Stochastic Stability of Perturbed Learning Automata in Positive-Utility Games

Georgios C. Chasparis

***Abstract*—This paper considers a class of reinforcement-based learning (namely, *perturbed learning automata*) and provides a stochastic-stability analysis in repeatedly-played, positive-utility, strategic-form games. Prior work in this class of learning dynamics primarily analyzes asymptotic convergence through stochastic approximations, where convergence 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 asymptotic convergence 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 computational 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*- or *utility*-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].

Given the repeated fashion of the involved strategic interactions in such formulations, several questions naturally emerge: a) Can agents "learn" to asymptotically select optimal decisions/actions?, b) What information should agents share with each other?, and c) What is the computational complexity of the learning process? Under the scope of engineering applications, it is usually desirable that each agent shares minimum amount of information with other agents, while the computational complexity of the learning process is small. A class of learning dynamics that achieves small communication and computational complexity is the so-called *payoff-based* learning. Under such class of learning dynamics, each agent *only* receives measurements of its own utility function, without the need to know the actions selected by other agents, or the details of its own utility function (i.e., its dependencies on other agents' actions).

In such repeatedly-played strategic-form games, a popular objective for payoff-based learning is to guarantee convergence (in some sense) to Nash equilibria. Convergence to Nash equilibria may be desirable, especially when the set of optimal centralized solutions coincides with the set of Nash equilibria.

Reinforcement-based 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 [6], [7] or approximate policy iteration or Q-learning [8]. In all these classes of learning dynamics, 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 it will be discussed in detail in the forthcoming Section II).

In the present paper, we consider a class of reinforcement-based learning introduced in [9] that is closely related to both discrete-time replicator dynamics and learning automata. We will refer to this class of dynamics as *perturbed 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 conclusions over the asymptotic behavior of the dynamics that is based upon 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 in coordination games, where convergence to mixed strategy profiles may only be excluded

An earlier version of parts of this paper appeared in [1]. This work has been partially supported by the European Union grant EU H2020-ICT-2014-1 project RePhrase (No. 644235). It has also been partially supported by the Austrian Ministry for Transport, Innovation and Technology, the Federal Ministry of Science, Research and Economy, and the Province of Upper Austria in the frame of the COMET center SCCH.

G.C. Chasparis is with the Department of Data Analysis Systems, Software Competence Center Hagenberg GmbH, Softwarepark 21, A-4232 Hagenberg, Austria, E-mail: georgios.chasparis@scch.at.

under strong conditions in the payoff matrix (e.g., existence of a potential function). Furthermore, we provide a methodology for computing the set of stochastically stable states in all positive-utility games. We illustrate this methodology in the context of coordination games and provide a simulation study in distributed network formation.

In the remainder of the paper, Section II presents the investigated class of learning dynamics, related work and the main contributions. Section III provides a simplification in the characterization of stochastic stability, while Section IV presents its technical derivation. This result is utilized for computing the stochastically stable states in positive-utility games in Section V. In Section VI, we present an illustration of the proposed methodology in the context of coordination games, together with a simulation study in distributed network formation. Finally, Section VII presents concluding remarks.

Notation:

- For a Euclidean topological space $\mathcal{Z} \subset \mathbb{R}^n$, let $\mathcal{N}_\delta(x)$ denote the δ-neighborhood of $x \in \mathbb{R}^n$, i.e.,
$$\mathcal{N}_\delta(x) \doteq \{y \in \mathcal{Z} : |x - y| < \delta\},$$
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.,
$$\mathbb{I}_A(x) \doteq \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{else.} \end{cases}$$
- For a finite set A, $|A|$ denotes its cardinality.
- For a finite set A and any probability distribution $\sigma \in \Delta(|A|)$, the random selection of an element of A will be denoted by $\mathrm{rand}_\sigma[A]$. If $\sigma = (1/|A|, ..., 1/|A|)$, the random selection will be denoted by $\mathrm{rand}_{\mathrm{unif}}[A]$.
- $\boldsymbol{\delta}_x$ denotes the Dirac measure at x.
- $\log(\cdot)$ denotes the natural logarithm.

II. Perturbed Learning Automata

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 i have a finite set of actions $\mathcal{A}_i$. 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 $\langle \mathcal{I}, \mathcal{A}, \{u_i(\cdot)\}_i \rangle$.

TABLE I
Perturbed Learning Automata

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:
$$\alpha_i(t+1) = \begin{cases} \mathrm{rand}_{x_i(t)}[\mathcal{A}_i], & \text{with probability } 1 - \lambda, \\ \mathrm{rand}_{\mathrm{unif}}[\mathcal{A}_i], & \text{with probability } \lambda, \end{cases} \qquad (1)$$
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 $x_i \in \Delta(|\mathcal{A}_i|)$ as follows:
$$\begin{aligned} & x_i(t+1) \\ & \quad = \quad x_i(t) + \epsilon \cdot u_i(\alpha(t+1)) \cdot [e_{\alpha_i(t+1)} - x_i(t)] \\ & \quad \doteq \quad \mathcal{R}_i(\alpha(t+1), x_i(t)), \end{aligned} \qquad (2)$$
for some constant step size $\epsilon > 0$.

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 $\alpha \in \mathcal{A}$, $u_i(\alpha) > 0$.

This property is rather generic and applies to a large family of games. For example, games at which some form of alignment of interests exists between agents (e.g., *coordination games* [10] or *weakly-acyclic games* [11]), can be designed to satisfy this property, since agents' utilities/preferences are rather close to each other at any given action profile. However, in the forthcoming analysis, we do not impose any structural constraint other than the Property 2.1.

B. Perturbed Learning Automata

We consider a form of reinforcement-based learning that belongs to the general class of *learning automata* [7]. In learning automata, each agent updates a finite probability distribution $x_i \in \mathcal{X}_i \doteq \Delta(|\mathcal{A}_i|)$ 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 learning model.

The proposed 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 $e_{\alpha_i(t)} \equiv e_j$. To better see how the strategies evolve, let us consider the following toy example. Let the current strategy of player i be $x_i(t) = \begin{pmatrix} 1/2 & 1/2 \end{pmatrix}^{\mathrm{T}}$, i.e., player i has two actions, each assigned probability $1/2$. 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:

$$x_i(t+1) = 1/2 \begin{pmatrix} 1 + \epsilon u_i(\alpha(t+1)) \\ 1 - \epsilon u_i(\alpha(t+1)) \end{pmatrix}.$$

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.

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 also change their actions. This dynamic element of the size of reinforcement is the factor that complicates the convergence analysis, as it will become clear in the forthcoming related work.

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 learning automata*.

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* (cf., [12]). It has been used in different forms, depending primarily on the step-size sequence ϵ in Table I. For example, Arthur [5] considered a similar rule, with $\lambda = 0$ and step size of each agent i defined as $\epsilon_i(t) = 1/(ct^{\nu} + u_i(\alpha(t+1))$, for some positive constant c and for $\nu \in (0,1)$ (in the place of the constant step size ϵ of (2)). A comparative model is also used by Hopkins and Posch in [13], 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 the urn process of Erev-Roth [14]. Some similarities are also shared with the Cross' learning model of [15], where $\epsilon(t) = 1$ and $u_i(\alpha(t)) \leq 1$, and its modification presented by Leslie in [16], where $\epsilon(t)$, instead, is assumed decreasing with time.

The main difference of the proposed dynamics of Table I lies in the perturbation parameter $\lambda > 0$ which was first introduced and analyzed in [9]. A state-dependent perturbation term has also been investigated in [17]. The perturbation parameter may serve as an equilibrium selection mechanism, since *it excludes convergence to non-Nash action profiles* [9]. It resolved one of the main issues of discrete-time replicator dynamics, that is the positive probability of convergence to action profiles that are not Nash equilibria (briefly, *non-Nash action profiles*).

Although excluding convergence to non-Nash action profiles can be guaranteed by using sufficiently small $\lambda > 0$, establishing convergence to action profiles that are Nash equilibria (*pure Nash equilibria*) may still be an issue. This is desirable in the context of coordination games [18], where Pareto-efficient outcomes are usually pure Nash equilibria (see, e.g., the definition of a coordination game in [10]). As presented in [17], convergence to pure Nash equilibria can be guaranteed only under strong conditions in the payoff matrix. For example, as shown in [17, Proposition 8], and under the ODE-method for stochastic approximations, it requires a) the existence of a potential function, and b) conditions over the Jacobian matrix of the potential function. Even if a potential function does exist, verifying condition (b) is practically infeasible for games of more than 2 players [17].

On the other hand, an important side-benefit of using this class of dynamics is the indirect "filtering" on the utility-function measurements (through the formulation of the strategy vectors in (2)). This is demonstrated, for example, in [13] for the Erev-Roth model [14], where the robustness of convergence/non-convergence asymptotic results is presented under the presence of noise in the utility measurements.

Learning automata: Learning automata, as first introduced by [6], have attracted attention with respect to the control of complex and distributed systems due to their simple structure and low computational complexity (cf., [7, 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. An example of such stochastic learning automata is the *linear reward-inaction scheme* described in [7, Chapter 4]. Comparing it with the reinforcement rule of (2), the linear reward-inaction scheme accepts a utility function of the form $u_i(\alpha) \in \{0, 1\}$, 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 [7], [19]. In identical interest games, convergence analysis has been derived for 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., [7, 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]. Similarly to the discrete-time replicator dynamics, 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 [17, Proposition 2] has shown.) Furthermore, deriving conditions for excluding convergence to mixed strategy profiles in coordination games 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 (pure) 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.

Q-learning: Similar questions of convergence to Nash equilibria also appear in alternative reinforcement-based learning formulations, such as approximate dynamic programming and Q-learning. Usually, under 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-Q learning algorithm of [8], 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 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 Q-value needs to be assigned for each joint action.

When the evaluation of the Q-values is totally independent, as in the individual 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 exist no convergence results in multi-player games. This is a main drawback for Q-learning dynamics in strategic-form games as also pointed out in [24]. To overcome this drawback, in the context of stochastic dynamic games, reference [24] employs an additional feature (motivated by [11]), namely *exploration phases*. In any such *exploration phase*, *all* agents use constant policies, something that allows the accurate computation of the optimal Q-factors. We may argue that the introduction of common exploration phases for all agents partially destroys the distributed nature of the dynamics, since it requires synchronization between agents.

Aspiration-based learning: 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 [11] for convergence to Nash equilibria in weakly-acyclic games, the trial-and-error learning [25] for convergence to Nash equilibria in generic games, the mood-based dynamics of [26] for maximizing welfare in generic games or the aspiration learning in [10] for convergence to efficient outcomes in coordination games). We will refer to such approaches as *aspiration-based learning*. For these types of dynamics, convergence to Nash equilibria or efficient outcomes can be established without requiring any strong monotonicity properties (as in the multi-player weakly-acyclic games in [11]).

The case of noisy utility measurements, which are present in many engineering applications, has not currently been addressed through aspiration-based learning. The only exception is reference [11], under benchmark-based dynamics, where (synchronized) *exploration phases* are introduced through which each agent plays a fixed action for the duration of the exploration phase. If such exploration phases are large in duration (as required by the results in [11]), this may reduce the robustness of the dynamics to dynamic changes in the environment (e.g., changes in the utility function). One reason that such robustness analysis is currently not possible in this class of dynamics is the fact that decisions are taken directly based on the measured performances (e.g., by comparing the currently measured performance with the benchmark performance in [11]).

D. Contributions

The aforementioned literature in payoff-based learning dynamics in strategic-form games can be grouped into two main categories, namely *reinforcement-based learning* (including discrete-time replicator dynamics, learning automata and Q-learning) and *aspiration-based learning*. Summarizing their main advantages/disadvantages, we may argue the following high-level observations.

(O1) *Strong asymptotic convergence guarantees* for large number of players, even for generic games, are currently possible under aspiration-based learning. Similar results in reinforcement-based learning are currently restricted to games of small number of players and under strong structural assumptions (e.g., the existence of a potential function). See, for example, the discussion on discrete-time replicator dynamics or learning automata in [17], or the discussion on Q-learning in [24].

(O2) *Noisy observations* can be "handled" through reinforcement-based learning due to the indirect *filtering* of the observation signals (e.g., through the strategy-vector formulation in the model of Table I or in the formulation of the Q factors in Q-learning). This is demonstrated, for example, in the convergence/non-convergence asymptotic results presented in [13] for a variation of the proposed learning dynamics of Table I (with $\lambda = 0$ and decreasing ϵ) and under the presence of noise. Similar effects in aspiration-based learning can currently be achieved only through the introduction of *synchronized exploration phases*, as discussed in Section II-C.

Motivated by these two observations (O1)–(O2), and the obvious inability of reinforcement-based learning to provide strong convergence guarantees in large games, this paper advances asymptotic convergence guarantees for a class of reinforcement-based learning described in Table I (closely related to both discrete-time replicator dynamics and learning automata, as discussed in Section II-C). Our goal is to go beyond common restrictions of small number of players and strong assumptions in the game structure (such as the existence of a potential function).

The proposed dynamics (also *perturbed learning automata*) were first introduced in [9] to resolve stability issues in the boundary of the domain appearing in prior schemes [5], [13]. This was achieved through the introduction of the perturbation factor λ of Table I. However, strong convergence guarantees (e.g., w.p.1 convergence to Nash equilibria or efficient outcomes) is currently limited to small number of players and under strict structural assumptions, e.g., the existence of a potential function and additional conditions on its Jacobian matrix [17].

In this paper, we drop the assumption of a decreasing step-size sequence, and instead we consider the case of a *constant* step size $\epsilon > 0$. Such selection increases the adaptivity of the dynamics to varying conditions (e.g., the number of agents or the utility function). Furthermore, we provide a stochastic-stability analysis that provides a detailed characterization of the invariant probability measure of the induced Markov chain with no restrictions on the number of players. In particular, our contributions are the following:

(C1) 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 and it applies to any strategic-form game with the Positive-Utility Property 2.1 (*Theorem 3.1*).

(C2) We capitalize on this simplification and provide a methodology for computing stochastically stable states in positive-utility strategic-form games (*Theorem 5.1*).

(C3) 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-based learning for the reasons explained in observation (O1). We have to note that the illustration result in coordination games (contribution (C3) above) is of independent interest. To the best of our knowledge, it is the first convergence result in the context of reinforcement-based learning in repeatedly-played strategic-form games with the following features: a) a completely distributed setup (i.e., with no information exchange), b) more than two players, and c) a set of weakly-acyclic games that do not require the strong condition of the existence of a potential function.

This paper is an extention over an earlier version appeared in [1], which only focused on contribution (C1) above.

III. Stochastic Stability

In this section, we provide a characterization of the *invariant probability measure* μ_λ of the induced Markov chain P_λ of the dynamics of Table I. The importance lies in an equivalence relation (established through a weak-convergence argument) of μ_λ 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 μ_λ will be the first important step for providing specialized results for stochastic stability in strategic-form games.

A. Terminology and notation

Let $\mathcal{Z} \doteq \mathcal{A} \times \mathcal{X}$, where $\mathcal{X} \doteq \mathcal{X}_1 \times \ldots \times \mathcal{X}_n$, i.e., pairs of joint actions α and strategy profiles x. We will denote the elements of the state space $\mathcal{Z}$ by z.

The set $\mathcal{A}$ is endowed with the discrete topology, $\mathcal{X}$ 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* (p.m.) on $\mathfrak{B}(\mathcal{Z})$ endowed with the Prohorov topology, i.e., the topology of weak convergence. The learning algorithm of 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 $P : \mathcal{Z} \times \mathfrak{B}(\mathcal{Z}) \to [0,1]$ 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 $Z \doteq \{Z_t : t \geq 0\}$. Let $\Omega \doteq \mathcal{Z}^\infty$ denote the canonical path space, i.e., an element $\omega \in \Omega$ is a sequence $\{\omega(0), \omega(1), \dots\}$, 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 p.m. induced by the unperturbed process P on the product σ-field 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-field of $\mathcal{Z}^\infty$ 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., [27, 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.

Proof. See Appendix A. □

The measure $\mu_\lambda \in \mathfrak{P}(\mathcal{Z})$ is called an *invariant probability measure* (i.p.m.) 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 i.p.m., denoted μ and μ_λ, respectively [27, 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 i.p.m. $\{\mu_\lambda \in \mathfrak{P}(\mathcal{Z}) : \mu_\lambda P_\lambda = \mu_\lambda, \lambda > 0\}$ 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 i.p.m.'s $\{\mu_\lambda \in \mathfrak{P}(\mathcal{Z}) : \mu_\lambda P_\lambda = \mu_\lambda, \lambda > 0\}$, the following hold:

(a) $\lim_{\lambda \to 0} \mu_\lambda(\cdot) = \hat{\mu}(\cdot) \doteq \sum_{s \in \mathcal{S}} \pi_s \boldsymbol{\delta}_s(\cdot)$, 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 $s, s' \in \mathcal{S}$,

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

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.

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, π. In particular, by Birkhoff's individual ergodic theorem [27, Theorem 2.3.4], the weak convergence of μ_λ to $\hat{\mu}$, and the fact that μ_λ is ergodic, we have that the expected percentage of time that the process spends in any $O \in \mathcal{B}(\mathcal{Z})$ such that $\partial O \cap \mathcal{S} \neq \varnothing$ is given by $\hat{\mu}(O)$ as the experimentation probability λ approaches zero and time increases, i.e.,

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

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 [28], [10], 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 $\mathcal{S}$ (Definition 3.1). 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 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 V, 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 $\mathcal{S}$.

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_∞ *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 $i \in \mathcal{I}$. Then, the following hold:

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

Proof. See Appendix B. □

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 $\mathcal{S}$, which is established by the following proposition.

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

(a) for μ-a.e. $z \in \mathcal{Z}$, $\Pi(z, \cdot)$ is an i.p.m. 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 i.p.m. for Π;
(d) the support[1] of Π is on $\mathcal{S}$ for all $z \in \mathcal{Z}$.

Proof. The state space $\mathcal{Z}$ 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 [27, Theorem 5.2.2 (a), (b), (e)].

(d) Let us assume that the support of Π includes points in $\mathcal{Z}$ other than the pure strategy states in $\mathcal{S}$. Then, there exists an open set $O \in \mathfrak{B}(\mathcal{Z})$ such that $O \cap \mathcal{S} = \varnothing$ and $\Pi(z^*, O) > 0$ for some $z^* \in \mathcal{Z}$. According to (b), P^t converges weakly to Π. Thus, from Portmanteau theorem (cf., [27, 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. □

Proposition 4.2 states that the limiting unperturbed t.p.f. converges weakly to a t.p.f. Π which accepts the same i.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. Invariant probability measure (i.p.m.) of perturbed process

According to the definition of perturbed learning automata of Table I, 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 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. 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^* as the t.p.f. where *at least two players tremble*. Then, we may write:

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

where $\psi(\lambda) \doteq 1 - \frac{n\lambda(1-\lambda)^{n-1}}{1-(1-\lambda)^n}$ 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_λ by Q as the perturbation factor λ 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:

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

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.

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

Proposition 4.3 (i.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 i.p.m. of $Q\Pi$.

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

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

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

$$\begin{aligned} & \varphi(\lambda) \sum_{t=T}^{\infty} (1 - \varphi(\lambda))^t \|P^t f - \Pi f\|_\infty \\ & \leq \quad (1 - \varphi(\lambda))^T \sup_{t\geq T} \|P^t f - \Pi f\|_\infty . \end{aligned}$$

[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 $F \cap O \neq \varnothing$.

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 δ for all $t \geq T$. Thus, the sequence

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

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

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

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

$$\lim_{\lambda\downarrow 0}\|R_\lambda f - \Pi f\|_\infty \leq \|A - A_T\|_\infty, \text{ for all } T > 0,$$

which concludes the proof.

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

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

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_λ, 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_λ, μ_λ, 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 μ_λ is also an i.p.m. of P_λ^L.

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

$$\begin{aligned}&\|P_\lambda f - Pf\|_\infty\\ &\geq \|\mu_\lambda(P_\lambda f - Pf)\|_\infty\\ &= \|(\mu_\lambda - \hat{\mu})(I-P)[f]\|_\infty.\end{aligned}$$

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

$$\begin{aligned}&\hat{\mu}[f] - \hat{\mu}Q\Pi f\\ &= (\hat{\mu}[f] - \mu_\lambda[f]) + \mu_\lambda[P_\lambda^L f - Q\Pi f] +\\ &\quad(\mu_\lambda[Q\Pi f] - \hat{\mu}[Q\Pi f]).\end{aligned}$$

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 μ_λ 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. μ_λ 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 μ_λ 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

$$\hat{\mu}(\cdot) = \sum_{s\in\mathcal{S}}\pi_s\boldsymbol{\delta}_s(\cdot) \tag{6}$$

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., $\pi = \pi\hat{P}$.

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

For any two distinct $s, s' \in \mathcal{S}$, note that $\mathcal{N}_\delta(s')$, $\delta > 0$, is a continuity set of $Q\Pi(s,\cdot)$, i.e., $Q\Pi(s, \partial\mathcal{N}_\delta(s')) = 0$. Thus, from Portmanteau theorem, given that $QP^t \Rightarrow Q\Pi$,

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

If we also define $\pi_s \doteq \hat{\mu}(\mathcal{N}_\delta(s))$, then

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

which shows that π 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 $\mathcal{S}$ of pure strategy states is isomorphic with the set $\mathcal{A}$ of action profiles. If agent i trembles (as t.p.f. Q dictates), then all actions in $\mathcal{A}_i$ have positive probability of being selected, i.e., $Q(\alpha, (\alpha_i', \alpha_{-i})) > 0$ for all $\alpha_i' \in \mathcal{A}_i$ and $i \in \mathcal{I}$. It follows by Proposition 4.1 that $Q\Pi(\alpha, (\alpha_i', \alpha_{-i})) > 0$ for all $\alpha_i' \in \mathcal{A}_i$ and $i \in \mathcal{I}$. Finite induction then shows that $(Q\Pi)^n(\alpha, \alpha') > 0$ for all $\alpha, \alpha' \in \mathcal{A}$. It follows that if we restrict the domain of $Q\Pi$ to $\mathcal{S}$, 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.

V. Stochastically Stable States

In this section, we capitalize on Theorem 3.1 and we further simplify the computation of the stochastically stable states in games satisfying Property 2.1.

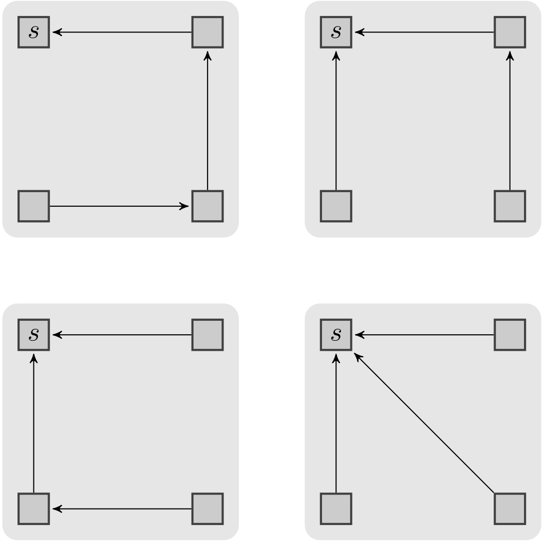


Fig. 1. Examples of s-graphs in case $\mathcal{S}$ contains four states.

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 [29]. 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 $\mathcal{S}$ be a finite set of states, whose elements will be denoted by s_k, s_ℓ, etc., and let a subset $\mathcal{W}$ of $\mathcal{S}$.

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

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

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

$$\varpi(g) \doteq \prod_{(s_k \to s_\ell) \in g} \hat{P}_{s_k s_\ell}.$$

The following Lemma holds:

Lemma 5.1 (Lemma 6.3.1 in [29]): Let us consider a Markov chain with a finite set of states $\mathcal{S}$ and transition probabilities $\{\hat{P}_{s_k s_\ell}\}$ 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

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

where $R_s \doteq \sum_{g \in \mathcal{G}\{s\}} \varpi(g)$.

In other words, in order to compute the weight that the stationary distribution assigns to a state $s \in \mathcal{S}$, 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 $\mathcal{S}$ since this will allow for explicitly computing the stationary distribution π of Theorem 3.1. Based on the definition of the t.p.f. $Q\Pi$, 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 $D \subset \mathcal{Z}$. Denote the minimum hitting time of a set $D \subset \mathcal{Z}$ as $\tau_s^*(D)$ when the process starts from state $s \in \mathcal{S}$. Let us also define the set

$$D_{i,\ell}(\alpha) \doteq \left\{(\alpha, x) \in \mathcal{Z} : x_{i\alpha_i} > 1 - H_i(\alpha)^\ell\right\},$$

where $H_i(\alpha) \doteq 1 - \epsilon u_i(\alpha)$. 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 $Q\Pi$ and plays action α_i for ℓ consecutive times.

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 α and α', respectively. Let also $z = (\alpha', x')$, where $x'_j \doteq e_{\alpha_j} + \epsilon u_j(\alpha')(e_{\alpha'_j} - e_{\alpha_j})$, which corresponds to the state after agent j perturbed once starting from s and played α'_j. Define also $\breve{P}_{ss'}(\delta) \doteq \mathbb{P}_z[\tau(\mathcal{N}_\delta(s')) < \infty]$ which corresponds to the probability that the process transits from the perturbed state z to a δ-neighborhood of s' in finite time. For sufficiently small ϵ such that $0 < \epsilon u_j(\alpha') < 1$, the following hold:

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

$$\hat{P}_{ss'} = \gamma_j \cdot \lim_{\delta \downarrow 0} \breve{P}_{ss'}(\delta), \tag{8}$$

where $\gamma_j \doteq 1/(n\,|\mathcal{A}_j|)$ corresponds to the probability that agent j trembled and selected action α'_j, given that only one player trembles (under t.p.f. Q).

(b) Along any sample path that reaches the set $\mathcal{N}_\delta(s')$, action profile α' is played at least $\tau_s^*(\mathcal{N}_\delta(s'))$ times.

(c) $\breve{P}_{ss'}(\delta)$ corresponds to the probability of the *shortest path*, i.e.,

$$\breve{P}_{ss'}(\delta) = \mathbb{P}_z\left[\alpha(t+1) = \alpha', t < \tau_s^*(\mathcal{N}_\delta(s'))\right].$$

(d) There exists positive constant $C_0(\delta)$, such that for any transition step $s \to s'$ (with the above properties) and as $\epsilon \downarrow 0$,

$$\breve{P}_{ss'}(\delta) \approx \exp\left(-\frac{C_0(\delta)}{\epsilon u_j(\alpha')}\right). \qquad (9)$$

Proof. See Appendix C. □

Note that for sufficiently small ϵ, *the larger the destination utility $u_j(\alpha')$, the larger the transition probability to s'*. In a way, the inverse of the destination utility at s' represents a measure of "resistance" of the process to transit to s'. 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.

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 $Q\Pi$ t.p.f.. By definition of $Q\Pi$, this approximation is based upon the observation that for the computation of the quantities R_s of Lemma 5.1, it suffices to consider only those paths in $\mathcal{G}\{s\}$ which involve *one-step* transitions as defined in the previous section.

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

Lemma 5.3 (Approximation of stationary distribution): The stationary distribution of the finite Markov chain $\{\hat{P}_{s_k s_\ell}\}$, $\pi = [\pi_s]$, satisfies

$$\pi_s = \lim_{\delta \downarrow 0} \frac{\breve{R}_s(\delta)}{\sum_{s_i \in \mathcal{S}} \breve{R}_{s_i}(\delta)}, \qquad s \in \mathcal{S}, \qquad (10)$$

where $\breve{R}_s(\delta) \doteq \sum_{g \in \mathcal{G}^{(1)}\{s\}} \breve{\varpi}(g;\delta)$, and

$$\breve{\varpi}(g;\delta) \doteq \bar{\gamma}_g \prod_{(s_k \to s_\ell) \in g} \breve{P}_{s_k s_\ell}(\delta), \qquad (11)$$

for some constant $\bar{\gamma}_g \in (0,1)$.

Proof. According to Lemma 5.1, for any $s \in \mathcal{S}$, we have $\pi_s = R_s / \sum_{s_i \in \mathcal{S}} R_{s_i}$. Given the definition of the t.p.f. Q, where only one player trembles, we should only consider one-step transition probabilities (as defined in Lemma 5.2). Thus,

$$R_s = \sum_{g \in \mathcal{G}^{(1)}\{s\}} \varpi(g) = \sum_{g \in \mathcal{G}^{(1)}\{s\}} \prod_{(s_k \to s_\ell) \in g} \hat{P}_{s_k s_\ell}.$$

According to Lemma 5.2 and Equation (8), we have

$$\begin{aligned} R_s &= \lim_{\delta \downarrow 0} \sum_{g \in \mathcal{G}^{(1)}\{s\}} \prod_{(s_k \to s_\ell) \in g} \gamma_{j(s_k,s_\ell)} \breve{P}_{s_k s_\ell}(\delta) \\ &= \lim_{\delta \downarrow 0} \sum_{g \in \mathcal{G}^{(1)}\{s\}} \bar{\gamma}_g \prod_{(s_k \to s_\ell) \in g} \breve{P}_{s_k s_\ell}(\delta) \end{aligned}$$

where $j(s_k, s_\ell)$ denotes the single player whose action changes from s_k to s_ℓ, and $\bar{\gamma}_g \doteq \prod_{(s_k \to s_\ell) \in g} \gamma_{j(s_k,s_\ell)} \in (0,1)$. Thus, the conclusion follows. □

Note that Lemma 5.3 provides a simplification to Theorem 3.1, since it suffices to compute the transition probabilities of the $\mathcal{W}$-graphs consisting solely of one step transitions. Furthermore, the transition probability of any such graph, $\breve{\varpi}(g;\delta)$, 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. δ-resistance

We have shown in Lemma 5.2, that the order of the one-step transition probability $\breve{P}_{ss'}(\delta)$ increases as the destination utility increases. Informally, *the inverse destination utility at s' represents a form of "resistance" in approaching state s'*. In this section, we will formalize this notion and we will relate it to the stationary distribution π.

Definition 5.2 (δ-resistance): For a pure strategy state $s \in \mathcal{S}$, let us consider any graph $g \in \mathcal{G}^{(1)}\{s\}$. For any $\delta > 0$, the δ-resistance associated with $s \in \mathcal{S}$ in graph g, is defined as follows:

$$\varphi_\delta(s|g) \doteq \sum_{(s_k \to s_\ell) \in g} \frac{1}{\epsilon u_j(\alpha^{(\ell)})}. \qquad (12)$$

In other words, the δ-resistance of a state s along a graph corresponds to the sum of the inverse utilities of the destination states along that graph. We further denote by $\varphi^*_\delta(s)$ the minimum δ-resistance, i.e., $\varphi^*_\delta(s) \doteq \min_{g \in \mathcal{G}^{(1)}\{s\}} \varphi_\delta(s|g)$ 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.

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

$$\overline{\Phi}_\delta(\mathcal{S}^*) \doteq \max_{s^* \in \mathcal{S}^*} \varphi^*_\delta(s^*) < \min_{s \in \mathcal{S} \backslash \mathcal{S}^*} \varphi^*_\delta(s) \doteq \underline{\Phi}_\delta(\mathcal{S} \backslash \mathcal{S}^*). \quad (13)$$

Proof. By Lemmas 5.2–5.3, for any state $s \in \mathcal{S}$ and for any graph $g \in \mathcal{G}^{(1)}\{s\}$, we have that, as $\epsilon \downarrow 0$,

$$\breve{\varpi}(g;\delta) = \bar{\gamma}_g \prod_{(s_k \to s_\ell) \in g} \breve{P}_{s_k s_\ell}(\delta) \approx \bar{\gamma}_g \exp\left(-C_0(\delta)\varphi_\delta(s|g)\right),$$

and, $\breve{R}_s(\delta) = \sum_{g \in \mathcal{G}^{(1)}\{s\}} \bar{\gamma}_g \exp\left(-C_0(\delta)\varphi_\delta(s|g)\right)$. Thus, for the states in $\mathcal{S} \backslash \mathcal{S}^*$, and as $\epsilon \downarrow 0$, we have

$$\sum_{s \in \mathcal{S} \backslash \mathcal{S}^*} \breve{R}_s(\delta) = e^{-C_0(\delta)\underline{\Phi}(\mathcal{S} \backslash \mathcal{S}^*)}.$$

$$\sum_{s \in \mathcal{S} \backslash \mathcal{S}^*} \sum_{g \in \mathcal{G}^{(1)}\{s\}} \bar{\gamma}_g e^{-C_0(\delta)(\varphi_\delta(s|g) - \underline{\Phi}(\mathcal{S} \backslash \mathcal{S}^*))},$$

which approaches 0 as $\epsilon \downarrow 0$, since $\varphi_\delta(s|g) \geq \underline{\Phi}(\mathcal{S}\backslash\mathcal{S}^*)$ for each $s \in \mathcal{S}\backslash\mathcal{S}^*$. Analogously, for the states in $\mathcal{S}^*$, we have

$$\sum_{s\in\mathcal{S}^*} \breve{R}_s(\delta) = e^{-C_0(\delta)\overline{\Phi}_\delta(\mathcal{S}^*)} \cdot \sum_{s\in\mathcal{S}^*} \sum_{g\in\mathcal{G}^{(1)}\{s\}} \bar{\gamma}_g e^{-C_0(\delta)(\varphi_\delta(s|g)-\overline{\Phi}_\delta(\mathcal{S}^*))}.$$

Thus, as $\epsilon \downarrow 0$,

$$\frac{\sum_{s\in\mathcal{S}\backslash\mathcal{S}^*} \breve{R}_s(\delta)}{\sum_{s\in\mathcal{S}^*} \breve{R}_s(\delta)} = e^{-C_0(\delta)(\underline{\Phi}(\mathcal{S}\backslash\mathcal{S}^*)-\overline{\Phi}_\delta(\mathcal{S}^*))} \cdot \frac{\sum_{s\in\mathcal{S}\backslash\mathcal{S}^*} \sum_{g\in\mathcal{G}^{(1)}\{s\}} \bar{\gamma}_g e^{-C_0(\delta)(\varphi_\delta(s|g)-\underline{\Phi}(\mathcal{S}\backslash\mathcal{S}^*))}}{\sum_{g\in\mathcal{G}^{(1)}\{s\}} \bar{\gamma}_g e^{-C_0(\delta)(\varphi_\delta(s|g)-\overline{\Phi}_\delta(\mathcal{S}^*))}}.$$

Given that $\underline{\Phi}(\mathcal{S}\backslash\mathcal{S}^*) - \overline{\Phi}_\delta(\mathcal{S}^*) > 0$, the first part of the above ratio approaches 0 as $\epsilon \downarrow 0$. The same holds for the numerator of the second part, due to the definition of $\underline{\Phi}_\delta(\mathcal{S}^*)$. On the other hand, the denominator approaches ∞ as $\epsilon \downarrow 0$. Thus, we conclude that

$$\frac{\sum_{s\in\mathcal{S}\backslash\mathcal{S}^*} \breve{R}_s(\delta)}{\sum_{s\in\mathcal{S}^*} \breve{R}_s(\delta)} \xrightarrow{\epsilon\downarrow 0} 0. \tag{14}$$

Denote by $\pi_{\mathcal{S}^*}$ the probability assigned by the invariant probability distribution π to the subset $\mathcal{S}^*$ of $\mathcal{S}$. Then, according to (10), we have:

$$\begin{aligned} &\lim_{\epsilon\downarrow 0} \pi_{\mathcal{S}^*} \\ &= \lim_{\epsilon\downarrow 0}\lim_{\delta\downarrow 0} \frac{\sum_{s^*\in\mathcal{S}^*} \breve{R}_{s^*}(\delta)}{\sum_{s\in\mathcal{S}} \breve{R}_s(\delta)} = \lim_{\delta\downarrow 0}\lim_{\epsilon\downarrow 0} \frac{\sum_{s^*\in\mathcal{S}^*} \breve{R}_{s^*}(\delta)}{\sum_{s\in\mathcal{S}} \breve{R}_s(\delta)} \\ &= \lim_{\delta\downarrow 0}\lim_{\epsilon\downarrow 0} \frac{1}{1+\sum_{s\in\mathcal{S}\backslash\mathcal{S}^*} \breve{R}_s(\delta)/\sum_{s^*\in\mathcal{S}^*} \breve{R}_{s^*}(\delta)}. \end{aligned}$$

Note that that the interchange of limits in the second equality is valid due to the fact that, by Lemma 5.3, the transition probabilities $\breve{P}_{s_k s_\ell} \to 0$, either when $\epsilon \downarrow 0$ or when $\delta \downarrow 0$. Given (14), we conclude that $\lim_{\epsilon\downarrow 0} \pi_{\mathcal{S}^*} = 1$. Conversely, $\lim_{\epsilon\downarrow 0} \pi_{\mathcal{S}\backslash\mathcal{S}^*} = 0$. Thus, the stochastically stable states may only be contained in $\mathcal{S}^*$. □

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

VI. Illustration in Coordination Games

A. Stochastic stability

In this section, 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:

$$\mathrm{BR}_i(\alpha) \doteq \arg\max_{a\in\mathcal{A}_i} u_i(a, \alpha_{-i}).$$

Definition 6.2 (Nash equilibrium): An action profile $\alpha^* = (\alpha_i^*, \alpha_{-i}^*)$ is a Nash equilibrium, if for every player i,

$$\alpha_i^* \in \mathrm{BR}_i(\alpha^*).$$

A best-response of a player i to the current action profile will often be denoted by α_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 satisfying the positive-utility property (Property 2.1) is a coordination game if, for every action profile α and player i, $u_j(\alpha_i', \alpha_{-i}) \geq u_j(\alpha_i, \alpha_{-i})$ for any $\alpha_i' \in \mathrm{BR}_i(\alpha)$.

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.

In order to address stochastic stability, we will further need to introduce the notion of the best-BR (briefly, BBR).

Definition 6.4 (Best-BR): Let $i^* : \mathcal{A} \to \mathcal{I}$ be defined as:

$$i^*(\alpha) \doteq \arg\max_{i\in\mathcal{I}} \left\{ u_i(\alpha_i, \alpha_{-i}) : \alpha_i \in \mathrm{BR}_i(\alpha) \right\}.$$

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

In other words, $\mathrm{BBR}(\alpha)$ corresponds to the one-step transition, where the player which changes its action receives the largest utility among all possible one-step transitions from α.

Lemma 6.1: Let $\mathcal{S}_{\mathrm{NE}}$ be the set of p.s.s.'s which correspond to the set of pure Nash equilibria. In any coordination game, the $\{\mathcal{S}_{\mathrm{NE}}\}$-graph that attains the minimum δ-resistance is: $g^*(\mathcal{S}_{\mathrm{NE}}) = \left\{ (s_k \to s_\ell) : \alpha^{(\ell)} \in \mathrm{BBR}(\alpha) \right\}$.

Proof. Under the coordination property, and starting from any state $s \notin \mathcal{S}_{\mathrm{NE}}$, we can construct a path starting from s and leading to $\mathcal{S}_{\mathrm{NE}}$ that consists only of one-step best-BR's. Such a path will include no cycles (since the utility of all players may not decrease along such path). Furthermore, such path of best-BR's may only terminate at a Nash equilibrium.

By Definition 5.1 of a $\{\mathcal{S}_{\mathrm{NE}}\}$-graph, a state $s \notin \mathcal{S}_{\mathrm{NE}}$ is the source of exactly one arrow. Among the possible arrows with source s, the one that corresponds to a best-BR is the one with the minimum δ-resistance (since it provides the maximum possible destination utility). We conclude that the $\{\mathcal{S}_{\mathrm{NE}}\}$-graph(s) consisting only of best-BR's provide the minimum δ-resistance. Thus, the conclusion follows. □

In other words, Lemma 6.1 shows that the $\{\mathcal{S}_{\mathrm{NE}}\}$-graph of minimum δ-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 in any coordination game.

Theorem 6.1 (Stochastic stability in coordination games): In any coordination game of Definition 6.3, as $\epsilon \downarrow 0$ and $\lambda \downarrow 0$, the stochastically-stable pure-strategy states satisfy $\mathcal{S}^* \subseteq \mathcal{S}_{\mathrm{NE}}$.

Proof. It suffices to show that all p.s.s.'s outside $\mathcal{S}_{\mathrm{NE}}$ provide a δ-resistance which is higher than the δ-resistance of any Nash equilibrium in $\mathcal{S}_{\mathrm{NE}}$ (as Theorem 5.1 dictates).

Consider an action profile α which is not a Nash equilibrium and the corresponding p.s.s. s. Consider the part of the optimal $\mathcal{S}_{\mathrm{NE}}$-graph which leads to s, i.e.,

$$g^*(s|\mathcal{S}_{\mathrm{NE}}) \doteq \{(s_k \to s_\ell) \in g^*(\mathcal{S}_{\mathrm{NE}}) : \exists \text{ path from } s_\ell \text{ to } s\}\,.$$

In other words, $g^*(s|\mathcal{S}_{\mathrm{NE}})$ corresponds to the part of the minimum-resistance graph $g^*(\mathcal{S}_{\mathrm{NE}})$ whose arrows lead to s. This graph might be empty if s is not a recipient of any arrow in $g^*(\mathcal{S}_{\mathrm{NE}})$. For the remainder of the proof, define the graphs: $g_1 \doteq g^*(\mathcal{S}_{\mathrm{NE}})\backslash g^*(s|\mathcal{S}_{\mathrm{NE}})$, $g_2 \doteq g^*(s)\backslash g^*(s|\mathcal{S}_{\mathrm{NE}})$. Note that, $g^*(s|\mathcal{S}_{\mathrm{NE}}) \subset g^*(s)$, i.e., the graph that leads to s through the minimum resistance graph of $\mathcal{S}_{\mathrm{NE}}$ is also part of the minimum resistance graph of s. By construction, we also have $g^*(s|\mathcal{S}_{\mathrm{NE}}) \subset g^*(\mathcal{S}_{\mathrm{NE}})$. Thus, the exact same arrows (i.e., the ones in $g^*(s|\mathcal{S}_{\mathrm{NE}})$) are subtracted from $g^*(\mathcal{S}_{\mathrm{NE}})$ and $g^*(s)$ to define the graphs g_1 and g_2, respectively.

By definition of the $\{\mathcal{S}_{\mathrm{NE}}\}$-graphs, a node within the set $\{\mathcal{S}_{\mathrm{NE}}\}$ cannot be the source of any arrow in g_1. Similarly, node s may not be the source of any arrow in g_2. Since $|\mathcal{S}_{\mathrm{NE}}| \geq 1$, and the fact that only a single arrow may stem from any given node, we conclude that $|g_1| \leq |g_2|$, 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 $s' \notin \mathcal{S}_{\mathrm{NE}}$ with the following property: $(s' \to s'') \in g_1$ such that $\alpha'' \in \mathrm{BBR}(\alpha')$, and $(s' \to s''') \in g_2$ such that $\alpha''' \notin \mathrm{BBR}(\alpha')$. This is due to the fact that any path in g_2 should eventually lead to $s \notin \mathcal{S}_{\mathrm{NE}}$.

Thus, we conclude that g_2 contains at least as many arrows as g_1, and g_2 contains arrows which are not best-BR steps. 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 δ-resistance as compared to the minimum δ-resistance of $g^*(\mathcal{S}_{\mathrm{NE}})$. □

B. Simulation study in distributed network formation

In this section, we perform a simulation study of the perturbed learning automata in network formation games, analyzed in [30]. We consider n nodes deployed on the plane and assume that the set of actions of each agent i, $\mathcal{A}_i$, contains all possible combinations of neighbors of i, denoted $\mathcal{N}_i$, with which a link can be established, i.e., $\mathcal{A}_i = 2^{\mathcal{N}_i}$. Links are considered unidirectional, 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.,

$$(j \to i) = \{j = j_0, (j_0, j_1), j_1, \ldots, (j_{m-1}, j_m), j_m = i\}$$

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

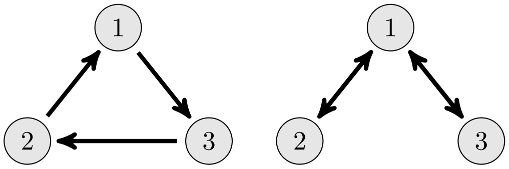


Fig. 2. Nash networks in case of $n = 3$ agents and $0 < \nu < 1$.

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

$$u_i(\alpha) \doteq \sum_{j \in \mathcal{I}\backslash\{i\}} \chi_\alpha(j \to i) - c\,|\alpha_i|\,, \tag{15}$$

where $|\alpha_i|$ denotes the number of links corresponding to α_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_α denotes the graph induced by joint action α. As it was shown in Proposition 4.2 in [30], *a network G^* is a Nash equilibrium 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 Nash equilibria for $n = 3$ agents and unconstrained neighborhoods are shown in Fig. 2.

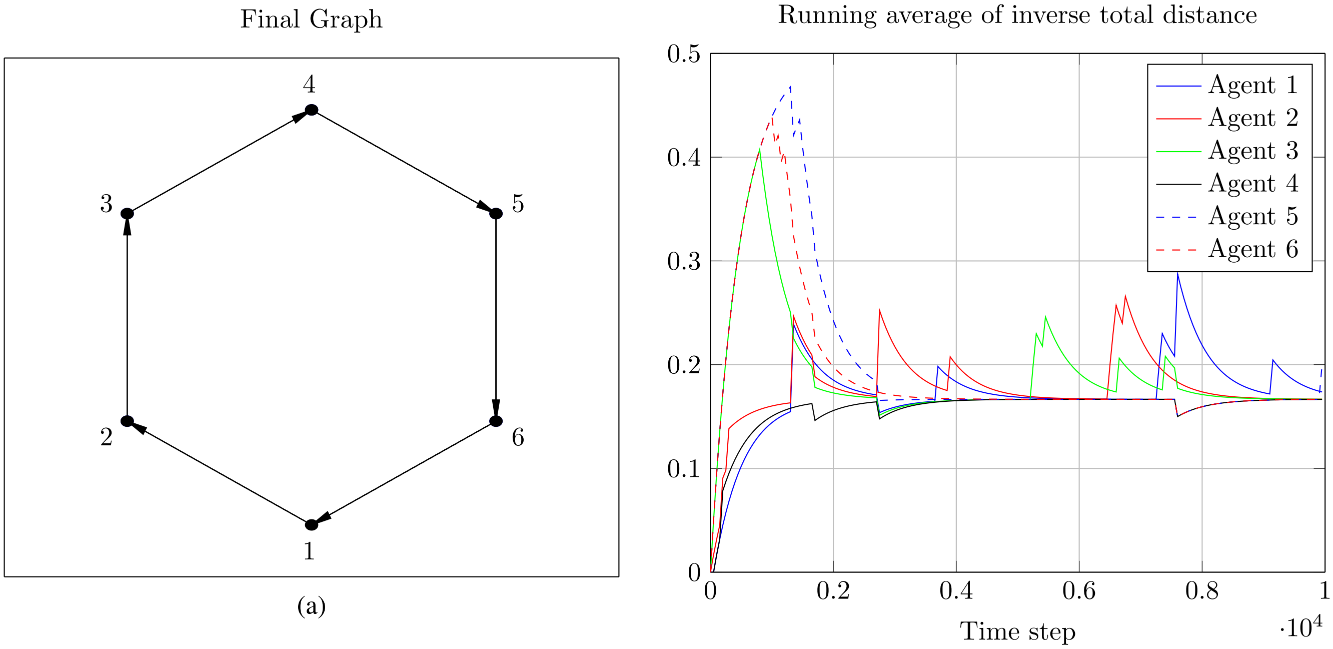
Proposition 6.1: The network formation game defined by (15) is a coordination game.

Proof. First, note that any network formation game with the utility of (15) satisfies the positive-utility property. This is due to the fact that for any single link of cost $c \in (0,1)$, an agent receives utility of at least 1.

For a joint action $\alpha \notin \mathcal{A}^*$ 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 depicts the response of the learning dynamics 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 λ and ϵ need to be sufficiently small. We choose: $\epsilon = \lambda = 0.005$, and $c = 1/2$.

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



(b)

Fig. 3. (a) Final graph and (b) running-average inverse total distance with time under reinforcement-learning.

structure (and only under this structure), the inverse total distance to the neighboring nodes is equal to $1/1+5 = 1/6 \approx 0.167$. The wheel structure is among the Nash equilibria of this game (as shown in [30]) and the unique payoff-dominant equilibrium (i.e., every node receives its maximum utility). The wheel structure is the emergent structure under perturbed learning automata as shown in Fig. 3.

The simulation of Fig. 3 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 with no restrictions in the number of players and without requiring the existence of a potential function. Thus, it opens up new possibilities for the use of reinforcement-based learning in distributed control of multi-agent systems.

Appendix A
Proof of Proposition 3.1

Let us consider the perturbed process P_λ. The proof for the unperturbed process will be directly implied by employing $\lambda = 0$.

Let us also consider any sequence $\{z^{(k)} = (\alpha^{(k)}, x^{(k)})\}$ such that $z^{(k)} \to z = (\alpha, x) \in \mathcal{Z}$. For any open set $O \in \mathfrak{B}(\mathcal{Z})$, the following holds:

$$\begin{aligned}
&P_\lambda(z^{(k)} = (\alpha^{(k)}, x^{(k)}), O) \\
&\quad = \sum_{\alpha \in \mathcal{P}_{\mathcal{A}}(O)} \Big\{ \prod_{i \in \mathcal{I}} \tilde{x}_{i\alpha_i}^{(k)} \cdot \prod_{i \in \mathcal{I}} \mathbb{P}_{z^{(k)}}[\mathcal{R}_i(\alpha, x_i^{(k)}) \in \mathcal{P}_{\mathcal{X}_i}(O)] \Big\} \\
&\quad = \sum_{\alpha \in \mathcal{P}_{\mathcal{A}}(O)} \Big\{ \prod_{i=1}^{n} \mathbb{I}_{\mathcal{P}_{\mathcal{X}_i}(O)}(\mathcal{R}_i(\alpha, x_i^{(k)})) \tilde{x}_{i\alpha_i}^{(k)} \Big\},
\end{aligned}$$

where $\mathcal{P}_{\mathcal{X}_i}(O)$ and $\mathcal{P}_{\mathcal{A}}(O)$ are the *canonical projections* defined by the product topology, and $\tilde{x}_{i\alpha_i}^{(k)} \doteq (1-\lambda) x_{i\alpha_i}^{(k)} + \lambda/|\mathcal{A}_i|$. Similarly, we have:

$$P_\lambda(z, O) = \sum_{\alpha \in \mathcal{P}_{\mathcal{A}}(O)} \Big\{ \prod_{i=1}^{n} \mathbb{I}_{\mathcal{P}_{\mathcal{X}_i}(O)} \left(\mathcal{R}_i \left(\alpha, x_i \right) \right) \tilde{x}_{i\alpha_i} \Big\}.$$

To investigate the limit of $P_\lambda(z^{(k)}, O)$ as $k \to \infty$, it suffices to investigate the behavior of the sequence

$$y_i^{(k)} \doteq \mathbb{I}_{\mathcal{P}_{\mathcal{X}_i}(O)}(\mathcal{R}_i(\alpha, x_i^{(k)})).$$

We distinguish the following (complementary) cases:

(a) $\mathcal{R}_i(\alpha, x_i) \notin \mathcal{P}_{\mathcal{X}_i}(O)$ and $\mathcal{R}_i(\alpha, x_i) \notin \partial\mathcal{P}_{\mathcal{X}_i}(O)$: 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 $\mathcal{R}_i(\alpha, \cdot)$, we have that $y_i^{(k)} \to y_i \doteq \mathbb{I}_{\mathcal{P}_{\mathcal{X}_i}(O)}(\mathcal{R}_i(\alpha, x_i)) \equiv 0$.

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

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

In either one of the above (complementary) cases, we have that $\liminf_{k\to\infty} y_i^{(k)} \geq y_i$. 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 $z^{(k)} \to z$,

$$\liminf_{k\to\infty} P_\lambda(z^{(k)}, O) \geq P_\lambda(z, O).$$

By [27, Proposition 7.2.1], we conclude that P_λ 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 $\tilde{x}_{i\alpha_i}$, 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 α is selected up to time t, then the strategy of agent i satisfies:

$$x_i(t) = e_{\alpha_i} - (1 - \epsilon u_i(\alpha))^t (e_{\alpha_i} - x_i(0)). \tag{16}$$

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

$$\mathbb{P}_z[B_\infty] = \lim_{t\to\infty} \mathbb{P}_z[B_t] = \lim_{t\to\infty} \prod_{k=0}^{t} \prod_{i=1}^{n} x_{i\alpha_i}(k).$$

Note that $\mathbb{P}_z[B_\infty] > 0$ if and only if

$$\sum_{t=0}^{\infty} \log(x_{i\alpha_i}(t)) > -\infty, \text{ for all } i \in \mathcal{I}. \tag{17}$$

Let us introduce the variable $y_i(t) \doteq 1 - x_{i\alpha_i}(t)$, which corresponds to the probability of agent i selecting any action other than α_i. Condition (17) is equivalent to

$$-\sum_{t=0}^{\infty} \log(1 - y_i(t)) < \infty, \text{ for all } i \in \mathcal{I}. \tag{18}$$

Note that $y_i(t+1)/y_i(t) = 1 - \epsilon u_i(\alpha) < 1$, which implies (by the Ratio test, cf., [31, Theorem 6.2.4]) that the series of positive terms $\sum_{t=1}^{\infty} y_i(t)$ is convergent. This further implies that $\lim_{t\to\infty} y_i(t) = 0$. Thus, from L'Hospital's rule (cf., [32, Theorem 5.13]),

$$\lim_{t\to\infty} \frac{-\log(1 - y_i(t))}{y_i(t)} = \lim_{t\to\infty} \frac{1}{1 - y_i(t)} = 1 > 0. \tag{19}$$

From the Limit Comparison Test (cf., [31, Theorem 6.2.2]), we conclude that condition (18) holds, which equivalently implies that $\mathbb{P}_z[B_\infty] > 0$. Lastly, due to (16), $\mathbb{P}_z[B_\infty]$ is continuous with respect to $x(0)$ which takes values in a bounded and closed set $\mathcal{X}$. Thus, by [31, Theorem 3.2.2], we conclude that $\inf_{z\in\mathcal{Z}} \mathbb{P}_z[B_\infty] > 0$.

(b) Define the event

$$C_t \doteq \left\{\exists \alpha' \neq \alpha(t) : x_{i\alpha_i'}(t) > 0, \text{ for all } i \in \mathcal{I}\right\},$$

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 $A_t^c \subseteq C_t$, since A_t^c may only occur if there is some action profile $\alpha' \neq \alpha(t)$ for which the strategy assigns positive probability. This further implies that $\mathbb{P}_z[A_t^c] \leq \mathbb{P}_z[C_t]$. Then, we have:

$$\begin{aligned}
&\mathbb{P}_z[A_{t+1}|A_t^c] \\
&\quad= \mathbb{P}_z[A_{t+1} \cap A_t^c]/\mathbb{P}_z[A_t^c] \\
&\quad\geq \mathbb{P}_z[A_{t+1} \cap A_t^c]/\mathbb{P}_z[C_t] \\
&\quad\geq \mathbb{P}_z[A_{t+1} \cap A_t^c \cap C_t]/\mathbb{P}_z[C_t] \\
&\quad= \mathbb{P}_z[A_{t+1} \cap A_t^c | C_t] \\
&\quad= \mathbb{P}_z[A_{t+1} | A_t^c \cap C_t] \cdot \mathbb{P}_z[A_t^c | C_t] \\
&\quad= \mathbb{P}_z[Z(\theta_{t+1}(\omega)) \in B_\infty | A_t^c \cap C_t] \cdot \mathbb{P}_z[A_t^c | C_t],
\end{aligned}$$

where $\theta : \Omega \to \Omega$ denotes the *shift operator*, such that $Z(\theta_t(\omega)) = \{Z_t, Z_{t+1}, ...\}$. We have

$$\begin{aligned}
&\inf_{z\in\mathcal{Z}} \mathbb{P}_z[A_t^c|C_t] \geq \\
&\qquad \inf_{\{\alpha' \neq \alpha(t) : x_{i\alpha_i'}(t) > 0, \forall i\}} (1 - \epsilon u_i(\alpha)) \cdot x_{i\alpha_i'}(t) > 0,
\end{aligned}$$

which corresponds to the probability of switching action at time $t+1$ to $\alpha' \neq \alpha$. Furthermore, if we define the restricted domain

$$\mathcal{Z}_+(\alpha) \doteq \{(\alpha', x) \in \mathcal{Z} : \alpha' \neq \alpha, x_{i\alpha_i'} > 0, \forall i\},$$

then

$$\begin{aligned}
&\mathbb{P}_z[Z(\theta_{t+1}(\omega)) \in B_\infty | A_t^c \cap C_t] \\
&\quad= \inf_{z\in\mathcal{Z}_+(\alpha)} \mathbb{P}_z[Z(\theta_{t+1}(\omega)) \in B_\infty] \\
&\quad\geq \inf_{z\in\mathcal{Z}} \mathbb{P}_z[B_\infty]
\end{aligned}$$

due to the fact that $\mathcal{Z}_+(\alpha) \subseteq \mathcal{Z}$ and the Markov property. Thus, we conclude that $\inf_{z\in\mathcal{Z}} \mathbb{P}_z[A_{t+1}|A_t^c] > 0$ which further implies that $\sum_{t=0}^{\infty} \mathbb{P}_z[A_{t+1}|A_t^c] = \infty$. Hence, from the counterpart of the Borel-Cantelli Lemma (cf., [33, Section 3.3]) and the fact that $A_t \subseteq A_{t+1}$, we have $\sup_{z\in\mathcal{Z}} \mathbb{P}_z[A_\infty] = 1$.

APPENDIX C
PROOF OF LEMMA 5.3

(a) This is a direct implication of the definition of $Q\Pi$ t.p.f..

(b) Let us assume that along a sample path from s to $\mathcal{N}_\delta(s')$ and at iteration t, the strategy of agent j with respect to action α'_j is $x_{j\alpha'_j}(t) = \rho > 0$. If agent j selects action α'_j at time $t+1$, then

$$x_{j\alpha'_j}(t+1) = \rho + \epsilon u_j(\alpha')(1-\rho) = \epsilon u_j(\alpha') + H_j(\alpha')\rho \doteq x^*_{j\alpha'_j}.$$

If, instead, agent j selects action $\alpha_j \neq \alpha'_j$ at time $t+1$ and then α'_j at time $t+2$, i.e., it deviates from playing action α'_j, then the strategy evolves as follows:

$$\begin{aligned} x_{j\alpha'_j}(t+1) &= \rho + \epsilon u_j(\alpha)(-\rho) \\ &= H_j(\alpha)\rho, \\ x_{j\alpha'_j}(t+2) &= H_j(\alpha)\rho + \epsilon u_j(\alpha')(1 - H_j(\alpha)\rho) \\ &= (H_j(\alpha')\rho)H_j(\alpha) + \epsilon u_j(\alpha') \\ &< x^*_{j\alpha'_j}, \end{aligned}$$

since $\epsilon u_j(\alpha) < 1$. Informally, any single deviation from the shortest path to s' cannot recover the drop in the strategy at the next iteration. Thus, along any path to $\mathcal{N}_\delta(s')$, action α' will be played for at least $\tau^*(\mathcal{N}_\delta(s'))$ times.

(c) Observe that one possibility for realizing a transition from s to $\mathcal{N}_\delta(s')$ is to follow the shortest path, that is, the path of playing action α' consecutively. Thus,

$$\breve{P}_{ss'}(\delta) \geq \mathbb{P}_z\left[\alpha(t+1) = \alpha', \forall t < \tau^*(\mathcal{N}_\delta(s'))\right].$$

Let us denote by t_k, $k \in \mathbb{N}$, a subsequence of the iteration index t. Given (b), when the process reaches $D_{j,k}(\alpha')$ for the first time, action α' has been played for at least $\tau^*_s(\mathcal{N}_\delta(s'))$ times. Furthermore, when action profile α' has been played for the kth time (at time instance t_k+1), the state at time t_k may not have reached $D_{j,k}(\alpha')$ (by definition of the set $D_{j,k}(\alpha')$). Formally, we can write:

$$\begin{aligned} \breve{P}_{ss'}(\delta) &\leq \mathbb{P}_z\left[\exists\{t_k\} : \alpha(t_k+1) = \alpha', t_k < \tau(D_{j,k}(\alpha')),\right. \\ &\quad \left.\forall k < \tau^*_s(\mathcal{N}_\delta(s'))\right], \\ &\leq \mathbb{P}_z\left[\exists\{t_k\} : \alpha(t_k+1) = \alpha', Z_{t_k} \in D_{j,k}(\alpha')^c,\right. \\ &\quad \left.\forall k < \tau^*_s(\mathcal{N}_\delta(s'))\right]. \end{aligned}$$

The second inequality is due to the Markov property and the fact that, $t_k < \tau(D_{j,k}(\alpha'))$ implies that the previous state Z_{t_k} may only be within $D_{j,k}(\alpha')^c$. Using the properties of the conditional probability, we may also write:

$$\begin{aligned} \breve{P}_{ss'}(\delta) &\leq \mathbb{P}_z\left[\exists\{t_k\} : \alpha(t_k+1) = \alpha' \right| \\ &\quad \left. Z_{t_k} \in D_{j,k}(\alpha')^c, \forall k < \tau^*_s(\mathcal{N}_\delta(s'))\right]. \end{aligned}$$

By the Markov property,

$$\begin{aligned} \breve{P}_{ss'}(\delta) &\leq \prod_{k=0}^{\tau^*_s(\mathcal{N}_\delta(s'))-1} \mathbb{P}_z\left[\alpha(t_k+1) = \alpha' | Z_{t_k} \in D_{j,k}(\alpha')^c\right] \\ &\leq \prod_{t=0}^{\tau^*_s(\mathcal{N}_\delta(s'))-1} \sup_{Z_t \in D_{j,t}(\alpha')^c} \mathbb{P}_z\left[\alpha(t+1) = \alpha'\right] \\ &= \mathbb{P}_z\left[\alpha(t+1) = \alpha', t < \tau^*_s(\mathcal{N}_\delta(s'))\right]. \end{aligned}$$

Thus, the conclusion follows.

(d) The minimum hitting time of the set $\mathcal{N}_\delta(s')$ starting from s, satisfies:

$$\tau^*_s(\mathcal{N}_\delta(s')) = \left\lceil \frac{\log(\delta)}{\log(1 - \epsilon u_j(\alpha'))} \right\rceil \doteq T(\epsilon).$$

There exists correction function $c(\epsilon) : \mathbb{R}_+ \to [0,1)$, such that

$$T(\epsilon) = \frac{\log(\delta) + c(\epsilon)}{\log(1 - \epsilon u_j(\alpha'))} = \frac{\log(\delta) + c(\epsilon)}{\log(H_j(\alpha'))}, \tag{20}$$

where we recall that $H_j(\alpha') \doteq 1 - \epsilon u_j(\alpha')$. Due to (c), we can write:

$$\log\left(\breve{P}_{ss'}(\delta)\right) = \sum_{t=1}^{T(\epsilon)} \log\left(1 - H_j(\alpha')^t\right). \tag{21}$$

In the remainder of the proof, we will approximate the r.h.s. of (21). To simplify notation, denote $H \doteq H_j(\alpha')$. Note that

$$\lim_{\epsilon \downarrow 0} \log\left(H^{T(\epsilon)}\right) = \lim_{\epsilon \downarrow 0} \left\{ \frac{\log(\delta) + c(\epsilon)}{\log(H)} \log(H) \right\} = \log(\delta),$$

and due to the continuity of the natural logarithm,

$$\lim_{\epsilon \downarrow 0} H^{T(\epsilon)} = \delta. \tag{22}$$

Let us define the sequence $\Gamma_t \doteq -\log(1 - H^t)/H^t$. Note that

$$\lim_{\epsilon \downarrow 0} \Gamma_{T(\epsilon)} = -\frac{\log(1-\delta)}{\delta} \doteq \omega(\delta) \xrightarrow{\delta \downarrow 0} -1. \tag{23}$$

Thus, there exists $T_0 \in \mathbb{N}$, such that for any sufficiently small $\epsilon > 0$, for which $T_0 < T(\epsilon)$, we have

$$\omega(\delta) - \epsilon \leq \Gamma_t \leq \omega(\delta) + \epsilon \tag{24}$$

for any $0 < T_0 \leq t \leq T(\epsilon)$. Equivalently,

$$-(\omega(\delta) + \epsilon)H^t \leq \log(1 - H^t) \leq -(\omega(\delta) - \epsilon)H^t$$

for any $0 < T_0 \leq t \leq T(\epsilon)$. Thus,

$$\begin{aligned} &-(\omega(\delta) + \epsilon) \sum_{t=T_0}^{T(\epsilon)} H^t \\ &\leq \sum_{t=T_0}^{T(\epsilon)} \log(1 - H^t) \leq -(\omega(\delta) - \epsilon) \sum_{t=T_0}^{T(\epsilon)} H^t. \end{aligned}$$

Given

$$\sum_{t=T_0}^{T(\epsilon)} H^t = \frac{H^{T_0} - H^{T(\epsilon)}}{1 - H},$$

we conclude that

$$\lim_{\epsilon \downarrow 0} \left\{ (1 - H) \sum_{t=T_0}^{T(\epsilon)} \log(1 - H^t) \right\} = -\omega(\delta)(1 - \delta),$$

where, we have used the property (22) and the fact that $\lim_{\epsilon \downarrow 0} H^{T_0} = \lim_{\epsilon \downarrow 0} (1 - \epsilon u_j(\alpha'))^{T_0} = 1$. Thus,

$$\lim_{\epsilon \downarrow 0} \left\{ (1 - H) \sum_{t=1}^{T(\epsilon)} \log(1 - H^t) \right\} = c_0 - \omega(\delta)(1 - \delta), \tag{25}$$

where $c_0 \doteq \lim_{\epsilon\downarrow 0}(1-H)\sum_{t=1}^{T_0-1}\log(1-H^t) < 0$. The constant c_0 is independent of ϵ and δ, and it only depends on T_0. Furthermore, it is finite. In fact, by using the property of the natural logarithm $\log(x) \geq 1 - \frac{1}{x}$, $x > 0$ (which can be shown using the Mean Value Theorem), we have that, for each fixed $t > 0$

$$\lim_{\epsilon\downarrow 0}\left\{(1-H)\log(1-H^t)\right\} \geq \lim_{\epsilon\downarrow 0}\frac{-H^t+H^{t+1}}{1-H^t} = -\frac{1}{t},$$

where the last equality is derived using the L'Hospital's rule. Thus, $c_0 \geq -\sum_{t=0}^{T_0-1} {}^{1}\!/_{t} > -\infty$ for fixed T_0. From (21) and (25), and using the fact that $1 - H = \epsilon u_j(\alpha')$, we conclude that, as $\epsilon \downarrow 0$,

$$\log(\breve{P}_{ss'}(\delta)) \approx -\frac{C_0(\delta)}{\epsilon u_j(\alpha')},$$

where $C_0(\delta) \doteq -c_0 + \omega(\delta)(1-\delta) > 0$.

Since the selection of T_0 only depends on the utility function (which admits finite number of values), we may select T_0 so that condition (24) holds uniformly over all action profiles in $\mathcal{A}$. In this case, $C_0(\delta)$ can be selected uniformly for any transition step $s \to s'$.

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] J. Weibull, *Evolutionary Game Theory*. Cambridge, MA: MIT Press, 1997.

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

[14] 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.

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

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

[17] 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.

[18] D. Lewis, *Convention: A Philosophical Study*. Blackwell Publishing, 2002.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] G. Arslan and S. Yuksel, "Decentralized Q-Learning for Stochastic Teams and Games," *IEEE Transactions on Automatic Control*, vol. PP, no. 99, pp. 1–1, 2016.

[25] H. P. Young, "Learning by trial and error," *Games and Economic Behavior*, vol. 65, no. 2, pp. 626–643, Mar. 2009.

[26] J. R. Marden, H. P. Young, and L. Pao, "Achieving Pareto optimality through distributed learning," *SIAM J. Control Optim.*, vol. 52, no. 5, pp. 2753–2770, 2014.

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

[28] R. Karandikar, D. Mookherjee, and D. Ray, "Evolving aspirations and cooperation," *J. Econ. Theory*, vol. 80, pp. 292–331, 1998.

[29] M. I. Freidlin and A. D. Wentzell, *Random perturbations of dynamical systems*. New York, NY: Springer-Verlag, 1984.

[30] G. C. Chasparis and J. S. Shamma, "Network Formation: Neighborhood Structures, Establishment Costs, and Distributed Learning," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1950–1962, Dec. 2013.

[31] M. Reed, *Fundamental Ideas of Analysis*. John Wiley & Sons, Inc., 1998.

[32] W. Rudin, *Principles of Mathematical Analysis*. McGraw-Hill Book Company, 1964.

[33] L. Breiman, *Probability*. Philadelphia: SIAM, 1992.