Stochastic Stability Analysis of Perturbed Learning Automata with Constant Step-Size in Strategic-Form Games*

Georgios C. Chasparis[1]

Abstract— This paper considers a class of reinforcement-learning that belongs to the family of Learning Automata and provides a stochastic-stability analysis in strategic-form games. For this class of dynamics, convergence to pure Nash equilibria has been demonstrated only for the fine class of potential games. Prior work primarily provides convergence properties of the dynamics through stochastic approximations, where the asymptotic behavior can be associated with the limit points of an ordinary-differential equation (ODE). However, analyzing global convergence through the ODE-approximation requires the existence of a Lyapunov or a potential function, which naturally restricts the applicability of these algorithms to a fine class of games. To overcome these limitations, this paper introduces an alternative framework for analyzing stochastic-stability that is based upon an explicit characterization of the (unique) invariant probability measure of the induced Markov chain.

I. INTRODUCTION

Recently, multi-agent formulations have been utilized to tackle distributed optimization problems, since communication and computation complexity might be an issue in centralized optimization problems. In such formulations, decisions are usually taken in a repeated fashion, where agents select their next actions based on their *own* prior experience of the game.

The present paper discusses a class of reinforcement-learning dynamics, that belongs to the large family of Learning Automata [2], [3], within the context of (non-cooperative) strategic-form games. In this class of dynamics, agents are repeatedly involved in a game with a fixed payoff-matrix, and they need to decide which action to play next having only access to their *own* prior actions and payoffs. In Learning Automata, agents build their confidence over an action through repeated selection of this action and proportionally to the reward received from this action. Naturally, it has been utilized to analyze human-like (bounded) rationality [4].

Reinforcement learning has been applied in evolutionary economics, for modeling human and economic behavior [4], [5], [6], [7], [8]. It is also highly attractive to several

engineering applications, since agents do not need to know neither the actions of the other agents, nor their own utility function. It has been utilized for system identification and pattern recognition [9], distributed network formation and coordination problems [10].

In strategic-form games, the main goal is to derive conditions under which convergence to Nash equilibria can be achieved. In social sciences, deriving such conditions may be important for justifying emergence of certain social phenomena. In engineering, convergence to Nash equilibria may also be desirable in distributed optimization problems, when the set of optimal solutions coincides with the set of Nash equilibria.

In Learning Automata, deriving conditions under which convergence to Nash equilibria is achieved may not be a trivial task. In particular, there are two main difficulties: a) excluding convergence to pure strategies that are *not* Nash equilibria, and b) excluding convergence to mixed strategy profiles. As it will be discussed in detail in a forthcoming Section II, for some classes of (discrete-time) reinforcement-learning algorithms, convergence to non-Nash pure strategies may be achieved with positive probability. Moreover, excluding convergence to mixed strategy profiles may only be achieved under strong conditions in the utilities of the agents, (e.g., existence of a potential function).

In the present paper, we consider a class of (discrete-time) reinforcement-learning algorithms introduced in [10] that is closely related to existing algorithms for modeling human-like behavior, e.g., [4]. The main difference with prior reinforcement learning schemes lies in a) the step-size sequence, and b) the perturbation (or *mutations*) term. The step-size sequence is assumed constant, thus introducing a fading-memory effect of past experiences in each agent's strategy. On the other hand, the perturbation term introduces errors in the selection process of each agent. Both these two features can be used for designing a desirable asymptotic behavior.

We provide an analytical framework for deriving conclusions over the asymptotic behavior of the dynamics that is based on an explicit characterization of the invariant probability measure of the induced Markov chain. In particular, *we show that in all strategic-form games satisfying the Positive-Utility Property, the support of the invariant*

---

[1]Software Competence Center Hagenberg GmbH, Softwarepark 21, A-4232 Hagenberg, Austria georgios.chasparis@scch.at

probability measure coincides with the set of pure strategy profiles. This extends prior work where nonconvergence to mixed strategy profiles may only be excluded under strong conditions in the payoff matrix (e.g., existence of a potential function). A detailed discussion of the exact contributions of this paper is provided in the forthcoming Section II. At the end of the paper, we also provide a brief discussion over how the proposed framework can be further utilized to provide a more detailed characterization of the stochastically stable states (e.g., excluding convergence to non-Nash pure strategy profiles). Due to space limitations, this analysis is not presented in this paper.

In the remainder of the paper, Section II presents a class of reinforcement-learning dynamics, related work and the main contribution of this paper. Section III provides the main result of this paper (Theorem 3.1), where the set of stochastically stable states is characterized. A short discussion is also provided over the significance of this result and how it can be utilized to provide further conclusions. Finally, Section IV provides the technical derivation of the main result and Section V presents concluding remarks.

Notation:

− For a Euclidean topological space $\mathcal{X} \subset \mathbb{R}^n$, let $\mathcal{N}_\delta(x)$ denote the δ-neighborhood of $x \in \mathbb{R}^n$, i.e.,

$$\mathcal{N}_\delta(x) \doteq \{y \in \mathcal{X} : |x - y| < \delta\},$$

where $|\cdot|$ denotes the Euclidean distance.

− e_j denotes the *unit vector* in $\mathbb{R}^n$ where its jth entry is equal to 1 and all other entries is equal to 0.

− $\Delta(n)$ denotes the *probability simplex* of dimension n, i.e.,

$$\Delta(n) \doteq \left\{x \in \mathbb{R}^n : x \geq 0, \mathbf{1}^\mathrm{T} x = 1\right\}.$$

− For some set A in a topological space $\mathcal{Z}$, let $\mathbb{I}_A : \mathcal{Z} \to \{0, 1\}$ denote the index function, i.e.,

$$\mathbb{I}_A(x) \doteq \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else.} \end{cases}$$

− δ_x denotes the Dirac measure at x.

− Let A be a finite set and let any (finite) probability distribution $\sigma \in \Delta(|A|)$. The random selection of an element of A will be denoted $\mathrm{rand}_\sigma[A]$. If $\sigma = (1/|A|, ..., 1/|A|)$, i.e., it corresponds to the uniform distribution, the random selection will be denoted by $\mathrm{rand}_\mathrm{unif}[A]$.

II. Reinforcement Learning

A. Terminology

We consider the standard setup of finite strategic-form games. Consider a finite set of agents (or *players*) $\mathcal{I} = \{1, ..., n\}$, and let each agent have a finite set of actions $\mathcal{A}_i$.

At fixed time instances $t = 1, 2, ...$, and for each agent $i \in \mathcal{I}$, the following steps are executed recursively. Let $\alpha_i(t)$ and $x_i(t)$ denote the current action and strategy of agent i, respectively.

1) (*action update*) Agent $i \in \mathcal{I}$ selects a new action $\alpha_i(t + 1)$ as follows:

$$\alpha_i(t+1) = \begin{cases} \mathrm{rand}_{x_i(t)}[\mathcal{A}_i], & \text{with probability } 1 - \lambda, \\ \mathrm{rnad}_\mathrm{unif}[\mathcal{A}_i], & \text{with probability } \lambda, \end{cases} \quad (1)$$

for some small perturbation factor $\lambda > 0$.

2) (*evaluation*) Agent i applies its new action $\alpha_i(t+1)$ and retrieves a measurement of its utility function $u_i(\alpha(t + 1)) > 0$.

3) (*strategy update*) Agent i revises its strategy vector $x_i \in \Delta(|\mathcal{A}_i|)$ as follows:

$$\begin{aligned} x_i&(t + 1) \\ &= x_i(t) + \epsilon \cdot u_i(\alpha(t+1)) \cdot [e_{\alpha_i(t+1)} - x_i(t)] \\ &\doteq \mathcal{R}_i(\alpha(t + 1), x_i(t)), \end{aligned} \quad (2)$$

for some constant step-size $\epsilon > 0$.

TABLE I

PERTURBED REINFORCEMENT LEARNING.

Let $\alpha_i \in \mathcal{A}_i$ denote any such action of agent i. The set of *action profiles* is the Cartesian product $\mathcal{A} \doteq \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ and let $\alpha = (\alpha_1, ..., \alpha_n)$ be a representative element of this set. We will denote $-i$ to be the complementary set $\mathcal{I} \backslash i$ and often decompose an action profile as follows $\alpha = (\alpha_i, \alpha_{-i})$. The *payoff/utility function* of agent i is a mapping $u_i(\cdot) : \mathcal{A} \to \mathbb{R}$. A *strategic-form game* is defined by the triple $\langle \mathcal{I}, \mathcal{A}, \{u_i(\cdot)\}_i \rangle$.

For the remainder of the paper, we will be concerned with strategic-form games that satisfy the ***Positive-Utility Property***.

Property 2.1 (Positive Utility Property): For any agent $i \in \mathcal{I}$ and any action profile $\alpha \in \mathcal{A}$, $u_i(\alpha) > 0$.

B. Reinforcement-learning algorithm

We consider a form of reinforcement learning that belongs to the general class of *learning automata* [3]. In learning automata, each agent updates a finite probability distribution $x_i \in \Delta(|\mathcal{A}_i|)$ representing its beliefs with respect to the most profitable action. The precise manner in which $x_i(t)$ changes at time t, depending on the performed action and the response of the environment, completely defines the reinforcement learning model.

The proposed reinforcement learning model is described in Table I. At the first step, each agent i updates its action given its current strategy vector $x_i(t)$. Its selection is slightly perturbed by a perturbation (or *mutations*) factor $\lambda > 0$, such that, with a small probability λ agent i follows a uniform strategy (or, it *trembles*). At the second step, agent i evaluates its new selection by collecting a utility measurement, while in the last step, agent i updates its strategy vector given its new experience.

Here we identify actions $\mathcal{A}_i$ with vertices of the simplex,

$\{e_1, ..., e_{|\mathcal{A}_i|}\}$. For example, if agent i selects its jth action at time t, then $e_{\alpha_i(t)} \equiv e_j$. Note that by letting the step-size ϵ to be sufficiently small and since the utility function $u_i(\cdot)$ is uniformly bounded in $\mathcal{A}$, $x_i(t) \in \Delta(|\mathcal{A}_i|)$ for all t.

In case $\lambda = 0$, the above update recursion will be referred to as the *unperturbed reinforcement learning*.

C. Related work

Erev-Roth type dynamics: In prior reinforcement learning in games, analysis has been restricted to decreasing step-size sequences $\epsilon(t)$ and $\lambda = 0$. More specifically, in [4], the step-size sequence of agent i is $\epsilon_i(t) = 1/(ct^\nu + u_i(\alpha(t+1)))$ for some positive constant c and for $0 < \nu < 1$ (in the place of the constant step-size ϵ of (2)). A comparative model is also used by [7], with $\epsilon_i(t) = 1/(V_i(t) + u_i(\alpha(t+1)))$, where $V_i(t)$ is the accumulated benefits of agent i up to time t which gives rise to an urn process [6]. Some similarities are also shared with the Cross' learning model of [5], where $\epsilon(t) = 1$ and $u_i(\alpha(t)) \leq 1$, and its modification presented in [11], where $\epsilon(t)$, instead, is assumed decreasing.

The main difference of the proposed reinforcement-learning algorithm (Table I) lies in the perturbation parameter $\lambda > 0$ which was first introduced and analyzed in [10]. A state-dependent perturbation term has also been investigated in [12]. The perturbation parameter may serve as an equilibrium selection mechanism, since *it excludes convergence to non-Nash action profiles*. It resolved one of the main issues of several (discrete-time) reinforcement-learning algorithms, that is the positive probability of convergence to non-Nash action profiles under some conditions in the payoff function and the step-size sequence.

This issue has also been raised by [13], [7]. Reference [13] considered the model by [4] and showed that convergence to non-Nash pure strategy profiles can be excluded as long as $c > u_i(\alpha)$ for all $i \in \mathcal{I}$ and $\nu = 1$. On the other hand, convergence to non-Nash action profiles was not an issue with the urn model of [6] (as analyzed in [7]). However, the use of an urn-process type step-size sequence significantly reduces the applicability of the reinforcement learning scheme. In conclusion, the perturbation parameter $\lambda > 0$ may serve as a design tool for reinforcing convergence to Nash equilibria without necessarily employing an urn-process type step-size sequence. For engineering applications this is a desirable feature.

Although excluding convergence to non-Nash pure strategies can be guaranteed by using $\lambda > 0$, establishing convergence to pure Nash equilibria may still be an issue, since it further requires excluding convergence to mixed strategy profiles. As presented in [12], this can be guaranteed only under strong conditions in the payoff matrix. For example, as shown in [12, Proposition 8], excluding convergence to mixed strategy profiles requires a) the existence of a potential function, b) conditions over the second gradient of the potential function. Requiring the existence of a potential function considerably restricts the class of games where equilibrium selection can be described. Furthermore, condition (b) may not easily be verified in games of large number of players or actions.

Learning automata: Certain forms of learning automata have been shown to converge to Nash equilibria in some classes of strategic-form games. For example, in [3], and for a generalized nonlinear reward-inaction scheme, convergence to Nash equilibrium strategies can be shown in identical interest games. Similar are the results presented in [14] for a linear reward-inaction scheme. These convergence results are restricted to games of payoffs in $[0, 1]$. Extension to a larger class of games is possible if *absolute monotonicity* (cf., [3, Definition 8.1]) is shown (similarly to the discussion in [12, Proposition 8]).

Reference [15] introduced a class of linear reward-inaction schemes in combination with a coordinated exploration phase so that convergence to the efficient Nash equilibrium is achieved. However, coordination of the exploration phase requires communication between the players.

Recently, work by the author [10] has introduced a new class of learning automata (namely, perturbed learning automata) which can be applied in games with no restriction in the payoff matrix. Furthermore, a small perturbation factor also influences the decisions of the players, through which convergence to non-Nash pure strategy profiles can be excluded. However, to demonstrate global convergence, a monotonicity condition still needs to be established [12].

Q-learning: Similar questions of convergence to Nash equilibria also appear in alternative reinforcement learning formulations, such as approximate dynamic programming methodologies and Q-learning. However, this is usually accomplished under a stronger set of assumptions, which increases the computational complexity of the dynamics. For example, the Nash-Q learning algorithm of [16] addresses the problem of maximizing the discounted expected rewards for each agent by updating an approximation of the cost-to-go function (or Q-values). Alternative objectives may be used, such as the minimax criterion of [17]. However, it is indirectly assumed that agents need to have full access to the joint action space and the rewards received by the other agents.

More recently, reference [18] introduces a Q-learning scheme in combination with either adaptive play or better-reply dynamics in order to attain convergence to Nash equilibria in potential games [19] or weakly-acyclic games. However, this form of dynamics require that each player observes the actions selected by the other players, since a Q-value needs to be assigned in each joint action.

When the evaluation of the Q-values is totally indepen-

dent, as in the individual Q-learning in [20], then convergence to Nash equilibria has been shown only for 2-player zero-sum games and 2-player partnership games with countably many Nash equilibria. Currently, there are no convergent results in games in multi-player games.

Payoff-based learning: The aforementioned types of dynamics can be considered as a form of payoff-based learning dynamics, since adaptation is only governed by the perceived utility of the players. Recently, there have been several attempts for establishing convergence to Nash equilibria through alternative payoff-based learning dynamics, (see, e.g., the benchmark-based dynamics of [21], or the aspiration-based dynamics in [22]). For these type of dynamics, convergence to Nash equilibria can be established without requiring any strong monotonicity property (e.g., in multi-player weakly-acyclic games in [21]). However, an investigation is required with respect to the resulting convergence rates as compared to the dynamics incorporating policy iterations (e.g., the Erev-Roth type of dynamics or the learning automata discussed above).

D. Objective

This paper provides an analytical framework for analyzing convergence in multi-player strategic-form games when players implement a class of perturbed learning-automata. We wish to impose no strong monotonicity assumptions in the structure of the game (e.g., the existence of a potential function). We provide a characterization of the invariant probability measure of the induced Markov chain that shows that only the pure-strategy profiles belong to its support. Thus, we implicitly exclude convergence to any mixed strategy profile (including mixed Nash equilibria). This result imposes no restrictions in the payoff matrix other than the Positive-Utility Property.

III. CONVERGENCE ANALYSIS

A. Terminology and notation

Let $\mathcal{Z} \doteq \mathcal{A} \times \mathbf{\Delta}$, where $\mathbf{\Delta} \doteq \Delta(|\mathcal{A}_1|) \times \ldots \times \Delta(|\mathcal{A}_n|)$, i.e., pairs of joint actions α and nominal strategy profiles x. The set $\mathcal{A}$ is endowed with the discrete topology, $\mathbf{\Delta}$ with its usual Euclidean topology, and $\mathcal{Z}$ with the corresponding product topology. We also let $\mathfrak{B}(\mathcal{Z})$ denote the Borel σ-field of $\mathcal{Z}$, and $\mathfrak{P}(\mathcal{Z})$ the set of probability measures on $\mathfrak{B}(\mathcal{Z})$ endowed with the Prohorov topology, i.e., the topology of weak convergence. The algorithm introduced in Table I defines an $\mathcal{Z}$-valued Markov chain. Let $P_\lambda : \mathcal{Z} \times \mathfrak{B}(\mathcal{Z}) \to [0,1]$ denote its transition probability function (t.p.f.), parameterized by $\lambda > 0$. We refer to the process with $\lambda > 0$ as the *perturbed process*. Let also $P : \mathcal{Z} \times \mathfrak{B}(\mathcal{Z})$ denote the t.p.f. of the *unperturbed process*, i.e., when $\lambda = 0$.

We let $C_b(\mathcal{Z})$ denote the Banach space of real-valued continuous functions on $\mathcal{Z}$ under the sup-norm (denoted by $\| \cdot \|_\infty$) topology. For $f \in C_b(\mathcal{Z})$, define

$$P_\lambda f(z) \doteq \int_{\mathcal{Z}} P_\lambda(z, dy) f(y),$$

and

$$\mu[f] \doteq \int_{\mathcal{Z}} \mu(dx) f(z), \text{ for } \mu \in \mathfrak{P}(\mathcal{Z}).$$

The process governed by the unperturbed process P will be denoted by $\{Z_t : t \geq 0\}$. Let $\Omega \doteq \mathcal{Z}^\infty$ denote the canonical path space, i.e., an element $\omega \in \Omega$ is a sequence $\{\omega(0), \omega(1), \ldots \}$, with $\omega(t) = (\alpha(t), x(t)) \in \mathcal{Z}$. We use the same notation for the elements (α, x) of the space $\mathcal{Z}$ and for the coordinates of the process $Z_t = (\alpha(t), x(t))$. Let also $\mathbb{P}_z[\cdot]$ denote the unique probability measure induced by the unperturbed process P on the product σ-algebra of $\mathcal{Z}^\infty$, initialized at $z = (\alpha, x)$, and $\mathbb{E}_z[\cdot]$ the corresponding expectation operator. Let also $\mathfrak{F}_t, t \geq 0$, denote the σ-algebra generated by $\{Z_\tau, \tau \leq t\}$.

B. Stochastic stability

First, we note that both P and P_λ ($\lambda > 0$) satisfy the *weak Feller property* (cf., [23, Definition 4.4.2]).

Proposition 3.1: Both the unperturbed process P ($\lambda = 0$) and the perturbed process P_λ ($\lambda > 0$) have the weak Feller property.

The measure $\mu_\lambda \in \mathfrak{P}(\mathcal{Z})$ is called an *invariant probability measure* for P_λ if

$$(\mu_\lambda P_\lambda)(A) \doteq \int_{\mathcal{Z}} \mu_\lambda(dx) P_\lambda(z, A) = \mu_\lambda(A), \qquad A \in \mathfrak{B}(\mathcal{Z}).$$

Since $\mathcal{Z}$ defines a locally compact separable metric space and P, P_λ have the weak Feller property, they both admit an invariant probability measure, denoted μ and μ_λ, respectively [23, Theorem 7.2.3].

We would like to characterize the *stochastically stable states* $z \in \mathcal{Z}$ of P_λ, that is any state $z \in \mathcal{Z}$ for which any collection of invariant probability measures $\{\mu_\lambda \in \mathfrak{P}(\mathcal{Z}) : \mu_\lambda P_\lambda = \mu_\lambda, \lambda > 0\}$ satisfies $\liminf_{\lambda \to 0} \mu_\lambda(z) > 0$. As the forthcoming analysis will show, the stochastically stable states will be a subset of the set of *pure strategy states* (p.s.s.) defined as follows:

Definition 3.1 (Pure Strategy State): A pure strategy state is a state $s = (\alpha, x) \in \mathcal{Z}$ such that for all $i \in \mathcal{I}$, $x_i = e_{\alpha_i}$, i.e., x_i coincides with the vertex of the probability simplex $\Delta(|\mathcal{A}_i|)$ which assigns probability 1 to action α_i.

We will denote the set of pure strategy states by $\mathcal{S}$.

Theorem 3.1 (Stochastic Stability): There exists a unique probability vector $\pi = (\pi_1, ..., \pi_{|\mathcal{S}|})$ such that for any collection of invariant probability measures $\{\mu_\lambda \in \mathfrak{P}(\mathcal{Z}) : \mu_\lambda P_\lambda = \mu_\lambda, \lambda > 0\}$, the following hold:

(a) $\lim_{\lambda \to 0} \mu_\lambda(\cdot) = \hat{\mu}(\cdot) \doteq \sum_{s \in \mathcal{S}} \pi_s \delta_s(\cdot)$, where convergence is in the weak sense.

(b) The probability vector π is an invariant distribution of the (finite-state) Markov process $\hat{P}$, such that, for any $s, s' \in \mathcal{S}$,

$$\hat{P}_{ss'} \doteq \lim_{t \to \infty} QP^t(s, \mathcal{N}_\delta(s')), \qquad (3)$$

for any $\delta > 0$ sufficiently small, where Q is the t.p.f. corresponding to *only one player trembling* (i.e., following the uniform distribution of (1)).

The proof of Theorem 3.1 requires a series of propositions and will be presented in detail in Section IV.

C. Discussion

Theorem 3.1 establishes an important observation. That is, the "*equivalence*" (in a weak convergence sense) of the original (perturbed) learning process with a simplified process, where *agents simultaneously tremble at the first iteration and then they do not tremble*. This form of simplification of the dynamics has originally been exploited to analyze *aspiration learning* dynamics in [22], and it is based upon the fact that *under the unperturbed dynamics, agents' strategies will eventually converge to a pure strategy profile*.

Furthermore, the limiting behavior of the original (perturbed) dynamics can be characterized by the (*unique*) invariant distribution of a finite-state Markov chain $\{P_{ss'}\}$, whose states correspond to the pure-strategy states of the game. In other words, *we should expect that as the perturbation parameter λ approaches zero, the algorithm spends the majority of the time on pure strategy profiles*. The importance of this result lies on the fact that no constraints have been imposed in the payoff matrix of the game other than the Positive-Utility Property 2.1. Thus, it extends to games beyond the fine set of potential games.

This convergence result can further be augmented with an ODE analysis for stochastic approximations to exclude convergence to pure strategies that are not Nash equilibria (as derived in [12] for the case of diminishing step-size). Due to space limitations this analysis is not presented in this paper, however it can be the subject of future work.

IV. TECHNICAL DERIVATION

A. Unperturbed Process

For $t \geq 0$ define the sets

$$A_t \doteq \{\omega \in \Omega : \alpha(\tau) = \alpha(t), \text{ for all } \tau \geq t\},$$

$$B_t \doteq \{\omega \in \Omega : \alpha(\tau) = \alpha(0), \text{ for all } 0 \leq \tau \leq t\}.$$

Note that $\{B_t : t \geq 0\}$ is a non-increasing sequence, i.e., $B_{t+1} \subseteq B_t$, while $\{A_t : t \geq 0\}$ is non-decreasing, i.e., $A_{t+1} \supseteq A_t$. Let

$$A_\infty \doteq \bigcup_{t=0}^{\infty} A_t \text{ and } B_\infty \doteq \bigcap_{t=1}^{\infty} B_t.$$

In other words, *the set A_∞ corresponds to the event that agents eventually play the same action profile, while B_∞ corresponds to the event that agents never change their actions*.

Proposition 4.1 (Convergence to p.s.s.): Let us assume that the step-size $\epsilon > 0$ is sufficiently small such that $0 < \epsilon u_i(\alpha) < 1$ for all $\alpha \in \mathcal{A}$ and for all agents $i \in \mathcal{I}$. Then, the following hold:

(a) $\inf_{z \in \mathcal{Z}} \mathbb{P}_z[B_\infty] > 0$,

(b) $\inf_{z \in \mathcal{Z}} \mathbb{P}_z[A_\infty] = 1$.

The first statement of Proposition 4.1 states that *the probability that agents never change their actions is bounded away from zero*, while the second statement states that *the probability that eventually agents play the same action profile is one*.

The above proposition is rather useful in characterizing the support of any invariant measure of the unperturbed process, as the following proposition shows.

Proposition 4.2 (Limiting t.p.f. of unperturbed process): Let μ denote an invariant probability measure of P. Then, there exists a t.p.f. Π on $\mathcal{Z} \times \mathfrak{B}(\mathcal{Z})$ such that

(a) for μ-a.e. $z \in \mathcal{Z}$, $\Pi(z, \cdot)$ is an invariant probability measure for P;

(b) for all $f \in C_b(\mathcal{Z})$, $\lim_{t \to \infty} \|P^t f - \Pi f\|_\infty = 0$;

(c) μ is an invariant probability measure of Π;

(d) the support[1] of Π is on $\mathcal{S}$ for all $z \in \mathcal{Z}$.

Proposition 4.2 states that the limiting unperturbed t.p.f. converges weakly to a t.p.f. Π which accepts the same invariant p.m. as P. Furthermore, *the support of Π is the set of pure strategy states $\mathcal{S}$*. This is a rather important observation, since the limiting perturbed process can also be "related" (in a weak-convergence sense) to the t.p.f. Π, as it will be shown in the following section.

B. Decomposition of perturbed t.p.f.

We can decompose the t.p.f. of the perturbed process as follows:

$$P_\lambda = (1 - \varphi(\lambda))P + \varphi(\lambda)Q_\lambda$$

where $\varphi(\lambda) = 1 - (1 - \lambda)^n$ is the probability that at least one agent trembles (since $(1 - \lambda)^n$ is the probability that no agent trembles), and Q_λ corresponds to the t.p.f. induced by the one-step reinforcement-learning update when at least one agent trembles. Note that $\varphi(\lambda) \to 0$ as $\lambda \to 0$.

Define also Q to be the t.p.f. when *only one* players trembles, and Q^* is the t.p.f. where at least two players tremble. Then, we may write:

$$Q_\lambda = (1 - \psi(\lambda))Q + \psi(\lambda)Q^*, \qquad (4)$$

---

[1]The *support* of a measure μ on $\mathcal{Z}$ is the unique closed set $F \subset \mathfrak{B}(\mathcal{Z})$ such that $\mu(\mathcal{Z} \backslash F) = 0$ and $\mu(F \cap O) > 0$ for every open set $O \subset \mathcal{Z}$ such that $F \cap O \neq \varnothing$.

where $\psi(\lambda) \doteq 1 - \frac{n\lambda}{1-(1-\lambda)^n}$ corresponds to the probability that at least two players tremble given that at least one player trembles.

Let us also define the infinite-step t.p.f. when trembling only at the first step (briefly, *lifted* t.p.f.) as follows:

$$P_\lambda^L \doteq \varphi(\lambda) \sum_{t=0}^{\infty} (1-\varphi(\lambda))^t Q_\lambda P^t = Q_\lambda R_\lambda \qquad (5)$$

where $R_\lambda \doteq \varphi(\lambda) \sum_{t=0}^{\infty} (1-\varphi(\lambda))^t P^t$, i.e., R_λ corresponds to the *resolvent* t.p.f.

Proposition 4.3 (Invariant p.m. of perturbed process): The following hold:

(a) For $f \in C_b(\mathcal{Z})$, $\lim_{\lambda \to 0} \|R_\lambda f - \Pi f\|_\infty = 0$.
(b) For $f \in C_b(\mathcal{Z})$, $\lim_{\lambda \to 0} \|P_\lambda^L f - Q\Pi f\|_\infty = 0$.
(c) Any invariant distribution μ_λ of P_λ is also an invariant distribution of P_λ^L.
(d) Any weak limit point in $\mathfrak{P}(\mathcal{Z})$ of μ_λ, as $\lambda \to 0$, is an invariant probability measure of $Q\Pi$.

C. Invariant p.m. of one-step perturbed process

Define the finite-state Markov process $\hat{P}$ as in (3).

Proposition 4.4 (Unique invariant p.m. of $Q\Pi$): There exists a unique invariant probability measure $\hat{\mu}$ of $Q\Pi$. It satisfies

$$\hat{\mu}(\cdot) = \sum_{s \in \mathcal{S}} \pi_s \delta_s(\cdot) \qquad (6)$$

for some constants $\pi_s \geq 0$, $s \in \mathcal{S}$. Moreover, $\pi = (\pi_1, ..., \pi_{|\mathcal{S}|})$ is an invariant distribution of $\hat{P}$, i.e., $\pi = \pi\hat{P}$.

D. Proof of Theorem 3.1

Theorem 3.1(a)–(b) is a direct implication of Propositions 4.3–4.4.

V. Conclusions & Future Work

In this paper, we considered a class of reinforcement-learning algorithms that belong to the family of learning automata, and we provided an explicit characterization of the invariant probability measure of its induced Markov chain. Through this analysis, we demonstrated convergence (in a weak sense) to the set of pure-strategy states, overcoming prior restrictions necessary under an ODE-approximation analysis, such as the existence of a potential function. Thus, we opened up new possibilities for equilibrium selection through this type of algorithms that goes beyond the fine class of potential games.

Although the set of pure-strategy-states (which are the stochastically-stable states) may contain non-Nash pure strategy profiles, a follow-up analysis that excludes convergence to such pure-strategy-states may be performed (similarly to the analysis presented in [12] for diminishing step-size).

References

[1] G. C. Chasparis, "Stochastic Stability Analysis of Perturbed Learning Automata with Constant Step-Size in Strategic-Form Games," *arXiv:1702.08334 [cs]*, Feb. 2017, arXiv: 1702.08334.

[2] M. Tsetlin, *Automaton Theory and Modeling of Biological Systems*. Academic Press, 1973.

[3] K. Narendra and M. Thathachar, *Learning Automata: An introduction*. Prentice-Hall, 1989.

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

[5] T. Börgers and R. Sarin, "Learning through reinforcement and replicator dynamics," *J. Econ. Theory*, vol. 77, no. 1, pp. 1–14, 1997.

[6] I. Erev and A. Roth, "Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria," *Amer. Econ. Rev.*, vol. 88, pp. 848–881, 1998.

[7] E. Hopkins and M. Posch, "Attainability of boundary points under reinforcement learning," *Games Econ. Behav.*, vol. 53, pp. 110–125, 2005.

[8] A. Beggs, "On the convergence of reinforcement learning," *J. Econ. Theory*, vol. 122, pp. 1–36, 2005.

[9] M. Thathachar and P. Sastry, *Networks of Learning Automata: Techniques for Online Stochastic Optimization*. Kluwer Academic Publishers, 2004.

[10] G. Chasparis and J. Shamma, "Distributed dynamic reinforcement of efficient outcomes in multiagent coordination and network formation," *Dynamic Games and Applications*, vol. 2, no. 1, pp. 18–50, 2012.

[11] D. Leslie, "Reinforcement learning in games," Ph.D. dissertation, School of Mathematics, University of Bristol, 2004.

[12] G. C. Chasparis, J. S. Shamma, and A. Rantzer, "Nonconvergence to saddle boundary points under perturbed reinforcement learning," *Int. J. Game Theory*, vol. 44, no. 3, pp. 667–699, 2015.

[13] M. Posch, "Cycling in a stochastic learning algorithm for normal form games," *J. Evolutionary Econ.*, vol. 7, pp. 193–207, 1997.

[14] P. Sastry, V. Phansalkar, and M. Thathachar, "Decentralized learning of Nash equilibria in multi-person stochastic games with incomplete information," *IEEE Trans. Syst. Man Cybern.*, vol. 24, no. 5, pp. 769–777, 1994.

[15] K. Verbeeck, A. Now, J. Parent, and K. Tuyls, "Exploring selfish reinforcement learning in repeated games with stochastic rewards," *Autonomous Agents and Multi-Agent Systems*, vol. 14, no. 3, pp. 239–269, Apr. 2007.

[16] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Machine Learning Research*, vol. 4, no. Nov, pp. 1039–1069, 2003.

[17] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. Int. Conf. Machine Learning*. Morgan Kaufmann, 1994, pp. 157–163.

[18] A. C. Chapman, D. S. Leslie, A. Rogers, and N. R. Jennings, "Convergent Learning Algorithms for Unknown Reward Games," *SIAM J. Control Optim.*, vol. 51, no. 4, pp. 3154–3180, Jan. 2013.

[19] D. Monderer and L. Shapley, "Potential games," *Games Econ. Behav.*, vol. 14, pp. 124–143, 1996.

[20] D. Leslie and E. Collins, "Individual Q-Learning in Normal Form Games," *SIAM J. Control Optim.*, vol. 44, no. 2, pp. 495–514, Jan. 2005.

[21] J. R. Marden, H. P. Young, G. Arslan, and J. S. Shamma, "Payoff based dynamics for multi-player weakly acyclic games," *SIAM J. Control Optim.*, vol. 48, no. 1, pp. 373–396, 2009.

[22] G. Chasparis, A. Arapostathis, and J. Shamma, "Aspiration learning in coordination games," *SIAM J. Control and Optim.*, vol. 51, no. 1, 2013.

[23] O. Hernandez-Lerma and J. B. Lasserre, *Markov Chains and Invariant Probabilities*. Birkhauser Verlag, 2003.