Stochastic Stability of Reinforcement Learning in Positive-Utility Games

Georgios C. Chasparis

Abstract—This paper considers a class of discrete-time reinforcement-learning dynamics and provides a stochastic-stability analysis in repeatedly played positive-utility (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 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 an ODE-approximation requires the existence of a Lyapunov or a potential function, which naturally restricts the analysis to a fine class of games. To overcome these limitations, this paper introduces an alternative framework for analyzing convergence under reinforcement learning that is based upon an explicit characterization of the invariant probability measure of the induced Markov chain. We further provide a methodology for computing the invariant probability measure in positive-utility games, together with an illustration in the context of coordination games.

I. INTRODUCTION

Recently, multi-agent formulations have been utilized to tackle distributed optimization problems, since communication and computation complexity might be an issue under centralized schemes. In such formulations, decisions are usually taken in a repeated fashion, where agents select their next actions based on their own prior experience. Naturally, such multi-agent interactions can be designed as strategic-form games, where agents are repeatedly involved in a strategic interaction with a fixed payoff-matrix. Such framework finds numerous applications, including, for example, the problem of distributed overlay routing [2], distributed topology control [3] and distributed resource allocation [4].

In reinforcement learning, agents build their confidence over an action through repeated selection of this action and proportionally to the reward/payoff received from this action. Reinforcement learning has been applied in evolutionary economics, for modeling human and economic behavior [5], [6], [7], [8], [9]. 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 [10], distributed network formation and coordination problems [11].

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 centralized solutions coincides with the set of Nash equilibria.

Reinforcement learning has been utilized in strategic-form games in order for agents to gradually learn to play Nash equilibria. It may appear under alternative forms, including discrete-time replicator dynamics [5], learning automata [12], [13] or approximate policy iteration or Q-learning [14]. In all these classes of reinforcement learning, deriving conditions under which convergence to Nash equilibria is achieved may not be a trivial task especially in the case of large number of agents (as in most engineering applications). 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 [11] that is closely related to discrete-time replicator dynamics and learning automata. 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 conclu-
sions 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 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). Furthermore, we provide a methodology for computing the stochastically stable states in positive-utility games in the investigated class of reinforcement-learning dynamics, related to the last step, agent $i$ follows a uniform distribution, the random selection will be denoted by $\text{rand}_{\text{uni}}[A_i]$. If $\sigma = (1/|A|, \ldots, 1/|A|)$, i.e., it corresponds to the uniform distribution, the random selection will be denoted by $\text{rand}_{\text{uni}}[A_i]$. Property 2.1 (Positive-Utility Property): For any agent $i \in \mathcal{I}$ and any action profile $\alpha \in A_i$, $u_i(\alpha) > 0$.TABLE I

PERTURBED REINFORCEMENT LEARNING.

At fixed time instances $t = 1, 2, \ldots$, 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:

$$
\alpha_i(t+1) = \begin{cases} \text{rand}_{\text{uni}}[A_i], & \text{with probability } 1-\lambda, \\ \text{rand}_{\text{uni}}[A_i], & \text{with probability } \lambda, 
\end{cases}
$$

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

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

3) **(strategy update)** Agent $i$ revises its strategy vector $x_i \in \Delta(|A_i|)$ as follows:

$$
x_i(t+1) = x_i(t) + \epsilon \cdot \text{sup}_i [e_{\alpha_i(t+1)} - x_i(t)] = \mathcal{R}_i(\alpha(t+1), x_i(t)),
$$

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

**B. Reinforcement-learning algorithm**

We consider a form of reinforcement learning that belongs to the general class of learning automata [13]. In learning automata, each agent updates a finite probability distribution $x_i \in \mathcal{X}_i \triangleq \Delta(|A_i|)$ 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 $\lambda$, 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 $A_i$ with vertices of the simplex, $\{e_1, \ldots, e_{|A_i|}\}$. For example, if agent $i$ selects its $j$th action at time $t$, then $e_{\alpha_i(t)} = e_j$. To better see how the strategies evolve, let us consider the following toy example. Let the
current strategy of player $i$ be $x_i(t) = \left( \frac{1}{2}, \frac{1}{2} \right)^T$, i.e., player $i$ has two actions, each assigned probability $\frac{1}{2}$. Let also $\alpha_i(t + 1) = 1$, i.e., player $i$ selects the first action according to rule (1). Then, the new strategy vector for agent $i$, updated according to rule (2), is:

$$x_i(t + 1) = \frac{1}{2} \left( 1 + \epsilon u_i(\alpha(t + 1)) \right)$$

In other words, when player $i$ selects its first action, the strategy of this action is going to increase proportionally to the reward received from this action. We may say that such type of dynamics reinforce repeated selection, however the size of reinforcement depends on the reward received.

Note that, by playing a strategic-form game repeatedly over time, players do not always experience the same reward when selecting the same action, since other players may change their actions. This dynamic element of the size of reinforcement is the factor that complicates the analysis of its convergence properties, as it will become clear in the next section of related work.

Note that by letting the step-size $\epsilon$ to be sufficiently small and since the utility function $u_i(\cdot)$ is uniformly bounded in $A$, $x_i(t) \in \Delta(|A_i|)$ for all $t$.

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

### C. Related work

**Discrete-time replicator dynamics:** A type of learning dynamics which is quite closely related to the dynamics of Table I is the discrete-time version of replicator dynamics. This type of learning dynamics has been used in different variations, depending primarily on the step-size sequence considered. Moreover, analysis has primarily been restricted to decreasing step-size sequences $\epsilon(t)$ and $\lambda = 0$.

Let us consider the notation of the proposed dynamic update of rule (2). Arthur [5] considered a similar rule, with a step-size sequence of agent $i$ being defined as $\epsilon_i(t) = 1/(c t^\nu + u_i(\alpha(t + 1)))$ for some positive constant $c$ and for $0 < \nu < 1$ (in the place of the constant step size $\epsilon$ of (2)). A comparative model is also used by Hopkins and Posch in [8], 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 [7]. Some similarities are also shared with the Cross’ learning model of [6], where $\epsilon(t) = 1$ and $u_i(\alpha(t)) \leq 1$, and its modification presented by Leslie in [15], 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 [11]. A state-dependent perturbation term has also been investigated in [16]. The perturbation parameter may serve as an equilibrium selection tool for reinforcing convergence to Nash equilibria without necessarily imposing any constraints in the utility of the players (as in [17]) or by employing an urn-process type step-size sequence (as in [8]). 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 [16], this can be guaranteed only under strong conditions in the payoff matrix. For example, as shown in [16, Proposition 8], and under the ODE-method for stochastic approximations, excluding convergence to mixed strategy profiles requires a) the existence of a potential function, and 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, making any extension to games of more than 2 players and 2 actions practically impossible.

**Learning automata:** Learning automata, as first introduced by [12], has attracted attention with respect to the control of complex and distributed systems due to its simple structure and low computational complexity (cf., [13] Chapter 1). Variable-structure stochastic automata may incorporate a form of reinforcement of favorable actions. Therefore, such stochastic automata bear a lot of similarities to the discrete-time analogs of replicator dynamics discussed above.

To see an example of such stochastic automata, consider the linear reward-inaction scheme described in [13, Chapter 4]. Comparing it with the reinforcement rule of (2), the linear reward-inaction scheme accepts a utility function that accepts only two values, $u_i(\alpha) \in \{0, 1\}$, where 0 corresponds to an unfavorable response and 1 corresponds to a favorable...
one. More general forms can also be used when the utility function may accept discrete or continuous values in the unit interval \([0, 1]\). Analysis of learning automata in games has been restricted to zero-sum and identical-interest games \([13, 19]\). In particular, in identical interest games, convergence analysis has been derived for games of small number of players and actions, due to the difficulty in deriving conditions for absolute monotonicity, which corresponds to the property that the expected utility received by each player increases monotonically in time (cf., \([13, \text{Definition 8.1}]\)). Similar are the results presented in \([19]\).

The property of absolute monotonicity is closely related to the existence of a potential function, as in the case of potential games \([20]\). The previous analysis presented in \([13, 19]\) can easily be extended in the context of potential games (similarly to the discussion in \([16, \text{Proposition 8}]\)). As in the case of discrete-time replicator dynamics, special attention is required to the conditions under which convergence to non-Nash action profiles can be excluded. As we will show in a forthcoming section, convergence to non-Nash action profiles cannot be excluded when the step-size sequence is constant even if the utility function satisfies \(u_i(\alpha) \in [0, 1]\) as in the learning automata. (The behavior under decreasing step-size is different as \([16, \text{Proposition 2}]\) has shown.) Furthermore, deriving conditions for excluding convergence to mixed strategy profiles continues to be an issue for the case of learning automata, as in the case of discrete-time replicator dynamics.

Recognizing these issues, reference \([21]\) 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, an approach that does not fit to the distributed nature of dynamics pursued here.

\textit{Q-learning:} Similar questions of convergence to Nash equilibria also appear in alternative reinforcement learning formulations, such as approximate dynamic programming and \textit{Q}-learning. Usually, under \textit{Q}-learning, players keep track of the discounted running average reward received by each action, based on which optimal decisions are made (see, e.g., \([22]\)). Convergence to Nash equilibria can be accomplished under a stronger set of assumptions, which increases the computational complexity of the dynamics. For example, in the Nash-\textit{Q} learning algorithm of \([14]\), it is indirectly assumed that agents need to have full access to the joint action space and the rewards received by other agents.

More recently, reference \([23]\) introduced a \textit{Q}-learning scheme in combination with either adaptive play or better-reply dynamics in order to attain convergence to Nash equilibria in potential games \([20]\) or weakly-acyclic games. However, this form of dynamics requires that each player observes the actions selected by the other players, since a \textit{Q}-value needs to be assigned for each joint action.

When the evaluation of the \textit{Q}-values is totally independent, as in the individual \textit{Q}-learning in \([22]\), 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 convergence results in multi-player games. This is a main drawback for \textit{Q}-learning dynamics in strategic-form games as also pointed out in \([24]\) when the decision making process is according to stochastic fictitious play. To overcome this drawback, in the context of stochastic dynamic games, reference \([24]\) employs an additional feature (motivated by \([25]\), namely \textit{exploration phases}. In any such \textit{exploration phase}, all agents use constant policies, something that allows the accurate computation of the optimal \textit{Q}-factors. We may argue that the introduction of a common exploration phase for all agents partially destroys the distributed nature of the dynamics, since it requires a synchronization between agents’ exploration phases.

\textit{Payoff-based learning:} The aforementioned types of dynamics can also 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 to establish convergence to Nash equilibria through alternative payoff-based learning dynamics, (see, e.g., the benchmark-based dynamics of \([26]\) for convergence to Nash equilibria in weakly-acyclic games, the benchmark-based dynamics of \([27]\) for maximizing welfare or the aspiration-based dynamics in \([28]\) for convergence to efficient outcomes in coordination games). For these types of dynamics, convergence to Nash equilibria can be established without requiring any strong monotonicity property (as in the multi-player weakly-acyclic games in \([25]\)). In fact, stronger convergence results can be derived that go beyond Nash equilibria.

However, noisy reward observations, which are present in many engineering applications, cannot be dealt properly through benchmark-based dynamics. For example, in reference \([26]\), under benchmark-based dynamics and noisy reward observations, \textit{exploration phases} need to be introduced through which an agent plays a fixed action with a small experimentation probability. If such exploration phases are large in duration (as required by the results in \([26]\)), this reduces the robustness of the dynamics to dynamic changes in the environment (e.g., changes in the utility function).

In the proposed dynamics of Table I decisions are not directly based on the current utility received (as in the benchmark-based dynamics). Instead, decisions are based on strategies \(x\) which are updated through a discounted averaging of the payoffs received. Thus, there is an indirect filtering of noisy observations that does not influence the speed of decision making. This benefit of reinforcement-learning dynamics has been shown in the context of adaptive system identification for bilinear systems presented in \([29]\).

Furthermore, stronger convergence results of reinforcement
learning dynamics can be derived usually through additional features with respect to processing local performance measurements, as in the feedback-based variation introduced in [11] for achieving convergence to the efficient Nash equilibrium. Although those results are local, they demonstrated the potential of reinforcement learning for efficient equilibrium selection. The present work is a contribution towards this direction, as clarified in detail in the following section.

D. Objective and contribution

This paper provides a framework for analyzing convergence in multi-player and multi-action positive-utility strategic-form games where players implement a class of perturbed reinforcement learning dynamics. Under a fully distributed setup, where agents have only access to their own actions and performance measurements, the contributions of this paper are as follows:

1) We provide an equivalent finite-dimensional characterization of the infinite-dimensional induced Markov chain of the dynamics, that simplifies significantly the characterization of its invariant probability measure. This simplification is based upon a weak-convergence result (Theorem 5.1).

2) We capitalize on this simplification and provide a methodology for computing stochastically stable states in positive-utility strategic-form games (Theorem 5.7).

3) We illustrate the utility of this methodology in establishing stochastic stability in a class of coordination games with no restriction on the number of players or actions (Theorem 6.1).

These contributions significantly extend the utility of reinforcement learning for distributed learning, since current convergence results in discrete-time replicator dynamics or learning automata in games only apply to the fine class of potential games. These restrictions are overcome through the proposed analytical framework by directly analyzing the invariant probability measure of the induced Markov chain. This also provides an alternative to Q-learning dynamics, where strong convergence guarantees in a fully distributed setup are currently restricted to 2-player games.

We have to note that the illustration result in coordination games (contribution (3) above) constitutes the first convergence result in the context of reinforcement learning in repeatedly played strategic-form games with the following features: a) a fully distributed setup, b) more than two players, and c) a set of games that is larger than potential games.

The derived convergence results may not be as strong as the ones derived under alternative payoff-based learning dynamics, as discussed in Section II-C. However, we should not neglect the fact that in comparison to these classes of dynamics, reinforcement learning better incorporates noisy observations, which can be a highly attractive feature for engineering applications. Furthermore, additional features can allow for stronger convergence guarantees, as presented in [11]. Given the significantly simplified analytical framework presented here, the prospects of even stronger convergence guarantees are in fact open-ended.

This paper is an extension over an earlier version appeared in [11], which only focused on contribution (1) above.

III. Stochastic Stability

In this section, we provide a characterization of the invariant probability measure \( \mu_\lambda \) of the induced Markov chain \( P_\lambda \) of the dynamics of Table I. The importance lies in an equivalence relation (established through a weak-convergence argument) of \( \mu_\lambda \) with an invariant distribution of a finite-state Markov chain. Characterization of the stochastic stability of the dynamics will follow directly due to the Birkhoff’s individual ergodic theorem.

This simplification in the characterization of \( \mu_\lambda \) will be the first important step for providing specialized results for stochastic stability in strategic-form games.

A. Terminology and notation

Let \( Z \triangleq A \times X \), where \( X \triangleq X_1 \times \ldots \times X_n \), i.e., pairs of joint actions \( \alpha \) and strategy profiles \( x \). We will denote the elements of the state space \( Z \) by \( z \).

The set \( A \) is endowed with the discrete topology, \( X \) with its usual Euclidean topology, and \( Z \) with the corresponding product topology. We also let \( \mathcal{B}(Z) \) denote the Borel \( \sigma \)-field of \( Z \), and \( \mathcal{F}(Z) \) the set of probability measures (p.m.) on \( \mathcal{B}(Z) \) endowed with the Prohorov topology, i.e., the topology of weak convergence. The reinforcement learning algorithm of Table I defines an \( Z \)-valued Markov chain. Let \( P_\lambda : Z \times \mathcal{B}(Z) \to [0,1] \) 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 \( P : Z \times \mathcal{B}(Z) \to [0,1] \) denote the t.p.f. of the unperturbed process, i.e., when \( \lambda = 0 \).

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

\[
P_\lambda f(z) \triangleq \int_Z P_\lambda(z,dy)f(y),
\]

and

\[
\mu[f] \triangleq \int_Z \mu(dx)f(z), \quad \text{for } \mu \in \mathcal{F}(Z).
\]

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

B. Stochastic stability

First, we note that both $P$ and $P_\lambda (\lambda > 0)$ satisfy the weak Feller property (cf., [30, Definition 4.4.2]).

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

Proof. See Appendix A. □

The measure $\mu_\lambda \in \mathfrak{B}(Z)$ is called an invariant probability measure (i.p.m.) for $P_\lambda$ if

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

Since $Z$ defines a locally compact separable metric space and $P, P_\lambda$ have the weak Feller property, they both admit an i.p.m., denoted $\mu$ and $\mu_\lambda$, respectively [30, Theorem 7.2.3].

We would like to characterize the stochastically stable states $z \in Z$ of $P_\lambda$, that is any state $z \in Z$ for which any collection of i.p.m. $\{\mu_\lambda \in \mathfrak{B}(Z) : \mu_\lambda P_\lambda = \mu_\lambda, \lambda > 0\}$ 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 $s = (\alpha, x) \in Z$ such that for all $i \in I$, $x_i = e_{\alpha_i}$, i.e., $x_i$ coincides with the vertex of the probability simplex $\Delta(|A_i|)$ which assigns probability 1 to action $\alpha_i$.

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

Theorem 3.1 (Stochastic Stability): There exists a unique probability vector $\pi = (\pi_1, \ldots, \pi_{|S|})$ such that for any collection of i.p.m.’s $\{\mu_\lambda \in \mathfrak{B}(Z) : \mu_\lambda P_\lambda = \mu_\lambda, \lambda > 0\}$, 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 $\pi$ is an invariant distribution of the (finite-state) Markov process $\hat{P}$, such that, for any $s, s' \in \mathcal{S}$,

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

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 $\pi$).

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

Theorem 3.1 implicitly provides a stochastically stability argument. In fact, the expected asymptotic behavior of the dynamics can be characterized by $\hat{\mu}$ and, therefore, $\pi$. In particular, by Birkhoff’s individual ergodic theorem [30, Theorem 2.3.4], the weak convergence of $\mu_\lambda$ to $\hat{\mu}$, and the fact that $\mu_\lambda$ is ergodic, we have that the expected percentage of time that the process spends in any $O \in \mathcal{B}(Z)$ such that $\partial O \cap S \neq \emptyset$ is given by $\hat{\mu}(O)$ as the experimentation probability $\lambda$ approaches zero and time increases, i.e.,

$$\lim_{\lambda \to 0} \left(\lim_{t \to \infty} \frac{1}{t} \sum_{k=0}^{t-1} P_\lambda^k(x, O)\right) = \hat{\mu}(O).$$

C. Discussion

Theorem 3.1 establishes “equivalence” (in a weak convergence sense) of the original (perturbed) learning process with a simplified process, where only one player trembles at the first iteration and then no player trembles thereafter. This simplification in the analysis has originally been capitalized to analyze aspiration learning dynamics in [31, 28], and it is based upon the observation that under the unperturbed process, agents’ strategies will converge to a pure strategy state, as it will be shown in the forthcoming Section IV.

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 $S$ (Definition 3.1). In other words, we should expect that as the perturbation parameter $\lambda$ approaches zero, the algorithm spends the majority of the time on pure strategy states. 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.

In the forthcoming Section IV, we will use this result to provide a methodology for computing the set of stochastically stable states. This methodology will further be illustrated in the context of coordination games.

IV. TECHNICAL DERIVATION

In this section, we provide the main steps for the proof of Theorem 3.1. We begin by investigating the asymptotic behavior of the unperturbed process $P$, and then we characterize the i.p.m. of the perturbed process with respect to the p.s.s.’s $S$.

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, $A_\infty$ corresponds to the event that agents eventually play the same action profile, while $B_\infty$ 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 $i \in Z$. Then, the following hold:

(a) $\inf_{z \in Z} \mathbb{P}_z[B_\infty] > 0$,
(b) $\inf_{z \in Z} \mathbb{P}_z[A_\infty] = 1$.

Proof. See Appendix B $\square$

Statement (a) of Proposition 4.1 states that the probability that agents never change their actions is bounded away from zero, while statement (b) states that the probability that eventually agents play the same action profile is one. This also indicates that any invariant measure of the unperturbed process can be characterized with respect to the pure strategy states $S$, which is established by the following proposition.

Proposition 4.2 (Limiting t.p.f. of unperturbed process): Let $\mu$ denote an i.p.m. of $P$. Then, there exists a t.p.f. $\Pi$ on $Z \times \mathcal{B}(Z)$ with the following properties:

(a) for $\mu$-a.e. $z \in Z$, $\Pi(z, \cdot)$ is an i.p.m. for $P$;
(b) for all $f \in C_b(Z)$, $\lim_{t \to \infty} \|P^t f - \Pi f\|_\infty = 0$;
(c) $\mu$ is an i.p.m. for $\Pi$;
(d) the support of $\Pi$ is on $S$ for all $z \in Z$.

Proof. The state space $Z$ is a locally compact separable metric space and the t.p.f. of the unperturbed process $P$ admits an i.p.m. due to Proposition 3.1. Thus, statements (a), (b) and (c) follow directly from Proposition 3.1 Theorem 5.2.2 (a), (b), (e)].

(d) Let us assume that the support of $\Pi$ includes points in $Z$ other than the pure strategy states in $S$. Then, there exists an open set $O \in \mathcal{B}(Z)$ such that $O \cap S = \emptyset$ and $\Pi(z^*, O) > 0$ for some $z^* \in Z$. According to (b), $P^t$ converges weakly to $\Pi$. Thus, from Portmanteau theorem (cf., Proposition 3.1 Theorem 1.4.16), we have that $\liminf_{t \to \infty} P^t(z^*, O) \geq \Pi(z^*, O) > 0$. This is a contradiction of Proposition 4.1(b), which concludes the proof. $\square$

Proposition 4.2 states that the limiting unperturbed t.p.f. converges weakly to a t.p.f. $\Pi$ which accepts the same i.p.m. as $P$. Furthermore, the support of $\Pi$ is the set of pure strategy states $S$. 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. $\Pi$, as it will be shown in the following section.

B. Invariant probability measure (i.p.m.) of perturbed process

According to the definition of reinforcement learning of Table 1 when a player updates its action, there is a small probability $\lambda > 0$ that it “trembles,” i.e., it selects a new action according to a uniform distribution (instead of using its current strategy). Thus, we can decompose the t.p.f. induced by the one-step update of reinforcement learning 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 $Q_\lambda$ corresponds to the t.p.f. when at least one agent trembles. Note that $\varphi(\lambda) \to 0$ as $\lambda \downarrow 0$.

Define also $Q$ as the t.p.f. where only one player trembles, and $Q^* \equiv Q$ as the t.p.f. where at least two players tremble. Then, we may write:

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

where $\psi(\lambda) = 1 - \frac{n\lambda(1-\lambda)^{n-1}}{1-(1-\lambda)^n}$ corresponds to the probability that at least two players tremble given that at least one player trembles. It also satisfies $\psi(\lambda) \to 0$ as $\lambda \downarrow 0$, which establishes an approximation of $Q_\lambda$ by $Q$ as the perturbation factor $\lambda$ approaches zero.

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

$$P^L_\lambda = \varphi(\lambda) \sum_{t=0}^{\infty} (1 - \varphi(\lambda))^t Q_\lambda P^t = Q_\lambda R_\lambda$$

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

In the following proposition, we establish weak-convergence of the lifted t.p.f. $P^L_\lambda$ with $Q\Pi$ as $\lambda \downarrow 0$, which will further allow for explicit characterization of the weak limit points of the i.p.m. of $P_\lambda$.

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

(a) For $f \in C_b(Z)$, $\lim_{\lambda \to 0} \|R_\lambda f - \Pi f\|_\infty = 0$.
(b) For $f \in C_b(Z)$, $\lim_{\lambda \to 0} \|P^L_\lambda f - Q\Pi f\|_\infty = 0$.
(c) Any invariant distribution $\mu_\lambda$ of $P_\lambda$ is also an invariant distribution of $P^L_\lambda$.
(d) Any weak limit point in $\mathcal{B}(Z)$ of $\mu_\lambda$, as $\lambda \to 0$, is an i.p.m. of $Q\Pi$.

Proof. (a) For any $f \in C_b(Z)$, we have

$$\|R_\lambda f - \Pi f\|_\infty = \|\varphi(\lambda) \sum_{t=0}^{\infty} (1 - \varphi(\lambda))^t P^t f - \Pi f\|_\infty$$

$$= \|\varphi(\lambda) \sum_{t=0}^{\infty} (1 - \varphi(\lambda))^t (P^t f - \Pi f)\|_\infty$$

where we have used the property $\varphi(\lambda) \sum_{t=0}^{\infty} (1 - \varphi(\lambda))^t = 1$. Note that

$$\varphi(\lambda) \sum_{t=0}^{\infty} (1 - \varphi(\lambda))^t \|P^t f - \Pi f\|_\infty$$

$$\leq (1 - \varphi(\lambda))^T \sup_{t \geq T} \|P^t f - \Pi f\|_\infty.$$
From Proposition 4.2(b), we have that for any $\delta > 0$, there exists $T = T(\delta) > 0$ such that the r.h.s. is uniformly bounded by $\delta$ for all $t \geq T$. Thus, the sequence

$$A_T \doteq \varphi(\lambda) \sum_{i=0}^{T} (1 - \varphi(\lambda))^i (P^i f - \Pi f)$$

is Cauchy and therefore convergent (under the sup-norm). In other words, there exists $A \in \mathbb{R}$ such that $\lim_{T \to \infty} \|A_T - A\|_{\infty} = 0$. For every $T > 0$, we have

$$\|R_{\lambda} f - \Pi f\|_{\infty} \leq \|A_T\|_{\infty} + \|A - A_T\|_{\infty}.$$ 

Note that

$$\|A_T\|_{\infty} \leq \varphi(\lambda) \sum_{i=0}^{T} (1 - \varphi(\lambda))^i \|P^i f - \Pi f\|_{\infty}.$$ 

If we take $\lambda \downarrow 0$, then the r.h.s. converges to zero. Thus,

$$\|R_{\lambda} f - \Pi f\|_{\infty} \leq \|A - A_T\|_{\infty},$$

for all $T > 0$, which concludes the proof.

(b) For any $f \in C_b(Z)$, we have

$$\|P_{\lambda}^L f - Q\Pi f\|_{\infty} \leq \|Q\lambda(R_{\lambda} f - \Pi f)\|_{\infty} + \|Q\lambda\Pi f - Q\Pi f\|_{\infty} \leq \|R_{\lambda} f - \Pi f\|_{\infty} + \|Q\lambda\Pi f - Q\Pi f\|_{\infty}.$$ 

The first term of the r.h.s. approaches 0 as $\lambda \downarrow 0$ according to (a). The second term of the r.h.s. also approaches 0 as $\lambda \downarrow 0$ since $Q_{\lambda} \to Q$ as $\lambda \downarrow 0$.

(c) By definition of the perturbed t.p.f. $P_{\lambda}$, we have

$$P_{\lambda} R_{\lambda} = (1 - \varphi(\lambda)) PR_{\lambda} + \varphi(\lambda) Q_{\lambda} R_{\lambda}.$$ 

Note that $Q_{\lambda} R_{\lambda} = P_{\lambda}^L$ and $(1 - \varphi(\lambda)) PR_{\lambda} = R_{\lambda} - \varphi(\lambda) I$, where $I$ corresponds to the identity operator. Thus,

$$P_{\lambda} R_{\lambda} = R_{\lambda} - \varphi(\lambda) I + \varphi(\lambda) P_{\lambda}^L.$$ 

For any i.p.m. of $P_{\lambda}$, $\mu_{\lambda}$, we have

$$\mu_{\lambda} P_{\lambda} R_{\lambda} = \mu_{\lambda} R_{\lambda} - \varphi(\lambda) \mu_{\lambda} + \varphi(\lambda) \mu_{\lambda} P_{\lambda}^L,$$

which equivalently implies that $\mu_{\lambda} = \mu_{\lambda} P_{\lambda}^L$, since $\mu_{\lambda} P_{\lambda} = \mu_{\lambda}$. We conclude that $\mu_{\lambda}$ is also an i.p.m. of $P_{\lambda}^L$.

(d) Let $\hat{\mu}$ denote a weak limit point of $\mu_{\lambda}$ as $\lambda \downarrow 0$. To see that such a limit exists, take $\hat{\mu}$ to be an i.p.m. of $P_T$. Then,

$$\|P_T f - P_T f\|_{\infty} = \|\mu_{\lambda} (P_T f - P_T f)\|_{\infty} = \|\mu_{\lambda} (I - \hat{\mu})(P_T f)\|_{\infty}.$$ 

Note that the weak convergence of $P_{\lambda}$ to $P_T$ it necessarily implies that $\mu_{\lambda} \Rightarrow \hat{\mu}$. Note further that

$$\hat{\mu}[f] - \mu_{\lambda}[f] = \hat{\mu}[f] - \mu_{\lambda}[f] + \mu_{\lambda}[P_T^L f - Q\Pi f] + (\mu_{\lambda}[Q\Pi f] - \hat{\mu}[Q\Pi f]).$$

The first and the third term of the r.h.s. approaches 0 as $\lambda \downarrow 0$ due to the fact that $\mu_{\lambda} \Rightarrow \hat{\mu}$. The same holds for the second term of the r.h.s. due to part (b). Thus, we conclude that any weak limit point of $\mu_{\lambda}$ as $\lambda \downarrow 0$ is an i.p.m. of $Q\Pi$. □

Proposition 4.3 establishes convergence (in a weak sense) of the i.p.m. $\mu_{\lambda}$ of the perturbed process to an i.p.m. of $Q\Pi$. In the following section, this convergence result will allow for a more explicit characterization of $\mu_{\lambda}$ as $\lambda \downarrow 0$.

C. Equivalent finite-state Markov process

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

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

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

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

Proof. From Proposition 4.2(d), we know that the support of $\Pi$ is the set of pure strategy states $S$. Thus, the support of $Q\Pi$ is also on $S$. From Proposition 4.3 we know that $Q\Pi$ admits an i.p.m., say $\hat{\mu}$, whose support is also $S$. Thus, $\mu$ admits the form of (6), for some constants $\pi_s \geq 0$, $s \in S$.

For any two distinct $s, s' \in S$, note that $N_\delta(s', \delta > 0)$ is a continuity set of $Q\Pi(s, \cdot)$, i.e., $Q\Pi(s, \delta N_\delta(s')) = 0$. Thus, from Portmanteau theorem, given that $Q\Pi^p \Rightarrow Q\Pi$,

$$Q\Pi(s, N_\delta(s')) = \lim_{t \to \infty} Q\Pi^p(s, N_\delta(s')) = \hat{P}_{ss'}.$$ 

If we also define $\pi_s = \hat{\mu}(N_\delta(s))$, then

$$\pi_s' = \hat{\mu}(N_\delta(s')) = \sum_{s \in S} \pi_s Q\Pi(s, N_\delta(s')) = \sum_{s \in S} \pi_s \hat{P}_{ss'},$$

which shows that $\pi$ is an invariant distribution of $\hat{P}$, i.e., $\pi = \pi \hat{P}$.

It remains to establish uniqueness of the invariant distribution of $Q\Pi$. Note that the set $S$ of pure strategy states is isomorphic with the set $A$ of action profiles. If agent $i$ trembles (as t.p.f. $Q$ dictates), then all actions in $A_i$ have positive probability of being selected, i.e., $Q(\alpha, (\alpha'_i, \alpha_{-i})) > 0$ for all $\alpha'_i \in A_i$ and $i \in I$. It follows by Proposition 4.1 that $Q\Pi(\alpha, (\alpha'_i, \alpha_{-i})) > 0$ for all $\alpha'_i \in A_i$ and $i \in I$. Finite induction then shows that $(Q\Pi)^p(\alpha, \alpha'_i) > 0$ for all $\alpha, \alpha'_i \in A$. It follows that if we restrict the domain of $Q\Pi$ to $S$, it defines an irreducible stochastic matrix. Therefore, $Q\Pi$ has a unique i.p.m. □

D. Proof of Theorem 3.1

Theorem 3.1(a)–(b) is a direct implication of Propositions 4.3, 4.4.
Fig. 1. Examples of $s$-graphs in case $S$ contains four states.

V. STOCHASTICALLY STABLE STATES

In this section, we capitalize on Theorem 5.1(a) which establishes weak convergence of the i.p.m. of the process to the stationary distribution of a finite-state Markov process. This significantly simplifies the computation of the stochastically stable states as this section will demonstrate.

A. Background on finite Markov chains

In order to compute the invariant distribution of a finite-state, irreducible and aperiodic Markov chain, we are going to consider a characterization introduced by [32]. In particular, for finite Markov chains an invariant measure can be expressed as the ratio of sums of products consisting of transition probabilities. These products can be described conveniently by means of graphs on the set of states of the chain. In particular, let $S$ be a finite set of states, whose elements will be denoted by $s_k$, $s_\ell$, etc., and let a subset $W$ of $S$.

Definition 5.1: ($W$-graph) A graph consisting of arrows $s_k \to s_\ell$ ($s_k \in S \setminus W$, $s_\ell \in S$, $s_\ell \neq s_k$) is called a $W$-graph if it satisfies the following conditions:

1) every point $k \in S \setminus W$ is the initial point of exactly one arrow;
2) there are no closed cycles in the graph; or, equivalently, for any point $s_k \in S \setminus W$ there exists a sequence of arrows leading from it to some point $s_\ell \in W$.

Fig. 1 provides examples of $\{s\}$-graphs for some state $s \in S$ when $S$ contains four states. We will denote by $\mathcal{G}(W)$ the set of $W$-graphs and we shall use the letter $g$ to denote graphs. If $\tilde{P}_{s_k s_\ell}$ are nonnegative numbers, where $s_k, s_\ell \in S$, define also the transition probability along path $g$ as

$$\pi(g) = \prod_{(s_k \to s_\ell) \in g} \tilde{P}_{s_k s_\ell}. $$

The following Lemma holds:

Lemma 5.1 (Lemma 6.3.1 in [32]): Let us consider a Markov chain with a finite set of states $S$ and transition probabilities $\{\tilde{P}_{s_k s_\ell}\}$ and assume that every state can be reached from any other state in a finite number of steps. Then, the stationary distribution of the chain is $\pi = [\pi_s]$, where

$$\pi_s = \frac{R_s}{\sum_{s_k \in S} R_{s_k}}, \ s \in S \tag{7}$$

where $R_s = \sum_{g \in \mathcal{G}(s)} \pi(g)$.

In other words, in order to compute the weight that the stationary distribution assigns to a state $s \in S$, it suffices to compute the ratio of the transition probabilities of all $\{s\}$-graphs over the transition probabilities of all graphs.

B. Approximation of one-step transition probability

We wish to provide an approximation in the computation of the transition probabilities between states in $S$ since this will allow for explicitly computing the stationary distribution $\pi$ of Theorem 3.1. Based on the definition of the t.p.f. QII, and as $\lambda \downarrow 0$, a transition from $s$ to $s'$ influences the stationary distribution only if $s$ differs from $s'$ in the action of a single player. This observation will be capitalized by the forthcoming Lemmas 5.2 5.3, to approximate the transition probability from $s$ to $s'$.

Let $\tau(D)$ denote the first hitting time of the unperturbed process to the set $D \subset \mathbb{Z}$. Denote the minimum hitting time of a set $D \subset \mathbb{Z}$ as $\tau_*^\delta(D)$ when the process starts from state $s \in S$. Let us also define the set

$$D_{i,\ell}(\alpha) = \{(\alpha, x) \in \mathbb{Z} : x_{i\alpha_i} > 1 - H_i(\alpha)^\ell\},$$

where $H_i(\alpha) = 1 - e_{u_i}(\alpha)$. The set $D_{i,\ell}(\alpha)$ defines the unreachable set in the strategy space of agent $i$ when starting from $x_{i\alpha_i} = 0$ under QII. In particular, if $x_{i\alpha_i}(0) = 0$, and agent $i$ plays action $\alpha_i$ for $\ell$ consecutive times, then $x_{i\alpha_i}(\ell) = 1 - H_i(\alpha)^\ell$.

Lemma 5.2 (One-step transition probability): Consider any two action profiles $\alpha, \alpha' \in \mathcal{A}$ which differ in the action of a single player $j$. Let $s, s'$ define the corresponding pure strategy states associated with $\alpha$ and $\alpha'$, respectively. Let also $z = (\alpha', x')$, where $x' = e_{\alpha_i} + e_{u_j}(\alpha')(e_{\alpha_i} - e_{\alpha_i})$, which corresponds to the state after agent $j$ perturbed once starting from $s$ and played $\alpha'$. Define also $\tilde{P}_{ss'}(\delta) = \mathbb{P}_z[\pi(N_\delta(s')) < \infty]$ which corresponds to the probability that the process transits from the perturbed state $z$ to a $\delta$-neighborhood of $s'$ in finite time. Then, the following hold:

(a) The transition probability from $s$ to $s'$ under QII can be approximated as follows:

$$\tilde{P}_{ss'} = \lim_{\delta \to 0} \lambda_{[\mathcal{A}_j]} \tilde{P}_{ss'}(\delta) \tag{8}$$

(b) For sufficiently small $\epsilon > 0$, such that $e_{u_i}(\alpha) < 1$ for all $i \in \mathcal{T}$ and $\alpha \in \mathcal{A}$, we have

$$\tilde{P}_{ss'}(\delta) = \mathbb{P}_z[\alpha(t + 1) = \alpha', t < \tau_*^\delta(N_\delta(s'))],$$

where $N_{s'}(s)$ are the neighbors of $s'$ in the set $\{s\}$-graphs.
which corresponds to the probability of the shortest path from $z$ to $N_\delta(s')$. Furthermore,
\[ \mathbb{P}_z [\alpha(t + 1) = \alpha', t < \tau^*_\alpha(N_\delta(s'))] \sim \mathcal{O}\left(\epsilon^* (N_\delta(s'))\right). \]

**Proof.** (a) This is a direct implication of the definition of QII t.p.f..

(b) Observe that one possibility for realizing a transition from $s$ to $N_\delta(s')$ is to follow the shortest path, that is, the path of playing action $\alpha'$ consecutively. Thus, 
\[ \tilde{P}_{ss'}(\delta) \geq \mathbb{P}_z [\alpha(t + 1) = \alpha', t < \tau^*_\alpha(N_\delta(s'))]. \]

Let us denote by $t_k$, $k \in \mathbb{N}$, a subsequence of the iteration index $t$. Note that 
\[ \tilde{P}_{ss'}(\delta) \leq \mathbb{P}_z [\exists \{t_k\} : \alpha(t_k + 1) = \alpha', t_k < \tau(D_{j,k}(\alpha'))], \]
\[ \leq \mathbb{P}_z [\exists \{t_k\} : \alpha(t_k + 1) = \alpha', Z_{t_k} \in D_{j,k}(\alpha')^c, t_k < \tau^*_\alpha(N_\delta(s'))]. \]

For any sample path of the unperturbed process that reaches the set $N_\delta(s')$, action profile $\alpha'$ will be played for at least $\tau^*_\alpha(N_\delta(s'))$ iterations when starting from $z$ and $\epsilon$ is taken sufficiently small. Furthermore, when action profile $\alpha'$ has been played for the $k$th time (at time $t_k + 1$), the state at time $t_k$ may not have reached $D_{j,k}(\alpha')$ (by definition of the set $D_{j,k}(\alpha')$). These observations result in the first inequality. Equivalently, $t_k < \tau(D_{j,k}(\alpha'))$ implies that the previous state $Z_{t_k}$ may only be within $D_{j,k}(\alpha')^c$, by the Markov property. This observation implies the second inequality.

The last inequality can also be written as:
\[ \tilde{P}_{ss'}(\delta) \leq \mathbb{P}_z [\exists \{t_k\} : \alpha(t_k + 1) = \alpha', Z_{t_k} \in D_{j,k}(\alpha')^c, t_k < \tau^*_\alpha(N_\delta(s'))]. \]

By the Markov property,
\[ \tilde{P}_{ss'}(\delta) \leq \mathbb{P}_z [\exists \{t_k\} : \alpha(t_k + 1) = \alpha', Z_{t_k} \in D_{j,k}(\alpha')^c, k < \tau^*_\alpha(N_\delta(s'))]. \]

Let us assume that along a sample path from $s$ to $N_\delta(s')$ and at iteration $t$, the strategy of agent $j$ with respect to action $\alpha'_j$ is $x_{j\alpha'_j}(t) = \rho > 0$. If agent $j$ selects action $\alpha'_j$ at time $t + 1$, then
\[ x_{j\alpha'_j}(t + 1) = \rho + \epsilon u_j(\alpha')(1 - \rho) = \epsilon u_j(\alpha') + H_j(\alpha')\rho = x_{j\alpha'_j}. \]

If, instead, agent $j$ selects action $\alpha_j \neq \alpha'_j$ at time $t + 1$ and $\alpha'_j$ at time $t + 2$, i.e., it deviates from playing action $\alpha'_j$, then the strategy evolves as follows:
\[ x_{j\alpha'_j}(t + 1) = \rho + \epsilon u_j(\alpha)(-\rho) = H_j(\alpha)\rho, \]
\[ x_{j\alpha'_j}(t + 2) = H_j(\alpha)\rho + \epsilon u_j(\alpha')(1 - H_j(\alpha)\rho) = H_j(\alpha')H_j(\alpha)\rho + \epsilon u_j(\alpha') < x^*_{\alpha'_j}. \]

If $\epsilon u_j(\alpha) < 1$. Informally, any single deviation from the shortest path to $s'$ may not recover the drop in the strategy at the next iteration. Thus, along any path to $N_\delta(s')$, action $\alpha'$ will be played for at least $\tau^*_\alpha(N_\delta(s'))$ times.

\[ \tau^*_\alpha(N_\delta(s')) \leq \prod_{k=0}^{\tau^*_\alpha(N_\delta(s'))-1} \mathbb{P}_z [\alpha(t_k + 1) = \alpha' | Z_{t_k} \in D_{k}(\alpha')^c] \]
\[ \tau^*_\alpha(N_\delta(s')) \leq \prod_{t=0}^{\tau^*_\alpha(N_\delta(s'))-1} \sup_{z \in D_{k}(\alpha')^c} \mathbb{P}_z [\alpha(t + 1) = \alpha'] \]
\[ = \mathbb{P}_z [\alpha(t + 1) = \alpha', t < \tau^*_\alpha(N_\delta(s'))]. \]

Thus, the conclusion follows.

To compute the order of the transition probability of the shortest path, note first that
\[ \mathbb{P}_z [\alpha(t + 1) = \alpha', t < \tau^*_\alpha(N_\delta(s'))] \]
\[ = \prod_{t=0}^{\tau^*_\alpha(N_\delta(s'))-1} \{1 - H_j(\alpha')^t\} \]
\[ = \prod_{t=1}^{\tau^*_\alpha(N_\delta(s'))} \{1 - H_j(\alpha')^t\}. \]

By the binomial series expansion, we have
\[ (1 - \epsilon u_j(\alpha'))^t = 1 + \sum_{k=1}^{t} \binom{t}{k} (-1)^k (\epsilon u_j(\alpha'))^k, \]

which implies that
\[ \prod_{t=1}^{\tau^*_\alpha(N_\delta(s'))} \{1 - H_j(\alpha')^t\} \sim \mathcal{O} \left(\epsilon^{\tau^*_\alpha(N_\delta(s'))}\right). \]

Thus, 
\[ \tau^*_\alpha(N_\delta(s')) \prod_{t=1}^{\tau^*_\alpha(N_\delta(s'))} \{1 - H_j(\alpha')^t\} \sim \mathcal{O} \left(\epsilon^{\tau^*_\alpha(N_\delta(s'))}\right). \]

which concludes the proof. □

Lemma 5.2 provides a tool for simplifying the computation of stochastically stable pure strategy states as it will become apparent in the following section. The importance lies on the fact that the transition probability between any two pure strategy states can be computed by considering only the shortest path between these two pure strategy states.

### C. Approximation of stationary distribution

In this section, using Lemma 5.2 that approximates one-step transition probabilities, we provide an approximation of the invariant stationary distribution of the QII t.p.f. By definition of QII, this approximation is based upon the observation that for the computation of the quantities $R_g$ of Lemma 5.1, it suffices to consider only those paths in $G^{(1)}\{s\}$ which involve one-step transitions as defined in the previous section.

Define $G^{(1)}\{s\} \subseteq G\{s\}$ to be the set of $s$-graphs consisting solely of one-step transitions, i.e., for any $g \in G^{(1)}\{s\}$ and any arrow $(s_k \rightarrow s_l) \in g$, the associated action profiles, say $\alpha(k), \alpha(l)$, respectively, differ in a single action of a single
player. It is straightforward to check that $G\{s\} \neq \emptyset$ for any $s \in S$.

**Lemma 5.3** (Approximation of stationary distribution): The stationary distribution of the finite Markov chain $\{P_{s_k s}\}, \pi = [\pi_s]$, satisfies

$$\lim_{\lambda \downarrow 0} \pi_s = \lim_{\delta \downarrow 0} \frac{\tilde{R}_s(\delta)}{\sum_{s_i \in S} \tilde{R}_{s_i}(\delta)}, \quad s \in S \quad (9)$$

where

$$\tilde{R}_s(\delta) = \sum_{g \in G\{s\}} \tilde{\varphi}(g; \delta),$$

and

$$\tilde{\varphi}(g; \delta) = \prod_{(s_k \rightarrow s)} \tilde{P}_{s_k s}(\delta). \quad (10)$$

**Proof.** By definition of $G\{s\}$ graphs, any p.s.s. may only be the initial point of exactly one arrow, while there are no cycles in the graph. Thus, for any graph $g \in G\{s\}$, the number of transitions involved is $|S| - 1$. For any $g \in G\{s\}$, $\varphi(g) \sim O(|S|^{-1})$, since in any arrow of such graph exactly one agent perturbs its strategy. For any other graph $g \in G\{s\} \setminus G\{s\}$, $\varphi(g) \sim O(\kappa^x)$ for some $\kappa > |S| - 1$, since any such graph includes arrows at which more than one agent perturbs its strategy. Thus, in the computation of the stationary distribution, we have

$$\pi_s = \frac{R_s}{\sum_{s_i \in S} R_{s_i}} = \lim_{\lambda \downarrow 0} \frac{\lambda^{\mid S\mid - 1} \sum_{g \in G\{s\}} \tilde{\varphi}(g; \delta) + O(\lambda^{|S|})}{\sum_{s_i \in S} \sum_{g \in G\{s\}} \tilde{\varphi}(g; \delta) + O(\lambda^{|S|})},$$

where $O(\lambda^{|S|})$ denotes terms of order $\lambda^{|S|}$ or higher. It is evident that if we take the limit as $\lambda \downarrow 0$, the conclusion follows. □

Note that Lemma 5.3 provides a simplification of Theorem 3.1 since it suffices to compute the transition probabilities of the graphs consisting solely of one step transitions. Furthermore, the transition probability of any such graph, $\tilde{\varphi}(g; \delta)$, can be computed by Lemma 5.2 which provides an explicit formula for one-step transitions. In the following section, the computation of the stationary distribution will further be simplified and related to the order of the one-step transition probabilities.

**D. $\delta$-resistance**

We have shown in Lemma 5.2 that the order of the one-step transition probability $\tilde{P}_{s' \rightarrow s}(\delta)$ is determined by the minimum hitting time $\tau_{s'}(\mathcal{N}_s(s'))$. Informally, the minimum hitting time defines the level of resistance in approaching state $s'$. In this section, we will formalize this notion and we will relate it to the stationary distribution $\pi$.

**Definition 5.2** ($\delta$-resistance): For a pure strategy state $s \in S$, let us consider any graph $g \in G\{s\}$. For any $\delta > 0$, the $\delta$-resistance associated with $s \in S$ in graph $g$, is defined as follows:

$$\varphi_\delta(s|g) \equiv \lim_{(s_k \rightarrow s) \in g} \tau_{s_k}(\mathcal{N}_s(s_k)). \quad (11)$$

In other words, the $\delta$-resistance of a state $s$ along a graph corresponds to the sum of the minimum hitting times along this graph. We further denote by $\varphi_\delta^s(s)$ the minimum $\delta$-resistance, i.e.,

$$\varphi_\delta^s(s) = \min_{g \in G\{s\}} \varphi_\delta(s|g)$$

and by $g^*(s)$ the $\{s\}$-graph that attains this minimum resistance.

The stochastically stable states can be identified as the states of minimum resistance, as the following lemma demonstrates.

**Theorem 5.1:** As $\epsilon \downarrow 0$, the set of stochastically-stable p.s.s.’s $S^*$ is such that, for any $\delta > 0$

$$\max_{s^* \in S^*} \varphi_\delta^s(s^*) < \min_{s \in S \setminus S^*} \varphi_\delta^s(s). \quad (12)$$

**Proof.** By Lemma 5.2 the one-step transition probability from $s_k$ to a $\delta$-neighborhood of $s_\ell$ can be approximated by

$$\tilde{P}_{s_k s_\ell}(\delta) \sim O \left(\epsilon^\varphi(s|g)\right).$$

Thus, for any graph $g \in G\{s\}$, we have that

$$\tilde{\varphi}(g; \delta) = \prod_{(s_k \rightarrow s)} \tilde{P}_{s_k s}(\delta) \sim O \left(\epsilon^\varphi^s(s)\right).$$

we conclude that the $s$-graph for which the exponent of $\epsilon$ is the smallest defines the order of the terms $\tilde{\varphi}(g; \delta)$, i.e.,

$$\tilde{R}_s(\delta) \sim O \left(\epsilon^\varphi^s(s)\right).$$

Denote by $\pi_{S^*}$ the probability assigned by the invariant probability distribution $\pi$ to the subset $S^*$ of $S$. Then, according to the simplification of the stationary distribution in Equation (9), we have:

$$\lim_{\epsilon \downarrow 0} \lim_{\lambda \downarrow 0} \pi_{S^*} = \lim_{\epsilon \downarrow 0} \lim_{\delta \downarrow 0} \frac{\sum_{s \in S^*} \tilde{R}_s(\delta)}{\sum_{s \in S} \tilde{R}_s(\delta)} = \lim_{\delta \downarrow 0} \lim_{\epsilon \downarrow 0} \frac{\sum_{s \in S^*} \tilde{R}_s(\delta)}{\sum_{s \in S} \tilde{R}_s(\delta)} = \lim_{\delta \downarrow 0} \lim_{\epsilon \downarrow 0} \frac{1 + \sum_{s \in S \setminus S^*} \tilde{R}_s(\delta)/\sum_{s \in S} \tilde{R}_s(\delta)}{\sum_{s \in S^*} \tilde{R}_s(\delta)/\sum_{s \in S} \tilde{R}_s(\delta)}.$$
which further implies that, for any \( \delta > 0 \),
\[
\sum_{s \in S_1, s \neq s^*} \tilde{R}_s(\delta) / \sum_{s \in S^*} \tilde{R}_s(\delta) \to 0 \text{ as } \epsilon \downarrow 0,
\]
which further implies that \( \lim_{\epsilon \downarrow 0} \lim_{\lambda \downarrow 0} \pi_{S^*} = 1 \). Conversely, \( \lim_{\epsilon \downarrow 0} \lim_{\lambda \downarrow 0} \pi_{S^* \setminus S^*} = 0 \). Thus, the stochastically stable states may only be contained in \( S^* \). □

Theorem 5.1 provides a guidance in the computation of the stochastically stable states (through the computation of the minimum \( \delta \)-resistance). It further applies to any game that satisfies the positive-utility property. In the following sections, we illustrate the utility of Theorem 5.1 in computing the stochastically stable states in multi-player multi-action coordination games.

VI. ILLUSTRATION IN COORDINATION GAMES

A. Stochastic stability

In the forthcoming analysis, we will be using the notion of best response of a player \( i \) into an action profile \( \alpha = (\alpha_i, \alpha_{-i}) \), as well as the notion of Nash equilibrium. In particular, we define:

**Definition 6.1 (Best response):** The best response of a player \( i \) to an action profile \( \alpha = (\alpha_i, \alpha_{-i}) \) is defined as the following set of actions:
\[
\text{BR}_i(\alpha) = \{ a \in A_i \mid u_i(\alpha, a, \alpha_{-i}) = \max_{a \in A_i} u_i(\alpha, a, \alpha_{-i}) \}.
\]

**Definition 6.2 (Nash equilibrium):** An action profile \( \alpha^* = (\alpha_i^*, \alpha_{-i}^*) \) is a Nash equilibrium, if for every player \( i \),
\[
\alpha_i^* \in \text{BR}_i(\alpha^*).
\]

A best-response of a player \( i \) to the current action profile will often be denoted by \( \alpha_i^* \). Note that according to the above definition the best response of a player to an action profile is never empty. We also introduce the following notion of a coordination game.

**Definition 6.3 (Coordination game):** A strategic-form game of \( n \) players and \( m \) actions, satisfying the positive-utility property (Property 2.1) is a coordination game if, for every action profile \( \alpha \) and for every player \( i \),
\[
u_j(\alpha'_i, \alpha_{-i}) \geq u_j(\alpha_i, \alpha_{-i})
\]
for any \( \alpha'_i \in \text{BR}_i(\alpha) \).

In other words, a coordination game is such that at any action profile, if a player plays a best response to its current action profile, then no other player gets worse-off. This is satisfied by default when the current action profile corresponds to a Nash equilibrium, since a player’s best response is to play the same action. Such strategic-form games are often encountered in practical applications, such as in network-formation games (cf., [28]), as we shall see in the forthcoming simulation study.

In order to address stochastic stability in such higher-dimensional strategic-form games, we will further need to introduce the notion of the best-BR (briefly BBR).

**Definition 6.4 (Best-BR):** Let \( i^* : A \to I \) be defined as follows:
\[
i^*(\alpha) = \arg \max_{\pi \in I} \{ u_i(\alpha_i, \alpha_{-i}) : \alpha_i \in \text{BR}_i(\alpha) \}.
\]

The one-step transition \( \alpha = (\alpha_i^*, \alpha_{-i}^*) \) to \( (\alpha_i^*, \alpha_{-i}^*) \), where \( \alpha_{-i}^* \in \text{BR}_i(\alpha) \) is the best-BR to the current action profile \( \alpha \) and will briefly be denoted by BBR(\( \alpha \)).

In other words, the best-BR to an action profile corresponds to the one-step transition, where the player which changes its action experiences the largest reward among all one-step transitions.

**Lemma 6.1:** Let \( S_{\text{NE}} \) be the set of p.s.s.’s which correspond to the set of Nash equilibria. In any coordination game, the \( \{ S_{\text{NE}} \} \)-graph that attains the minimum \( \delta \)-resistance is:
\[
g^*(S_{\text{NE}}) = \{(s_k \to s_\ell) : \alpha(\ell) \in \text{BBR}(\alpha) \).
\]

**Proof.** In each \( \{ S_{\text{NE}} \} \)-graph, there exists a fixed number of arrows, equal to \( |S| - |S_{\text{NE}}| \) (since each action profile outside \( S_{\text{NE}} \) is the source of exactly one arrow according to Definition 5.1). Each such arrow also corresponds to a one-step transition, i.e., only a single player changes its action along this arrow (as Lemma 5.3 demonstrated).

Under the coordination property, any path consisting of one-step best-BR’s will include no cycles (since the utility of each player may not decrease along this path). Furthermore, such path may only terminate at a Nash equilibrium. Thus, under the coordination property, we may always create a \( \{ S_{\text{NE}} \} \)-graph consisting solely of best-BR’s. Such graph will have the minimum \( \delta \)-resistance, since the best-BR step of an action profile achieves the minimum \( \delta \)-resistance among the one-step transitions. Thus, the conclusion follows. □

In other words,Lemma 6.1 shows that the \( \{ S_{\text{NE}} \} \)-graph of minimum \( \delta \)-resistance is the graph consisting of the one-step best-BR’s starting from any non-Nash action profile. Using this property, we can show that the set of Nash equilibria are the stochastically stable states of any coordination game.

**Theorem 6.1 (Stochastic stability in coordination games):** In any coordination game of Definition 6.3 as \( \epsilon \downarrow 0 \) and \( \lambda \downarrow 0 \),
\[
S^* \subseteq S_{\text{NE}}.
\]

**Proof.** It suffices to show that all p.s.s.’s outside \( S_{\text{NE}} \) provide a \( \delta \)-resistance which is higher than the \( \delta \)-resistance of any Nash equilibrium in \( S_{\text{NE}} \) (as Property 12 dictates).

Consider an action profile \( \alpha \) which is not a Nash equilibrium and the corresponding p.s.s. \( s \). Consider the part of the optimal \( S_{\text{NE}} \)-graph which leads to \( s \), i.e.,
\[
g^*(S_{\text{NE}}) = \{(s_k \to s_\ell) : \exists \text{ a path from } s_\ell \text{ to } s \}.
\]
In other words, \( g^*(s|S_{NE}) \) corresponds to the part of the minimum-resistance graph \( g^*(S_{NE}) \) whose arrows lead to \( s \). This graph might be empty if \( s \) is not a recipient of any arrow in \( g^*(S_{NE}) \). For the remainder of the proof, define the graphs:

\[
g_1 \equiv g^*(S_{NE}) \setminus g^*(s|S_{NE}) \\
g_2 \equiv g^*(s) \setminus g^*(s|S_{NE})
\]

In other words, graph \( g_1 \) contains all arrows in \( g^*(S_{NE}) \) except for the best-BR steps leading to \( s \), and graph \( g_2 \) contains all arrows in \( g^*(s) \) but the best-BR steps leading to \( s \). Note that, \( g^*(s|S_{NE}) \subset g^*(s) \), i.e., the graph that leads to \( s \) through the minimum resistance graph of \( S_{NE} \) is also part of the minimum resistance graph of \( s \). By construction, we also have \( g^*(s|S_{NE}) \subset g^*(S_{NE}) \). Thus, graph \( g_1 \) involves the exact same nodes as graph \( g_2 \). However, since \( |S_{NE}| \geq 1 \), by definition of the \( S_{NE} \)-graphs, a node within the set \( S_{NE} \) may be the source of no arrow in \( g_1 \). Thus, in general, \( |g_1| \leq |g_2| \), i.e., \( g_2 \) contains at least as many arrows as \( g_1 \).

Furthermore, by construction of graphs \( g_1 \) and \( g_2 \), there exists at least one node \( s' \notin S_{NE} \) with the following property: \((s' \rightarrow s'') \in g_1 \) such that \( s'' \in BBR(\alpha') \), and \((s' \rightarrow s''') \in g_2 \) such that \( s''' \notin BBR(\alpha') \). This is due to the fact that \( s \notin S_{NE} \).

Since only best-BR transition steps achieve the minimum resistance, we conclude that \( \varphi(s|g_2) > \varphi(s|g_1) \), which implies that any \( \{s\} \)-graph may only have larger \( \delta \)-resistance as compared to the minimum \( \delta \)-resistance of \( g^*(S_{NE}) \). \( \square \)

B. Simulation study in distributed network formation

In this section, we perform a simulation study of reinforcement learning in network formation games.

To illustrate how a network formation game can be modeled as a coordination game, we consider a simple network formation game motivated by [33]. We consider \( n \) nodes deployed on the plane and assume that the set of actions of each agent \( i \), \( A_i \), contains all possible combinations of neighbors of \( i \), denoted \( N_i \), with which a link can be established, i.e., \( A_i = 2^{N_i} \). Links are considered undirectional, and a link established by node \( i \) with node \( j \), denoted \((j,i)\), starts at \( j \) with the arrowhead pointing to \( i \).

A graph \( G \) is defined as a collection of nodes and directed links. Define also a path from \( j \) to \( i \) as a sequence of nodes and directed links that starts at \( j \) and ends to \( i \) following the orientation of the graph, i.e.,

\[
(j \rightarrow i) = \{j = j_0, (j_0, j_1), j_1, \ldots, (j_{m-1}, j_m), j_m = i\}
\]

for some positive integer \( m \). In a connected graph, there is a path from any node to any other node.

Let us consider the utility function \( u_i : A \rightarrow \mathbb{R}, i \in \mathcal{I} \), defined by

\[
u_i(\alpha) = \sum_{j \in \mathcal{I} \setminus \{i\}} \chi_{\alpha}(j \rightarrow i) - c|\alpha_i|,
\]

where \( \chi_{\alpha}(j \rightarrow i) \) is an indicator function that takes the value 1 if \( \alpha \) is a path from \( j \) to \( i \) and 0 otherwise. This utility function reflects the gain a node receives from being connected to other nodes, with a penalty for the number of links it forms.
where $|\alpha_i|$ denotes the number of links corresponding to $\alpha_i$ and $c$ is a constant in $(0, 1)$. Also,

$$\chi_\alpha(j \to i) \doteq \begin{cases} 1 & \text{if } (j \to i) \subseteq G_\alpha, \\ 0 & \text{otherwise}, \end{cases}$$

where $G_\alpha$ denotes the graph induced by joint action $\alpha$. The resulting Nash equilibria are usually called Nash networks [34]. As it was shown in Proposition 4.2 in [35], a network $G^*$ is a Nash network if and only if it is critically connected, i.e., i) it is connected, and ii) for any $(s, i) \in G$, $(s \to i)$ is the unique path from $s$ to $i$. For example, the resulting Nash networks for $n = 3$ agents and unconstrained neighborhoods are shown in Fig. 3.

**Proposition 6.1:** The network formation game defined by (13) is a coordination game.

**Proof.** For a joint action $\alpha \not\in A^*$ suppose that a node $i$ picks its best response. Then no other agent becomes worse off, since a best response of any node $i$ always retains connectivity. Note that this is not necessarily true for any other change in actions. Thus, the coordination property of Definition 6.3 is satisfied. □

![Fig. 3. Nash networks in case of $n = 3$ agents and $0 < \nu < 1$.](image)

Fig. 2 depicts the response of reinforcement learning in the network formation game. We consider 6 nodes deployed on the plane, where the neighbors of each node are defined as the two immediate nodes (e.g., the neighbors of node 1 are \{2, 6\}). According to Theorem 6.1 in order for the average behavior to be observed $\lambda$ and $\epsilon$ need to be sufficiently small. We choose: $\epsilon = \lambda = 0.005$, and $\epsilon = 1/2$.

Given the large number of actions, we do not plot the strategy vector for each node. Instead, we plot the inverse total distance from each node to its neighboring nodes. In a wheel structure (and only under this structure), the inverse distance from each node to its neighboring nodes is equal to $1/\nu \approx 0.167$. The wheel structure is among the Nash equilibria of this game (as shown in [35]) and the unique payoff-dominant equilibrium (i.e., every node receives its maximum utility). The wheel structure is the emergent structure under reinforcement learning as shown in Fig. 2.

The simulation of Fig. 2 verifies Theorem 6.1 since convergence (in a weak sense) is attained to the set of Nash equilibria. However, it also demonstrates the potential of this class of dynamics for stronger convergence results, since the emergent Nash equilibrium is also payoff-dominant.

---

**VII. Conclusions & Future Work**

In this paper, we considered a class of reinforcement-learning dynamics that belongs to the family of discrete-time replicator dynamics and learning automata, and we provided an explicit characterization of the invariant probability measure of the induced Markov chain. Through this analysis, we demonstrated convergence (in a weak sense) to the set of pure-strategy states, overcoming prior limitations of the ODE-method for stochastic approximations, such as the existence of a potential function. Furthermore, we provided a simplified methodology for computing the set of stochastically-stable states, and we demonstrated its utility in the context of coordination games. This is the first result in this class of dynamics that demonstrates global convergence properties for a larger family of games than the fine class of potential games. Thus, it opens up new possibilities for the use of reinforcement learning in distributed control of multi-agent systems. However, stronger convergence results are possible, as the simulation study demonstrated, that go beyond the set of Nash equilibria.

**References**

[1] G. C. Chasparis, “Stochastic stability analysis of perturbed learning automata with constant step-size in strategic-form games,” in American Control Conference, Seattle, USA, 2017, pp. 4607–4612.

[2] B. G. Chan, R. Fonseca, I. Stoica, and J. Kubiatowicz, “Characterizing selfishly constructed overlay routing networks,” in Proc. of IEEE INFOCOM 04, Hong-Kong, 2004.

[3] R. Komali, A. B. MacKenzie, and R. P. Gilles, “Effect of selfish node behavior on efficient topology design,” IEEE Transactions on Mobile Computing, vol. 7, no. 9, pp. 1057–1070, 2008.

[4] G. Wei, A. V. Vasilakos, Y. Zheng, and N. Xiong, “A game-theoretic method of fair resource allocation for cloud computing services,” The Journal of Supercomputing, vol. 54, no. 2, pp. 252–269, Nov. 2010.

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

[6] T. Börjes and R. Sarin, “Learning through reinforcement and replicator dynamics,” J. Econ. Theory, vol. 77, no. 1, pp. 1–14, 1997.

[7] 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.

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

[9] A. Beggs, “On the convergence of reinforcement learning,” J. Econ. Theory, vol. 122, pp. 1–36, 2005.

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

[11] 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.

[12] M. Tsetlin, Automation Theory and Modeling of Biological Systems. Academic Press, 1973.

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

[14] 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.

[15] D. Leslie, “Reinforcement learning in games,” Ph.D. dissertation, School of Mathematics, University of Bristol, 2004.
We conclude that for any sequence of the perturbed strategy vector \( \lim \inf \) \( k \to \infty \) exists an open ball about the next strategy vector that does not share any common points with the canonical case, there exists an open ball about the next strategy vector that belongs to the canonical projection of \( O \), since \( O \in \mathcal{B}(Z) \). Due to the continuity of the function \( R_i(\cdot, \cdot) \), we have that \( y^{(i)} \to y_i \equiv \mathbb{P}_{X_i}(R_i(\alpha, x_i)) \equiv 0 \).

To investigate the limit of \( P_\lambda(Z^{(k)}, O) \) as \( k \to \infty \), it suffices to investigate the behavior of the sequence \( y^{(k)}_i \equiv \mathbb{P}_{X_i}(R_i(\alpha, x_i^{(k)})) \).

We distinguish the following (complementary) cases:

(a) \( R_i(\alpha, x_i) \notin \mathcal{P}_{X_i}(O) \) and \( R_i(\alpha, x_i) \notin \partial \mathcal{P}_{X_i}(O) \): In this case, there exists an open ball about the next strategy vector that does not share any common points with the canonical projection of \( O \). Due to the continuity of the function \( R_i(\cdot, \cdot) \), we have that \( y^{(k)}_i \to y_i \equiv \mathbb{P}_{X_i}(R_i(\alpha, x_i)) \equiv 0 \).

(b) \( R_i(\alpha, x_i) \in \mathcal{P}_{X_i}(O) \): In this case, there exists an open ball about the next strategy vector that belongs to the canonical projection of \( O \), since \( O \in \mathcal{B}(Z) \). Due to the continuity of the function \( R_i(\cdot, \cdot) \), we have that \( y^{(k)}_i \to y_i = 1 \).

(c) \( R_i(\alpha, x_i) \notin \mathcal{P}_{X_i}(O) \) and \( R_i(\alpha, x_i) \in \partial \mathcal{P}_{X_i}(O) \): In this case, \( y_i \equiv 0 \). We conclude that \( \lim \inf_{k \to \infty} y^{(k)}_i \geq y_i = 0 \), since \( y_i^{(k)} \in \{0, 1\} \).

In either one of the above (complementary) cases, we have that \( \lim \inf_{k \to \infty} y^{(k)}_i \geq y_i = 0 \). Finally, due to the continuity of the perturbed strategy vector \( \tilde{x}_{i\alpha_i} \) with respect to \( x_{i\alpha_i} \), we conclude that for any sequence \( Z^{(k)} \to Z \),

\[
\lim \inf_{k \to \infty} P_\lambda(Z^{(k)}, O) \geq P_\lambda(Z, O).
\]

By \( \textbf{[30]} \) Proposition \( 7.2.1 \), we conclude that \( P_\lambda \) satisfies the weak Feller property.

The above derivation can be generalized to any selection probability function \( f(x_{i\alpha_i}) \) in the place of \( x_{i\alpha_i} \), provided that it is a continuous function. Thus, the proof for the unperturbed process \( P \) follows the exact same reasoning by simply setting \( f(x_{i\alpha_i}) = x_{i\alpha_i} \).

**APPENDIX B**

**PROOF OF PROPOSITION 4.1**

(a) Let us consider an action profile \( \alpha = (\alpha_1, ..., \alpha_n) \in \mathcal{A} \), and an initial strategy profile \( x(0) = (x_1(0), ..., x_n(0)) \) such that \( x_{i\alpha_i}(0) > 0 \) for all \( i \in \mathcal{I} \). Note that if the same action profile \( \alpha \) is selected up to time \( t \), then the strategy of agent \( i \) satisfies:

\[
x_i(t) = e_{\alpha_i} - (1 - \epsilon u_i(\alpha))^t (e_{\alpha_i} - x_i(0)).
\]

Given that \( B_t \) is non-increasing, from continuity from above we have

\[
P_x(B_{\infty}) = \lim_{t \to \infty} P_x(B_t) = \lim_{t \to \infty} \prod_{i=0}^{t} P_x[B_i] = \lim_{t \to \infty} \prod_{i=0}^{t} x_{i\alpha_i}(k).
\]
Note that $\mathbb{P}[B_\infty] > 0$ if and only if
\[ \sum_{t=1}^{\infty} \log(x_{i\alpha_i}(t)) > -\infty. \] (15)

Let us introduce the variable $y_i(t) \equiv 1 - x_{i\alpha_i}(t)$, which corresponds to the probability of agent $i$ selecting any action other than $\alpha_i$. Condition (15) is equivalent to
\[ -\sum_{t=0}^{\infty} \log(1 - y_i(t)) < \infty, \text{ for all } i \in \mathcal{I}. \] (16)

We also have that
\[ \lim_{t \to \infty} -\frac{\log(1 - y_i(t))}{y_i(t)} = \lim_{t \to \infty} \frac{1}{1 - y_i(t)} > \rho \]
for some $\rho > 0$, since $0 \leq y_i(t) \leq 1$. Thus, from the Limit Comparison Test, we conclude that condition (16) holds if and only if $\sum_{t=1}^{\infty} y_i(t) < \infty$, for each $i \in \mathcal{I}$.

Lastly, note that $y_i(t+1)/y_i(t) = 1 - \epsilon u_i(\alpha)$ and
\[ t \left( \frac{y_i(t)}{y_i(t+1)} - 1 \right) = t \frac{\epsilon u_i(\alpha)}{1 - \epsilon u_i(\alpha)}. \]
Thus, if $\epsilon u_i(\alpha) < 1$ for all $\alpha \in \mathcal{A}$ and $i \in \mathcal{I}$, then $1 - \epsilon u_i(\alpha) > 0$ and $\lim_{t \to \infty} t(y_{i(t)}/y_{i(t+1)} - 1) > 1$, which implies (according to Raabe’s test) that the series of positive terms $\sum_{t=1}^{\infty} y_i(t)$ is convergent. Thus, we conclude that $\mathbb{P}_z[B_\infty] > 0$.

(b) Define the event
\[ C_t = \{ \exists \alpha' \neq \alpha(t) : x_{i\alpha_i'}(t) > 0, \text{ for all } i \in \mathcal{I} \}, \]
i.e., $C_t$ corresponds to the event that there exists an action profile different from the current action profile for which the nominal strategy assigns positive probability for all agents $i$.

Note that $A_t' \subseteq C_t$, since $A_t'$ may only occur if there is some action profile $\alpha' \neq \alpha(t)$ for which the strategy assigns positive probability. This further implies that $\mathbb{P}_z[A_t' \cap A_{t+1}'] \leq \mathbb{P}_z[C_t]$. Then, we have:
\[
\mathbb{P}_z[A_{t+1} | A_t'] = \frac{\mathbb{P}_z[A_{t+1} \cap A_t']}{\mathbb{P}_z[A_t']}
\geq \frac{\mathbb{P}_z[A_{t+1} \cap C_t]}{\mathbb{P}_z[C_t]}
= \mathbb{P}_z[\{ \alpha(\tau) = \alpha' \neq \alpha(t), \forall \tau > t \} | C_t]
\geq \inf_{\alpha' \neq \alpha} \prod_{t=1}^{\infty} x_{i\alpha_i'}(t) \prod_{k=1}^{\infty} \{ 1 - (1 - \epsilon u_i(\alpha'))^{k-1} c_i(\alpha') \}
\geq \inf_{\alpha' \neq \alpha} \prod_{t=1}^{\infty} x_{i\alpha_i'}(t) \prod_{k=0}^{\infty} \{ 1 - (1 - \epsilon u_i(\alpha'))^{k} c_i(\alpha') \}
\]
where $c_i(\alpha') = 1 - x_{i\alpha_i'}(t) \geq 0$. We have already shown in part (a) that the second part of the r.h.s. (which corresponds to the probability of playing the same action for all future times) is bounded away from zero. Therefore, we conclude that $\mathbb{P}_z[A_{t+1} | A_t'] > 0$. Thus, from the counterpart of the Borel-Cantelli Lemma and the fact that $A_t \subseteq A_{t+1}$, we have $\mathbb{P}_z[A_\infty] = 1$. 

