathbb{1}_{\{l_{i,t}=l\}}f_i^{(l)}(\mathbf{x}_t)+M_{i,t+1}^{(l)}\right)$$

By taking λ_t as the reference learning rate, the drift (expected change in one step) can be computed via

$$\lim_{\lambda_{i,t}\longrightarrow 0}\mathbb{E}\left(\frac{\mathbf{x}_{i,t+1}-\mathbf{x}_{i,t}}{\lambda_{i,t}}\mid\mathcal{F}_t\right)=\mathbb{P}(i\in\mathcal{B}^2(t))\left(\sum_{l\in\mathcal{L}}\omega_{i,l}f_i^{(l)}(\mathbf{x}_t)\right)$$

where we used the fact that $\mathbb{E}\left(M_{i,t+1}^{(l)}\mid\mathcal{F}_t\right)=0$. The drift has the form

$$g_{i,t}\sum_{l\in\mathcal{L}}\omega_{i,l}f_i^{(l)}(\mathbf{x}_t).$$

We check that the assumptions A1-A4 given in the Appendix are all met. The asymptotic pseudo-trajectory reduces to

$$\frac{d}{dt}\mathbf{x}_{i,t}=g_{i,t}\sum_{l\in\mathcal{L}}\omega_{i,l}f_i^{(l)}(\mathbf{x}_t).$$

For two time-scales CODIPAS-RL, we use the same lines as in [7, 31]. ■

In Table 14.2, we give the asymptotic pseudo-trajectory of the pure learning when the rate of payoff learning is faster than that of 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 14.2, 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

$$\dot{x}_j(a_j) = q_j \left(\frac{e^{\frac{1}{\epsilon}\overline{\mathbb{U}}_j(e_{a_j}, \mathbf{x}_{-j})}}{\sum_{a'_j} e^{\frac{1}{\epsilon}\overline{\mathbb{U}}_j(e_{a'_j}, \mathbf{x}_{-j})}} - x_j(a_j) \right).$$

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

$$\frac{d}{dt}\hat{u}_j(a_j) = \bar{q}_j x_j(a_j)(\overline{\mathbb{U}}_j(e_{a_j}, \mathbf{x}_{-j}) - \hat{u}_j(a_j)).$$

The imitative Boltzman-Gibbs dynamics are given by

$$\dot{x}_j(a_j) = q_j \left(\frac{x_j(a_j) e^{\frac{1}{\epsilon}\overline{\mathbb{U}}_j(e_{a_j}, \mathbf{x}_{-j})}}{\sum_{a'_j} x_j(a'_j) e^{\frac{1}{\epsilon}\overline{\mathbb{U}}_j(e_{a'_j}, \mathbf{x}_{-j})}} - x_j(a_j) \right).$$

14.4.3 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.

14.4.3.1 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

$$\dot{\alpha}_1 = Q_1(\alpha_1, \alpha_2),\ \dot{\alpha}_2 = Q_2(\alpha_1, \alpha_2), \tag{14.9}$$

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 [11] to establish a convergence result for (14.9).

Theorem 14.2 ([11]) *For an autonomous planar vector field as in (14.9), the Dulac's criterion states as follows: 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 14.1 *Consider a two-player two-action game. Assume that each of the players 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 Dulac's criterion.

Proof : We apply Theorem 14.2 with $\gamma(\cdot) \equiv 1$ and find the divergence to be equal to -2, which is strictly negative. Hence, the result follows. ■

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 one of four types: coordination, anti-coordination, prisoner's dilemma, hawk-and-dove [21, 27].

The following corollary now follows from Corollary 14.1.

Corollary 14.2 *Consider a two-player two-action game.*

(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 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

$$\begin{array}{l} \omega_{1,2}[(1-2\alpha_1)(u_1(e_{a_1^1},\alpha_2)-u_1(e_{a_1^2},\alpha_2))] \\ +\,\omega_{2,2}[(1-2\alpha_2)(u_2(\alpha_1,e_{a_2^1})-u_2(\alpha_1,e_{a_2^2}))] \end{array} \quad < \quad w_{1,1}+w_{2,2}$$

for all (α_1,α_2).

Remark 14.2 (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.*

14.4.3.2 Lyapunov games We say that a game Ξ is a *Lyapunov game* under a given hybrid dynamics if the resulting dynamics has an associated Lyapunov function $V(\mathbf{x}) : \Delta \subseteq \mathbb{R}^{\sum_i |\mathcal{A}_i|} \to \mathbb{R}_+$. Note that a Lyapunov function $V(\mathbf{x})$ is positive definite on $\mathbb{R}^{\sum_i |\mathcal{A}_i|}$ for every $\mathbf{x} \neq \mathbf{x}^* \in \Delta$, and its time-derivative is negative, $\frac{dV}{dt} < 0$, for all $\mathbf{x} \neq \mathbf{x}^*$, where x^* is a stationary point of the dynamics [13]. The Lyapunov function can also defined to be negative definite as in [12]; in this case, the time derivative will need to be positive.

Theorem 14.3 *Consider a Lyapunov game under the learning schemes $\mathcal{L}_1, \mathcal{L}_4$. Then, the learning procedure has convergence to the set of equilibria of the expected robust game for all interior initial conditions.*

Proof: Lyapunov function V provides the stability of the set of rest points. Since the dynamics are positively correlated for adjusted replicator dynamics [21, 27], the state-independent equilibria are rest points of the dynamics obtained from $\mathcal{L}_1$ and $\mathcal{L}_4$. The stability of any convex combination of these dynamics follows. ■

Note that Theorem 14.3 says that starting from interior initial points, the hybrid dynamics lead to one of the equilibria, which we do not know which one in advance. The set can have either a finite number or a continuum of equilibria. This result holds also for $n-$player stochastic games with random updates.

14.4.3.3 Potential games We say that the stochastic game Ξ is an *expected robust potential game* if the expected payoff derives from a potential function. Potential games are 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., $\mathbb{U}_i(e_{a_i}, \mathbf{x}_{-i}) = \frac{\partial\Phi(\mathbf{x})}{\partial x_i(a_i)}, i \in \mathcal{N}, a_i \in \mathcal{A}_i$. We use a Lyapunov approach to show the global convergence of hybrid learning for potential games.

Lemma 14.2 *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.*

Proof: Since the payoff matrix is bounded, we can study its strategically equivalent game [4], [19] by subtracting a certain offset from every matrix entries so that $U_i(\mathbf{a})$ is negative for every strategy pair $\mathbf{a}$, and hence $\mathbb{E}_{s,B^2}\mathbb{U}_i(e_{a_j}, x_{-i,t})$ is negative. Without loss of generality, we can assume the game payoff matrix or its strategically equivalent game payoff matrix is negative entry-wise. Therefore, $V^R = \Phi$ is negative. We take the time derivative of the Lyapunov function V^R as follows:

$$\frac{d}{dt}V^R(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}) = \sum_{i\in\mathcal{N}}\sum_{a_j\in\mathcal{A}_i}\left(\frac{dx_{i,t}(a_j)}{dt}\right)\left(\frac{\partial V^R}{\partial x_{i,t}(a_j)}\right),$$

which leads to the following set of inequalities

$$\begin{aligned}\frac{d}{dt}V^R(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}) = & \sum_{i\in\mathcal{N}} g_{i,t}\left\{\sum_{a_j\in\mathcal{A}_i} x_{i,t}(a_j)\left(\mathbb{E}_{s,B^2}\mathbb{U}_i(e_{a_j}, \mathbf{x}_{-i,t})\right)^2\right. \\ & \left. - \left(\sum_{a_j\in\mathcal{A}_i} x_{i,t}(a_j)\mathbb{E}_{s,B^2}\mathbb{U}_i(e_{a_j}, \mathbf{x}_{-i,t})\right)^2\right\} \\ \geq & \sum_{i\in\mathcal{N}} g_{i,t}\left\{\sum_{a_j\in\mathcal{A}_i} x_{i,t}(a_j)\left(\mathbb{E}_{s,B^2}\mathbb{U}_i(e_{a_j}, \mathbf{x}_{-i,t})\right)^2\right. \\ & \left. - \sum_{a_j\in\mathcal{A}_i} x_{i,t}^2(a_j)\left(\mathbb{E}_{s,B^2}\mathbb{U}_i(e_{a_j}, \mathbf{x}_{-i,t})\right)^2\right\} \geq 0.\end{aligned}$$

The last two inequalities result from Jensen's inequality and the positivity and the range of $\mathbf{x}_{i,t}$. We have $\frac{dV^R}{dt} \geq 0$ with equality only at the equilibrium. Hence, convergence to equilibria holds for all initial conditions in the interior of the simplex■

Lemma 14.3 *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*

$$V^B(\mathbf{x}_1, \mathbf{x}_2) = \Phi(\mathbf{x}_1, \mathbf{x}_2) + \epsilon_1 H_1(\mathbf{x}_1) + \epsilon_2 H_2(\mathbf{x}_2),$$

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.

Proof: Using the same argument as in the proof of Lemma 14.2, we can assume Φ or its strategic equivalent form is positive without loss of generality, and hence V^B is nonnegative. The Lyapunov function V^B has its critical points given by $\nabla_{\mathbf{x}_1} V^B = \nabla_{\mathbf{x}_2} V^B = 0$, i.e.,

$$\nabla_{\mathbf{x}_i} \Phi + \epsilon_i \nabla_{\mathbf{x}_i} H_i = 0, i = 1, 2. \tag{14.10}$$

The first-order condition (14.10) yields the perturbed equilibria of the B-G type of learning schemes. By taking the time derivative of V^B, we arrive at

$$\frac{dV^B}{dt} = \sum_{i \in \mathcal{N}} \sum_{a_j \in A_i} \frac{\partial x_i(a_j)}{\partial t} \frac{\partial \Phi}{\partial x_i(a_j)} + \epsilon_i \frac{\partial x_i(a_j)}{\partial t} \frac{\partial H_i}{\partial x_i(a_j)}.$$

Denote the perturbed payoff function by $\widetilde{\mathbb{U}}_i(\mathbf{x}_1, \mathbf{x}_2) := \mathbb{E}_{s,B^2} \mathbb{U}_i(s, B^2, \mathbf{x}_1, \mathbf{x}_2) + \epsilon_i H_i(\mathbf{x}_i)$. The first-order condition for a maximum satisfies, for every $a_j \in \mathcal{A}_i$,

$$\mathbb{E}_{s,B^2} \mathbb{U}_i(s, B^2, e_{a_j}, \mathbf{x}_{-i}) + \epsilon_i \frac{\partial H_i(\bar{\beta}_i(\mathbf{x}_{-i}))}{\partial x_i(a_j)} = 0 \tag{14.11}$$

where $\bar{\beta}_i(\cdot)$ is a type of B-G strategy that corresponds to the learning type. Since the game is assumed to be a potential game, we have

$$\frac{\partial \Phi}{\partial x_i(a_j)} = -\epsilon_i \frac{\partial H_i(\bar{\beta}_i(\mathbf{x}_{-i}))}{\partial x_i(a_j)}. \tag{14.12}$$

Hence, we obtain from (14.11) and (14.12),

$$\frac{dV^B}{dt} = \sum_{i\in\mathcal{N}}\sum_{a_j\in\mathcal{A}_i}\epsilon_i g_{i,t}\left(\frac{\partial H_i(\bar{\beta}(\mathbf{x}_{-i}))(a_j)}{\partial x_i(a_j)} - \frac{\partial H_i(x_i(a_j))}{\partial x_i(a_j)}\right) \cdot(\bar{\beta}_i\left(\mathbf{x}_{-i}\right)(a_j) - x_i(a_j))\,. \tag{14.13}$$

Due to the strict concavity of the perturbation functions H_i, we conclude that $\frac{dV^B}{dt} \leq 0$, with equality only at the equilibrium. Hence, the pure learning dynamics converge to the set of perturbed equilibria. ■

Theorem 14.4 *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 .*

Proof: The Lyapunov functions for replicator and G-B dynamics share the same term Φ. For hybrid learning between these two dynamics, we can pick Φ as a Lyapunov function. For small ϵ_i close to zero, the Lyapunov function Φ for pure learning $\mathcal{L}_2$ yields a strictly positive time derivative for non-equilibrium points due to continuity. Let $\mathbf{x}_\epsilon$ be a maximizer of $V^R(\mathbf{x}) = \Phi(\mathbf{x}) + \sum_i \epsilon_i H_i(\mathbf{x}_i)$. Then, there exists $\epsilon_h > 0$ such that $\tilde{V}(\mathbf{x}) = \Phi(\mathbf{x}) + \sum_i \epsilon_i' H_i(\mathbf{x}_i)$ is strictly positive in a neighborhood of the considered non-equilibrium point with $\epsilon_i' = \min(\epsilon_h, \frac{\epsilon_i}{M_i})$, where M_i is the maximum of H_i over $\mathcal{X}_i$. Since the maximizer of $\tilde{V}$ is an $\epsilon'-$ equilibrium where $\epsilon' = \max_i \epsilon_i'$, there exists a subsequence of $\mathbf{x}_\epsilon$ converging to $\mathbf{x}^*$ which is an equilibrium and $\mathbf{x}^*$ maximizes Φ and this holds for any convergent subsequence. This means the time derivative of $\tilde{V}$ is strictly positive in all the neighborhood of $\mathbf{x}^*$ and vanishes only at $\mathbf{x}^*$. Thus, when $\epsilon' = \max \epsilon_i'$ goes to zero, one gets an equilibrium. Hence, in view of Lemma 14.3, we can conclude that for sufficiently small $\epsilon_i > 0$, we have local convergence of the hybrid learning. ■

Note that the equilibrium $\mathbf{x}_1^*, \mathbf{x}_2^*$ in Theorem 14.4 may not be unique, which depends on the rest point of the nonlinear hybrid dynamics.

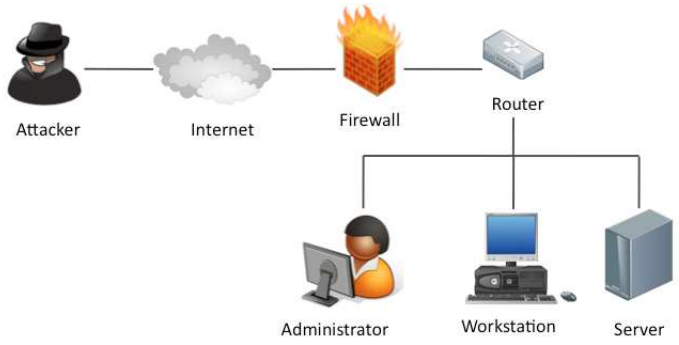

Figure 14.1 An illustration of the network security game scenario where an attacker attempts to breach the network security by compromising the servers and workstations whereas the network administrator monitors the network activity to prevent possible intrusions.

14.5 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. In Figure 14.1, we show a local network connected to the Internet where an attacker attempts to launch an internal denial-of-service attack to bring down a network server capture important data from a workstation. Let P1 and P2 denote the administrator and the intruder, respectively. 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:

$$\mathbf{M}_1 := \begin{bmatrix} & S & W \\ H & 1,-1 & 1,0 \\ L & -2,1 & 2,0 \end{bmatrix}, \quad \mathbf{M}_2 := \begin{bmatrix} H & 1 \\ L & 2 \end{bmatrix}. \tag{14.14}$$

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, and 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 varies in different environmental states 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 14.2 and 14.3, we show the payoffs and mixed strategies of both players when both players use the learning pattern $\mathcal{L}_1$. We can see that the replicator dynamics from $\mathcal{L}_1$ do not converge. However the time average strategies $\lim_{T\to\infty} \frac{1}{T} \int_0^T \mathbf{x}_{i,t} dt$ converge to $\mathbf{x}_1^*, \mathbf{x}_2^*$, respectively, and, the time average payoffs $\lim_{T\to\infty} \frac{1}{T} \int_0^T \hat{\mathbf{u}}_{i,t} dt$ converge to $\hat{\mathbf{u}}_1^*, \hat{\mathbf{u}}_2^*$, respectively.

In Figures 14.4 and 14.5, 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 14.6 and 14.7, we show the convergence of the heterogenous 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 $\bar{\hat{\mathbf{u}}}_1, \bar{\hat{\mathbf{u}}}_2$. We can see that the adoption of $\mathcal{L}_2$ by P2 in the heterogenous learning facilitates the convergence of the algorithm even though the learning exhibits high magnitude of oscillations at the beginning, which is mainly due to $\mathcal{L}_1$ learning pattern adopted by P1.

In Figures 14.8 and 14.9, 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. We can see that the hybrid mixture of $\mathcal{L}_1$ and $\mathcal{L}_2$ learning patterns leads to convergence

with smaller magnitude of oscillations in comparison to the ones shown in Figures 14.6 and 14.7.

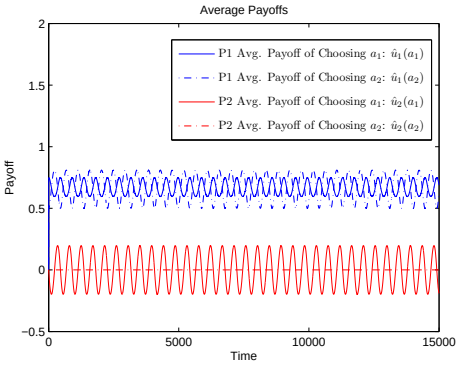

Figure 14.2 The payoffs to the players with both players using $\mathcal{L}_1$.

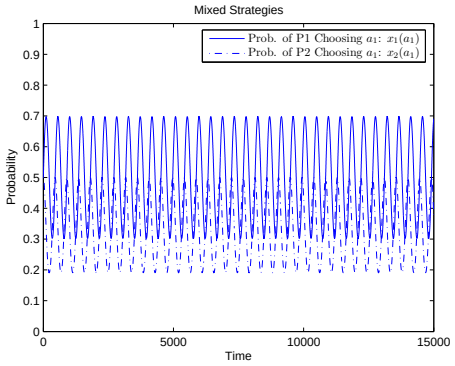

Figure 14.3 The mixed strategies of the players with both players using $\mathcal{L}_1$.

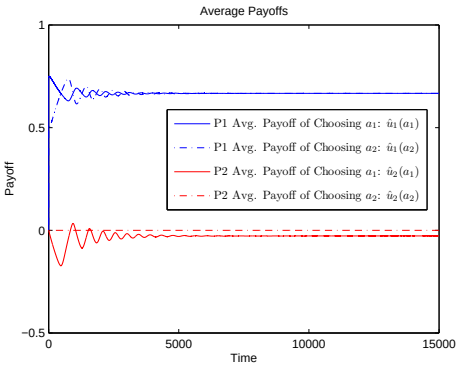

Figure 14.4 The payoffs to the players with both players using $\mathcal{L}_2$.

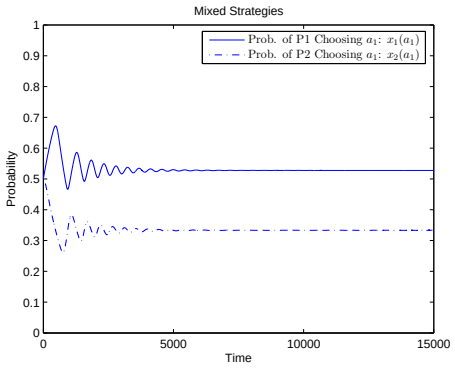

Figure 14.5 The mixed strategies of the players with both players using $\mathcal{L}_2$.

14.6 CONCLUSIONS AND FUTURE WORKS

We have presented distributed hybrid strategic learning algorithms for a class of two-person nonzero-sum stochastic games along with their general convergence and non-

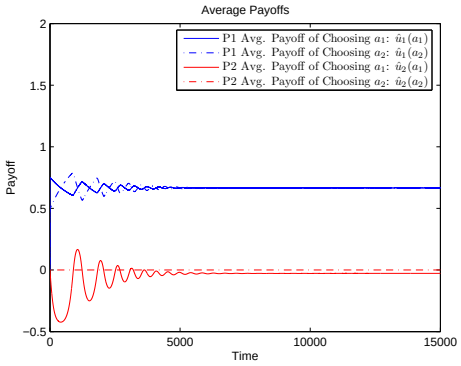

Figure 14.6 The payoffs to the heterogeneous players with P1 using $\mathcal{L}_1$ and P2 using $\mathcal{L}_2$.

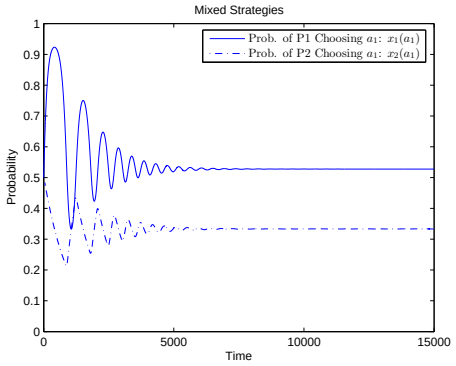

Figure 14.7 The mixed strategies of the heterogeneous players with P1 using $\mathcal{L}_1$ and P2 using $\mathcal{L}_2$.

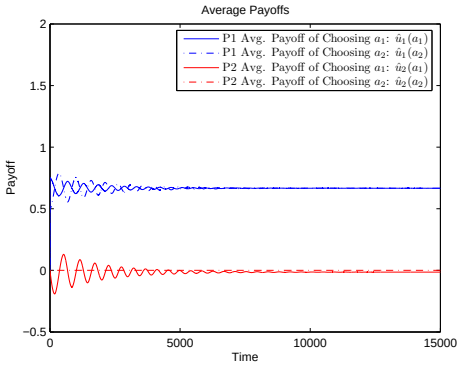

Figure 14.8 The payoffs to the players with both players using hybrid learning scheme with equal weights on $\mathcal{L}_1$ and $\mathcal{L}_2$.

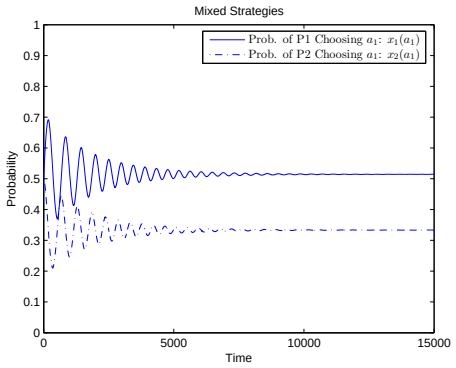

Figure 14.9 The mixed strategies of the players with both players using hybrid learning scheme with equal weights on $\mathcal{L}_1$ and $\mathcal{L}_2$.

convergence properties. The players are assumed to have information limitations in their knowledge not only of other players' payoff functions and strategies but also of their own. In addition, the interactions among the players occur at random times according to their modes. We have applied the framework to security games where the noncooperative behaviors between an attacker and a defender are well

characterized by the features of distributed hybrid learning. Interesting work that we leave for the future is to extend this learning framework to the case of an arbitrary (but still fixed) number of players, each of them adopting hybrid learning with a diffusion term leading to *stochastic differential equations*. It is also of our interest to capture the evolution of the players' rationality through a time-varying ordinary differential equation of the learning weights, which should be analyzed together with the hybrid learning dynamics.

Appendix: Assumptions for Stochastic Approximation

Consider the difference equation $\mathbf{x}_{t+1} = \mathbf{x}_t + \lambda_t(f(\mathbf{x}_t) + M_{t+1})$ in $\mathbb{R}^{\sum_i |\mathcal{A}_i|}$ and assume that

(A1) f is Lipschitz.

(A2) $\lambda_t \geq 0, \sum_{t\geq 0} \lambda_t = +\infty, \sum_{t\geq 0} \lambda_t^2 < \infty$.

(A3) M_{t+1} is a martingale difference sequence with respect to the increasing family of sigma-fields $\mathcal{F}_t = \sigma(\mathbf{x}_{t'}, \hat{\mathbf{u}}_{t'}, M_{t'},\ t' \leq t)$ i.e., $\mathbb{E}\left(M_{t+1} \mid \mathcal{F}_t\right) = 0$.

(A4) M_t is square integrable and there is a constant $c > 0$ such that

$$\mathbb{E}\left(\| \mathbb{M}_{t+1} \|^2 \ \mid \ \mathcal{F}_t\right) \leq c(1+ \| \mathbf{x}_t \|^2)$$

almost surely, for all $t \geq 0$.

(A5) $\sup_t \ \| \mathbf{x}_t \| < \infty$ almost surely.

Then, the asymptotic pseudo-trajectory of the difference equation is given by the ordinary differential equation (ODE) [7, 14], $\dot{\mathbf{x}}_t = f(\mathbf{x}_t)$, with $\mathbf{x}_0$ fixed.

REFERENCES

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

2. E. Altman, T. Boulogne, R. El-Azouzi, T. Jiménez, and L. Wynter. A survey on networking games in telecommunications. *Comput. Oper. Res.*, 33(2):286–311, 2006.

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

4. T. Başar and G. J. Olsder. Dynamic noncooperative game theory. *SIAM Series in Classics in Applied Mathematics, Philadelphia*, January 1999.

5. 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.

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

7. V. S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Cambridge University Press, 2008.

8. R. Bush and F. Mosteller. *Stochastic Models of Learning*. John Wiley and Sons, 1955.

9. 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.

10. D. Fudenberg and D. Levine. *Learning in Games*. MIT Press, Cambridge, MA, 1998.

11. J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, New York, 1983.

12. F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *The Journal of the Operational Research Society*, 49(3):237–252, 1998.

13. H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 3rd edition, 2001.

14. H. J. Kushner and D. S. Clark. Stochastic approximation methods for constrained and unconstrained systems. *Springer, New York*, 1978.

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

16. 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.

17. M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux. Game theory meets network security and privacy. *EPFL, Technical Report, No. EPFL-REPORT-151965*, 2010.

18. 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.

19. O. Morgenstern and J. Von Neumann. *Theory of Games and Economic Behavior*. Princeton University Press, 3rd edition, 1980.

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

21. W. H. Sandholm. Population games and evolutionary dynamics. *MIT Press*, 2010.

22. 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.

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

24. P. D. Taylor and L. B. Jonker. Evolutionarily stable strategies and game dynamics. *Mathematical Bioscience*, 40:145–156, 1978.

25. M. A. L. Thathachar, P.S. Sastry, and V.V. Phansalkar. Decentralized learning of Nash equilibria in multiperson stochastic games with incomplete information. *IEEE Transactions on System, Man, and Cybernetics*, 24(5):769–777, 1994.

26. C.J.C.H. Watkins. *Learning from Delayed Rewards*. PhD thesis, 1989.

27. J. Weibull. Evolutionary game theory. *MIT Press*, 1995.

28. Y. Xing and R. Chandramouli. Stochastic learning solution for distributed discrete power control game in wireless data networks. *IEEE/ACM Transactions on Networking*, 16(4):932–944, August 2008.

29. H. P. Young. Learning by trial and error. *Games and Economic Behavior*, 65(2):626–643, March 2009.

30. Q. Zhu and T. Başar. Dynamic policy-based IDS configuration. *Proceedings of the 48th IEEE Conf. on Decision and Control (CDC/CCC) 2009*, pages 8600 –8605, Dec. 2009.

31. 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*, pages 219–224, 2010.

32. Q. Zhu, H. Tembine, and T. Başar. Distributed strategic learning with application network security. *in Proceedings of American Control Conference (ACC)*, pages 4057–4062, 2011.