

# The Confluence of Networks, Games and Learning\*

## A game-theoretic framework for multi-agent decision making over networks

Tao Li<sup>†‡</sup>, Guanze Peng<sup>‡</sup>, Quanyan Zhu<sup>‡</sup>, Tamer Başar<sup>§</sup>

### Abstract

Recent years have witnessed significant advances in technologies and services in modern network applications, including smart grid management, wireless communication, cybersecurity as well as multi-agent autonomous systems. Considering the heterogeneous nature of networked entities, emerging network applications call for game-theoretic models and learning-based approaches in order to create distributed network intelligence that responds to uncertainties and disruptions in a dynamic or an adversarial environment. This paper articulates the confluence of networks, games and learning, which establishes a theoretical underpinning for understanding multi-agent decision-making over networks. We provide an selective overview of game-theoretic learning algorithms within the framework of stochastic approximation theory, and associated applications in some representative contexts of modern network systems, such as the next generation wireless communication networks, the smart grid and distributed machine learning. In addition to existing research works on game-theoretic learning over networks, we highlight several new angles and research endeavors on learning in games that are related to recent developments in artificial intelligence. Some of the new angles extrapolate from our own research interests. The overall objective of the paper is to provide the reader a clear picture of the strengths and challenges of adopting game-theoretic learning methods within the context of network systems, and further to identify fruitful future research directions on both theoretical and applied studies.

## 1 Introduction

Multi-agent decision making over networks has recently attracted an exponentially growing number of researchers from the systems and control community. The area has gained increasing momentum in various fields including engineering, social sciences, economics, urban

---

\*Prepared for IEEE control system magazine, as part of the special issue “Distributed Nash Equilibrium Seeking over Networks”.

<sup>†</sup>Corresponding author

<sup>‡</sup>Department of Electrical and Computer Engineering, New York University, NY, USA; Email: {t12636, gp1363, qz494}@nyu.edu.

<sup>§</sup>Department of Electrical and Computer Engineering & Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, IL, USA; Email: {basar1}@illinois.edu.

science, and artificial intelligence, as it serves as a prevalent framework for studying large and complex systems, and has been widely applied in tackling many problems arising in these fields, such as social networks analysis [1], smart grid management [2, 3], traffic control [4], wireless and communication networks [5–7], cybersecurity [8,9], as well as multi-agent autonomous systems [10].

Due to the proliferation of advanced technologies and services in modern network applications, solving the decision-making problems in multi-agent networks calls for novel models and approaches that can capture the following characteristics of emerging network systems and the design of autonomous controls:

1. the heterogeneous nature of the underlying network, where multiple entities, represented by the set of nodes, aim to pursue their own goals with independent decision-making capabilities;
2. the need for distributed or decentralized operation of the system, when the underlying network is of a complex topological structure and is too large to be managed in a centralized approach;
3. the need for creating network intelligence that is responsive to changes in the network and the environment, as the system oftentimes operates in a dynamic or an adversarial environment.

Game theory provides a natural set of tools and frameworks addressing these challenges, and bridging networks to decision making. It entails development of mathematical models that both qualitatively and quantitatively depict how the interactions of self-interested agents with different information and rationalities can attain a global objective or lead to emerging behaviors at a system level. Moreover, with the underlying network, game-theoretic models capture the impact of the topology on the process of distributed decision making, where agents plan their moves independently according to their goals and local information available to them, such as their observations of their neighbors.

In addition to game-theoretic models over networks, learning theory is indispensable when designing decentralized management mechanisms for network systems, in order to equip networks with distributed intelligence. Through the combination of game-theoretic models and associated learning schemes, such network intelligence allows heterogeneous agents to interact strategically with each other and learn to respond to uncertainties, anomalies, and disruptions, leading to desired collective behavior patterns over the network or an optimal system-level performance. The key feature of such network intelligence is that even though each agent’s own decision-making process is influenced by the others’ decisions, the agents reach an equilibrium state, that is, a Nash equilibrium as we elucidate later, in an online and decentralized manner. To equip networks with distributed intelligence, networked agents should adapt themselves to the dynamic environment with limited and local observations over a large network that may be unknown to them. Computationally, decentralized learning scales efficiently to large and complex networks, and requires no global information regarding the entire network, which is more practical compared with centralized control laws.

This paper articulates the confluence of networks, games and learning, which establishes a theoretical underpinning for understanding multi-agent decision-making over networks.

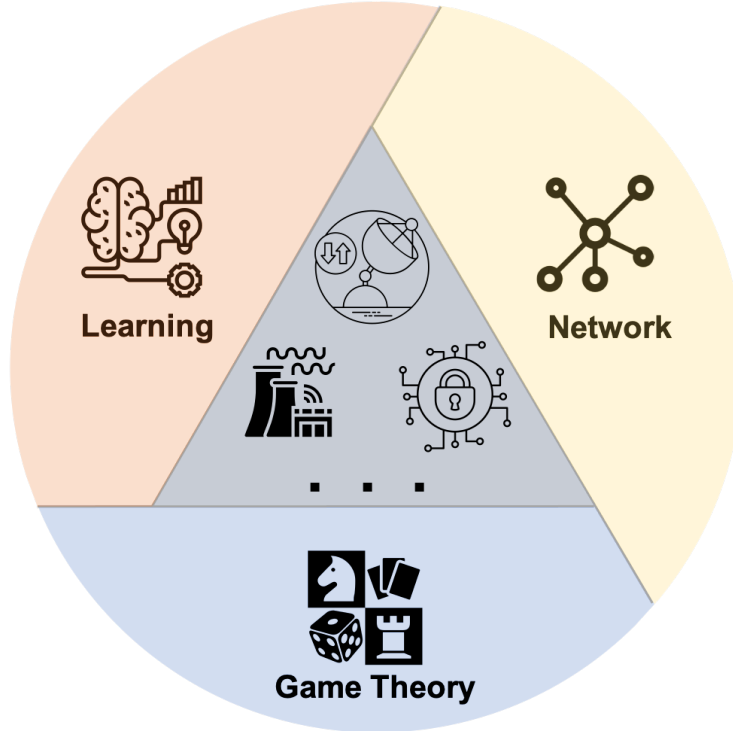


Figure 1: The confluence of networks, games and learning. The combination of game-theoretic modelling and learning theories leads to resilient and agile network controls for various networked systems.

We aim to provide a systematic treatment of game-theoretic learning methods and their applications in network problems, which meet the three requirements specified above. As shown in Figure 1, emerging network applications call for novel approaches, and thanks to the decentralized nature, game-theoretic models as well as associated learning methods provide an elegant approach for tackling network problems arising from various fields. Specifically, our objectives are threefold:

1. to provide a high-level introduction to game-theoretic models that apply to multi-agent decision making problems;
2. to present the key analytical tool based on stochastic approximation and Lyapunov theory for studying learning processes in games, and pinpoint some extensively studied learning dynamics;
3. to introduce various multi-agent systems and network applications that can be addressed through game-theoretic learning.

We aim to provide the reader a clear picture of the strengths and challenges of adopting novel game-theoretic learning methods within the context of network systems. Besides the highlighted contents, we also provide the reader with references for further reading. In this paper, complete-information games are the basis of the subject, for which we give a brief introduction to both static and dynamics games. More comprehensive treatments on this

Symbol	Meaning
$\mathcal{N}$	The set of players
$i, j \in \mathcal{N}$	Subscript index denoting players
$\mathcal{N}(i)$	The set of neighbors of player $i$
$\mathcal{A}_i$	The set of actions available to player $i$
$\Delta(\mathcal{A}_i)$	The set of Borel probability measures (The probability simplex in $\mathbb{R}^{\mathcal{A}_i}$ for finite action set $\mathcal{A}_i$ )
$s \in \mathcal{S}$	State variable
$u_i : \prod_{j \in \mathcal{N}} \mathcal{A}_j \rightarrow \mathbb{R}$	Player $i$ 's utility function
$a_i \in \mathcal{A}_i$	Action of player $i$
$a_{-i} \in \prod_{j \in \mathcal{N}, j \neq i} \mathcal{A}_j$	Joint actions of players other than $i$
$\mathbf{a} \in \prod_{i \in \mathcal{N}} \mathcal{A}_i$	Joint actions of all players
$\pi_i \in \Delta(\mathcal{A}_i)$	Strategy of player $i$
$\pi_{-i} \in \prod_{j \in \mathcal{N}, j \neq i} \Delta(\mathcal{A}_j)$	Joint strategy of players other than $i$
$\mathbf{u}_i(\pi_{-i})$ or $\mathbf{u}_i \in \mathbb{R}^{ \mathcal{A}_i }$	Player $i$ 's utility vector in finite games
$D_i(\mathbf{a})$	Individual payoff gradient of player $i$
$D(\mathbf{a})$	The concatenation of $\{D_i(\mathbf{a})\}_{i \in \mathcal{N}}$
$I_i^k$	Feedback of player $i$ at time $k$
$U_i^k \in \mathbb{R}$	The payoff feedback received by player $i$ at time $k$
$\hat{\mathbf{u}}_i^k \in \mathbb{R}^{ \mathcal{A}_i }$	Estimated utility vector at time $k$
$\hat{\mathbf{U}}_i^k \in \mathbb{R}^{ \mathcal{A}_i }$	Estimator of $\mathbf{u}_i(\pi_{-i}^k)$ at time $k$
$BR_i$	Best response mapping for player $i$
$QR^\epsilon$	Regularized best response or quantal response

Table 1: Table of Notations

topic as well as other game models, such as incomplete information games, can be found in [11–13]. As most of the network topologies can be characterized by the structure of the utility function of the game [1, 14], we do not articulate the influence of network topologies on the game itself. Instead, we focus on its influence on the learning process in games, where players' information feedback depends on the network structures, and we present representative network applications to showcase this influence. We refer the reader to [1, 14] for further reading on games over various networks.

We structure our discussions as follows. In Section 2, we introduce non-cooperative games and associated solution concepts, including Nash equilibrium and its variants, which capture the strategic interactions of self-interested players. Then, in Section 3, we move to the main focus of this paper: learning dynamics in games that converge to Nash equilibrium. Within the stochastic approximation framework, a unified description of various dynamics is provided, and the analytical properties can be studied by ordinary differential equation (ODE) methods. In Section 4, we discuss applications of these learning algorithms in networks, leading to distributed and learning-based controls for network systems. Finally, Section 5 concludes the paper. For the reader's convenience, we summarize the notations that are frequently used in Table 1.

## 2 Noncooperative Game Theory

Game theory constitutes a mathematical framework with two main branches: noncooperative game theory and cooperative game theory. Noncooperative game theory focuses on the strategic decision-making process of independent entities or players that aim to optimize their distinct objective functions, without any external enforcement of cooperative behaviors. The term noncooperative does not necessarily mean that players are not engaged in cooperative behaviors. As a matter of fact, induced cooperative or coordinated behaviors do arise in noncooperative circumstances, within the context of Nash equilibrium, a solution concept of noncooperative games. However, such coordination is self-enforcing and arises from decentralized decision-making processes of self-interested players, and will be further discussed in Section 4, where we introduce game-theoretic methods for distributed machine learning.

As briefly discussed above, noncooperative game theory naturally characterizes the decision-making process of heterogeneous entities acting independently over networks, which is the main focus of this paper. In the following, we introduce various game models and related solution concepts, including Nash equilibrium and its variants. Generally speaking, a game involves the following elements: decision makers (players); choices available to each player (actions); knowledge that a player acquires for making decisions (information) and each player's preference ordering among its actions, affected by also others' actions (utilities or cost). Below we provide a short list of these concepts that will be further discussed and explained in this section.

1. *Players* are participants in a game, where they compete for their own good. A player can be an individual or encapsulation of a set of individuals.
2. *Actions* of a player, in the terminology of control theory, are the implementations of the player's control.
3. *Information* in games refers to the structure regarding the knowledge players acquire about the game and its history when they decide on their moves. The information structure can vary considerably. For some games, the information is *static* and does not change during the play. While for other games, new information will be revealed after players' moves, as the "state" of the game, a concept to be elucidated later, is determined by players' actions during the play. In the latter case, the information is *dynamic*. We shall address both types of games in this paper.
4. A *strategy* is a mapping that associates a player's move with the information available to him at the time when he decides which move to choose.
5. A *utility or payoff* is oftentimes a real-valued function capturing a player's preference ordering among possible outcomes of the game. Using the terminology in control theory, this can also be viewed as a cost function for the player's controls.

The above list refers to elements of games in relatively imprecise common language terms, and more formal definitions are presented below. To facilitate this discussion, we categorize noncooperative games into two main classes: static and dynamic games, based on the nature of the information structure.

## 2.1 Static Games

Static games are one-shot, where players make decisions simultaneously based on the prior information on the games, such as sets of players' actions, and their payoffs. In such games, each player's knowledge about the game is static and does not evolve during the play. Mathematically speaking, a static noncooperative game is defined as follows.

**Definition 1 (Static Games)** *A static game is defined by a triple  $G := \langle \mathcal{N}, (\mathcal{A}_i)_{i \in \mathcal{N}}, (u_i)_{i \in \mathcal{N}} \rangle$ , where*

1.  $\mathcal{N} = \{1, 2, \dots, N\}$  is a finite set of players;
2.  $\mathcal{A}_i$  with some specified topology denotes the set of actions available to the player  $i \in \mathcal{N}$ ;
3.  $u_i : \prod_{j \in \mathcal{N}} \mathcal{A}_j \rightarrow \mathbb{R}$  defines player  $i$ 's utility, and  $u_i(a_i, a_{-i})$  gives the payoff of player  $i$  when taking action  $a_i$ , given other players' actions  $a_{-i} := (a_j)_{j \in \mathcal{N}, j \neq i}$ .

In static games, each player develops its strategy, a probability distribution over his action set, with the objective of maximizing the expected value of its own utility. If players have finite action sets, then such a static game is called a finite game. In this case, a strategy is a finite-dimensional vector in the probability simplex over the action set, that is,  $\pi_i \in \Delta(\mathcal{A}_i) := \{\pi \in \mathbb{R}^{|\mathcal{A}_i|} | \pi(a) \geq 0, \forall a \in \mathcal{A}_i, \sum_{a \in \mathcal{A}_i} \pi(a) = 1\}$ . If  $\pi_i$  is a unit vector  $e_a$ ,  $a \in \mathcal{A}_i$  with the  $a$ -th entry being 1 and 0 for others, then it is referred to a pure strategy, selecting action  $a$  with probability 1; otherwise, it is a mixed strategy, choosing actions randomly under the selected probability distribution. Similarly, for infinite action sets, the strategy is defined as a Borel probability measure over the action set, with Dirac measure being the pure strategy. By a possible abuse of notation, we denote the set of Borel probability measures over  $\mathcal{A}_i$  by  $\Delta(\mathcal{A}_i)$ . Unless specified otherwise, static games considered in this paper are all assumed to be finite, where the player set, and the action sets are all finite.

As a special case of games with infinite actions, the mixed extension of finite games is introduced in the sequel. Consider a two-player finite game  $G = \langle \mathcal{N}, (\mathcal{A}_i)_{i \in \mathcal{N}}, (u_i)_{i \in \mathcal{N}} \rangle$ , where  $\mathcal{N} = \{1, 2\}$ , and the action sets are finite  $|\mathcal{A}_i| < \infty, i \in \mathcal{N}$ . Given the mixed strategies of players,  $\pi_i \in \Delta(\mathcal{A}_i)$ , the expected utility of player  $i$  is  $\mathbb{E}_{a_1 \sim \pi_1, a_2 \sim \pi_2}[u_i(a_1, a_2)]$ . With a slight abuse of notation, we denote this expected utility by  $u_i(\pi_1, \pi_2) := \mathbb{E}_{a_1 \sim \pi_1, a_2 \sim \pi_2}[u_i(a_1, a_2)]$ . Then, studying the players' strategic interactions is equivalent to considering the following infinite game  $G^\infty = \langle \mathcal{N}, (\Delta(\mathcal{A}_i))_{i \in \mathcal{N}}, (u_i)_{i \in \mathcal{N}} \rangle$ , where  $u_i$  denotes the expected utility. In  $G^\infty$ , an action is a vector from the corresponding probability simplex, a convex and compact set with a continuum of elements. Similar to the notations used in the definition, for the mixed extension  $G^\infty$ , we denote the joint action of players other than  $i$  by  $\pi_{-i} := (\pi_j)_{j \in \mathcal{N}, j \neq i}$ . Furthermore, we let  $\mathbf{u}_i(\pi_{-i}) \in \mathbb{R}^{|\mathcal{A}_i|}$  be the utility vector of player  $i$ , given other players' strategy profiles,  $\pi_{-i}$ , whose  $a$ -th entry is defined as  $\mathbf{u}_i(\pi_{-i})(a) := u_i(e_a, \pi_{-i})$ . Due to the definition of expectation,  $u_i(\pi_i, \pi_{-i})$  can be expressed as an inner product  $\langle \pi_i, \mathbf{u}_i(\pi_{-i}) \rangle$ , which will be frequently used later when discussing learning algorithms in finite games. This mixed extension allows us to give a geometric characterization to Nash equilibria of finite games, based on variational inequalities, as discussed in Section 2.3. Meanwhile, this inner product expression connects learning theory in finite games with online linear optimization [15], where

the generic player's decision variable is  $\pi_i$  and the loss function specified by  $\langle \cdot, \mathbf{u}_i(\pi_{-i}) \rangle$  is linear in  $\pi_i$ .

Even though widely applied in modeling behaviors of self-interested players, the static game model is far from being sufficient to cover multi-agent decision making problems arising in different fields. For instance, when playing poker games, new information will be revealed during the game play, such as cards played at each round, based on which players can adjust their moves. There are many games where players' information about the game changes over time during the play, which cannot be suitably described by static games. Therefore, we must resort to another model for capturing the underlying dynamics.

## 2.2 Dynamic Games

To explicitly represent the dynamic nature of the decision-making process, we adopt system theory terminology and use the state of the game to describe its evolution over a period of time, which could be finite or infinite. Roughly speaking, the current state specifies the current situation of the dynamic game, including the set of players who are about to take actions, actions available to them and their utilities at this time. The fundamental difference between static games and dynamic games is that for the latter the game changes over time as players implement their sequences of actions during the play. Hence, players' knowledge regarding the game also evolves, as players can fully or partially observe the current state.

In the following, a subclass of Markov games is introduced as an example of dynamic games, which is a very popular game model for studies on multi-agent sequential decision making under uncertainties, such as multi-agent reinforcement learning [16].

**Definition 2 (Markov Games)** *An  $N$ -person discrete-time infinite horizon discounted Markov game consists of*

1. a player set  $\mathcal{N} = \{1, 2, \dots, N\}$ ;
2. a discrete time set  $\mathbb{N}_+ := \{1, 2, \dots\}$ , with actions by players taken at each  $k \in \mathbb{N}_+$ ;
3. a set  $\mathcal{A}_i$  with some specified topology, defined for each  $i \in \mathcal{N}$ , corresponding to the set of actions or controls available to player  $i$ ;
4. a set  $\mathcal{S}$  with some specified topology, denoting the state space of the game, where  $s^k \in \mathcal{S}, k \in \mathbb{N}_+$  represent the state of the game at time  $k$ ;
5. a transition kernel  $T : \mathcal{S} \times \prod_{i \in \mathcal{N}} \mathcal{A}_i \rightarrow \Delta(\mathcal{S})$ , according to which the next state is sampled, that is,  $s^{k+1} \sim T(s^k, \mathbf{a}^k)$ , where  $\mathbf{a}^k = (a_1^k, \dots, a_N^k)$  is the  $N$ -tuple of actions at time  $k \in \mathbb{N}_+$ , and  $s^1 \in \mathcal{S}$  has a given distribution;
6. an instantaneous payoff:  $u_i : \mathcal{S} \times \prod_i \mathcal{A}_i \rightarrow \mathbb{R}$ , defined for each  $i \in \mathcal{N}$  and  $k \in \mathbb{N}_+$ , determining the payoff  $u_i(s^k, \mathbf{a}^k)$  received by player  $i$  at time  $k$ ;
7. a discounting factor  $\gamma$ . Given  $\{s^1, \dots, s^k, \dots; \mathbf{a}^1, \dots, \mathbf{a}^k, \dots\}$ , the discounted cumulative payoffs for player  $i$  is  $\sum_{k=1}^{\infty} \gamma^k u_i(s^k, \mathbf{a}^k)$ .

The above definition only characterizes one special case of dynamic games. Based on this definition, we can derive many other game models. For example, we can make state transitions independent of players' actions as well as the current state, yielding a special case of stochastic games, which will be further discussed in another paper in this special issue [17]. We can also consider continuous-time dynamic games where the transition is described by a differential equation, leading to a differential game model. For an extensive coverage of dynamic game models, we refer the reader to [11].

With the full observation of states, we can consider the stationary strategy  $\pi_i : \mathcal{S} \rightarrow \Delta(\mathcal{A}_i)$ , by which players plan their moves only based on the current state  $s \in \mathcal{S}$ . In this case, we say the state variable  $s$  characterizes players' knowledge of the game, since the actions, utilities and next possible states are all determined by the current state. For dynamic games under partial observation and/or non-Markovian transition, we refer the reader to [11], since these topics are beyond the scope of this paper.

## 2.3 Solution Concepts

The solution or outcome of any given game is more or less a matter of understanding game rules and relations between players. However, besides these concrete matters, there exist general principles, which dictate players' behaviors and apply to all games. Here, we argue that these principles revolve around the notion of rationality, based on which we introduce the solution concept of Nash equilibrium and some of its variants. Mathematically speaking, a solution to an  $N$ -person game is a collection of all players' strategies, that has attractive properties expressed in terms of payoffs received by the players. In addition, players can admit different strategies depending on how the game is defined and, in particular, on the information that players acquire. We start with static games, where the information structure is relatively simple.

Compared with single-agent optimization problems, the analysis of games is more involved, as each player's utility is determined not only by its own decision but also by others' moves. Hence, when a player takes an action, it must take into account possible moves of the other players, which leads to the notion of best response. To introduce "best response", for clarity, but without any loss of conceptual generality, let us focus on games with two players. For player 1, given the other player's strategy  $\pi_2$ , the optimal choice is

$$\pi_1 \in BR_1(\pi_2) := \arg \max_{\pi \in \Delta(\mathcal{A}_1)} \{\langle \pi, \mathbf{u}_1(\pi_2) \rangle\}, \quad (1)$$

which is referred to as a best response of player 1 to player 2's strategy  $\pi_2$ , and  $BR_1(\cdot)$  is called the best response set of player 1. Similarly, given player 1's strategy  $\pi_1$ , a best response of player 2 is  $\pi_2 \in BR_2(\pi_1) := \arg \max_{\pi \in \Delta(\mathcal{A}_2)} \{\langle \pi, \mathbf{u}_2(\pi_1) \rangle\}$ . Therefore, we can define a point-to-set mapping  $BR : \Delta(\mathcal{A}_1) \times \Delta(\mathcal{A}_2) \rightarrow 2^{\Delta(\mathcal{A}_1) \times \Delta(\mathcal{A}_2)}$ , which is the concatenation of  $BR_1$  and  $BR_2$ . Given a joint strategy profile  $\pi = (\pi_1, \pi_2)$ ,  $BR(\pi)$  is defined as

$$BR(\pi) := \{(\pi'_1, \pi'_2) | \pi'_1 \in BR_1(\pi_2), \pi'_2 \in BR_2(\pi_1)\}. \quad (2)$$

If we can find  $\pi^* = (\pi_1^*, \pi_2^*)$ , a fixed point of this best-response mapping, that is,  $\pi^* \in BR(\pi^*)$ , then when both players adopt the corresponding strategy in this profile, they could do no



better by unilaterally deviating from current strategy. In other words, this fixed point corresponds to an equilibrium outcome of the game, which further leads to the definition of Nash equilibrium, which we introduce below for the general  $N$ -player game.

**Definition 3 (Nash Equilibrium)** *For a static game  $\langle \mathcal{N}, (\mathcal{A}_i)_{i \in \mathcal{N}}, (u_i)_{i \in \mathcal{N}} \rangle$ , Nash equilibrium is a strategy profile  $\pi^* = (\pi_i^*, \pi_{-i}^*)$  with the property that for all  $i \in \mathcal{N}$ ,*

$$u_i(\pi_i^*, \pi_{-i}^*) \geq u_i(\pi_i, \pi_{-i}^*), \quad (3)$$

where  $\pi_i$  is an arbitrary strategy of player  $i$  and  $\pi_{-i}^* = (\pi_j^*)_{j \in \mathcal{N}, j \neq i}$  denotes the joint strategy profile of the other players. If the inequality holds strictly for all  $\pi_i \neq \pi_i^*$ , then it is referred to as a strict Nash equilibrium.

Note that the preceding definition naturally carries over to games with infinite action sets, and we refer the reader to [11, Chapter 4] for more details. Furthermore, for infinite games, if we impose some topological structures on the action sets and regularity conditions on the utility functions, we can come up with a geometric interpretation of Nash equilibrium derived from the inequality in (3). Toward that end, we consider a (static) game with compact and convex action sets  $(\mathcal{A}_i)_{i \in \mathcal{N}}$  and smooth concave utilities:

$$u_i(a_i, a_{-i}) \text{ is concave in } a_i \text{ for all } a_{-i} \in \prod_{j \in \mathcal{N}, j \neq i} \mathcal{A}_j, i \in \mathcal{N}.$$

In such a game, the number of actions to each player is a continuum, and the utility function is continuous; such games is referred to as continuous-kernel games or continuous games. In this case, a pure strategy Nash equilibrium  $\mathbf{a}^* = (a_i^*, a_{-i}^*) \in \prod_{i \in \mathcal{N}} \mathcal{A}_i$  is defined by the following inequality,

$$u_i(a_i^*, a_{-i}^*) \geq u_i(a_i, a_{-i}^*), \quad \text{for all } a_i \in \mathcal{A}_i \text{ and all } i \in \mathcal{N}. \quad (4)$$

Further assuming that  $u_i(a_i, a_{-i})$  is continuously differentiable in  $a_i \in \mathcal{A}_i$ , for all  $a_{-i}$ , by the first order condition, Nash equilibrium in (4) can be characterized by

$$\langle D_i(\mathbf{a}^*), a_i - a_i^* \rangle \leq 0, \quad \text{for all } a_i \in \mathcal{A}_i, i \in \mathcal{N},$$

where  $D_i(\mathbf{a}) := \nabla_{a_i} u_i(a_i, a_{-i})$  denotes the individual payoff gradient of player  $i$ , and  $\nabla_{a_i} u_i(a_i, a_{-i})$  denotes differentiation with respect to the variable  $a_i$ . By rewriting the inequality above in a more compact form, we obtain the following variational characterization of Nash equilibrium

$$\langle D(\mathbf{a}^*), \mathbf{a} - \mathbf{a}^* \rangle \leq 0, \quad \text{for all } \mathbf{a} \in \prod_{i \in \mathcal{N}} \mathcal{A}_i, \quad (5)$$

where  $D(\mathbf{a})$  is the concatenation of  $\{D_i(\mathbf{a})\}_{i \in \mathcal{N}}$ , that is,  $D(\mathbf{a}) = (D_1(\mathbf{a}), \dots, D_N(\mathbf{a}))$ . Geometrically speaking, (5) states that for concave games,  $\mathbf{a}^*$  is a Nash equilibrium if and only if  $D(\mathbf{a}^*)$  lies within the polar cone of the set  $\prod_{i \in \mathcal{N}} \mathcal{A}_i - \mathbf{a}^* := \{\mathbf{a} - \mathbf{a}^* | \mathbf{a} \in \prod_{i \in \mathcal{N}} \mathcal{A}_i\}$ , as shown in Fig 2.

In addition to concave games, such variational inequality characterization has been studied in much broader contexts, such as monotone games [18], which bridges the gap between

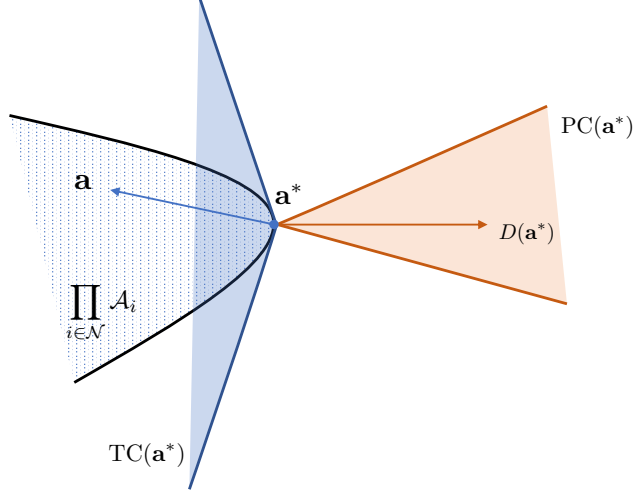


Figure 2: Variational characterization of a Nash equilibrium  $\mathbf{a}^*$  in concave games.  $\text{TC}(\mathbf{a}^*)$  and  $\text{PC}(\mathbf{a}^*)$  denote, respectively, the tangent and the polar cone of  $\prod_{i \in \mathcal{N}} \mathcal{A}_i - \mathbf{a}^*$ . According to the variational inequality (5),  $\mathbf{a}^*$  is a Nash equilibrium if and only if  $D(\mathbf{a}^*)$  lies in the polar cone.

the theory of monotone operators and Nash equilibrium seeking. For a detailed discussion, we refer the reader to another paper in this special issue [19]. The variational inequality (5) is referred to as the Stampacchia-type inequality in the literature [20], and a similar variational inequality of this type can also be derived in the context of the mixed extension. As a special case of continuous games, the mixed extension of finite games also satisfies the regularity conditions: the action spaces are probability simplex regions, which are compact and convex, and the utility function, due to its linearity with respect to any player's mixed strategy, is naturally smooth and concave. Therefore, the mixed strategy Nash equilibrium can be characterized by variational inequality as well. Thanks to the inner product expression of the utility in the mixed extension, the individual payoff gradient is simply  $\mathbf{u}_i(\pi_{-i})$ , and we denote the concatenation of  $\{\mathbf{u}_i\}_{i \in \mathcal{N}}$  by  $\mathbf{u}(\pi) := [\mathbf{u}_1(\pi_{-1}(t)), \mathbf{u}_2(\pi_{-2}(t)), \dots, \mathbf{u}_N(\pi_{-N}(t))]$ , which we also refer to as the joint utility vector under the strategy profile  $\pi$ . In the same spirit of (5), a strategy profile  $\pi^*$  is Nash equilibrium of the underlying finite game if and only if the following Stampacchia-type inequality holds

$$\langle \mathbf{u}(\pi^*), \pi - \pi^* \rangle \leq 0, \quad \text{for all } \pi \in \prod_{i \in \mathcal{N}} \Delta(\mathcal{A}_i). \quad (\text{SVI})$$

As we will later see in Section 3.4.2, this variational characterization of Nash equilibrium bridges the equilibrium concept of games and the equilibrium concept of dynamical systems induced by learning algorithms.

In the same spirit of (3), Nash equilibrium in dynamic games can also be defined accordingly. For Markov games, given players' stationary strategy profile  $\pi$ , the cumulative expected utility of player  $i$ , starting from the initial state  $s^1 = s$ , is

$$V_i^\pi(s) := \mathbb{E}_{s^{k+1} \sim T, \mathbf{a}^k \sim \pi} \left[ \sum_{k=1}^{\infty} \gamma^k u_i(s^k, \mathbf{a}^k) \mid s^1 = s \right], \quad (6)$$

which is referred to as state-value function in Markov decision process [21]. If we view  $V_i^\pi$  as a function of the strategy profile, following (3), we can define Nash equilibrium for the Markov game, where the inequality holds for every state. In other words, regardless of previous play, as long as players follow  $\pi^*$  from the current state  $s$ , they achieve the best outcome for the rest of the game, and no player has any incentive to deviate from the strategy dictated by  $\pi^*$ . Hence, this kind of Nash equilibrium is referred to as subgame perfect Nash equilibrium (SPNE), which is widely used in the study of dynamic games [22, 23].

The Nash equilibrium serves as a building block for noncooperative games. One of its major advantages is that it characterizes a stable state of a noncooperative game, in which no rational player has the incentive to move unilaterally. This stability idea will be further discussed when we focus on learning in games, which relates stability theory of differential equations to the convergence of learning algorithms in Nash equilibrium seeking.

### 3 Learning in Games

Learning in games refers to a long-run non-equilibrium process of learning, adaptation, and/or imitation that leads to some equilibrium [24]. Different from pure equilibrium analysis based on the definition, learning in games accounts for how players behave adaptively during repeated game play under uncertainties and partial observations. Computationally speaking, computing NE based on equilibrium analysis is challenging due to the computational complexity [25], and this hardly accounts for the decision-making process in practice, where players have limited computation power and information. Hence, learning models are needed to describe how less than fully rational players behave in order to reach the equilibrium. Equilibrium seeking or computation motivates learning in games [23].

If we view the learning process as a dynamical system, then the learning model can predict how each player adjusts its behavior in response to other players over time to search for strategies that will lead to higher payoffs. From this perspective, a Nash equilibrium can also be interpreted as the steady state of the learning process, which serves as a prediction of the limiting behavior of the dynamical system induced by the learning model. This viewpoint has been widely adopted in the study of population biology and evolutionary game theory, as we shall see more clearly when we discuss later reinforcement learning and replicator dynamics [26].

In this section, various learning dynamics are presented in the context of infinitely repeated games for Nash equilibrium seeking. We consider a number of players repeatedly playing the game  $\langle \mathcal{N}, (\mathcal{A}_i)_{i \in \mathcal{N}}, (u_i)_{i \in \mathcal{N}} \rangle$  infinitely many times. At time  $k$ , players determine their moves based on their observations up to time  $k - 1$ . Then, they receive feedback from the environment, which provides information on the past actions. For example, in finite games, based on the information available to it, player  $i$  constructs a mixed strategy  $\pi_i^k \in \Delta(\mathcal{A}_i)$ , from which it samples an action  $a_i^k$  and implements it. Then it will receive a payoff feedback related to  $u_i(a_i^k, a_{-i}^k)$ , which evaluates the performance of  $a_i^k$  and helps the player shape its strategy for future plays. In such a repeated game, the amount of information that players acquire in repeated plays directly determines how players plan their moves at each round and further influences the resulting learning dynamics. Besides being of theoretical importance, the information feedback in the learning process, such as players' ob-

servations of their opponents' moves, is also of vital importance in designing learning-based methods for solving network problems. As we shall see more clearly in Section 4, in many network applications, networked agents only observe their surroundings, without any access to the global information regarding the whole network. Therefore, due to its significance in learning processes, we first present existing feedback structures that are of wide use in learning, before moving to the details of learning algorithms.

### 3.1 Feedback Structures in Learning

The feedback structure for a player in a repeated game includes its observations regarding the game and the repeated plays, which is a subset of every player's histories of plays and payoffs. To make our discussion more concrete, we introduce the following notation. Let  $I_i^k$  be the feedback of player  $i$  up to time  $k$ . Denote the payoff received by player  $i$  at the  $k$ -th round by  $u_i^k := u_i(a_i^k, a_{-i}^k)$ , and the sequence of payoffs received up to time  $k$  by  $u_i^{1:k} := \{u_i^1, \dots, u_i^k\}$ .

The simplest feedback structure is called the *perfect global feedback*, where

$$I_i^k = \{\{u_j^{1:k}\}_{j \in \mathcal{N}}, \{a_j^{1:k}\}_{j \in \mathcal{N}}\},$$

indicating the completeness of the feedback from both the temporal and the spatial sense. Furthermore, we can also consider the noisy feedback of payoffs,  $U_i^k$ , defined as

$$U_i^k = u_i(a_i^k, a_{-i}^k) + \xi_i^k,$$

where  $\xi_i^k$  is a zero-mean martingale noise process with finite second moment, that is  $\mathbb{E}[\xi_i^k | \mathcal{F}^{k-1}] = 0$ ,  $\mathbb{E}[(\xi_i^k)^2 | \mathcal{F}^{k-1}]$  is less than a constant, and the expectation is taken with respect to the  $\sigma$ -field  $\mathcal{F}^{k-1}$  generated by the history of play up to time  $k-1$ . Simply put, the noisy feedback  $U_i^k$  is a conditionally unbiased estimator of  $u_i^k$  with respect to the history, which is a standing assumption when dealing with the convergence of learning dynamics in games. For noisy feedback in general, or equivalently  $\xi_i^k$  being a generic random variable, the discussion will be carried out in a different context. In that case, a system state should be introduced, which accounts for the uncertainty in the environment, and the learning problem becomes Nash equilibrium seeking in stochastic games (see Definition 2). For more detailed discussions, we refer the reader to another paper in this special issue [17].

The perfect global feedback is of limited use in practice when designing learning algorithms, as the global information is difficult or even impossible to acquire for individuals in large-scale network systems. For example, in distributed or decentralized learning over heterogeneous networks, players may have no access to others' utilities due to physical limitations. Therefore, we are interested in the scenario where players only have direct or indirect access to their own utilities as well as their neighbors', and hence players' feedback can be dependent on the topological structure of the underlying network that connects them.

Consider a repeated game over a graph  $\mathcal{G} := (\mathcal{N}, \mathcal{E})$ , where  $\mathcal{N} = \{1, 2, \dots, N\}$  is the set of nodes, representing the players in the game, who are connected via the edges in  $\mathcal{E} = \{(i, j) | i, j \text{ are connected}\}$ . To simplify the exposition, we assume that the graph is undirected. Note that the direction of the edges does not affect our discussion as long as the neighborhood is properly defined. For example, in a directed graph, when in-neighbors or

out-neighbors specify to which player(s) the player in question can pass information, then the following characterizations of feedback structures still apply. For a more comprehensive treatment of games over networks, we refer the reader to [14].

Each player is allowed to exchange payoff feedback with its neighbors through the edges and observe their actions during the repeated play, whereas the information regarding the rest is hidden from him. In this case, the feedback structure for player  $i$  is

$$I_i^k = \{\{u_j^{1:k}\}_{j \in \{i\} \cup \mathcal{N}(i)}, \{a_j^{1:k}\}_{j \in \{i\} \cup \mathcal{N}(i)}\}, \quad \mathcal{N}(i) := \{j | (i, j) \in \mathcal{E}\}.$$

Note that the player's feedback regarding the payoffs and actions may not be consistent. For example, in a multi-agent robotic system where only the sensors network is effective, each agent can only observe its neighbors' movements through sensors. In this case, without any information of others' utilities, the information feedback of agent  $i$  reduces to  $I_i^k = \{\{u_i^{1:k}\}, \{a_j^{1:k}\}_{j \in \{i\} \cup \mathcal{N}(i)}\}$ . To sum up, if the players can only receive feedback from their neighbors, then players' feedback structures are related to the underlying topology, leading to what is referred to as the *local feedback*. In accordance with this, the extreme case of local feedback is one where the player is isolated in the network, and no information other than its own payoff feedback and actions is available to it. We refer to this extreme case as *individual feedback*, which is a typical information feedback considered in fully decentralized learning and will be further elaborated on when discussing specific learning dynamics later in this section.

In addition to the refinements from the spatial side, we can also consider feedback with various temporal structures. If the player has perfect recall of previous plays, the resulting feedback is said to be *perfect*, and those we have introduced above all fall within this class. Otherwise, players have access to *imperfect feedback*, and we discuss two common cases of imperfect information feedback in the following, namely windowed and delayed feedback.

For the sake of simplicity, we use perfect feedback  $I_i^k = \{u_i^{1:k}, a_i^{1:k}\}$  as a baseline to illustrate that different missing parts of  $I_i^k$  lead to different kinds of imperfect feedback. If the head of  $u_i^{1:k}$  and/or  $a_i^{1:k}$  is not available to the player, that is, there exists a window  $0 < m < k$  such that the player only recalls  $u_i^{(k-m):k}, a_i^{(k-m):k}$ , then the corresponding feedback  $I_i^{(k-m):k} = \{u_i^{(k-m):k}, a_i^{(k-m):k}\}$  is referred to as the windowed feedback with a window size  $m$ . Similarly, if the tail of  $u_i^{1:k}$  and/or  $a_i^{1:k}$  is not available, that is, the player only recalls  $u_i^{1:(k-m)}, a_i^{1:(k-m)}$ , then the imperfect information feedback is  $I_i^{1:(k-m)} = \{u_i^{1:(k-m)}, a_i^{1:(k-m)}\}$ , which is called  $m$ -step delayed feedback.

For learning in games, each player learns to select actions by updating the strategy based on the available feedback at each round. To describe this in mathematical terms, let  $F_i^k$  the strategy learning policy of player  $i$ . The learning policy produces a new strategy  $\pi_i^{k+1}$  for the next play according to

$$\pi_i^{k+1} = (1 - \lambda_i^k) \pi_i^k + \lambda_i^k F_i^k(I_i^k), \quad (7)$$

where  $\lambda_i^k$  is the learning rate, indicating the player's capabilities of information retrieval. Different feedback structures lead to different learning dynamics in repeated games. Under the global or the local feedback structure, each player's feedback is influenced by its opponents' actions and/or payoffs, which makes the players' learning processes coupled, as shown in Figure 3.

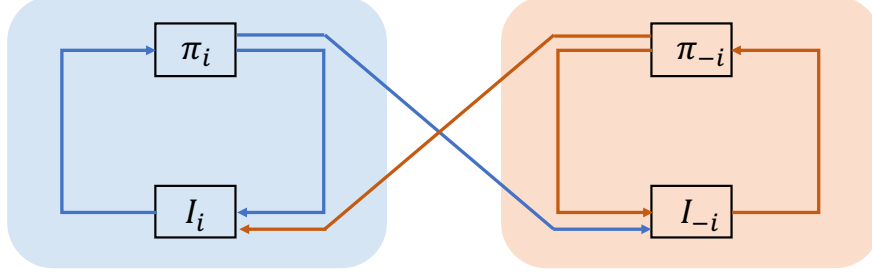


Figure 3: Player’s strategy learning with the corresponding feedback. Under the global or the local feedback structure, players’ learning processes are coupled, as their feedback is influenced by their opponents’ moves. By contrast, players learn to play the game independently under the individual feedback.

In the case of fully decentralized learning under individual information feedback, players learn to play the game independently, and such a learning process is said to be uncoupled. Uncoupled learning processes are of great significance in both theoretical studies [27] and practical applications. Theoretically, learning with such limited information feedback is much more transferable in the sense that learning algorithms under this feedback also apply to online optimization problems, where the online decision-making process is viewed as a repeated game played between a player and the nature [15].

Considering its theoretical importance, we focus on learning with individual feedback in the sequel, and we refer the reader to [28] for a survey on learning methods under other kinds of feedback. We first present reinforcement learning for finite games, where the learning algorithms are characterized into two main classes, due to their distinct nature in exploration. Then, we proceed to gradient play for infinite games, and elaborate on its connection to reinforcement learning. The convergence results of presented algorithms are discussed in Section 3.4 based on stochastic approximation [29, 30] and Lyapunov stability theory.

### 3.2 Reinforcement Learning

Reinforcement learning has been studied in many disciplines and has become a catch-all term for learning in sequential decision making processes where the players’ future choices of actions are shaped by the feedback. In general, reinforcement learning consists of two functions, one of which is the *score function*, evaluating the performance of actions, and the other one is the *choice mapping*, determining the next move. Note that in the machine learning literature [31], the score function and the choice mapping are also called the critic and the actor, respectively. Different score functions and choice mappings lead to different reinforcement learning algorithms. We first provide a generic description of the score function and choice mapping in reinforcement learning from a dynamic system viewpoint, and then we give a characterization of various reinforcement learning algorithms based on different natures in choice mappings. Finally, at the end of this subsection, relations among introduced reinforcement learning algorithms are discussed.

To begin with, we show how the score function can be constructed using the information feedback recursively. Since the player has no direct access to its utility function in this case,

it can construct an estimator  $\hat{\mathbf{u}}_i^k \in \mathbb{R}^{|\mathcal{A}_i|}$  based on  $I_i^k$  to evaluate actions  $a \in \mathcal{A}_i$ . By using this estimator, the player can compare its actions and choose the one that can achieve higher payoffs in the next round. In mathematical terms, the estimator (score function) is given by the following discrete-time dynamical system

$$\hat{\mathbf{u}}_i^{k+1} = (1 - \mu_i^k)\hat{\mathbf{u}}_i^k + \mu_i^k G_i^k(\pi_i^k, \hat{\mathbf{u}}_i^k, U_i^k, a_i^k), \quad (8)$$

where  $G_i^k : \Delta(\mathcal{A}_i) \times \mathbb{R}^{|\mathcal{A}_i|} \times \mathbb{R} \times \mathcal{A}_i \rightarrow \mathbb{R}^{|\mathcal{A}_i|}$  is the learning policy for utility learning,  $\pi_i^k$  is the policy employed at time  $k$ , and  $\mu_i^k$  is the learning rate. Based on the score function, the player can modify its strategy accordingly in the sense that better actions shall be played more frequently in the future. With slight abuse of notations, the strategy update is

$$\pi_i^{k+1} = (1 - \lambda_i^k)\pi_i^k + \lambda_i^k F_i^k(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}, U_i^k, a_i^k), \quad (9)$$

where  $F_i^k : \Delta(\mathcal{A}_i) \times \mathbb{R}^{|\mathcal{A}_i|} \times \mathbb{R} \times \mathcal{A}_i \rightarrow \Delta(\mathcal{A}_i)$  is the learning policy for strategy learning, yielding a new policy for the next play. Compared with (7), the above discrete-time systems (8) (9) explicitly show how the feedback shapes the player's future play. According to (8), the player recursively updates its estimate of the utility function based on the feedback it receives after playing  $\pi_i^k$ , and then the player determines its move in the next round, following (9). Intuitively, we can view  $(\pi_i^k, \hat{\mathbf{u}}_i^{k+1})$  as the information extracted from  $I_i^k$  for updating the player's strategy.

In reinforcement learning, the choice mapping plays an important role in achieving the balance between exploitation and exploration. On one hand, the player would like to choose the best action that is supposed to incur the highest payoff based on the score function. However, this pure exploitation oftentimes leads to myopic behaviors, as the score function may return a poor estimate of the utility function at the beginning of the learning process. Hence, to gather more information for a better estimator, the player also needs some experimental moves for exploration, where suboptimal actions are implemented. To sum up, the trade-off between exploitation and exploration is of vital importance to the success of reinforcement learning, and it depends on the construction of the choice mapping. Different choice mappings result in different reinforcement learning algorithms. Based on their distinct natures in exploration, the algorithms can be categorized into two main classes: *exploitative reinforcement learning* and *exploratory reinforcement learning*.

Recall that in the strategy learning (9), the next strategy produced by the corresponding choice mapping is

$$\pi_i^{k+1} = (1 - \lambda_i^k)\pi_i^k + \lambda_i^k F_i^k(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}, U_i^k, a_i^k),$$

where  $(1 - \lambda_i^k)\pi_i^k$  is referred to as the *cognitive inertia* or simply *inertia*, describing the player's tendency to repeat previous choices independently of the outcome. When determining its next move  $\pi_i^{k+1}$ , the player takes into account both its previous strategy  $\pi_i^k$  and the increment update using the strategy learning policy  $F_i^k$ . Therefore, players' exploration at  $(k + 1)$ -th round stems either from this inertia or the strategy learning policy  $F_i^k$ . The former is called *passive exploration*, as it relies on the player's tendency to repeat previous choices, while the latter one is referred to as *active exploration*, as the player deliberately tries actions based on what he has learned from previous plays.

As the new strategy is a convex combination of the inertia term  $\pi_i^k$  and the learned incremental update  $F_i^k(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}, U_i^k, a_i^k)$ , there is no clear-cut boundary between passive and

active exploration. In fact, reinforcement learning is a continuum of learning algorithms. In the following, we illustrate such a continuum by three prominent learning schemes. The first one is the best response dynamics, located on the left endpoint, which is an example of exploitative reinforcement learning. Solely relying on the inertia for passive exploration, the best response dynamics adopts a purely exploitative learning policy: the best response mapping in (1). On the contrary to the exploitative one, we present dual averaging as an example of exploratory reinforcement learning, which only leverages the learning policy for exploring suboptimal actions without any cognitive inertia. In between, there lies the smoothed best response dynamics, where both the inertia and the strategy learning policy come into play for achieving the balance between exploration and exploitation.

### 3.2.1 Exploitative Reinforcement Learning

For exploitative reinforcement learning, the strategy learning policy always outputs the best strategy based on the score function, which can be viewed as a natural extension of the best response idea in the context of Nash equilibrium (1). In the repeated play scenario, given the opponent's strategy at the  $k$ -th round,  $\pi_{-i}^k$ , from player  $i$ 's standpoint, the best it can do is to choose the best response  $BR_i(\pi_{-i}^k) := \arg \max_{\pi \in \Delta(\mathcal{A}_i)} \{\langle \pi, \mathbf{u}_i(\pi_{-i}^k) \rangle\}$ , which is purely exploitative. In this case, the strategy learning scheme becomes

$$\pi_i^{k+1} \in (1 - \lambda_i^k) \pi_i^k + \lambda_i^k BR_i(\pi_{-i}^k). \quad (10)$$

In general, the best response mapping is a point-to-set mapping, and to analyze the associated learning dynamics, differential inclusion theory [30] is needed, which make the convergence analysis more involved as discussed in Section 3.4.2.

Under the noisy feedback  $I_i^k = \{U_i^{1:k}, a_i^{1:k}\}$ , the score function of player  $i$  is the estimated utility  $\hat{\mathbf{u}}_i^k$ , which is updated according to the following moving average scheme [32]

$$\hat{\mathbf{u}}_i^{k+1}(a) = (1 - \mu_i^k) \hat{\mathbf{u}}_i^k(a) + \mu_i^k \frac{\mathbb{1}_{\{a=a_i^k\}}}{\pi_i^k(a)} U_i^k, \quad a \in \mathcal{A}_i, \quad (11)$$

where  $\mathbb{1}_{\{\cdot\}}$  is an indicator function. Note that in (11), the importance sampling technique, which is common in bandit algorithms [15], is utilized to construct an unbiased estimator of  $\mathbf{u}_i(\pi_{-i}^k)$ . To see this, define a vector  $\hat{\mathbf{U}}_i^k \in \mathbb{R}^{|\mathcal{A}_i|}$ , whose  $a$ -th entry is  $\hat{\mathbf{U}}_i^k(a) := \mathbb{1}_{\{a=a_i^k\}} U_i^k / \pi_i^k(a)$ ; and we then obtain  $\mathbb{E}[\hat{\mathbf{U}}_i^k(a) | \mathcal{F}^{k-1}] = u_i(a, \pi_{-i}^k)$ . Hence, (11) can be rewritten as

$$\hat{\mathbf{u}}_i^{k+1} = (1 - \mu_i^k) \hat{\mathbf{u}}_i^k + \mu_i^k \hat{\mathbf{U}}_i^k, \quad (12)$$

and  $\hat{\mathbf{u}}_i^{k+1}(a)$  gives the averaged payoff incurred by  $a$  in the first  $k$  rounds. This importance sampling technique can be viewed as compensating for the fact that actions played with a low probability do not receive frequent updates of the corresponding estimates, so that when they are played, any estimation error  $U_i^k - \hat{\mathbf{u}}_i^k(a_i^k)$  must have greater influence on the estimated value than if frequent updates occur. We refer the reader to [15, 24] for more details on importance sampling and its use in learning processes.



With a slight abuse of the notation of best response mapping in (2), we define the corresponding best response under the noisy feedback as

$$BR_i(\hat{\mathbf{u}}_i^k) := \arg \max_{\pi \in \Delta(\mathcal{A}_i)} \{\langle \pi, \hat{\mathbf{u}}_i^k \rangle\}. \quad (13)$$

Then, we obtain the following strategy learning scheme [24]

$$\pi_i^{k+1} \in (1 - \lambda_i^k) \pi_i^k + \lambda_i^k BR_i(\hat{\mathbf{u}}_i^k). \quad (14)$$

The resulting dynamical system under the noisy feedback is a coupled system as shown below

$$\begin{aligned} \hat{\mathbf{u}}_i^{k+1} &= (1 - \mu_i^k) \hat{\mathbf{u}}_i^k + \mu_i^k \hat{\mathbf{U}}_i^k, \\ \pi_i^{k+1} &\in (1 - \lambda_i^k) \pi_i^k + \lambda_i^k BR_i(\hat{\mathbf{u}}_i^k). \end{aligned} \quad (\text{BR-d})$$

Originally proposed as a computational method for Nash equilibrium seeking [32, 33], the best response dynamics (BR-d) is directly built upon the best response idea and has been widely applied to evolutionary game problems [34]. One prominent example of best response dynamics is fictitious play [35], where a player's empirical play follows (BR-d); and more details are included in Appendix A. As shown above, best response dynamics adopts passive exploration, and the best response mapping  $BR_i(\cdot)$  encourages greedy actions that might be myopic. As a result, exploitative reinforcement learning may fail to converge [24, 36].

### 3.2.2 Exploratory Reinforcement Learning

In contrast to the inertia-based passive exploration in (BR-d), dual averaging, as introduced in this subsection, only relies on the strategy learning policy  $F_i^k$  for exploring suboptimal actions, in order to avoid myopic behaviors due to the poor estimates of the utility function. In dual averaging, given the player's utility vector  $\mathbf{u}_i$ , the strategy learning policy is a regularized best response [37], defined as

$$QR^\epsilon(\mathbf{u}_i) := \arg \max_{\pi_i \in \Delta(\mathcal{A}_i)} \{\langle \pi_i, \mathbf{u}_i \rangle - \epsilon h(\pi_i)\}, \quad (15)$$

where  $h(\cdot)$  is a penalty function or regularizer and  $\epsilon$  is the regularization parameter. According to [38], a proper regularizer  $h(\cdot)$  defined on the probability simplex should be continuous over the simplex and smooth on the relative interior of every face of the simplex. Besides,  $h$  should be a strongly convex function, and these assumptions ensure that  $QR^\epsilon(\cdot)$  always returns a unique maximizer. The mapping  $QR^\epsilon$  is referred to as a quantal response mapping [39], which allows players to choose suboptimal actions with positive probability. To see how this regularization contributes to active exploration, consider the entropy regularizer  $h(x) = \sum_{x_i} x_i \log x_i$ . In this case,  $QR^\epsilon$  is

$$QR^\epsilon(\mathbf{u}_i)(a) := \frac{\exp(\frac{1}{\epsilon} u_i(a, \pi_{-i}))}{\sum_{a' \in \mathcal{A}_i} \exp(\frac{1}{\epsilon} u_i(a', \pi_{-i}))}, \quad a \in \mathcal{A}_i, \quad (16)$$

which is also known as the Boltzmann-Gibbs strategy mapping [40] or the soft-max function parameterized by  $\epsilon > 0$ . On the one hand, the Boltzmann-Gibbs mapping produces a

strategy that assigns more weight to the actions leading to higher payoffs, that is, the larger  $\mathbf{u}_i(a) = u_i(a, \pi_{-i})$  is, the larger  $QR^\epsilon(\mathbf{u}_i)(a)$  becomes. On the other hand, it always retains positive probabilities for every action, when  $\epsilon > 0$ . Note that  $QR^\epsilon$  can induce different levels of exploration by adjusting the parameter  $\epsilon$ . When  $\epsilon$  tends to 0, the strategy (16) simply returns the action that yields the highest payoff, implying that  $QR^\epsilon$  reduces to the best response mapping  $BR_i(\cdot)$  in (2). As  $\epsilon$  gets larger,  $1/\epsilon$  tends to 0, and the strategy does not distinguish among actions, leading to equal weights for all actions.

Similar to the previous argument, with the noisy feedback, we replace  $\mathbf{u}_i$  by the estimator  $\hat{\mathbf{u}}_i^k$ , and the definition of quantal response mapping is then modified accordingly as

$$QR^\epsilon(\hat{\mathbf{u}}_i^k)(a) := \frac{\exp(\frac{1}{\epsilon}\hat{\mathbf{u}}_i^k(a))}{\sum_{a' \in \mathcal{A}_i} \exp(\frac{1}{\epsilon}\hat{\mathbf{u}}_i^k(a'))}, \quad a \in \mathcal{A}_i.$$

Due to the active exploration brought up by  $QR^\epsilon$ , we can consider an inertia-free reinforcement learning scheme, where the choice map is simply the strategy learning policy  $QR^\epsilon$ . The corresponding strategy learning scheme is then as

$$\pi_i^{k+1} = QR^\epsilon(\hat{\mathbf{u}}_i^{k+1}),$$

where the score function  $\hat{\mathbf{u}}_i^k$  is updated according to the following [41]

$$\hat{\mathbf{u}}_i^{k+1} = \hat{\mathbf{u}}_i^k + \mu_i^k \hat{\mathbf{U}}_i^k. \quad (17)$$

To recap, the learning algorithm operates in the following fashion: at each time  $k$ , an unbiased estimator  $\hat{\mathbf{U}}_i^k$  is constructed as introduced in (11), using importance sampling, and the score function is updated according to (17). Then, the next strategy is produced by the mapping  $QR^\epsilon$ , acting on the score function  $\hat{\mathbf{u}}_i^{k+1}$ , as shown below

$$\begin{aligned} \hat{\mathbf{u}}_i^{k+1} &= \hat{\mathbf{u}}_i^k + \mu_i^k \hat{\mathbf{U}}_i^k, \\ \pi_i^{k+1} &= QR^\epsilon(\hat{\mathbf{u}}_i^{k+1}), \end{aligned} \quad (\text{DA-d})$$

(DA-d) is also referred to as dual averaging, pioneered by Nesterov [41], which was originally proposed as a variant of gradient methods for solving convex programming problems. We elucidate the term “dual averaging” later when we discuss the relation between dual averaging and gradient play, where we demonstrate that (DA-d) can be viewed as a gradient-based algorithm in finite games with  $\hat{\mathbf{u}}_i^k$  being the gradient. Finally, as a remark, we note that in (DA-d), the score function is updated in a manner different than in best response dynamics (BR-d). However, this is merely a matter of presentation, and by selecting a proper  $\epsilon$ , the moving averaging scheme (12) is essentially the same as the discounted accumulation (17), for which we refer the reader to [41, 42]. By adopting the discounted accumulation (17), we later can draw a connection between dual averaging and gradient play.

Apparently, the discrete-time system (DA-d) does not depict how  $\pi_i(t)$  evolves in  $\Delta(\mathcal{A}_i)$ , and it is not straightforward to tell how those good actions bringing up higher payoffs are “reinforced” in the sense that probabilities of choosing them are increasing as the learning process proceeds. In Appendix B, we present that when choosing entropy regularization, (DA-d) is equivalent to the replicator dynamics, one of the well-known evolutionary dynamics

[43–45], which explicitly displays a gradual adjustment of strategies based on the quality of each action. Meanwhile, with an example of population games, we show that this connection brings learning in games to the broader context of evolutionary game theory [34, 44].

As we have mentioned, reinforcement learning is a continuum of learning algorithms, and the best response dynamics (BR-d) and dual averaging (DA-d) are the two endpoints of the continuum. Naturally, we can consider reinforcement learning methods with a blend of both passive and active exploration, where the exploration stems from both the inertia term and the strategy learning policy, as we present in the following.

Instead of choosing actions greedily, we replace the best response  $BR_i(\cdot)$  in (14) by  $QR^\epsilon(\cdot)$ , the quantal response for active exploration, and then we obtain the following strategy learning scheme [24]

$$\pi_i^{k+1} = (1 - \lambda_i^k)\pi_i^k + \lambda_i^k QR^\epsilon(\hat{\mathbf{u}}_i^k).$$

Similar to the best response dynamics in (BR-d), if utility learning follows the moving average scheme in (11), the resulting reinforcement learning has the following discrete-time learning dynamics

$$\begin{aligned}\hat{\mathbf{u}}_i^{k+1} &= (1 - \mu_i^k)\hat{\mathbf{u}}_i^k + \mu_i^k \hat{\mathbf{U}}_i^k, \\ \pi_i^{k+1} &= (1 - \lambda_i^k)\pi_i^k + \lambda_i^k QR^\epsilon(\hat{\mathbf{u}}_i^k).\end{aligned}\tag{SBR-d}$$

Considering its similarity to best response dynamics, (SBR-d) is referred to as smoothed best response dynamics in the literature [24, 46]. Specifically, if the entropy regularizer is adopted, the resulting learning process is called Boltzmann-Gibbs reinforcement learning [47] or entropic reinforcement learning, which has been extensively studied in the context of Markov decision processes [48].

### 3.2.3 Relations among Reinforcement Learning Algorithms

Before wrapping up our presentation on reinforcement learning in finite games, we discuss the relations among the introduced learning algorithms. We reiterate that reinforcement learning corresponds to a continuum of learning algorithms, where one algorithm can be converted to the other by adjusting the learning rate  $\lambda_i^k$  in strategy learning (7) and/or the exploration parameter  $\epsilon$ , and a diagram of such conversion is presented in Figure 4. Our discussion will revolve around the learning rate  $\lambda_i^k$  and the exploration parameter  $\epsilon$ . For simplicity, we suppress the subscript and the superscript of the learning rate and simply denote it by  $\lambda$ .

We begin the discussion with the learning rate  $\lambda$ . Different from dual averaging (DA-d), the best response dynamics (BR-d) and the smoothed best response dynamics (SBR-d) are in fact actor-critic learning [32, 49, 50] due to a positive learning rate  $\lambda > 0$ . Under the actor-critic framework such as (BR-d)(SBR-d), the player maintains two recursive schemes for updating the estimated utility vector and the strategy, respectively. The recursive schemes lead to coupled dynamical systems of  $\hat{\mathbf{u}}_i^k$  and  $\pi_i^k$ . In contrast, even though dual averaging (DA-d) also consists of both updating schemes for estimated utility vector and the strategy, since the learning rate is zero, there is only one effective dynamical system: the one induced by the estimation of utility vector (17). Another way to see the difference between actor-critic

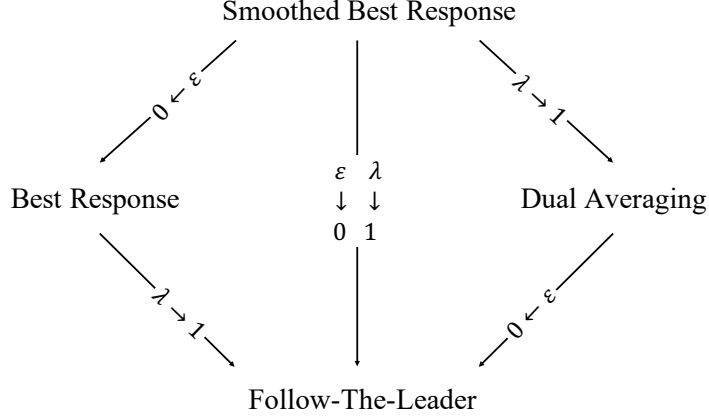


Figure 4: Relationships of reinforcement learning algorithms. For  $0 < \lambda < 1$  and  $\epsilon > 0$ , we obtain the exploratory reinforcement learning: smoothed best response dynamics (SBR-d), where exploration arises from both the inertia and the learning policy. If the active exploration vanishes as  $\epsilon$  goes to zero, smoothed best response reduces to best response dynamics (BR-d), an example of exploitative reinforcement learning. By contrast, we obtain dual averaging (DA-c), if  $\lambda$  tends to 1. Finally, if  $\epsilon$  goes to zero while  $\lambda$  tends to 1, players always choose their actions greedily according to follow-the-leader policy.

learning (BR-d)(SBR-d) and dual averaging (DA-d) is through the corresponding continuous-time learning dynamics in Section 3.4.1.

Even though (DA-d) is not an actor-critic learning, its trajectory is closely related to that of (BR-d)(SBR-d). Intuitively speaking, dual averaging only differs from the smoothed best response in that (DA-d) does not acquire an inertia term, as the learning rate is zero. Hence,  $\pi_i^k$  in (SBR-d) can be seen as the moving average of  $QR^\epsilon(\hat{\mathbf{u}}_i^k)$  in (DA-d). Therefore, it is reasonable to expect that the time average of the trajectory produced by (DA-d) is related to the one produced by the smoothed best response. This intuition has been verified in [38, 51], where it has been shown that the time averaged trajectory of (DA-d) follows (SBR-d) with a time-dependent perturbation  $\epsilon(t)$ .

Apart from the difference in the learning rates, learning algorithms also display distinct asymptotic behavior due to the difference in the exploration parameter. The exploration parameter  $\epsilon$  has less drastic consequence under (DA-d) than under the actor-critic learning (BR-d)(SBR-d). As observed in [38], adding a positive  $\epsilon$  is equivalent to rescaling the regularizer, that is, replacing  $h(\cdot)$  with  $\epsilon h(\cdot)$ . As long as  $\epsilon > 0$ , the regularization  $\epsilon h(\cdot)$  is still proper (see (15) and the following discussion). This implies that even though the choice of  $\epsilon$  affects the speed at which (DA-d) evolves, the qualitative results remain the same. We refer the reader to [38, 52] for a detailed discussion. When  $\epsilon = 0$ , there is no exploration nor inertia for dual averaging, and in this case, players always choose their actions greedily according to the best response mapping

$$\pi_i^{k+1} = \arg \max_{\pi \in \Delta(\mathcal{A}_i)} \langle \pi, \hat{\mathbf{u}}_i^k \rangle, \quad (\text{FTL})$$

where  $\hat{\mathbf{u}}_i^k$  is the score function of player  $i$ , based on its history of play up to round  $k$ , and it

can be updated following (11) or (17). In the online learning literature [15], (FTL) is known as *follow-the-leader (FTL)* policy, which can also be obtained by eliminating the inertia term in the best response dynamics (BR-d). Due to lack of exploration, (FTL) is too aggressive and can be exploited by the adversary, resulting in a positive, non-diminishing regret [15]. The regret is a measure of the performance gap between the cumulative payoffs of current policy (FTL) and that of the best policy in hindsight.

The exploration parameter plays a more important role in the actor-critic learning which balances exploration and exploitation [31]. The smoothed best response (SBR-d), which is a perturbed version of the best response, can only use the regularization  $\epsilon h(\cdot)$  for encouraging active exploration. Thanks to the positive exploration parameter, the smoothed best response (SBR-d) enjoys an  $\epsilon$ -no-regret property, a weak form of external consistency studied in [51, 53], which is desired in an adversarial environment [15]. In contrast, the best response dynamics (BR-d), due to the myopic nature of the best response mapping (2), does not possess similar properties.

### 3.3 Gradient Play

Heretofore, we have limited our discussions to learning processes in finite games, where the score function (8) and the choice mapping (9) act on finite-dimensional vectors. For continuous-kernel games, it is not straightforward to extend reinforcement learning, since a suitable score function is required to evaluate a continuum of actions, and constructing such a score function can be very challenging. Even though function approximators, such as linear [54, 55] or nonlinear [56] ones can be of some help, we present here a mathematically more elegant way of leveraging the reinforcement idea based on gradients of utility functions. In other words, instead of seeking the maximizers, we seek for a better response by searching along the gradient direction. Such gradient-based learning algorithms, referred to as gradient play, are popular in a variety of multi-agent settings due to their versatility, ease of implementation, and dependence on local information.

For the sake of simplicity, we restrict our discussion to pure strategy Nash equilibrium in continuous games (see (4) for the definition and (5) for its variational characterization), in order to avoid measure-theoretic issues when studying the mixed strategy case. We further assume that utilities are smooth functions and perfect feedback is available to players, implying that each player can compute the gradient of the utility function given current iterates:  $D_i^k = \nabla_{a_i} u_i(a_i^k, a_{-i}^k)$ . Even though the perfect feedback is assumed here, it is purely for the simplicity of exposition. It is viable for players to estimate the gradient based on the realized payoff under noisy individual feedback by simultaneous perturbation stochastic approximation [57, 58]. Based on this gradient, players update their actions according to the following

$$\begin{aligned} a_i^{k+1} &= \text{proj}_{\mathcal{A}_i}[a_i^k + \mu_i^k D_i^k], \\ &:= \arg \min_{a \in \mathcal{A}_i} \{\|a_i^k + \mu_i^k D_i^k - a\|_2^2\} \end{aligned} \quad (\text{GD})$$

where  $\text{proj}_{\mathcal{A}_i}(\cdot)$  is the Euclidean projection operator, and (GD) is called online gradient descent or projected gradient descent [42]. One extensively studied variant of (GD) [42, 59]



where  $QR^\epsilon$  is the quantal response mapping in the context of the continuous game, defined as

$$QR^\epsilon(Y) = \arg \max_{a \in \mathcal{A}_i} \{\langle Y, a \rangle - \epsilon h(a)\}.$$

When we choose the Euclidean norm as the regularizer, that is,  $h(x) = \frac{1}{2}\|x\|_2^2$  and  $\epsilon = 1$ ,  $QR^\epsilon$  reduces to the projection operator  $\text{proj}_{\mathcal{A}_i}$ . Geometrically, the gradient search step is performed in the dual space, and then the primal update is produced by the mapping  $QR^\epsilon$ . Since  $QR^\epsilon$  “mirrors” the gradient update in the dual space back to the primal space, it is also referred to as the mirror map in the online optimization literature [15].

### 3.3.1 Mirror Descent as Reinforcement Learning in Continuous Games

Mirror descent (MD) and the reinforcement learning (DA-d) share the same choice map, and they are closely connected. We demonstrate in the following that as a gradient-based algorithm, mirror descent can also be cast as a reinforcement learning scheme in continuous games, with  $Y_i^k$  being the “score function”.

To evaluate a certain action  $a \in \mathcal{A}_i$  at time  $k$ , consider  $\sum_{\tau=1}^k u_i(a, a_{-i}^\tau)$ , the counterfactual outcome had player  $i$  implemented  $a$  all the time in the past. The higher the sum is, the better action is  $a$ , since it could have brought up higher payoffs. Hence, the player can choose the next action that is optimal in hindsight:

$$a_i^{k+1} = \arg \max_{a \in \mathcal{A}_i} \left\{ \sum_{\tau=1}^k u_i(a, a_{-i}^\tau) - \epsilon h(a) \right\}, \quad (\text{FTRL})$$

where  $\epsilon h(\cdot)$  is the regularization introduced in (15), encouraging exploration in the learning process. Based on the optimality in hindsight, this action selection (FTRL) is known as *follow-the-regularized-leader* (FTRL) [60]. Moreover, if  $u_i$  is well-behaved in the sense that it can be approximated by the first-order Taylor expansion, that is,  $u_i(a, a_{-i}^\tau) \approx u_i(a_i^\tau, a_{-i}^\tau) + \langle D_i(\mathbf{a}^\tau), a - a_i^\tau \rangle$ , then (FTRL) is equivalent to

$$\begin{aligned} a_i^{k+1} &= \arg \max_{a \in \mathcal{A}_i} \left\{ \sum_{\tau=1}^k \langle D_i(\mathbf{a}^\tau), a \rangle - \epsilon h(a) \right\} \\ &= \arg \max_{a \in \mathcal{A}_i} \left\{ \left\langle \sum_{\tau=1}^k D_i(\mathbf{a}^\tau), a \right\rangle - \epsilon h(a) \right\} \\ &= QR^\epsilon \left( \sum_{\tau=1}^k D_i(\mathbf{a}^\tau) \right), \end{aligned}$$

which is exactly the mirror descent scheme in (MD), despite using an auxiliary variable  $Y_i^k$  to aggregate these gradients weighted by the learning rates  $\mu_i^k$ . In other words, by the first-order expansion, the sum of gradients living in the dual space serves a linear functional for evaluating the quality of the actions. Hence, the sum or equivalently  $Y_i^k$  can be treated as a “score function”, based on which the mirror map outputs a better action in hindsight, yielding a reinforcement procedure.

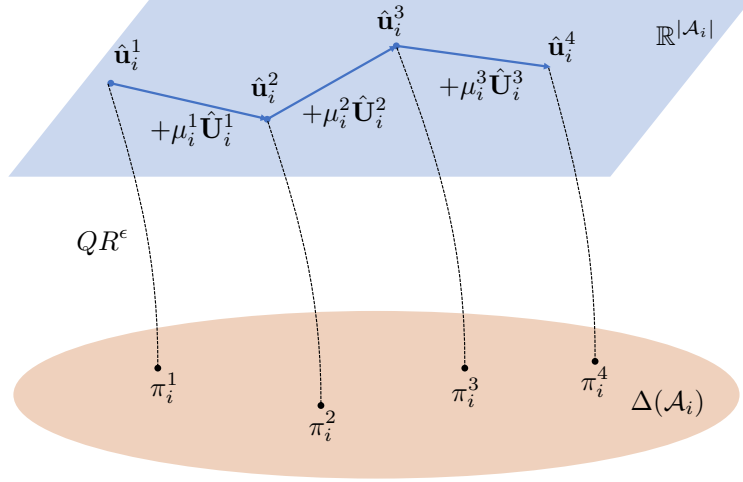


Figure 6: Schematic representation of dual averaging (DA-d). There is no explicit dynamics in the primal space  $\Delta(\mathcal{A}_i)$ . Instead, the dual variables  $\hat{\mathbf{U}}_i^k$  are first aggregated within the dual space  $\mathbb{R}^{|\mathcal{A}_i|}$ , and then are “mirrored” back to the primal space via the mirror mapping  $QR^\epsilon$ .

### 3.3.2 Reinforcement Learning as Mirror Descent in Finite Games

In the above discussion, we interpreted the mirror descent scheme (MD) as a “reinforcement learning” in continuous games. In this subsection, we further show that the idea of mirror descent can also be employed in finite games, and the resulting learning dynamics is in fact the exploratory reinforcement learning scheme (DA-d).

In finite games, the utility function is not differentiable with respect to the action, since actions sets are finite. In order to leverage the gradient play, we consider the mixed extension of the finite games. Consider the expected utility  $u_i(\pi_i, \pi_{-i}) = \langle \pi_i, \mathbf{u}_i(\pi_{-i}) \rangle$ , then the gradient of the expected utility with respect to player  $i$ ’s strategy  $\pi_i$  is given by  $\mathbf{u}_i(\pi_{-i})$ . Naturally, we can apply the mirror descent scheme (MD) to this mixed extension without difficulty. Furthermore, if the gradient is not directly available, for example, learning under the noisy feedback, we rely on the unbiased estimator of  $\mathbf{u}_i(\pi_{-i}^k)$ ,  $\hat{\mathbf{U}}_i^k$ , which can be viewed as an estimator of the payoff gradient  $D_i$  in (MD). It can be easily seen that the mirror descent scheme for this induced continuous game reduces to the exploratory reinforcement learning in (DA-d). Consequently, the learning scheme (DA-d) is called dual averaging: the dual variables, the gradients  $\hat{\mathbf{U}}_i^k$ , are aggregated first within the dual space and then are “mirrored” back to the primal space by the mirror mapping [41]. A schematic representation of dual averaging is provided in Figure 6.

## 3.4 Convergence of Learning in Games

This subsection examines the asymptotic behavior of learning algorithms introduced above, with the focus on the convergence results of the introduced learning algorithms. Due to the close connection between gradient play in continuous games and reinforcement learning in finite games, we limit our scope to reinforcement learning algorithms in finite games, while



pointing the reader to [37, 58, 61–63] for the treatment in continuous games. The discussion in this subsection is primarily based on stochastic approximation theory and Lyapunov stability theory [30, 64], and a generic procedure of applying such analytical tools consists of three steps: 1) develop the mean-field continuous-time dynamics using stochastic approximation theory; 2) study the continuous-time learning dynamics using ODE methods, relating its Lyapunov stability to Nash equilibria of the underlying game; 3) derive the convergence results of discrete-time algorithms using asymptotic convergence of corresponding continuous-time dynamics. Since the third step is direct corollary of the results of the first and second steps, we articulate the first two steps in analyzing the asymptotic behaviors of reinforcement learning in the sequel. We refer the reader to Appendix C and references therein for details on the relation between discrete-time trajectory and its continuous counterpart.

### 3.4.1 Learning Dynamics and Stochastic Approximation

With proper  $F_i^k$  and  $G_i^k$ , learning algorithms allow the players to reach the Nash equilibrium of the game in the limit. Hence, the problem boils down to analyzing the limiting behavior of the discrete-time systems (BR-d), (DA-d), (SBR-d), that is, whether its global attractor comprises equilibria. Direct investigations into such learning dynamics are challenging, as stochasticity enters the updating rules. For example, the action at time  $k$ ,  $a_i^k$  is sampled from the strategy  $\pi_i^k$ , and the payoff feedback  $U_i^k$  also incurs randomness.

Thanks to the celebrated stochastic approximation theory, we can turn to the continuous counterpart of the discrete-time dynamics: an ordinary differential equation (ODE), whose trajectory enjoys the same asymptotic property. From a technical standpoint, the continuous-time dynamics often produce a more comprehensible picture for analysis with fruitful tools available at our disposal. One of the most powerful tools is Lyapunov stability theory. Besides, such a continuous-time framework also allows us to connect learning theory with the extensive literature on game dynamics in biology and evolutionary theory [24], where the time interval between two repetitions of the game is infinitesimally small.

Recall that reinforcement learning adopts two coupled discrete-time dynamical systems: one for the score function (8) and the other one for the choice mapping (9).

$$\begin{aligned}\hat{\mathbf{u}}_i^{k+1} &= (1 - \mu_i^k)\hat{\mathbf{u}}_i^k + \mu_i^k G_i^k(\pi_i^k, \hat{\mathbf{u}}_i^k, U_i^k, a_i^k), \\ \pi_i^{k+1} &= (1 - \lambda_i^k)\pi_i^k + \lambda_i^k F_i^k(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}, U_i^k, a_i^k).\end{aligned}$$

In the following, the continuous-time dynamics associated with (8) and (9) is obtained via stochastic approximation, which paves the way for the ODE-based convergence analysis. We begin with a generic description of the learning dynamics under reinforcement learning, and then we specify the learning dynamics corresponding to (BR-d)(DA-d)(SBR-d). For more details regarding stochastic approximation, we refer the reader to Appendix C and the references therein.

For the sake of simplicity in exposition, we assume that learning policies in (8) and (9) are time-invariant, denoted by  $F_i$  and  $G_i$ , respectively. When the learning policies are time-variant, stochastic approximation theory still applies, and we refer the reader to [47] for more details. Let the mean-field components of (8) and (9) be denoted by  $f_i(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}) = \mathbb{E}[F_i(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}, U_i^k, a_i^k) | \mathcal{F}^{k-1}]$  and  $g_i(\pi_i^k, \hat{\mathbf{u}}_i^k) = \mathbb{E}[G_i(\pi_i^k, \hat{\mathbf{u}}_i^k, U_i^k, a_i^k) | \mathcal{F}^{k-1}]$ , respectively. We

can then write down the following coupled differential equations

$$\begin{aligned}\frac{d\hat{\mathbf{u}}_i(t)}{dt} &= g_i(\pi_i(t), \hat{\mathbf{u}}_i(t)), \\ \frac{d\pi_i(t)}{dt} &= f_i(\pi_i(t), \hat{\mathbf{u}}_i(t)),\end{aligned}$$

which are closely related to (8) and (9). By stochastic approximation theory (see Appendix C), the linear interpolations of the sequences  $\{\pi_i^k\}$  and  $\{\hat{\mathbf{u}}_i^k\}$  are the perturbed solutions to the differential equations above, which are arbitrarily close to the true solution as time goes to infinity. In other words, the convergence results of (8) and (9) can be obtained by studying the limiting behavior of the associated differential equations.

Following the same argument, the learning dynamics of the best response (BR-d) can be written as

$$\begin{aligned}\frac{d\hat{\mathbf{u}}_i(t)}{dt} &= \mathbf{u}_i(\pi_{-i}(t)) - \hat{\mathbf{u}}_i(t), \\ \frac{d\pi_i(t)}{dt} &\in BR_i(\hat{\mathbf{u}}_i(t)) - \pi_i(t).\end{aligned}\tag{BR-c}$$

If the best response dynamics is adopted by every player, we can consider the continuous-time dynamics of the strategy profile of all players  $\pi(t) = [\pi_1(t), \pi_2(t), \dots, \pi_N(t)]$  under best response. Denote the joint utility vector by  $\mathbf{u}(\pi(t)) := [\mathbf{u}_1(\pi_{-1}(t)), \mathbf{u}_2(\pi_{-2}(t)), \dots, \mathbf{u}_N(\pi_{-N}(t))]$ , and similarly, the joint estimated utility vector by  $\hat{\mathbf{u}}(t) := [\hat{\mathbf{u}}_1(t), \hat{\mathbf{u}}_2(t), \dots, \hat{\mathbf{u}}_N(t)]$ . Then, for the strategy profile  $\pi(t)$ , the continuous-time learning dynamics under the best response algorithm is

$$\frac{d\hat{\mathbf{u}}(t)}{dt} = \mathbf{u}(\pi(t)) - \hat{\mathbf{u}}(t),\tag{18}$$

$$\frac{d\pi(t)}{dt} \in BR(\hat{\mathbf{u}}(t)) - \pi(t).\tag{19}$$

From its associated learning dynamics, we can see that the best response algorithm (BR-d) or equivalently its continuous-time mean-field dynamics (BR-c) is in fact an actor-critic learning [31], where the approximation  $\hat{\mathbf{u}}(t)$  given by (18) serves as the actor, evaluating the performance of the current strategy profile, while the strategy update (19) is the critic that improves the strategy.

As observed in the literature [31], the performance of the actor-critic learning relies on the quality of evaluation from the actor. One approach to obtain a satisfying actor in learning is to leverage the two-timescale idea [29], according to which (18) should operate at a faster timescale than (19). Intuitively speaking, in order to obtain a  $\hat{\mathbf{u}}(t)$  that can approximately evaluate the current strategy profile  $\pi(t)$ , the player must wait until  $\hat{\mathbf{u}}(t)$  nearly converges before it updates the strategy using (19). To analyze the convergence of the two-timescale dynamics, one can study its equivalent single-timescale dynamics. Since the actor (18) runs at a faster timescale, the system (18) and (19) can be “decoupled” in the following way: by fixing  $\pi(t) = \pi$ , the faster timescale update (18) converges to  $\mathbf{u}(\pi)$ , where  $\pi$  is viewed as a parameter. Then, after the convergence of the fast dynamics to an equilibrium  $\mathbf{u}(\pi)$ , the

slow dynamics (19) is set into motion, where  $\hat{\mathbf{u}}(t)$  is replaced by its equilibrium point  $\mathbf{u}(\pi(t))$  and the resulting learning dynamics is

$$\frac{d\pi(t)}{dt} \in BR(\pi(t)) - \pi(t). \quad (20)$$

As we illustrate in Appendix C, the coupled dynamics (18)(19) and the single-timescale (20) share similar asymptotic behaviors. Hence, we can focus on the much simplified one (20) for the derivation of the convergence results. For more details about the two-timescale learning and the derivation of the equivalent dynamics, we refer the reader to Appendix C and references therein.

Applying the same argument to the smoothed best response (SBR-d), we obtain

$$\begin{aligned} \frac{d\hat{\mathbf{u}}_i(t)}{dt} &= \mathbf{u}_i(\pi_{-i}(t)) - \hat{\mathbf{u}}_i(t), \\ \frac{d\pi_i(t)}{dt} &= QR^\epsilon(\hat{\mathbf{u}}_i(t)) - \pi_i(t), \end{aligned} \quad (\text{SBR-c})$$

and its equivalent dynamics regarding the joint strategy profile is

$$\frac{d\pi(t)}{dt} = QR^\epsilon(\mathbf{u}(\pi(t))) - \pi(t). \quad (21)$$

Different from the best response (BR-d) and the smoothed best response (SBR-d), dual averaging (DA-d) does not belong to the class of actor-critic methods. To see this, let us write down its continuous-time dynamics

$$\begin{aligned} \frac{d\hat{\mathbf{u}}_i(t)}{dt} &= \mathbf{u}_i(\pi_{-i}(t)), \\ \pi_i(t) &= QR^\epsilon(\hat{\mathbf{u}}_i(t)). \end{aligned} \quad (\text{DA-c})$$

Similar to the previous argument, the learning dynamics for the strategy profile is

$$\begin{aligned} \frac{d\hat{\mathbf{u}}(t)}{dt} &= \mathbf{u}(\pi(t)), \\ \pi(t) &= QR^\epsilon(\hat{\mathbf{u}}(t)), \end{aligned} \quad (\text{DA})$$

where the dynamics regarding  $\hat{\mathbf{u}}(t)$  does not produce an approximation of  $\mathbf{u}(\pi(t))$ . Instead, it gives the cumulative payoff:  $\hat{\mathbf{u}}(t) = \int_0^t \mathbf{u}(\pi(\tau))d\tau + \hat{\mathbf{u}}(0)$ . It is straightforward to see that as there is only one differential equation in (DA), the resulting autonomous dynamical system is only related to  $\hat{\mathbf{u}}(t)$ . Hence, there is no additional dynamics regarding the strategy update, which makes (DA) fundamentally different from (BR-c) and (SBR-c).

### 3.4.2 Nash Equilibrium and Lyapunov Stability

Since the various learning algorithms belong to different classes, the discussion regarding the convergence results of the introduced learning dynamics are organized in the following way. We begin with dual averaging (DA), a type of gradient-based dynamics, and then proceed to the best response dynamics (BR-c) and the smoothed best response (SBR-c).

**Dual Averaging** Consider the learning dynamics of the joint strategy profile and the estimated utility vector under dual averaging

$$\begin{aligned}\frac{d\hat{\mathbf{u}}(t)}{dt} &= \mathbf{u}(\pi(t)), \\ \pi(t) &= QR^\epsilon(\hat{\mathbf{u}}(t)).\end{aligned}\tag{DA}$$

This compact form implies that (DA) is an autonomous system evolving in the dual space. Here, similar to the discussion in Section 3.3, we adopt the terminology in [41, 42], where the gradient  $\mathbf{u}(\pi(t))$  is the dual variable and the corresponding space is termed the dual space. As shown in [38], (DA) is a well-posed dynamical system in the dual space in that it admits a unique global solution for every initial  $\hat{\mathbf{u}}(0)$ . Furthermore, it can be shown that the dynamics of  $\pi(t)$  on the game's strategy space induced by (DA) under steep regularizers is also well-posed [38, 52]. However, the well-posedness of the induced dynamics under generic regularizers remains unclear [38]. The reason lies in the fact that under steep regularizers, such as the entropy regularizer, the projected dynamics regarding  $\pi(t)$  evolves within the interior of the simplex, and the resulting ODE is also well posed in the primal space, which need not hold for nonsteep regularizers. For more generic choices of  $QR$  and related stability analysis, we refer the reader to [38].

Even though studying the stability of the induced dynamics in the primal space may not be viable due to the well-posedness issue, the asymptotic behavior of  $\pi(t)$  can be characterized by investigating its dual  $\hat{\mathbf{u}}(t)$ . Toward that end, we call  $\pi(t) = QR^\epsilon(\hat{\mathbf{u}}(t))$  the induced orbit of (DA) or simply orbit, and we introduce the following notions regarding the stability and stationarity of  $\pi(t)$ , which is adapted from [38].

**Definition 4** Denote by  $\text{im}(QR^\epsilon)$  the image of  $QR^\epsilon$ . For  $\pi(t) = QR(\mathbf{u}(t))$ , an orbit of (DA), we say that a fixed  $\pi^* \in \prod_{i \in \mathcal{N}} \Delta(\mathcal{A}_i)$  is

1. *stationary*, if  $\pi(t) = \pi^* \in \text{im}(QR^\epsilon)$  for all  $t \geq 0$ , whenever  $\pi(0) = \pi^*$ ;
2. *Lyapunov stable*, if for every neighborhood  $U$  of  $\pi^*$ , there exists a neighborhood  $U'$  of  $\pi^*$  such that  $\pi(t) \in U$  for all  $t \geq 0$  whenever  $\pi_0 \in U' \cap \text{im}(QR^\epsilon)$ ;
3. *attracting*, if there exists a neighborhood  $U$  such that  $\pi(t) \rightarrow \pi^*$  as  $t \rightarrow \infty$  whenever  $\pi_0 \in U \cap \text{im}(QR^\epsilon)$ ;
4. *globally attracting*, if  $\pi^*$  is attracting with the attracting basin being the entire image  $\text{im}(QR^\epsilon)$ ;
5. *asymptotically stable*, if  $\pi^*$  is both attracting and Lyapunov stable;
6. *globally asymptotically stable*, if  $\pi^*$  is both globally attracting and Lyapunov stable.

Similar to the Folk Theorem of evolutionary game theory [34], there is an equivalence between the stationary points of (DA) and the Nash equilibria [34, 38]: any stationary point is a Nash equilibrium and conversely, every Nash equilibrium that is within the image of the mirror map (15) is a stationary point. In addition to the relation between Nash equilibrium and the stationary point, another important question is the following:

*Are Nash equilibria of the underlying game (globally) asymptotically stable under (DA)?*

To answer this question, we shall revisit the variational characterization of Nash equilibrium, which bridges the equilibrium concepts associated with two different mathematical models: games and dynamical systems. Recall that the Nash equilibrium is equivalent to the solution of the variational inequality

$$\langle \mathbf{u}(\pi^*), \pi - \pi^* \rangle \leq 0, \quad \text{for all } \pi \in \prod_{i \in \mathcal{N}} \Delta(\mathcal{A}_i). \quad (\text{SVI})$$

Since the utility function  $u_i(\pi_i, \pi_{-i})$  is linear in  $\pi_i$ , the Stampacchia-type variational inequality (SVI) is equivalent to the following Minty-type variational inequality

$$\langle \mathbf{u}(\pi), \pi - \pi^* \rangle \leq 0, \quad \text{for all } \pi \in \prod_{i \in \mathcal{N}} \Delta(\mathcal{A}_i), \quad (\text{MVI})$$

which implies that the Nash equilibrium  $\pi^*$  is the solution to (MVI) [20]. Then, to answer the question of interest, it suffices to investigate whether the solution to (MVI) is attracting under (DA). As discussed in [61], the answer is negative: not every Nash equilibrium of  $N$ -player general-sum game is attracting. To ensure the convergence of (DA), an additional condition has to be imposed on (MVI).

**Definition 5 (Variational Stability [38])**  $\pi^*$  is said to be *variationally stable* if there exists a neighborhood  $U$  of  $\pi^*$  such that

$$\langle \mathbf{u}(\pi), \pi - \pi^* \rangle \leq 0, \quad \text{for all } \pi \in U, \quad (\text{VS})$$

where equality holds if and only if  $\pi^* = \pi$ . In particular, if  $U = \prod_{i \in \mathcal{N}} \Delta(\mathcal{A}_i)$ ,  $\pi^*$  is said to be *globally variationally stable*.

The definition of variational stability (VS) can be extended to sets [38]. Let a subset  $\Pi^* \subset \prod_{i \in \mathcal{N}} \Delta(\mathcal{A}_i)$  be closed and nonempty.  $\Pi^*$  is said to be *variationally stable* if there exists a neighborhood  $U$  of  $\Pi^*$  such that

$$\langle \mathbf{u}(\pi), \pi - \pi^* \rangle \leq 0, \quad \text{for all } \pi \in U, \pi^* \in \Pi^*, \quad (22)$$

where equality holds for a given  $\pi^* \in \Pi^*$  if and only if  $\pi \in \Pi^*$ .

The notion “variational stability” is proposed in [38] as a relaxation of the monotonicity condition of the pseudo-gradient mapping of the game, e.g.,  $\mathbf{u}(\pi)$  in the mixed extension of finite games, or  $D(\mathbf{a})$  in continuous games. Variational stability alludes to the seminal notion of evolutionary stability introduced in [43], and the introduced definition is in a similar spirit to the variational characterization of evolutionarily stable state studied in [34]. An equivalent notion is developed in the line of works on gradient-based learning [61], named *locally asymptotically stable Nash equilibria (LASNE)*, and as its name suggests, Nash equilibria satisfying the variational stability (VS) are asymptotically stable under gradient-based dynamics. Likewise, Nash equilibria satisfying global variational stability are globally asymptotically stable (GASNE). We refer the reader to [61] and references therein for more details about this characterization of Nash equilibria.

What has been presented above provides a generic criterion for examining the convergence of gradient-based dynamics (DA), and in the following, based on the notion of variational stability, we discuss some concrete cases, where the learning dynamics converges either locally or globally to Nash equilibria. As shown in [37], for any finite games, every strict Nash equilibrium satisfies (VS) and hence is a LASNE. Therefore, every strict Nash equilibrium in finite games is locally attracting. On the other hand, to ensure global convergence, the underlying Nash equilibrium has to be GASNE or equivalently satisfy the global variational stability. For finite games, the existence of a potential implies monotonicity, which further implies the existence of globally variationally stable Nash equilibria [37]. Hence, for potential games [38, 65] and monotone games [37, 66], regardless of the initial points, the orbit of (DA) always converges to the set of Nash equilibria. We summarize our discussions in the following, where 1) and 2) are direct extensions of the folk theorem of evolutionary dynamics [34], while 3)-5) are corollaries of variational characterization of Nash equilibria in [38] and [61].

For every finite game, we have the following characterization of Nash equilibrium using the language of Lyapunov stability [38, 52]. For a fixed  $\pi^* \in \prod_{i \in \mathcal{N}} \Delta(\mathcal{A}_i)$ ,

1. if  $\pi^*$  is stationary, it is a Nash equilibrium;
2. if  $\pi^*$  is Lyapunov stable, then  $\pi^*$  is a Nash equilibrium;
3. if  $\pi^*$  is a Nash equilibrium and it falls within the image of the mirror map, then it is stationary;
4. if  $\pi^*$  is a strict Nash equilibrium, it is asymptotically stable;
5. if  $\pi^*$  is a Nash equilibrium of a potential game or a monotone game, it is globally asymptotically stable.

**Best response dynamics** The analysis of the best response dynamics (20) is more involved than that of dual averaging (DA). The theoretical challenge is mainly due to the discontinuous, set-valued nature of the best response mapping (2). In general, as a differential inclusion, (20) typically admits non-unique solutions through every initial point [30]. Early works have established the convergence results on (20) for games with special structures: best response dynamics converges to Nash equilibrium in zero-sum games, where the Nash equilibrium is essentially a saddle point [33, 62, 67], in two-player strictly supermodular games [44] and in finite potential games [30, 33]. However, we note that these research works, even though most of them still rely on the Lyapunov argument [30, 33, 62, 67], do not directly reveal any generic relation between Lyapunov stability and Nash equilibrium in general multi-player non-zero sum games, and they are more or less on an ad hoc basis.

Recent endeavors on the study of the best response dynamics have helped shed some light on the asymptotic behavior of best response dynamics by relating the best response vector field  $BR(\pi) - \pi$  to the gradient field  $\mathbf{u}(\pi)$ , which renders the best response dynamics in some potential games [68, 69] as an approximation of the gradient-based dynamical system [68]. For the finite potential games considered in [68], additional regularity conditions are imposed, which are closely related to the notion of variational stability introduced above. Therefore, the variational characterization of Nash equilibrium and variational stability becomes relevant

under the best response dynamics. Following this line of reasoning, it is shown in [68], in regular potential games, that the best response dynamics is well-posed for almost every initial condition, and converges to the set of Nash equilibria.

**Smoothed Best Response** As we can see from the explicit expression, smoothed best response dynamics (21) only differs from the best response dynamics (20) in the operator  $QR^\epsilon(\cdot)$ , which serves as a perturbed best response [70], and the perturbation is determined by  $\epsilon$  [51]. Hence, if  $\epsilon$  tends to zero, it is straightforward to see that the smoothed best response (21) will enjoy the same asymptotic property as the best response (20), which implies that identical results should also be achievable for smoothed best response with vanishing exploration. This intuition has been verified in [46, 63], where smoothed best response (21) is shown to converge in zero-sum games, potential games and supermodular games.

On the other hand, with a constant  $\epsilon$ , it is not realistic to expect the smoothed best response, essentially a fixed point iteration, to always converge to exact Nash equilibrium. Hence, a new equilibrium concept has been introduced in the literature, which is termed perturbed Nash equilibrium in [71, 72] or Nash distribution in [32, 50]. The new equilibrium is defined as the fixed point of the smoothed best response. We do not carry out detailed discussion on that in this paper, since the convergence analysis still rests on the standard Lyapunov argument, and the epistemic justification of such equilibrium [24, 33] is beyond the scope of this paper. We refer the reader to [24, 50, 63, 72] for a rigorous treatment of the smoothed best response.

### 3.5 Beyond Stochastic Approximation

In addition to stochastic approximation and related ODE methods, another class of widely applied learning algorithms is built upon Markov Chain theory [73], which is termed learning by trial and error (LTE) [74]. Even though the name of the proposed learning suggests its similarity to reinforcement learning, the learning process is quite different in the sense that there are no explicit score functions or choice mappings in the proposed method. In LTE, there are two basic rules: 1) players occasionally experiment with alternative strategies, keeping the new strategy if, and only if, it leads to a strict increase in payoff; 2) if the player experiences a payoff decrease due to a strategy change by someone else, it starts a random search for a new strategy, eventually settling on one with a probability that increases monotonically with its realized payoff. In words, the “error” part relies on the realized payoff, and no advanced device is needed, such as score functions like Q-functions or estimated utilities, while the “trial” part is a random search procedure implemented according to the two basic rules. A novel feature of the process is that different search procedures are triggered by different psychological states or *moods*, where mood changes are induced by the relationship between a player’s realized payoffs and his current payoff expectations. To be specific, there are four moods: *Content*( $C$ ), *Hopeful*( $H$ ), *Watchful*( $W$ ) and *Discontent*( $D$ ), and different moods lead to different random search procedures. Briefly, players will explore new strategies with high probabilities when in  $W$  and  $D$ , while sticking to the current one, with high probabilities, if the mood is  $C$  or  $H$ . Details can be found in the original paper, and a concise summary is provided in [75].

This mood-based trial and error is different from reinforcement learning introduced in the previous subsection, where the exploration is not determined explicitly by the score function and the choice mapping. Hence, LTE does not fit the stochastic approximation framework introduced above, and instead, the associated convergence proof relies on perturbed Markov Chain theory [73, 76]. It is shown in [74] that in a two-player finite game, if there at least exists a pure Nash equilibrium, then LTE guarantees that pure Nash equilibrium is played at least  $1 - \epsilon$  of the time, where  $\epsilon$  is the probability of exploring new strategies. For an  $N$ -player finite game, if the game is *interdependent* [74] and there at least exists one pure Nash equilibrium, the same theoretical guarantee for the two-player case also holds. It is not surprising that LTE does not achieve convergence in conventional ways, that is, almost sure convergence and convergence in the mean, since players will always explore new strategies with positive probability at least  $\epsilon$ . The proposed learning method and its variants have also been applied to learning efficient equilibrium [77] (Pareto dominant, maximizing social welfare), learning efficient correlated equilibria [78], achieving the Pareto optimality [79] and other related works in engineering applications, especially in cognitive radio problems [28].

The idea of trial and error in LTE leads to many important variants, such as sample experimentation dynamics in [76] and optimal dynamical learning [75, 79], which also rely on perturbed Markov processes for equilibrium seeking. Even though the convergence results of these algorithms all rest on Markov Chain (MC) theory [73], the analysis of their performance remains unclear, due to the computation complexity of the inherent MC generated by these algorithms. To circumvent the dimensionality issue regarding the number of states in the original MC, an approximation-based dimension reduction method is proposed in [75], which allows numerical convergence analysis for LTE and its variants based on Monte Carlo simulations. Besides, we also note that a much simplified trial-and-error algorithm has been theoretically analyzed in [80], where the optimal exploration rate is identified and the associated convergence rate is discussed. It is not unrealistic to expect a similar argument may apply to LTE and its variants, but the technical challenges regarding the dimensionality should not be downplayed.

### 3.6 Resurgence of Learning in Games

With machine learning (ML) algorithms being increasingly deployed in real-world applications, there has been a resurgence in research endeavors on multi-agent learning and learning in games [81]. In addition to the line of research driven by evolutionary dynamics dating back to 1950s [34, 44], the current wave of learning theory development is mainly driven by a desire to better understand and improve the performance of ML algorithms in a competitive environment. In general, there are two possible roles that game theoretic methods can play in ML study: 1) Game-theoretic methods is an add-on for improving the performance of ML algorithms. 2) Certain ML problems manifest the game features, which calls for game-theoretic tools. For supervised learning, the recent interest in adversarial learning techniques serves as an example to show how game-theoretic models and learning methods can be used to robustify machine learning [82, 83], where potential attacks or disturbance are viewed as strategic moves of an opponent. On the other hand, there are problems in unsupervised learning where game-theoretic models are no longer tools for solving the problem but the problem itself. Generative Adversarial Networks (GAN) [84], is an approach to generative



modeling using deep learning methods, involving automatically discovering and learning the patterns of input data in such a way that newly generated examples output by the generative model (generator) cannot be distinguished from the input. In game-theoretic language, the training process of GAN is essentially a learning process in a zero-sum game between the generator and the discriminator, where the generator tries to generate new samples that plausibly could have been drawn from the original dataset, while the discriminator tries to pick those fake ones produced by the generator. We do not intend to provide a comprehensive survey for these machine learning applications, instead we refer the reader to [81, 82].

Despite different contexts under which the learning theory is studied, recent research efforts mainly revolve around the following three aspects:

1. learning dynamics in general multi-player repeated games;
2. learning dynamics in repeated games with acceleration design;
3. learning dynamics in dynamic games in a decentralized manner.

The first research direction is a natural follow-up to the study of evolutionary dynamics [34, 44], which aims to bring learning in games to a broad range of ML applications, since in ML, the game structure is specified by the underlying data and may not enjoy any desired properties. Recall that convergence results and asymptotic behaviors regarding the three dynamics (BR-c)(SBR-c)(DA-c) are discussed with the assumption that the underlying game acquires special structures, such as potential games, supermodular games and zero-sum games. However, for games with fewer assumptions on the utility function, there is still a lack of understanding of the dynamics and the limiting behavior of learning algorithms. One of the central questions of this direction is *what the relations between Nash equilibria and stationary points as well as attracting sets under the learning dynamics are*. Recent attempts try to answer this question from a variational perspective [85], and provide various characterizations of Nash equilibria with desired properties under gradient-based dynamics [52, 61, 86]. Furthermore, considering its applications in ML problems, learning algorithms in stochastic settings are of great significance in recent studies, and we refer the reader to [61, 87] for more details as well as to [17] for an introduction to stochastic Nash equilibrium seeking.

The second research direction, which attracts attention from the ML community, the optimization community as well as the control community, is directly related to the design of ML algorithms. The goal is to develop acceleration techniques that improve the performance of learning algorithms. Based on the understanding of first-order gradient-based dynamics games such as (GD)(LGD), recent research efforts have focused on high-order gradient methods, which can be dated back to Nesterov’s momentum idea [42], and researchers endeavor to propose a general framework that generalize the momentum for the generation of accelerated gradient-based algorithms [85]. On account of the close relationships among Nash equilibrium, variational problems and dynamical systems [20], one approach for developing acceleration is to generalize the concept momentum by formulating the equilibrium seeking as a variational (optimization) problem [20, 88], and then investigate acceleration methods within the optimization context using, for example, variational analysis [85], extra-gradient [88] and differential equation [89]. In addition to these mentioned research works, we refer the reader to [19] for a review on the optimization-based approach. On the other

hand, as depicted in Figure 3, a learning process in general is a feedback system, and it is not surprising that control theory can play a part in designing the acceleration. For example, recent studies on reinforcement learning demonstrate that passivity-based control theory can be leveraged in designing high-order learning algorithms [66, 90], where the learning rule is treated as the control law to be designed. Another paper [91] promotes the use of memory in best response maps to accelerate convergence in Nash seeking, and demonstrates substantial improvements in doing so. In addition to the mentioned references, we further refer the reader to [92] for a review on control-theoretic approaches on distributed Nash equilibrium seeking, and to [93] for the use of extreme seeking in the learning process.

The recent advance on the third research direction is in part driven by multi-agent reinforcement learning and its applications such as multi-agent robotic control [10, 94, 95]. Different from the first two directions where the learning dynamics is primarily studied in the context of repeated games, the third research direction focuses on games with dynamic information (see Section 2.2). In this context, the appropriate learning objective, out of practical consideration [16], is to obtain stationary strategies that are subgame perfect [96] (see Section 2.2 for the definition of subgame perfectness). Different from the first two where the change to payoffs resulted from a certain action completely comes from the opponents' move, in dynamic games, the feedback each player receives not only depends on other players' moves but also the dynamic environment. Moreover, when making decisions at each state, players have to trade off current stage payoff for estimated future payoffs while forming predictions on the opponent's strategies. Dynamic trade-off makes the analysis of learning in stochastic games potentially challenging [97].

Earlier works on seeking for such Markov perfect Nash equilibrium are largely based on dynamic programming [98, 99], which requires a global information feedback, a restrictive assumption in practice. Recent efforts focus on various approaches to lessen this requirement. Currently, there are mainly three lines of research regarding learning in dynamic games. The first approach is to extend learning dynamics in repeated games to dynamic games. Built upon similar ideas in best response dynamics (BR-d), two-timescale best response dynamics for zero-sum Markov games have been considered in [97, 100], meanwhile the gradient play has been investigated in linear-quadratic dynamic zero-sum games [61, 101, 102]. The key challenge in the approach, particularly in the case of Markov games, is to properly construct the score function, which balance current stage payoffs and the future payoffs, and we refer the reader to the mentioned references for more details and to [81] for an overview. The second approach is to extend learning methods in single-agent Markov decision process to Markov games. However, the direct extension of methods such as Q-learning [103], policy gradient [31] and actor-critic [49] often fail to deliver desired results due to the non-stationarity issue [104]. One natural way to overcome the non-stationarity issue is to allow players to exchange information with neighbors [105, 106], by which enables players to jointly identify the non-stationarity created by the dynamic environment. For more details regarding this approach, we refer the reader to recent reviews [81, 104]. Finally, the third approach is about a unilateral viewpoint of dynamic games. Different from the first two approaches where learning processes are still investigated in a competitive environment, the third one interprets learning in Markov games as an online optimization problem [107, 108], where players independently make decisions based on the received feedback. This approach accounts for the fully decentralized learning, where from each player's perspective, other players are

considered as part of the environment. The key idea of this approach is to leverage the regret minimization technique [15], which has led to many successes in solving extensive form games of incomplete information [109]. Despite recent advances regarding the first two approaches [61, 81, 97, 100, 110] and positive results for the last one [107, 108, 111], we still lack a unified framework and a thorough understanding regarding the learning process in general Markov games, which remains an open area for researchers from diverse communities.

## 4 Game-Theoretic Learning over Networks

Learning in games is not only intellectually interesting but also practically useful. When combined with game-theoretic modeling, such learning methods, thanks to their decentralized and adaptive nature, provide a comprehensive tool kit for designing resilient, agile, and computationally efficient controls or mechanisms for diverse applications of networks.

In this section, we demonstrate that such a combination of game-theoretic models and associated learning dynamics, referred to as game-theoretic learning, has become indispensable for modern network problems. On the one hand, these networks often admit complex topological structures and heterogeneous nodes, resulting in large-scale complex systems, making centralized controls or mechanisms either impractical or costly. By contrast, game-theoretic models treat each node in the network as a rational and self-interested player, and the heterogeneous nature is captured by players' distinct utilities and action sets as well as information available to them, leading to a bottom-up approach for designing decentralized and scalable mechanisms and controls. On the other hand, modern networked systems, such as wireless communication networks and the smart grid, operate in a dynamic or an adversarial environment, calling for learning-based mechanisms that are responsive to changes in the environment or malicious attacks from adversaries. As shown in the last section, game-theoretic learning provides a self-adaptive procedure for each player in the system, according to which players adjust their moves based on feedback from the environment, resulting in desired collective behaviors.

Thanks to its advantages over the centralized approach, game-theoretic learning has gained much popularity among researchers working on multi-agent systems and network applications. There have been numerous encouraging successes in many fields, ranging from wireless and IoT communication networks [112–116], the smart grid and power networks [2, 3, 117, 118], infrastructure systems [119–122], to cybersecurity applications [123–125, 125–127]. In the following, some representative works in these fields are presented. To be specific, the focus of this section is on the applications of learning methods in wireless communications, the smart grid, and distributed machine learning, while other related applications will be briefly discussed at the end of the section.

### 4.1 Next-Generation Wireless Networks

The next-generation wireless communication technologies offer an accommodating and adaptive solution that meets the requirements of a diverse range of use cases within a common network infrastructure, providing the necessary flexibility for service heterogeneity and compatibility [7]. Such architecture, as pointed out in [128], aims to meet following demands:

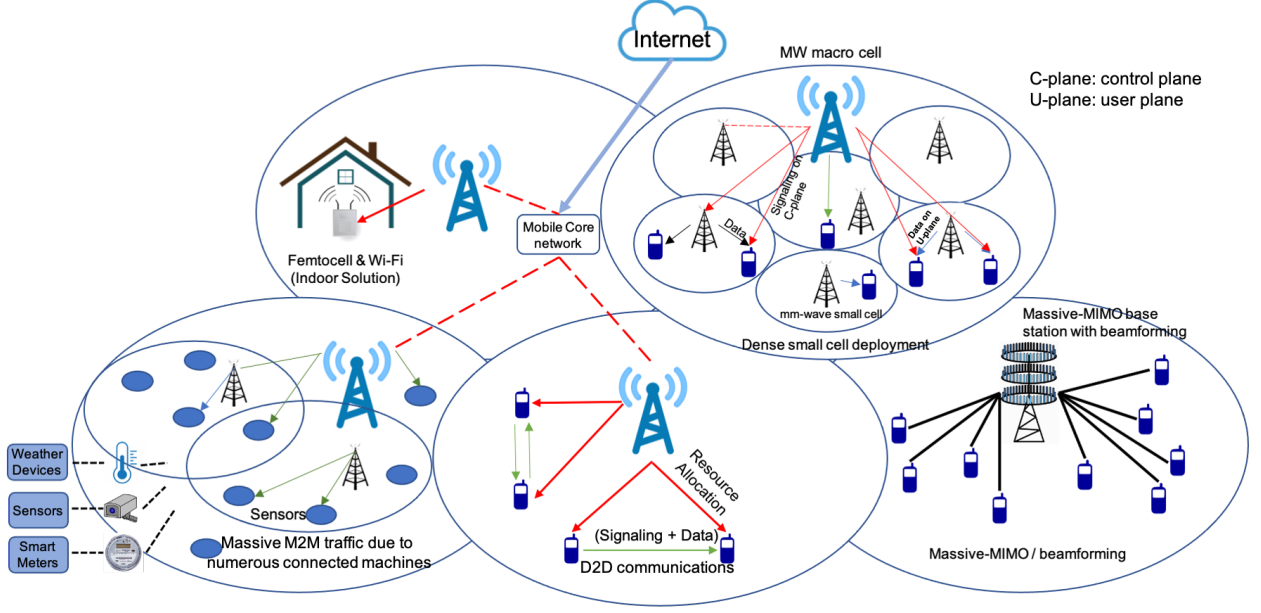


Figure 7: The next generation of communication network: macrocells (bands  $< 3$  GHz); small cells (millimeter-wave); femtocells and Wi-Fi (millimeter-wave); massive multiple-input, multiple-output with beamforming; and device-to-device (D2D) and machine-to-machine (M2M) communications. Solid arrows indicate wireless links, whereas the dashed arrows indicate backhaul links.

- increased indoor and small cell/hotspot traffic, which will make up the majority of mobile traffic volume, leading to complex network structures;
- higher numbers of connected heterogeneous devices stemming from the Internet of Things (IoT), which will support massive machine-to-machine (M2M) communications and applications;
- improved energy consumption or efficient power control for reducing carbon footprint.

From a system science perspective, these requirements impose a large-scale, time-variant, and heterogeneous network topology on modern wireless communication systems, as shown in Figure 7. Hence, it is impractical to manage/secure the wireless communications network centrally. Game-theoretic learning provides a scalable distributed solution with adaptive attributes to deal with this challenge. In the following, we take the dynamic secure routing mechanism as an example to illustrate how game-theoretic learning contributes to a resilient and agile communication system.

Security of routing in a distributed cognitive network (CR) is a prime issue, as the routing may be compromised by unknown attacks, malicious behaviors, and unintentional misconfigurations, which makes it inherently fragile. Even with appropriate cryptographic techniques, routing in CR networks is still vulnerable to attacks in the physical layer, which can critically compromise performance and reliability. Most of the existing work focuses on the resource allocation perspective, which fails to capture the user's lack of knowledge of the attacker due to the distributed mechanism. To address these issues, [113] provides a

learning-based secure scheme, which allows the network to defend against unknown attacks with a minimum level of deterioration in performance.

Consider  $\mathcal{G}_w := (\mathcal{N}_w, \mathcal{E}_w)$ , which is a topology graph for a multi-hop CR network, where  $\mathcal{N}_w = \{n_1, n_2, \dots, n_N\}$  is a set of secondary users, and  $\mathcal{E}_w$  is a set of links connecting these users. The system state  $s$  indicates whether the primary users occupy nodes. The objective of the secondary user is to find an optimal path to its destination. In multi-hop routing, a secondary user  $n_i$  starts with exploring neighboring nodes that are not occupied and then chooses a node among them to which the user routes data. The selected node initializes another exploration process for discovering the next node, and the same process is repeated until the destination is reached.

Let  $\mathcal{P}_i(0, L_i) := \{(n_i, l_i), l_i \in \{0, 1, 2, \dots, L_i\}\}$  be the multi-hop path from the node  $n_i$  to its destination, where  $L_i$  is the total number of explorations until it reaches its destination. Suppose there are  $J$  jammers in the network, the set of which is given by  $\mathcal{J} := \{1, 2, \dots, J\}$ . Let  $\mathcal{R}_j$ ,  $j \in \mathcal{J}$ , be the set of nodes under the influence of jammer  $j$ . Denote the joint action of the jammers by  $\mathbf{r} = [r_j]_{j \in \mathcal{J}}$ , where  $r_j \in \mathcal{R}_j$ . A zero-sum game formulation is proposed in [113], where the secondary users aim to find an optimal routing path by selecting  $\mathcal{P}_i(0, L_i)$ , while the jammers aim to compromise the data transmission by choosing  $\mathbf{r}$ . The expected utility function is

$$\mathbb{E}_s[u_i(s, \mathcal{P}_i(0, L_i), \mathbf{r})] = -\mathbb{E}_s \left[ \sum_{l_i=1}^{L_i} \left( \ln q_{(n_i, l_i-1)}^{(n_i, l_i)} + \lambda \tau_{(n_i, l_i-1)}^{(n_i, l_i)} \right) \right],$$

where  $q_{(n_i, l_i-1)}^{(n_i, l_i)}$  is the probability of successful transmissions from node  $(n_i, l_i - 1)$  to node  $(n_i, l_i)$ , and  $\lambda \tau_{(n_i, l_i-1)}^{(n_i, l_i)}$  is the transmission delay between these two nodes. Here, the expectation  $\mathbb{E}_s[\cdot]$  is taken over all the possible system states.

Due to a lack of complete knowledge of adversaries and payoff structures, Boltzmann-Gibbs reinforcement learning (SBR-d) is utilized to find the optimal path because of its capability of estimating the expected utility. The resulting secure routing algorithm can spatially circumvent jammers along the routing path and learn to defend against malicious attackers as the state changes. As shown in Figure 8, the routing path generated from the proposed routing algorithm in [113, 129] can avoid the nodes compromised by the jammers. Thus, the routing algorithm stemming from the proposed game-theoretic formulation provides more resilience, security, and agility than the ad-hoc on-demand distance vector (AODV) algorithm, as AODV fails to dynamically adjust the routing path in the case of a malicious attack. Moreover, the proposed routing algorithm can reduce the delay time incurred by the attack due to its adaptive and dynamic feature, and hence, is more efficient than AODV.

## 4.2 The Smart Grid

Gradual replacements of conventional energies with renewable energies greatly help with the reduction of greenhouse gases and the mitigation of climate change. More and more microgrids are being integrated with the main power grid, which are green systems that rely on renewable distributed resources such as wind turbines and fuel cells. As shown in Figure 9,

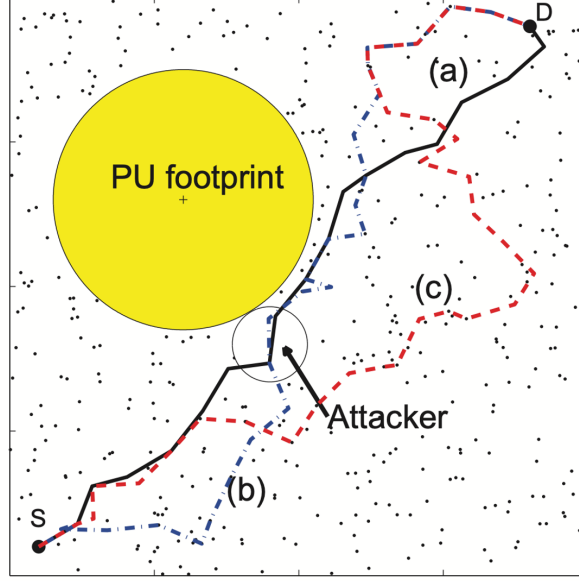


Figure 8: Illustration of a random network topology for 500 secondary users with a source (S) and a destination (D), and routes of AODV and the proposed secure routing algorithm in 2 km by 2 km area. The PU footprint denotes the set of nodes unavailable to secondary users. Without an attacker, AODV establishes the route path (a), described by the solid line, while the route path (b), the blue dashed line, is generated by the Boltzmann-Gibbs learning method. Even though the AODV path is the shortest path between the source and the destination, it is disrupted by malicious attacks. By contrast, the learning method can develop a new route path (c) that circumvents jammers, leading to a resilient routing mechanism.

the integration of microgrids can enhance the stability, resiliency, and reliability of the power system, as they can operate independently from the main power grid autonomously. Such integration, together with smart meters and appliances, produces the so-called smart grid, a modern infrastructure for the reliable delivery of electricity.

The future smart grid is envisioned as a large-scale cyber-physical system comprising advanced power, communications, control, and computing technologies. To accommodate these technologies employed by different parties in the grid and to ensure an efficient and robust operation of such heterogeneous and large-scale cyber-physical systems, game-theoretic methods have been widely employed in smart grid management problems. In the grid, microgrids are modeled as self-interested players who can operate, communicate, and interact autonomously to deliver power and electricity to their consumers efficiently. Here, we discuss a microgrid management mechanism developed in [117], built on game-theoretic learning, enabling autonomous management of renewable resources.

The system model considered in [117] includes the generators, microgrids, and communications. As shown in Figure 10, generators in the upper layer determine the amount of power to be generated, along with the electricity price, and send them to the bottom layer. A microgrid can generate renewable energies and make decisions by responding to the strategies of the generators and other microgrids to optimize their payoffs, specified in the following

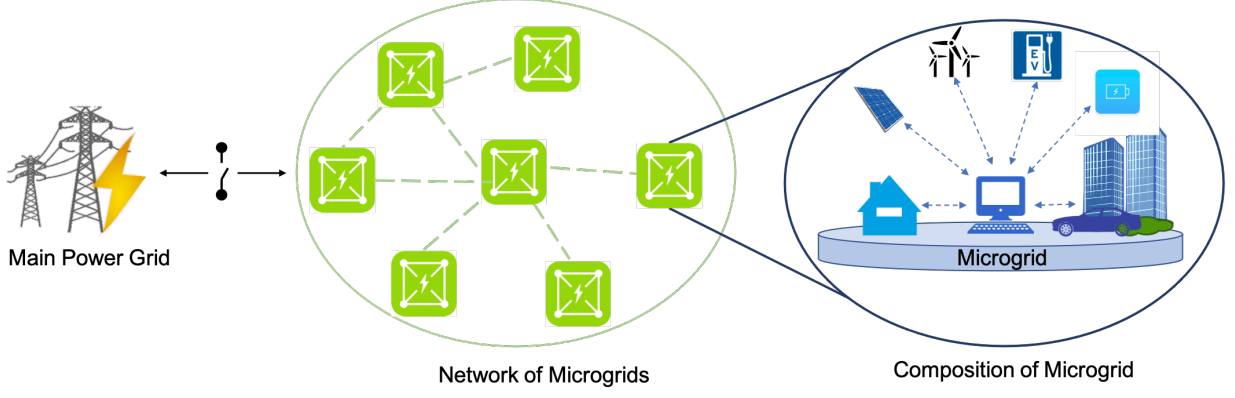


Figure 9: The integration of microgrids. A microgrid consists of a controller, consumers, generators, and energy storage. In the grid, microgrids can either be connected to the main grid or other microgrids, and these networked microgrids can operate, communicate, and interact autonomously to deliver power and electricity to their consumers efficiently.

game-theoretic model.

Let  $\mathcal{N}_d = \{r, 1, 2, \dots, N_d\}$  be the set of  $N_d + 1$  buses in a power grid, where  $r$  denotes the slack bus. Assume that a smart grid is composed of load buses and generator buses and let  $p_i^g$ ,  $p_i^l$  and  $\theta_i$  be, respectively, the power generation, power load, and voltage angle at the  $i$ -th bus. Note that the active power injection at the  $i$ -th bus satisfies

$$p_i = p_i^g - p_i^l, \quad \forall i \in \mathcal{N}_d,$$

while the balance of the grid gives  $\sum_{i \in \mathcal{N}_d} p_i^g = \sum_{i \in \mathcal{N}_d} p_i^l$ . Let  $\mathcal{N} := \{1, 2, \dots, N\} \subseteq \mathcal{N}_d$  be the set of  $N$  buses that can generate renewable energies, such as wind power, solar power, etc.

In the game considered in [117], the utility function of the  $i$ -th bus measures not only economic factors related to power generation but also the efficiency of the microgrids. Before giving the mathematical definition of the utility function, we first introduce the following notations. Let  $c_i$  be the unit cost of generated power for the  $i$ -th player, and  $c$  the unit price of renewable energy for sale defined by the power market.  $c_i, c$  are quantities relevant to the profit gained by the bus. For the efficiency part, denote by  $r_i$  a weighting parameter that measures the importance of regulations of voltage angle at the  $i$ -th bus. Further,  $[s_{ij}]_{i,j \in \mathcal{N}_d} = -[b_{ij}]_{i,j \in \mathcal{N}_d}^{-1}$ , where  $b_{ij}$  is the imaginary part of the element  $(i, j)$  in the admittance matrix of the power grid. Moreover, each microgrid has a maximum generation, denoted by  $\bar{p}_i^g$ . Finally, we note that as a physical constraint,  $[s_{ij}]$  and  $[p_i]$  satisfy (23) due to the power flow equation [117]

$$\sum_{j \in \mathcal{N}_d \setminus \mathcal{N}} s_{ij} p_j + \sum_{j \neq i \in \mathcal{N}} s_{ij} p_j = \theta_i - s_{ii} p_i, \quad \forall i \in \mathcal{N}, \quad (23)$$

where  $\theta_i$  is the voltage angle of the  $i$ -th bus. With all the notations above, the utility function of the  $i$ -th bus is defined as

$$u_i(p_i^g, p_{-i}^g) := -c_i p_i^g - c(p_i^l - p_i^g) - \frac{1}{2} r_i^2 \left( \sum_{j \in \mathcal{N}_d} s_{ij} p_j \right), \quad 0 \leq p_i^g \leq \bar{p}_i^g, \quad i \in \mathcal{N}.$$

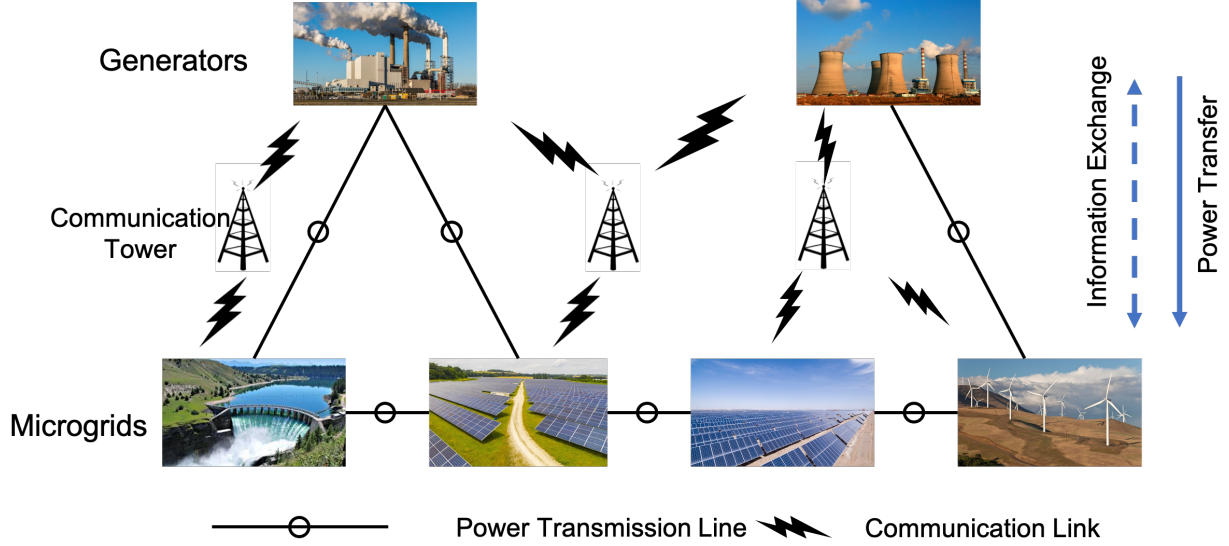


Figure 10: Smart grid hierarchy model. The upper layer containing conventional generators forms a generator network, and the distributed renewable energy generators in the bottom layer constitute the microgrid network; the information exchange, such as the electricity market price and the amount of power generation, between the two layers are through the communication network layer in the middle.

Three learning methods are proposed in the paper to seek the Nash equilibrium, all based on best response dynamics (10). The first two algorithms are parallel-update algorithm (PUA) and random-update algorithm (RUA) studied in [112]. PUA is essentially the best response algorithm we represent in (10), with the learning rate  $\lambda_i^k$  being zero for all  $i$ , and all players update their strategies in parallel. As its name suggests, RUA incorporates randomness into the best response algorithm, resulting in an  $\epsilon$ -greedy best response algorithm: players update their strategies according to (10) with probability  $1 - \epsilon$ , with  $\epsilon \in (0, 1)$  and retain their previous strategies otherwise. When  $\epsilon = 0$ , players constantly update their strategies in every round; in this case, RUA reduces to PUA.

However, as special cases of (10), PUA and RUA require global information regarding the grid, including the specific generated power of generators and other players' active power injections, which are assumed to be private in practice. Hence, to implement these algorithms, communication networks are needed to broadcast information to players, which is costly and not confidential. As a possible remedy, we can consider incorporating utility estimation and using smoothed best response dynamics (SBR-d) as in the wireless setting. Another more straightforward approach, as shown in the paper, is to modify the best response algorithm by using the power flow equations in the smart grid. Based on a phasor measurement unit (PMU), the third algorithm, termed PMU-enabled distributed algorithm (PDA), enables each player to compute the aggregation of others' actions, and the only information needed is the player's voltage angle  $\theta_i$ . Therefore, by taking into account the power flow equation (23), a player does not need other players' private information of active power injection when using PDA, as shown in Figure 11. Compared with the other two, PDA requires much less information and is more self-dependent as players only need their current voltage angles  $\theta_i$ ,



and the common knowledge of the electricity price.

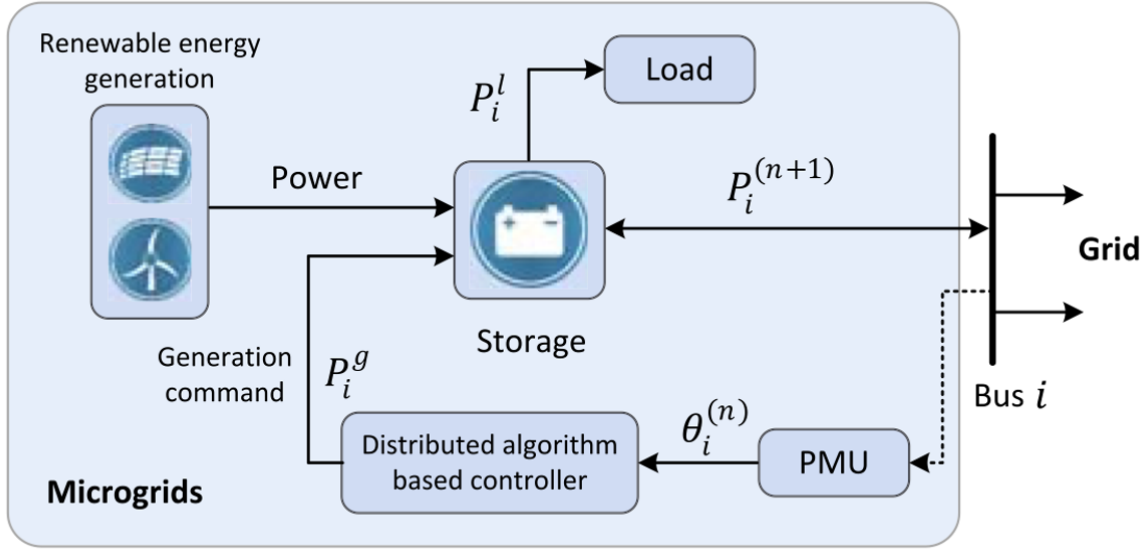


Figure 11: The framework to implement the PMU-enabled distributed algorithm. PMU measures the voltage angle at the bus, and the controller generates a command regarding the amount of microgrid renewable energy injection from the local storage to the grid based on the received voltage angle.

As indicated in [117], the effectiveness and resiliency of the algorithm have been validated via case studies based on the IEEE 14-bus system: the game-theory-based distributed algorithm not only can converge to the unique Nash equilibrium but also provides strong resilience against fault models (generator breakdown, microgrid turn-off, and open-circuit of the transmission line, etc.) and attack models (data injection attacks, unavailability of PMU data and jamming attacks, etc.). The strong resilience enables the microgrids to operate appropriately in unanticipated situations. Moreover, the distributed algorithm enables autonomous management of renewable resources and the plug-and-play feature of the smart grid. The proposed learning algorithm only requires the players to have common knowledge without revealing their private information, which increases security and privacy and reduces communication overhead.

### 4.3 Distributed Machine Learning over Networks

The rise of Big Data has led to new demands for large-scale machine learning systems that promise adequate capacity to digest massive data sets and offer powerful predictive analytics. With the unrestrainable growth of data, large-scale machine learning needs to address new challenges regarding the scalability and efficiency of learning algorithms concerning computational and memory resources. Compared with classical machine learning approaches that are designed to learn from a single integrated data set, one of the promising research lines of large-scale machine learning is distributed machine learning over networks (DMLON), which

aims to develop efficient and scalable algorithms with reasonable requirements of memory computation resources, by allocating the learning processes among several networked computing units with distributed data sets.

The key feature of DMLON is that data sets are stored and processed locally on these computing units, which enables distributed and parallel computing schemes in large-scale machine learning systems. Compared with centralized approaches, distributed machine learning avoids maintaining and mining a central data set and preserves data privacy, as these networked units exchange knowledge about the learned models without exchanging raw private data.

Based on the idea of “local learning and global integration,” DMLON can utilize different learning processes to train several models from distributed data sets and then produce an integration of learning models that can increase the possibility of achieving higher accuracy, especially on a large-size domain. For example, in federated learning [130], the global integration is created by a third-party coordinator other than computing units, which makes networked computing units collaboratively train a machine learning model using their data in security. On the other hand, as indicated in [131], such a global integration can also stem from the collective patterns of local learning without external enforcement. The key behind this bottom-up integration is that each computing unit is modeled as a self-interested player who learns the learning model based on the local data set and the feedback from its neighbors. It has been shown in the paper that by modeling DMLON as a noncooperative game, game-theoretic learning methods lead to a communication-efficient distributed machine learning, where the global outcome is characterized by the Nash equilibrium, resulting from players’ self-adaptive behaviors.

Specifically, the networked system of computing units is described by a graph with the set of nodes  $\mathcal{N}_m := \{1, 2, \dots, N\}$  representing these units. Each node  $i \in \mathcal{N}_m$  possesses local data that cannot be transferred to other nodes. In the game model considered in [131, 132], instead of fixing the network topology, nodes can determine the network’s connectivity based on their attributes when they perform learning tasks, resulting in a network formation game. In mathematical terms, the action of node  $i$  consists of two components: the learning parameter  $\theta_i \in \mathbb{R}^d$ , and the network formation parameter  $e_i \in \mathbb{R}^{N-1}$ . The first component  $\theta_i$  corresponds to the weights or parameters of the machine learning model, which captures the local learning process at node  $i$ , and the corresponding empirical loss, given the local data, is denoted by  $L_i(\theta_i)$ . In addition to this learning parameter  $\theta_i$ , the network formation parameter  $e_i$  plays an important role in bringing up the global integration. The parameter  $e_i := (e_i^j)_{j \neq i, j \in \mathcal{N}} \in [0, 1]^{N-1}$  denotes concatenation of weights on the directed edges from node  $i$  to other nodes, where  $e_i^j$  can be interpreted as the attention node  $i$  pays to the local learning at node  $j$ , and this further influence the communication between the nodes. Each node can communicate with its neighbors during the distributed learning process to exchange learning parameters if their objectives are aligned. Otherwise, the corresponding edge weight  $e_i^j$  is set to zero. For node  $i$ , the communication cost is  $C_i(\theta_i, \theta_{-i}, e_i)$ . In the game considered in [131], each node aims to maximize its utility function, defined as

$$u_i(\theta_i, \theta_{-i}, e_i, e_{-i}) := -L_i(\theta_i) - C_i(\theta_i, \theta_{-i}, e_i),$$

In this definition, the first term  $L_i(\theta_i)$  captures the local learning process at node  $i$ , whereas

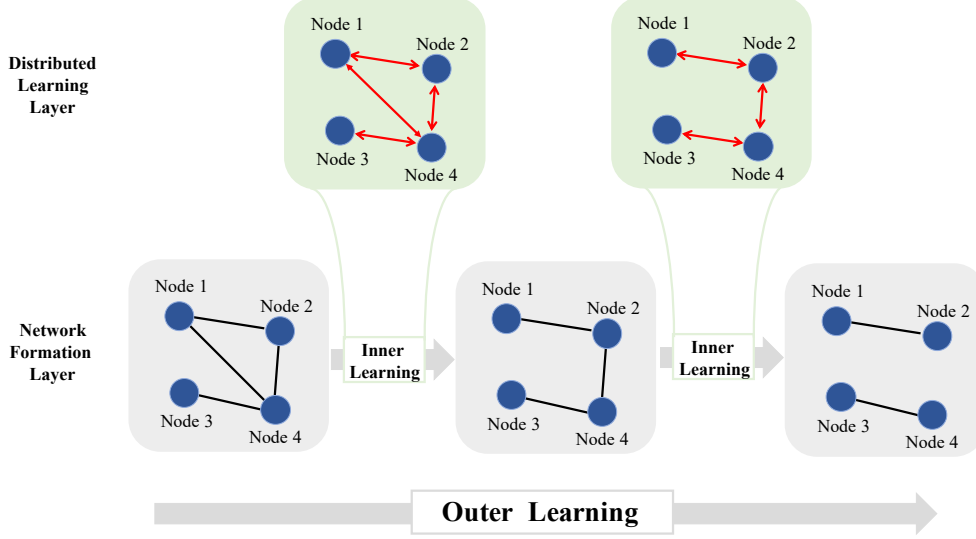


Figure 12: A schematic representation of two-layer learning. The directed red lines stand for the communication between nodes. In the network formation layer, the nodes learn to eliminate/establish links with other nodes to achieve efficient communication. In the distributed machine learning layer, the nodes communicate their parameters with their neighbors and perform their learning tasks.

the second term  $C_i(\theta_i, \theta_{-i}, e_i)$  depicts the interactions among nodes. The objective of each node is to improve the performance of learning while reducing the communication overhead.

A two-layer learning approach is proposed in [131] to find the Nash equilibrium of the game, and a schematic representation is provided in Figure 12. The outer layer corresponds to network formation learning, where each node decides its network formation parameter  $e_i$  with the learning parameter fixed, and the joint parameters of all nodes  $e = (e_i)_{i \in \mathcal{N}_m}$  give rise to new network topology, leading to efficient communication. In network formation learning, each node decides their optimal parameter  $e_i$  by gradient play (GD), and computing the individual payoff gradient  $\nabla_{e_i} u_i(\theta_i, \theta_{-i}, e_i, e_{-i})$  relies on the stabilized learning parameters  $\theta_i, \theta_{-i}$  given by the inner layer: distributed learning layer. In this inner learning, the network formation parameter is fixed, and each node implements online mirror descent (MD) for seeking the Nash equilibrium with the local feedback under the current network topology, as the networked nodes can exchange information with their neighbors.

Compared with existing works on distributed machine learning, the game-theoretic method studied in [131] enables distributed machine learning over strategic networks. On the one hand, the global outcome characterized by the Nash equilibrium is self-enforcing, resulting from the coordinated behaviors of independent computing compared with the external enforcing one in federated learning. This bottom-up approach scales efficiently when additional computing units are introduced into the system. On the other hand, the strategic interactions over the network, described by the network formation decision of each node, create a network intelligence that allows each computing unit to adaptively adjust the underlying topology, resulting in a desired distributed learning pattern that minimizes communication costs during the learning process.

## 4.4 Emerging Network Applications

From the examples above, game-theoretic learning provides a natural scalable design framework to create network intelligence for autonomous control, management, and coordination of large-scale complex network systems with heterogeneous parties. In the following, we offer some thoughts regarding various applications of game-theoretic learning in a broader context, showing that such a design framework is pervasive for diverse network problems.

Interdependent infrastructure networks, including wireless communication networks and the smart grid, play a significant role in modern society, where Internet-of-Things (IoT) devices are massively deployed and interconnected. These devices are connected to cellular/cloud networks, creating multi-layer networks, referred to as networks-of-networks [133]. The smart grid is one prominent example, where wireless sensors collect the data of buses and power transmission lines, forming a sensor network built on the power networks for grid monitoring and decision planning purposes [134]. Besides, the networks-of-networks model has also been extensively studied in other infrastructure networks. For instance, in an intelligent transportation network, apart from vehicle-to-vehicle (V2V) communications, vehicles can also communicate with roadside infrastructures or units belonging to one or several service providers to exchange various types of data related to different applications, such as GPS navigation. In this case, the vehicles form one network while the infrastructure nodes form another network, and the interconnections between the two networks lead to the intelligent management and operation of modern transportation networks.

Due to interdependent networks' heterogeneous and multi-tier features, the required management mechanisms or controls can vary for different networks. For example, the connectivity of sensor networks in smart grids or V2V communication networks requires higher security levels than the infrastructure networks, as cyberspace is more likely to be targeted by adversaries [135]. Therefore, to manage and secure interdependent infrastructure networks, game-theoretic learning methods, especially heterogeneous learning [40, 47], can be used to design decentralized and resilient mechanisms that are responsive to attacks and adaptive to the dynamic environment, as different parties in interdependent infrastructure networks may acquire different information. For further readings on this topic, we refer the reader to [47, 133] and references therein.

Similar to distributed optimization and machine learning based on game-theoretic learning, the control of autonomous mobile robots can also be cast as a Nash equilibrium seeking problem over networks, where the equilibrium is viewed as the desired coordination of all robots [94, 95]. For applications of this kind, where the nature of robot movements determines the network topologies, dynamic games over networks are considered, and corresponding learning algorithms are employed. Based on their observations of the surroundings, robots rely on game-theoretic learning, for example, reinforcement learning, for developing self-rule policies, leading to a need for decentralized and scalable control laws for multi-agent robotic systems. Moreover, reinforcement learning has proven effective for real-world multi-agent robotic control when combined with powerful function approximators, such as deep neural networks. This area of research, termed deep multi-agent reinforcement learning [81, 136], is growing rapidly and attracting the attention of researchers from machine learning, robotics as well as control communities.

In addition to these prescriptive mechanisms in engineering practices, game-theoretic

learning also provides a descriptive model for studying human decision-making and strategic interactions in epidemiology and social sciences, where the Nash equilibrium represents a stable state of the underlying noncooperative game. For example, a differential game model has been proposed in [137] to study the virus or diseases spreading over the network. Authors have developed a decentralized mitigation mechanism for controlling the spreading. Such an approach has been further explored in [138], where an optimal quarantining strategy of suppressing two interdependent epidemics spreading over complex networks has been proposed and proven robust against random changes in network connections.

## 5 Summary

This article provides a comprehensive overview of game theory basics and related learning theories, which serve as building blocks for systematically treating multi-agent decision-making over networks. We have elaborated on the game-theoretic learning methods for network applications drawn from spanning emerging areas such as the next-generation wireless networks, the smart grid, and networked machine learning. In each area, we have identified the main technical challenges and discussed how game theory can be applied to address them in a bottom-up approach.

From the surveyed works, we conclude that noncooperative game theory is the cornerstone of decentralized mechanisms for large-scale complex networks with heterogeneous entities, where each node is modeled as an independent decision-maker. The resulting collective behaviors of these rational decision-makers over the network can be mathematically depicted by the solution concept: Nash equilibrium. In addition to various game models, learning in games is of great significance for creating distributed network intelligence, which enables each entity in the network to respond to unanticipated situations, such as malicious attacks from adversaries in cyber-physical systems [134]. Under local or individual feedback, the introduced learning dynamics lead to a decentralized and self-adaptive procedure, resulting in desired collective behavior patterns without external enforcement.

Beyond the existing successes of game-theoretic learning, which mainly focuses on learning in static repeated games, it is also of interest to investigate dynamic game models and associated learning dynamics, in order to better understand the decision-making process in dynamic environments. The motivation for studying dynamic models and related learning theory stems, on the one hand, from the pervasive presence of time-varying network structures, such as generation and demand in the smart grid [117]. On the other hand, by defining auxiliary state variables, the problem of decision-making under uncertainties can be modeled as a dynamic game, where the state of the game includes the hidden information players do not have access to when making decisions. For example, the state variable can capture the uncertainty of the environment, as we have discussed in the context of the dynamic routing problem [113], or it can describe the global status of the entire system, as we have shown in the example of distributed optimization [139]. The dynamic game models not only simplify the construction of players' utilities and actions, providing a clear picture of the strategic interactions under uncertainties in the dynamic environment but can also offer a scalable design framework for prescribing players' self-adaptive behaviors that lead to equilibrium states under various feedback structures.

To recap, this article has presented a comprehensive overview of game-theoretic learning and its potential for tackling the challenges emerging from network applications. The combination of game-theoretic modeling and related learning theories constitutes a powerful tool for designing future data-driven network systems with distributed intelligent entities, which serve as the bedrock and a key enabler for resilient and agile control of large-scale artificial intelligence systems in the near future.

## References

- [1] M. O. Jackson, *Social and Economic Networks*. Princeton, NJ: Princeton University Press, 2010.
- [2] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Başar, “Dependable demand response management in the smart grid: A Stackelberg game approach,” *IEEE Transactions on Smart Grid*, vol. 4, no. 1, pp. 120–132, 2013.
- [3] Q. Zhu, Z. Han, and T. Başar, “A differential game approach to distributed demand side management in smart grid,” in *2012 IEEE International Conference on Communications (ICC)*, pp. 3345–3350, 2012.
- [4] N. Groot, B. De Schutter, and H. Hellendoorn, “Toward system-optimal routing in traffic networks: A reverse Stackelberg game approach,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 1, pp. 29–40, 2014.
- [5] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge University Press, 2012.
- [6] Q. Zhu, Z. Yuan, J. B. Song, Z. Han, and T. Başar, “Interference aware routing game for cognitive radio multi-hop networks,” *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 10, pp. 2006–2015, 2012.
- [7] Z. Han, D. Niyato, W. Saad, and T. Başar, *Game Theory for Next Generation Wireless and Communication Networks: Modeling, Analysis, and Design*. Cambridge University Press, 2019.
- [8] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Başar, and J.-P. Hubaux, “Game theory meets network security and privacy,” *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, pp. 1–39, 2013.
- [9] Q. Zhu and T. Başar, “Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: games-in-games principle for optimal cross-layer resilient control systems,” *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 46–65, 2015.
- [10] P. Stone and M. Veloso, “Multiagent systems: a survey from a machine learning perspective,” *Autonomous Robots*, vol. 8, no. 3, pp. 345–383, 2000.
- [11] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory, 2nd Edition*. Society for Industrial and Applied Mathematics, 1998.
- [12] D. Fudenberg and J. Tirole, *Game Theory*. Cambridge, MA: MIT Press, 1991.
- [13] M. Maschler, E. Solan, and S. Zamir, *Game Theory*. Cambridge University Press, 2013.

- [14] M. O. Jackson and Y. Zenou, “Chapter 3 Games on Networks,” *Handbook of Game Theory with Economic Applications*, vol. 4, pp. 95–163, 2015.
- [15] S. Shalev-Shwartz, “Online learning and online convex optimization,” *Foundations and Trends® in Machine Learning*, vol. 4, no. 2, pp. 107–194, 2011.
- [16] M. L. Littman, “Markov games as a framework for multi-agent reinforcement learning,” in *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, ICML’94, (San Francisco, CA, USA), p. 157–163, Morgan Kaufmann Publishers Inc., 1994.
- [17] J. Lei and U. V. Shanbhag, “Stochastic Nash equilibrium problems: Models, analysis, and algorithms,” *submitted as part of CSM special issue*, 2020.
- [18] J. B. Rosen, “Existence and uniqueness of equilibrium points for concave N-person games,” *Econometrica*, vol. 33, no. 3, pp. 520–534, 1965.
- [19] G. Belgioioso, P. Yi, S. Grammatico, and L. Pavel, “Distributed generalized nash equilibrium seeking: An operator theoretic perspective,” *submitted as part of CSM special issue*, 2020.
- [20] E. Cavazzuti, M. Pappalardo, and M. Passacantando, “Nash equilibria, variational inequalities, and dynamical systems,” *Journal of Optimization Theory and Applications*, vol. 114, no. 3, pp. 491–506, 2002.
- [21] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. USA: John Wiley & Sons, 1st ed., 1994.
- [22] R. Selten, “Reexamination of the perfectness concept for equilibrium points in extensive games,” *International Journal of Game Theory*, vol. 4, no. 1, pp. 25–55, 1975.
- [23] T. Başar, “Time consistency and robustness of equilibria in non-cooperative dynamic games,” in *Dynamic Policy Games in Economics*, vol. 181 of *Contributions to Economic Analysis*, pp. 9 – 54, Elsevier, 1989.
- [24] D. Fudenberg, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.
- [25] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, “The complexity of computing a Nash equilibrium,” *SIAM Journal on Computing*, vol. 39, no. 1, pp. 195–259, 2009.
- [26] P. D. Taylor and L. B. Jonker, “Evolutionary stable strategies and game dynamics,” *Mathematical Biosciences*, vol. 40, no. 1-2, pp. 145–156, 1978.
- [27] S. Hart and A. Mas-Colell, “Uncoupled dynamics do not lead to Nash equilibrium,” *The American Economic Review*, vol. 93, no. 5, pp. 1830–1836, 2003.
- [28] J. R. Marden and J. S. Shamma, “Chapter 16 Game Theory and Distributed Control,” *Handbook of Game Theory with Economic Applications*, vol. 4, pp. 861–899, 2015.



- [29] V. S. Borkar, *Stochastic Approximation, A Dynamical Systems Viewpoint*, vol. 48 of *Springer*. Springer, 2008.
- [30] M. Benaïm, J. Hofbauer, and S. Sorin, “Stochastic approximations and differential inclusions,” *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.
- [31] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 2018.
- [32] D. S. Leslie and E. J. Collins, “Convergent multiple-timescales reinforcement learning algorithms in normal form games,” *The Annals of Applied Probability*, vol. 13, pp. 1231–1251, 11 2003.
- [33] C. Harris, “On the Rate of Convergence of Continuous-Time Fictitious Play,” *Games and Economic Behavior*, vol. 22, no. 2, pp. 238–259, 1998.
- [34] J. Hofbauer and K. Sigmund, “Evolutionary game dynamics,” *Bulletin of the American Mathematical Society*, vol. 40, no. 4, pp. 479–519, 2003.
- [35] G. W. Brown, “Iterative solution of games by fictitious play,” *Activity Analysis of Production and Allocation*, vol. 13, no. 1, pp. 374–376, 1951.
- [36] V. Krishna and T. Sjöström, “On the convergence of fictitious play,” *Mathematics of Operations Research*, vol. 23, no. 2, pp. 479–511, 1998.
- [37] P. Mertikopoulos and Z. Zhou, “Learning in games with continuous action sets and unknown payoff functions,” *Mathematical Programming*, vol. 173, no. 1-2, pp. 465–507, 2018.
- [38] P. Mertikopoulos and W. H. Sandholm, “Learning in games via reinforcement and regularization,” *Mathematics of Operations Research*, vol. 41, no. 4, pp. 1297–1324, 2016.
- [39] R. D. McKelvey and T. R. Palfrey, “Quantal response equilibria for normal form games,” *Games and Economic Behavior*, vol. 10, no. 1, pp. 6–38, 1995.
- [40] Q. Zhu, H. Tembine, and T. Başar, “Heterogeneous learning in zero-sum stochastic games with incomplete information,” *49th IEEE Conference on Decision and Control (CDC)*, pp. 219–224, 2010.
- [41] Y. Nesterov, “Primal-dual subgradient methods for convex problems,” *Mathematical Programming*, vol. 120, no. 1, pp. 221–259, 2009.
- [42] Y. Nesterov, “Introductory lectures on convex optimization, a basic course,” *Applied Optimization*, 2004.
- [43] J. M. Smith and G. R. Price, “The logic of animal conflict,” *Nature*, vol. 246, no. 5427, pp. 15–18, 1973.

- [44] W. H. Sandholm, *Population Games and Evolutionary Dynamics*. Cambridge, MA: MIT Press, 2010.
- [45] R. Cressman and Y. Tao, “The replicator equation and other game dynamics,” *Proceedings of the National Academy of Sciences*, vol. 111, no. Supplement 3, pp. 10810–10817, 2014.
- [46] D. S. Leslie and E. Collins, “Generalised weakened fictitious play,” *Games and Economic Behavior*, vol. 56, no. 2, pp. 285–298, 2006.
- [47] Q. Zhu, H. Tembine, and T. Başar, “Hybrid learning in stochastic games and its application in network security,” in *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*, pp. 303–329, John Wiley & Sons, Ltd, 2012.
- [48] G. Neu, A. Jonsson, and V. Gómez, “A unified view of entropy-regularized markov decision processes,” *arXiv preprint arXiv:1705.07798*, 2017.
- [49] V. R. Konda and V. S. Borkar, “Actor-critic-type learning algorithms for Markov decision processes,” *SIAM Journal on Control and Optimization*, vol. 38, no. 1, pp. 94–123, 1999.
- [50] D. S. Leslie and E. J. Collins, “Individual Q-learning in normal form games,” *SIAM Journal on Control and Optimization*, vol. 44, no. 2, pp. 495–514, 2005.
- [51] J. Hofbauer, S. Sorin, and Y. Viossat, “Time average replicator and best-reply dynamics,” *Mathematics of Operations Research*, vol. 34, no. 2, pp. 263–269, 2009.
- [52] P. Mertikopoulos and W. H. Sandholm, “Riemannian game dynamics,” *Journal of Economic Theory*, vol. 177, pp. 315–364, 2018.
- [53] M. Benaïm, J. Hofbauer, and S. Sorin, “Stochastic approximations and differential inclusions, Part II: Applications,” *Mathematics of OR*, vol. 31, pp. 673–695, 11 2006.
- [54] A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, and J. P. How, “A tutorial on linear function approximators for dynamic programming and reinforcement learning,” *Foundations and Trends® in Machine Learning*, vol. 6, no. 4, pp. 375–451, 2013.
- [55] T. Li and Q. Zhu, “On convergence rate of adaptive multiscale value function approximation for reinforcement learning,” *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2019.
- [56] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

- [57] J. C. Spall, “A one-measurement form of simultaneous perturbation stochastic approximation,” *Automatica*, vol. 33, no. 1, pp. 109–112, 1997.
- [58] M. Bravo, D. S. Leslie, and P. Mertikopoulos, “Bandit learning in concave N-person games,” in *Advances in Neural Information Processing Systems 31*, Advances in Neural Information Processing Systems, Curran Associates, Inc., 2018.
- [59] L. Xiao, “Dual averaging methods for regularized stochastic learning and online optimization,” *J. Mach. Learn. Res.*, vol. 11, p. 2543–2596, Dec. 2010.
- [60] P. Mertikopoulos, C. Papadimitriou, and G. Piliouras, “Cycles in adversarial regularized learning,” in *Proceedings of the 2018 Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2703–2717, 2018.
- [61] E. Mazumdar, L. J. Ratliff, and S. S. Sastry, “On gradient-based learning in continuous games,” *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 1, pp. 103–131, 2020.
- [62] J. Hofbauer and S. Sorin, “Best response dynamics for continuous zero-sum games,” *Discrete & Continuous Dynamical Systems - B*, vol. 6, no. 1, pp. 215–224, 2006.
- [63] J. Hofbauer and W. H. Sandholm, “On the global convergence of stochastic fictitious play,” *Econometrica*, vol. 70, no. 6, pp. 2265–2294, 2002.
- [64] S. Perkins and D. S. Leslie, “Asynchronous stochastic approximation with differential inclusions,” *Stochastic Systems*, vol. 2, no. 2, pp. 409–446, 2012.
- [65] A. Heliou, J. Cohen, and P. Mertikopoulos, “Learning with bandit feedback in potential games,” in *Advances in Neural Information Processing Systems 30*, pp. 6369–6378, Curran Associates, Inc., 2017.
- [66] B. Gao and L. Pavel, “On passivity, reinforcement learning, and higher order learning in multiagent finite games,” *IEEE Transactions on Automatic Control*, vol. 66, no. 1, pp. 121–136, 2019.
- [67] E. N. Barron, R. Goebel, and R. R. Jensen, “Best response dynamics for continuous games,” *Proceedings of the American Mathematical Society*, vol. 138, no. 03, pp. 1069–1069, 2010.
- [68] B. Swenson, R. Murray, and S. Kar, “On best-response dynamics in potential games,” *SIAM Journal on Control and Optimization*, vol. 56, no. 4, pp. 2734–2767, 2018.
- [69] B. Swenson, R. Murray, and S. Kar, “Regular potential games,” *Games and Economic Behavior*, vol. 124, pp. 432–453, 2020.
- [70] M. Benaïm, J. Hofbauer, and S. Sorin, “Perturbations of set-valued dynamical systems, with applications to game theory,” *Dynamic Games and Applications*, vol. 2, no. 2, pp. 195–205, 2012.

- [71] J. C. Harsanyi, “Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points,” *International Journal of Game Theory*, vol. 2, no. 1, pp. 1–23, 1973.
- [72] J. Hofbauer and E. Hopkins, “Learning in perturbed asymmetric games,” *Games and Economic Behavior*, vol. 52, no. 1, pp. 133–152, 2005.
- [73] H. P. Young, “The evolution of conventions,” *Econometrica*, vol. 61, no. 1, pp. 57–84, 1993.
- [74] H. P. Young, “Learning by trial and error,” *Games and Economic Behavior*, vol. 65, no. 2, pp. 626–643, 2009.
- [75] J. Gaveau, C. J. Le Martret, and M. Assaad, “Performance analysis of trial and error algorithms,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 6, pp. 1343–1356, 2020.
- [76] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, “Payoff-based dynamics for multiplayer weakly acyclic games,” *SIAM Journal on Control and Optimization*, vol. 48, no. 1, pp. 373–396, 2009.
- [77] B. S. Pradelski and H. P. Young, “Learning efficient Nash equilibria in distributed systems,” *Games and Economic Behavior*, vol. 75, no. 2, pp. 882–897, 2012.
- [78] J. R. Marden, “Selecting efficient correlated equilibria through distributed learning,” *Games and Economic Behavior*, vol. 106, pp. 114–133, 2017.
- [79] J. R. Marden, H. P. Young, and L. Y. Pao, “Achieving Pareto optimality through distributed learning,” *SIAM Journal on Control and Optimization*, vol. 52, no. 5, pp. 2753–2770, 2014.
- [80] Z. Hu, M. Zhu, P. Chen, and P. Liu, “On convergence rates of game theoretic reinforcement learning algorithms,” *Automatica*, vol. 104, pp. 90–101, 2019.
- [81] K. Zhang, Z. Yang, and T. Başar, “Multi-agent reinforcement learning: A selective overview of theories and algorithms,” *arXiv preprint arXiv:1911.10635*, 2019.
- [82] Y. Zhou, M. Kantarcioglu, and B. Xi, “A survey of game theoretic approach for adversarial machine learning,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 3, 2019.
- [83] K. Zhang, B. Hu, and T. Basar, “On the stability and convergence of robust adversarial reinforcement learning: A case study on linear quadratic systems,” in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, eds.), vol. 33, pp. 22056–22068, 2020.
- [84] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

- [85] A. Wibisono, A. C. Wilson, and M. I. Jordan, “A variational perspective on accelerated methods in optimization,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 47, pp. E7351–E7358, 2016.
- [86] E. V. Mazumdar, M. I. Jordan, and S. S. Sastry, “On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games,” *arXiv*, 2019.
- [87] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, “On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points,” *arXiv*, 2019.
- [88] J. Diakonikolas, C. Daskalakis, and M. Jordan, “Efficient methods for structured nonconvex-nonconcave min-max optimization,” in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (A. Banerjee and K. Fukumizu, eds.), vol. 130 of *Proceedings of Machine Learning Research*, pp. 2746–2754, PMLR, 13–15 Apr 2021.
- [89] W. Su, S. Boyd, and E. J. Candès, “A differential equation for modeling Nesterov’s accelerated gradient method: theory and insights,” *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.
- [90] D. Gadjov and L. Pavel, “A passivity-based approach to Nash equilibrium seeking over networks,” *IEEE Transactions on Automatic Control*, vol. 64, no. 3, pp. 1077–1092, 2017.
- [91] T. Başar, “Relaxation techniques and asynchronous algorithms for on-line computation of non-cooperative equilibria,” *Journal of Economic Dynamics and Control*, vol. 11, no. 4, pp. 531–549, 1987.
- [92] G. Hu, Y. Pang, C. Sun, and Y. Hong, “Distributed Nash equilibrium seeking: continuous-time control-theoretic approaches,” *submitted as part of CSM special issue*, 2020.
- [93] P. Frihauf, M. Krstic, and T. Basar, “Nash equilibrium seeking in noncooperative games,” *IEEE Transactions on Automatic Control*, vol. 57, no. 5, pp. 1192–1207, 2012.
- [94] G. A. Kaminka, D. Erusalimchik, and S. Kraus, “Adaptive multi-robot coordination: a game-theoretic perspective,” in *2010 IEEE International Conference on Robotics and Automation*, pp. 328–334, 2010.
- [95] W. Inujima, K. Nakano, and S. Hosokawa, “Multi-robot coordination using switching of methods for deriving equilibrium in game theory,” in *2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, pp. 1–6, 2013.
- [96] W. He and Y. Sun, “Stationary Markov perfect equilibria in discounted stochastic games,” *Journal of Economic Theory*, vol. 169, pp. 35–61, 2017.

- [97] M. O. Sayin, F. Parise, and A. Ozdaglar, “Fictitious play in zero-sum stochastic games,” *arXiv*, 2020.
- [98] R. Bellman, “The theory of dynamic programming,” *Bulletin of the American Mathematical Society*, vol. 60, no. 6, pp. 503–515, 1954.
- [99] J. Hu and M. P. Wellman, “Nash Q-learning for general-sum stochastic games,” *Journal of machine learning research*, vol. 4, no. Nov, pp. 1039–1069, 2003.
- [100] D. S. Leslie, S. Perkins, and Z. Xu, “Best-response dynamics in zero-sum stochastic games,” *Journal of Economic Theory*, vol. 189, p. 105095, 2020.
- [101] J. Bu, L. J. Ratliff, and M. Mesbahi, “Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games,” *arXiv*, 2019.
- [102] K. Zhang, X. Zhang, B. Hu, and T. Başar, “Derivative-free policy optimization for risk-sensitive and robust control design: implicit regularization and sample complexity,” *arXiv*, 2021.
- [103] P. Dayan and C. J. Watkins, “Q-Learning,” *Machine Learning*, vol. 8, no. 3-4, p. 279 292, 1992.
- [104] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. d. Cote, “A survey of learning in multiagent environments: Dealing with non-stationarity,” *arXiv*, 2017.
- [105] H.-T. Wai, Z. Yang, Z. Wang, and M. Hong, “Multi-agent reinforcement learning via double averaging primal-dual optimization,” *arXiv*, 2018.
- [106] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, “Fully decentralized multi-agent reinforcement learning with networked agents,” in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, (Stockholmsmässan, Stockholm Sweden), pp. 5872–5881, PMLR, 2018.
- [107] I. A. Kash, M. Sullins, and K. Hofmann, “Combining no-regret and Q-learning,” in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS ’20*, (Richland, SC), p. 593–601, International Foundation for Autonomous Agents and Multiagent Systems, 2020.
- [108] T. Li, G. Peng, and Q. Zhu, “Blackwell online learning for Markov decision processes,” *arXiv preprint arXiv:2012.14043*, 2020.
- [109] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione, “Regret minimization in games with incomplete information,” in *Advances in Neural Information Processing Systems 20* (J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds.), pp. 1729–1736, Curran Associates, Inc., 2008.
- [110] K. Zhang, S. M. Kakade, T. Başar, and L. F. Yang, “Model-based multi-agent RL in zero-sum markov games with near-optimal sample complexity,” *arXiv*, 2020.

- [111] V. Hakami and M. Dehghan, “Learning stationary correlated equilibria in constrained general-sum stochastic games,” *IEEE Transactions on Cybernetics*, vol. 46, no. 7, pp. 1640–1654, 2016.
- [112] T. Alpcan, T. Başar, R. Srikant, and E. Altman, “CDMA uplink power control as a noncooperative game,” *Wireless Networks*, vol. 8, no. 6, pp. 659–670, 2002.
- [113] Q. Zhu, J. B. Song, and T. Başar, “Dynamic secure routing game in distributed cognitive radio networks,” in *2011 IEEE Global Telecommunications Conference-GLOBECOM 2011*, pp. 1–6, IEEE, 2011.
- [114] Q. Zhu, C. Fung, R. Boutaba, and T. Başar, “A game-theoretical approach to incentive design in collaborative intrusion detection networks,” in *2009 International Conference on Game Theory for Networks*, pp. 384–392, IEEE, 2009.
- [115] M. J. Farooq and Q. Zhu, “On the secure and reconfigurable multi-layer network design for critical information dissemination in the internet of battlefield things (IoBT),” *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2618–2632, 2018.
- [116] M. J. Farooq and Q. Zhu, “Modeling, analysis, and mitigation of dynamic botnet formation in wireless IoT networks,” *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 9, pp. 2412–2426, 2019.
- [117] J. Chen and Q. Zhu, “A game-theoretic framework for resilient and distributed generation control of renewable energies in microgrids,” *IEEE Transactions on Smart Grid*, vol. 8, no. 1, pp. 285–295, 2016.
- [118] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, and T. Başar, “Demand response management in the smart grid in a large population regime,” *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 189–199, 2015.
- [119] J. Chen, C. Touati, and Q. Zhu, “A dynamic game approach to strategic design of secure and resilient infrastructure network,” *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 462–474, 2019.
- [120] L. Huang, J. Chen, and Q. Zhu, “A large-scale Markov game approach to dynamic protection of interdependent infrastructure networks,” in *International Conference on Decision and Game Theory for Security*, pp. 357–376, Springer, 2017.
- [121] J. Chen and Q. Zhu, “Interdependent network formation games with an application to critical infrastructures,” in *2016 American Control Conference (ACC)*, pp. 2870–2875, IEEE, 2016.
- [122] J. Chen, C. Touati, and Q. Zhu, “Heterogeneous multi-layer adversarial network design for the IoT-enabled infrastructures,” in *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pp. 1–6, IEEE, 2017.

- [123] Z. Xu and Q. Zhu, “A game-theoretic approach to secure control of communication-based train control systems under jamming attacks,” in *Proceedings of the 1st International Workshop on Safe Control of Connected and Autonomous Vehicles*, pp. 27–34, 2017.
- [124] Q. Zhu, W. Saad, Z. Han, H. V. Poor, and T. Başar, “Eavesdropping and jamming in next-generation wireless networks: A game-theoretic approach,” in *2011-MILCOM 2011 Military Communications Conference*, pp. 119–124, IEEE, 2011.
- [125] Q. Zhu and T. Başar, “Game-theoretic approach to feedback-driven multi-stage moving target defense,” in *International Conference on Decision and Game Theory for Security*, pp. 246–263, Springer, 2013.
- [126] L. Huang and Q. Zhu, “A dynamic games approach to proactive defense strategies against advanced persistent threats in cyber-physical systems,” *Computers & Security*, vol. 89, p. 101660, 2020.
- [127] Q. Zhu and S. Rass, “On multi-phase and multi-stage game-theoretic modeling of advanced persistent threats,” *IEEE Access*, vol. 6, pp. 13958–13971, 2018.
- [128] N. Al-Falahy and O. Y. Alani, “Technologies for 5G networks: challenges and opportunities,” *IT Professional*, vol. 19, no. 1, pp. 12–20, 2017.
- [129] J. B. Song and Q. Zhu, “Performance of dynamic secure routing game,” in *Game Theory for Networking Applications*, pp. 37–56, Springer, 2019.
- [130] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, “Federated optimization: Distributed machine learning for on-device intelligence,” *arXiv preprint arXiv:1610.02527*, 2016.
- [131] S. Liu, T. Li, and Q. Zhu, “Communication-efficient distributed machine learning over strategic networks: A two-layer game approach,” *arXiv preprint arXiv:2011.01455*, 2020.
- [132] S. Liu, T. Li, and Q. Zhu, “Game-theoretic distributed empirical risk minimization with strategic network design,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 9, pp. 542–556, 2023.
- [133] J. Chen and Q. Zhu, “A game- and decision-theoretic approach to resilient inter-dependent network analysis and design,” *SpringerBriefs in Electrical and Computer Engineering*, pp. 75–102, 2019.
- [134] Q. Zhu, “Multilayer cyber-physical security and resilience for smart grid,” in *Smart Grid Control*, pp. 225–239, Springer, 2019.
- [135] M. J. Farooq and Q. Zhu, “On the secure and reconfigurable multi-layer network design for critical information dissemination in the internet of battlefield things (IoBT),” *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2618–2632, 2018.



- [136] J. Foerster, I. A. Assael, N. De Freitas, and S. Whiteson, “Learning to communicate with deep multi-agent reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 29, pp. 2137–2145, 2016.
- [137] Y. Huang and Q. Zhu, “A differential game approach to decentralized virus-resistant weight adaptation policy over complex networks,” *IEEE Transactions on Control of Network Systems*, vol. 7, no. 2, pp. 944–955, 2020.
- [138] J. Chen, Y. Huang, R. Zhang, and Q. Zhu, “Optimal quarantining strategy for interdependent epidemics spreading over complex networks,” *arXiv preprint arXiv:2011.14262*, 2020.
- [139] N. Li and J. R. Marden, “Designing games for distributed optimization,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 2, pp. 230–242, 2013.
- [140] H. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, vol. 35. Springer Science & Business Media, 2003.

## A Fictitious Play

Consider the repeated play between two players, with each player knowing his own utility function. Further, each player is able to observe the actions of the other player and choose an optimal action based on the empirical frequency of these actions.

In fictitious play, from player 1’s viewpoint, player 2’s strategy at time  $k$  can be estimated as  $\pi_2^k(a) = \sum_{s=1}^k \mathbb{1}_{\{a_2^s=a\}}/k$ ,  $a \in \mathcal{A}_2$ , which is the empirical frequency of actions player 2 has implemented up to that point.  $\pi_2^k$  can be computed by a moving average scheme:

$$\pi_2^k = (1 - \frac{1}{k})\pi_2^{k-1} + \frac{1}{k}e_{a_2^k}.$$

Using this, player 1 chooses the best response:  $a_1^{k+1} = \arg \max_{a \in \mathcal{A}_1} u_1(a, \pi_2^k)$  for the next play. Then, the empirical frequency of player 1’s implemented actions is updated according to

$$\pi_1^{k+1} = (1 - \frac{1}{k+1})\pi_1^k + \frac{1}{k+1}e_{a_1^{k+1}},$$

where  $e_{a_1^{k+1}} \in \Delta(\mathcal{A}_1)$  is exactly given by  $BR_1(\pi_2^k)$  and the equation is the same as the one in (10), with the learning rate being  $\lambda_1^k = \frac{1}{k+1}$ . Hence, we conclude that in fictitious play, a player’s empirical play follows best response dynamics. Furthermore, if we replace the best response mapping  $BR$  with the quantal response  $QR^\epsilon$ , we then obtain an important variant: stochastic fictitious play [24].

## B Replicator Dynamics

Recall that continuous-time learning dynamics under dual averaging is

$$\begin{aligned}\frac{d\hat{\mathbf{u}}_i(t)}{dt} &= \mathbf{u}_i(\pi_{-i}(t)), \\ \pi_i(t) &= QR^\epsilon(\hat{\mathbf{u}}_i(t)).\end{aligned}\tag{DA-c}$$

We now consider the entropy regularizer  $h(x) = \sum x_i \log x_i$  and let  $\epsilon = 1$  for simplicity. Differentiate the strategy  $\pi_i(t)$  with respect to time variable in (DA-c), arriving at

$$\begin{aligned}\frac{d\pi_{i,a}(t)}{dt} &= \frac{1}{(\sum_{a'} e^{\hat{\mathbf{u}}_{i,a}(t)})^2} \left( \frac{d\hat{\mathbf{u}}_{i,a}(t)}{dt} e^{\hat{\mathbf{u}}_{i,a}(t)} \sum_{a'} e^{\hat{\mathbf{u}}_{i,a'}(t)} - e^{\hat{\mathbf{u}}_{i,a}(t)} \sum_{a'} e^{\hat{\mathbf{u}}_{i,a'}(t)} \frac{d\hat{\mathbf{u}}_{i,a'}(t)}{dt} \right) \\ &= \pi_{i,a}(t) \left( \frac{d\hat{\mathbf{u}}_{i,a}(t)}{dt} - \sum_{a'} \pi_{i,a'}(t) \frac{d\hat{\mathbf{u}}_{i,a'}(t)}{dt} \right) \\ &= \pi_{i,a}(t) [u_i(a, \pi_{-i}(t)) - u_i(\pi_i(t), \pi_{-i}(t))].\end{aligned}\tag{RD}$$

From the equation above, we can see that for a certain action  $a$ , if its outcome  $u_i(a, \pi_{-i}(t))$  is above the average  $u_i(\pi_i(t), \pi_{-i}(t))$ , then it will be “reinforced” in the sense that the probability of choosing  $a$  gets higher as time evolves. The above equation (RD) is referred to as replicator dynamics, and has been widely used in evolutionary game theory to understand natural selection and population biology. We consider a two-population system and we reinterpret the elements in the two-player game using population biology language. For population 1, there are  $|\mathcal{A}_1|$  types and each type is specified by an element  $a \in \mathcal{A}_1$ . We let  $\pi_{1,a}(t)$  be the percentage of type  $a$  in population 1 at time  $t$ , and assume here that  $\pi_1(t)$  is differentiable with respect to time  $t$ , as the population, which is infinitely large, interacts with the other population in a continuous-time manner.

For population 2, we have similar notions. If individuals from the two population meet randomly, then they engage in a competition or a game with payoff dependent on their types. For example, if type  $a_1$  from population 1 competes with type  $a_2$  from population 2, then the payoffs for the two types are given by  $u_1(a_1, a_2)$  and  $u_2(a_1, a_2)$ , respectively. For population  $i$ , if we assume that the per capita rate of growth is given by the difference between the payoff for type  $a$  and the average payoff in the population, a rule studied in [43], then the percentage of different types within a population is precisely described by

$$\frac{1}{\pi_{i,a}} \frac{d\pi_{i,a}(t)}{dt} = u_i(a, \pi_i(t)) - u_i(\pi_i(t), \pi_{-i}(t)),$$

which is exactly the replicator dynamics (RD). In addition, as shown in [38], different regularizers lead to different learning dynamics, which display different asymptotic behavior accounts for the evolutionary process under different circumstances.

With replicator dynamics and other related evolutionary dynamics, biologists can predict the evolutionary outcome of the multi-population system by examining the Nash equilibrium of the underlying game, which brings strategic reasoning into population biology and has a profound influence in evolutionary game theory [44, 45]. Moreover, the Nash equilibrium in this population game, characterized by the limiting behavior of the dynamics

under proper conditions [45], represents an evolutionarily stable state of the population, which is an important refinement of Nash equilibrium. When this stable state is reached, natural selection alone is sufficient to prevent the population from being influenced by mutation [34, 44]. For more details on this refinement and its application in biology, we refer the reader to [11, 34, 44, 45].

## C Stochastic Approximation Theory

Following the multiple timescale stochastic approximation framework developed in [29, 140], one can write (8) and (9) using discrete-time stochastic approximation

$$\begin{aligned}\pi_i^{k+1} - \pi_i^k &= \bar{\lambda}_i^k (f_i(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}) + M_i^{k+1}), \\ \hat{\mathbf{u}}_i^{k+1} - \hat{\mathbf{u}}_i^k &= \bar{\mu}_i^k (g_i(\pi_i^k, \hat{\mathbf{u}}_i^k) + \Gamma_i^{k+1}),\end{aligned}\tag{C1}$$

where  $f_i(\pi_i^k, \hat{\mathbf{u}}_i^{k+1})$  and  $g_i(\pi_i^k, \hat{\mathbf{u}}_i^k)$  are the mean-field components of (8) and (9), respectively, and are defined as

$$\begin{aligned}f_i(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}) &= \mathbb{E}[F_i(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}, U_i^{k+1}, a_i^{k+1}) | \mathcal{F}^{k-1}], \\ g_i(\pi_i^k, \hat{\mathbf{u}}_i^k) &= \mathbb{E}[G_i(\pi_i^k, \hat{\mathbf{u}}_i^k, U_i^{k+1}, a_i^{k+1}) | \mathcal{F}^{k-1}].\end{aligned}$$

With the mean-field part defined as above,  $M_i^{k+1} = F_i(\pi_i^k, \hat{\mathbf{u}}_i^{k+1}, U_i^{k+1}, a_i^{k+1}) - f_i(\pi_i^k, \hat{\mathbf{u}}_i^{k+1})$  and  $\Gamma_i^{k+1}$  takes a similar form.  $\bar{\lambda}_i^k, \bar{\mu}_i^k$  are time-scaling factors dependent on the learning rates  $\lambda_i^k, \mu_i^k$ , which account for the adjustment of the original step sizes in asynchronous schemes [29, 64], and in synchronous cases, the time scaling factors coincide with the original step sizes. Similar to our discussion in the main text (see (18) and (19)), we consider the dynamical system of the joint strategy profile  $\pi^k$  and utility vector  $\hat{\mathbf{u}}^k$

$$\begin{aligned}\pi^{k+1} - \pi^k &= \bar{\lambda}^k (f(\pi^k, \hat{\mathbf{u}}^{k+1}) + M^{k+1}), \\ \hat{\mathbf{u}}^{k+1} - \hat{\mathbf{u}}^k &= \bar{\mu}^k (g(\pi^k, \hat{\mathbf{u}}^k) + \Gamma^{k+1}),\end{aligned}\tag{DSA}$$

where  $f$  and  $g$  are concatenations of  $\{f_i\}_{i \in \mathcal{N}}$  and  $\{g_i\}_{i \in \mathcal{N}}$ , respectively.  $\bar{\lambda}^k, \bar{\mu}^k$  and  $M^k, \Gamma^k$  take similar forms.

As we have discussed in ‘‘Convergence of Learning in Games’’, in order to obtain an approximately accurate score function, the two coupled dynamical systems in (DSA) should operate on different timescales: the score function  $\hat{\mathbf{u}}^k$  should be updated sufficiently many times until near-convergence before updating the strategy. This two-timescale iteration can be achieved by adjusting the time-scaling factors:  $\bar{\lambda}^k$  and  $\bar{\mu}^k$  are chosen so that  $\lim_{k \rightarrow \infty} \bar{\lambda}^k / \bar{\mu}^k = 0$ . To understand this timescale system, it is instructive to consider a coupled continuous-time dynamical system, as suggested in [29]

$$\begin{aligned}\frac{d\pi(t)}{dt} &= f(\pi(t), \hat{\mathbf{u}}(t)), \\ \frac{d\hat{\mathbf{u}}(t)}{dt} &= \frac{1}{\varepsilon} g(\pi(t), \hat{\mathbf{u}}(t)),\end{aligned}\tag{C2}$$

where in the limit  $\varepsilon$  tends to zero. Hence,  $\hat{\mathbf{u}}(t)$  is fast transient while  $\pi(t)$  is slow. Then, we can analyze the long-run behavior of the above coupled system as if the fast process is always fully calibrated to the current value of the slow process. This suggests investigating the ODE

$$\frac{d\hat{\mathbf{u}}(t)}{dt} = g(\pi, \hat{\mathbf{u}}(t)), \quad (\text{C3})$$

where  $\pi$  is held fixed as a constant parameter. Suppose (C3) has a globally asymptotically stable equilibrium  $\Lambda(\pi)$ , where the mapping  $\Lambda(\cdot)$  satisfies regularity conditions specified in [30, 64]. Then, it is reasonable to expect  $\hat{\mathbf{u}}(t)$  given by (C3) to closely track  $\Lambda(\pi)$ . In turn, this suggests that the investigation into the coupled system (C2) is equivalent to the study of the single-timescale one

$$\frac{d\pi(t)}{dt} = f(\pi(t), \Lambda(\pi(t))), \quad (\text{C4})$$

which would capture the long-run behavior of  $\pi(t)$  in (C2) to a good approximation [29].

Informally speaking, to study the convergence of (DSA), we can relate its discrete-time trajectory to that of (C2), which is further equivalent to  $(\pi(t), \Lambda(\pi(t)))$  specified by (C4). Therefore, we can apply Lyapunov stability theory to (C4), in order to derive the convergence results of the original discrete-time algorithm. We begin with the linear interpolation process of the discrete-time trajectory, which connects the discrete-time system (DSA) and its continuous-time counterpart (C2), (C4). Under some regularity conditions [30], for  $\{\pi^k\}$ , the sequence generated by (DSA), we can construct the following continuous time process  $\bar{\pi}(t) : \mathbb{R}_+ \rightarrow \Delta(\mathcal{A})$ , based on the linear interpolation of  $\{\pi^k\}$ . Letting  $\tau^0 = 0$  and  $\tau^k = \sum_{s=1}^k \bar{\lambda}^s$ , we define

$$\bar{\pi}(t) := \pi^k + (t - \tau^k) \frac{\pi^{k+1} - \pi^k}{\tau^{k+1} - \tau^k}, \quad t \in [\tau^k, \tau^{k+1}).$$

Similarly, we can define a continuous-time process  $\bar{\mathbf{u}}(t)$  corresponding to  $\{\hat{\mathbf{u}}^k\}$ .

As shown in [30, 64], such a linearly interpolated process  $(\bar{\pi}(t), \bar{\mathbf{u}}(t))$  is closely related to the flow of the following differential equations:

$$\begin{aligned} \frac{d\pi(t)}{dt} &= f(\pi(t), \hat{\mathbf{u}}(t)), \\ \frac{d\hat{\mathbf{u}}(t)}{dt} &= g(\pi(t), \hat{\mathbf{u}}(t)). \end{aligned} \quad (\text{C5})$$

We note that (C5) is defined for ease of presentation, and the actual differential inclusion systems involves rearrangement of several terms; which we refer the reader to [64] for more details. Further, we denote the flow of (C5) by

$$\Phi_t(\pi^0, \mathbf{u}^0) := \{(\pi(t), \hat{\mathbf{u}}(t)) | (\pi(0), \hat{\mathbf{u}}(0)) \text{ is a solution to (C5), with } \pi(0) = \pi^0, \hat{\mathbf{u}}(0) = \mathbf{u}^0\}.$$

The key of stochastic approximation theory lies in the fact that in the presence of a global attractor for (C5), the continuous-time process  $(\bar{\pi}(t), \bar{\mathbf{u}}(t))$  asymptotically tracks the flow with arbitrary accuracy over windows of arbitrary length [30],

$$\lim_{t \rightarrow \infty} \sup_{s \in [0, T]} \text{dist}\{(\bar{\pi}(t+s), \bar{\mathbf{u}}(t+s)), \Phi_s(\bar{\pi}(t), \bar{\mathbf{u}}(t))\} = 0,$$

where  $\text{dist}\{\cdot, \cdot\}$  denotes a distance measure on  $\Delta(\mathcal{A}) \times \mathbb{R}^A$ . We refer to  $(\bar{\pi}(t), \bar{\mathbf{u}}(t))$  as an asymptotic pseudo-trajectory (APT) of the dynamics (C5). In other words, in order to study the convergence of (DSA), we can resort to the convergence analysis of (C5), which can be addressed by Lyapunov stability theory as shown in [30, 64], where the key conclusion is that if there is a global attractor  $A$  for (C4). Then the interpolated process  $(\bar{\pi}(t), \bar{\mathbf{u}}(t))$  or simply  $(\pi^k, \bar{\mathbf{u}}^k)$  converges almost surely to  $(A, \Lambda(A))$ .