

Aspiration-based Perturbed Learning Automata

Georgios C. Chasparis

Abstract—This paper introduces a novel payoff-based learning scheme for distributed optimization in repeatedly-played strategic-form games. Standard reinforcement-based learning schemes exhibit several limitations with respect to their asymptotic stability. For example, in two-player coordination games, payoff-dominant (or efficient) Nash equilibria may not be stochastically stable. In this work, we present an extension of perturbed learning automata, namely aspiration-based perturbed learning automata (APLA) that overcomes these limitations. We provide a stochastic stability analysis of APLA in multi-player coordination games. We further show that payoff-dominant Nash equilibria are the only stochastically stable states.

I. INTRODUCTION

Multi-agent formulations can be used to tackle distributed optimization problems due to their reduced communication and computational complexity. In such formulations, agents make their *own* decisions repeatedly over time trying to maximize their *own* utility/performance function. However, due to the interdependencies among agents' utility functions, *local* (or *distributed*) optimization does not necessarily imply *global* (or *centralized*) optimization. The problem becomes even more challenging when the utility function of each agent is unknown, and only measurements of this function (possibly corrupted by noise) are available. For this reason, there have been several efforts towards the design of *distributed payoff-based learning dynamics* for convergence to globally optimal outcomes.

Naturally, several such distributed optimization problems can be formulated as strategic-form games. A rather common objective is then to derive conditions under which convergence to *efficient Nash equilibria* can be achieved, i.e., locally stable outcomes that also maximize a centralized objective. One large class of payoff-based learning dynamics that has been utilized for convergence to Nash equilibria is *reinforcement-based learning*. It may appear under alternative forms, including *discrete-time replicator dynamics* [2], *learning automata* [3], [4] or *approximate policy iteration* or *Q-learning* [5]. It is highly attractive to several engineering applications, since agents do not need to know neither the actions of other agents, nor their own utility function. For example, it has been utilized for system identification and

For the sake of brevity, all proofs have been omitted and can be found 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.

pattern recognition [6], distributed network formation and resource-allocation problems [7].

In reinforcement-based learning, deriving conditions for convergence to Nash equilibria may not be a trivial task especially in the case of large number of agents. Especially in the context of *coordination games* (e.g., [8]), two main difficulties are encountered: a) excluding convergence to pure strategies that are *not* Nash equilibria, and b) excluding convergence to mixed strategy profiles. Recent work by the author in *perturbed learning automata* [9], overcame these limitations by directly characterizing the stochastically stable states of the induced Markov chain (independently of the number of players or actions). This type of analysis allowed for acquiring convergence guarantees in multi-player coordination games (thus, extending previous results in reinforcement-based learning restricted only to potential games).

Although Nash equilibria are stochastically stable in coordination games under perturbed learning automata, not all Nash equilibria may be desirable. An example may be drawn from the classical Stag-Hunt coordination game of Table I. In this game, the first player selects the row of the payoff

TABLE I
THE STAG-HUNT GAME

	A	B
A	5, 5	1, 3
B	3, 1	4, 4

matrix and the second player selects the column. The first element of the selected entry determines the reward of the row player, and the second element determines the reward of the column player. This game has two pure Nash equilibria, which correspond to the symmetric plays (A, A) and (B, B) . Ideally, we would prefer that agents eventually learn to play (A, A) which corresponds to the payoff-dominant (or Pareto-efficient) equilibrium. However, existing results in perturbed learning automata [7], [10], [9] demonstrate that (B, B) may prevail asymptotically with positive probability. The reason lies in the cost that an agent experiences when the other agent deviates from a Nash equilibrium, which captures the notion of *risk dominance* (cf., [11]). In fact, (B, B) is the *risk-dominant* equilibrium in the Stag-Hunt game of Table I.

In the present paper, we extend the perturbed learning automata dynamics presented in [9] to incorporate agents' satisfaction levels, namely *aspiration-based perturbed learning automata* (APLA). In this extended version, an agent reinforces an action based on both repeated selection and its satisfaction level. We provide a stochastically stability

analysis of the proposed dynamics in multi-player coordination games. Furthermore, we show that payoff-dominant Nash equilibria are the only stochastically-stable states, as opposed to standard learning automata.

This paper (also in combination with [9]) provides an analytical framework that significantly expands the utility of reinforcement-based learning in strategic-form games. Note though that several classes of *aspiration-based learning* also guarantee convergence to efficient outcomes in coordination games. For example, the *baseline-based* dynamics of [12], the *mode-based* dynamics of [13], the *trial-and-error* dynamics of [14], and the *aspiration-learning* dynamics of [8] also guarantee convergence to efficient Nash equilibria in certain classes of coordination and weakly-acyclic games. However, existing analysis does not take into account the possibility of noisy observations (with the exception of [12] and through the introduction of sufficiently large *exploration phases*). In comparison with these learning dynamics, learning automata can naturally incorporate noisy observations, as demonstrated in the robustness convergence analysis of [15], due to the indirect filtering of measurement noise in the formulation of the agents' strategies.

In the remainder of the paper, Section II introduces coordination games and Section III presents the *aspiration-based perturbed learning automata* dynamics. Section IV presents the main weak-convergence result and Section V its technical derivation. Section VI provides a refinement of stochastically stable states together with a simulation study. Finally, Section VII presents concluding remarks.

Notation:

- For a Euclidean topological space, $|\cdot|$ denotes the Euclidean distance.
- e_j denotes the *unit vector* in \mathbb{R}^n where its j th entry is equal to 1 and all other entries is equal to 0.
- $\Delta(n)$ denotes the *probability simplex* of dimension n , i.e., $\Delta(n) \doteq \{x \in \mathbb{R}^n : x \geq 0, \mathbf{1}^T x = 1\}$.
- δ_x denotes the Dirac measure at x .
- For a finite set A , $|A|$ denotes its cardinality.
- Let $\sigma \in \Delta(|A|)$ be a finite probability distribution for some finite set A . The random selection of an element of A will be denoted $\text{rand}_\sigma[A]$. If $\sigma = (1/|A|, \dots, 1/|A|)$, i.e., it corresponds to the uniform distribution, the random selection will be denoted by $\text{rand}_{\text{unif}}[A]$.

II. COORDINATION GAMES

We consider the standard setup of finite strategic-form games. There is a finite set of *agents* or *players*, $\mathcal{I} = \{1, 2, \dots, n\}$, and each agent has a finite set of actions, denoted by \mathcal{A}_i . The set of action profiles is the Cartesian product $\mathcal{A} \doteq \mathcal{A}_1 \times \dots \times \mathcal{A}_n$; $\alpha_i \in \mathcal{A}_i$ denotes an *action* of agent i ; and $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathcal{A}$ denotes the *action profile* or *joint action* of all agents. The *payoff/utility function* of player i is a mapping $u_i : \mathcal{A} \rightarrow \mathbb{R}$. An action profile $\alpha^* \in \mathcal{A}$ is a (*pure*) *Nash equilibrium* if, for each $i \in \mathcal{I}$,

$$u_i(\alpha_i^*, \alpha_{-i}^*) \geq u_i(\alpha_i', \alpha_{-i}^*) \quad (1)$$

for all $\alpha_i' \in \mathcal{A}_i$, where $-i$ denotes the complementary set $\mathcal{I} \setminus \{i\}$. We denote the set of pure Nash equilibria by \mathcal{A}^* .

Before defining coordination games, we first need to define the notion of *best response*.

Definition 2.1 (Best Response): The best response of agent $i \in \mathcal{I}$ to an action profile $\alpha = (\alpha_i, \alpha_{-i}) \in \mathcal{A}$ is a set valued map $\text{BR}_i : \mathcal{A} \rightarrow 2^{\mathcal{A}_i}$ such that

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

A *coordination game* is defined as follows:

Definition 2.2 (Coordination game): A game of two or more agents is a coordination game if the following conditions are satisfied:

- (a) for any $\alpha \in \mathcal{A} \setminus \mathcal{A}^*$, there exist $i \in \mathcal{I}$ and action $\alpha_i' \in \text{BR}_i(\alpha)$ such that

$$u_j(\alpha_i', \alpha_{-i}) \geq u_j(\alpha_i, \alpha_{-i}), \text{ for all } j \neq i; \quad (2)$$

- (b) for any $\alpha^* \in \mathcal{A}^*$, there exist an agent i and an action $\tilde{\alpha}_i \in \mathcal{A}_i$ such that

$$u_j(\tilde{\alpha}_i, \alpha_{-i}^*) < u_j(\alpha^*), \text{ for all } j \in \mathcal{I}. \quad (3)$$

The conditions of a coordination game establish a weak form of “coincidence of interests” among players. For example, condition (2) states that there always exists a best response of a player that can make no other player worse off. Due to this condition, a pure Nash equilibrium always exists. Furthermore, condition (3) states that at a pure Nash equilibrium, there exists an action profile that can make every player worse off. It is straightforward to show that coordination games are weakly acyclic games (cf., [16]). For example, the Stag-Hunt game of Table I satisfy the properties of Definition 2.2. Alternative examples can be found, e.g., the *network formation games* and *common-pool games* presented in [8].

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

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

III. ASPIRATION-BASED PERTURBED LEARNING AUTOMATA (APLA)

In this section, we present a novel reinforcement-based learning algorithm, namely *aspiration-based perturbed learning automata* (APLA).

The proposed dynamics is presented in Table II and extends the recently developed *perturbed learning automata* [7], [9]. At the first step, each agent i selects an action according to a finite probability distribution (i.e., *strategy*) $x_i(t) \in \mathcal{X}_i \doteq \Delta(|\mathcal{A}_i|)$ (capturing its beliefs about the most rewarding action). 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 third step, agent i updates its strategy vector given its new experience. Finally, each agent i updates its discounted

TABLE II
ASPIRATION-BASED PERTURBED LEARNING AUTOMATA (APLA)

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

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

$$\alpha_i(t+1) = \begin{cases} \text{rand}_{x_i(t)}[\mathcal{A}_i], & \text{with probability } 1 - \lambda, \\ \text{rand}_{\text{unif}}[\mathcal{A}_i], & \text{with probability } \lambda, \end{cases} \quad (4)$$

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:

$$\begin{aligned} x_i(t+1) &= x_i(t) + \epsilon \cdot u_i(\alpha(t+1)) \cdot [e_{\alpha_i(t+1)} - x_i(t)] \cdot \\ &\quad \phi(u_i(\alpha(t+1)) - \rho_i(t)) \\ &\doteq \mathcal{R}_i(\alpha(t+1), x_i(t), \rho_i(t)), \end{aligned} \quad (5)$$

for some constant step size $\epsilon > 0$, where

$$\phi(y) \doteq \begin{cases} 1, & \text{if } y \geq 0, \\ \max(h, 1 + y/h), & \text{if } y < 0. \end{cases} \quad (6)$$

for some positive constant $h > 0$.

- 4) (**aspiration-level update**) Agent i revises its aspiration level $\rho_i \in [\underline{\rho}, \bar{\rho}]$ as follows:

$$\begin{aligned} \rho_i(t+1) &= \rho_i(t) + \nu \cdot [u_i(\alpha(t+1)) - \rho_i(t)] \\ &\doteq \mathcal{K}_i(\alpha(t+1), \rho_i(t)). \end{aligned} \quad (7)$$

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

running average performance, namely *aspiration level*, $\rho_i \in [\underline{\rho}, \bar{\rho}]$, for some $\underline{\rho}, \bar{\rho} > 0$.

Here we identify actions \mathcal{A}_i with vertices of the simplex, $\{e_1, \dots, e_{|\mathcal{A}_i|}\}$. For example, if agent i selects its j th action at time t , then $e_{\alpha_i(t)} \equiv e_j$. To better clarify how the strategies evolve, consider the following toy example. Let the current strategy of player i be $x_i(t) = (1/2 \ 1/2)^T$, 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 (4). Then, the new strategy vector for agent i , updated according to rule (5), is:

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

If $y \doteq u_i(\alpha(t+1)) - \rho_i(t) \geq 0$, i.e., player i receives a satisfactory performance, then $\phi(y) = 1$, and the strategy of the selected action is going to increase proportionally to the reward received from this action. If, instead, $y < 0$, then $\phi(y) = \max\{h, 1 + y/h\} < 1$, and the strategy of the selected action is going to increase proportionally to both the reward received and $\phi(y)$. By adjusting $h > 0$, we may control the increase in the strategy of a dissatisfactory action, since, if the current reward is far below the current aspiration level, then $\phi(y) = h$.

The introduction of the level of satisfaction captured by $\phi(\cdot)$ in the strategy update is the main contribution of this paper as compared to the original *perturbed learning automata* (PLA) introduced in [7], [9], where $h \equiv 1$.

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 .

We also deliberately set the step size of the aspiration-level update $\nu > 0$ to be different than the step size of the strategy vector update $\epsilon > 0$. In general, and for reasons that will become more clear in the forthcoming Section V, we would like ν to be sufficiently larger than ϵ in order for the aspiration level to evolve at a faster rate than the strategy vector. **For the remainder of the paper**, we will assume the following design property.

Property 3.1: Given $\epsilon > 0$, we set the step size $\nu = \nu(\epsilon)$ sufficiently small such that, for any $\delta > 0$,

$$\sup_{i \in \mathcal{I}} \frac{\log(\delta / |\bar{\rho} - \underline{\rho}|)}{\log(1 - \nu)} \leq \inf_{\alpha \in \mathcal{A}} \inf_{i \in \mathcal{I}} \frac{\log(\delta)}{\log(1 - \epsilon u_i(\alpha))} \quad (8)$$

The ratio of the l.h.s. represents the minimum number of steps that the aspiration-level update of a player i needs to reach a δ -neighborhood of the utility $u_i(\alpha)$ starting from any other action profile. The ratio of the r.h.s. represents the minimum number of steps that the strategy update of a player i needs to reach a δ -neighborhood of the pure strategy vector corresponding to α , when playing α continuously. It is clear that for a finite number of actions and bounded utilities, and for ν sufficiently larger than ϵ , property (8) will be satisfied.

IV. STOCHASTIC STABILITY

A. Terminology and notation

Let $\mathcal{Z} \doteq \mathcal{A} \times \mathcal{X} \times \mathcal{W}$, where $\mathcal{X} \doteq \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ and $\mathcal{W} \doteq [\underline{\rho}, \bar{\rho}]^n$, i.e., tuples of joint actions α , strategy profiles $x = (x_1, \dots, x_n)$ and aspiration-level profiles $\rho = (\rho_1, \dots, \rho_n)$. 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} and \mathcal{W} with the 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 dynamics of Table II defines an \mathcal{Z} -valued Markov chain. Let $P_\lambda : \mathcal{Z} \times \mathfrak{B}(\mathcal{Z}) \rightarrow [0, 1]$ denote its transition probability function (t.p.f.), parameterized by $\lambda > 0$. We will refer to this process as the *perturbed process*.

Note that under the perturbed process P_λ one or more agents may *tremble* (i.e., select randomly an action according to the uniform distribution). Define also the process $P : \mathcal{Z} \times \mathfrak{B}(\mathcal{Z}) \rightarrow [0, 1]$ where *at most one agent may tremble*. We will refer to this process as the *unperturbed process*.

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

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

The process governed by the unperturbed process P will be denoted by $Z = \{Z_t : t \geq 0\}$. 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 $\omega(t) = (\alpha(t), x(t), \rho(t)) \in \mathcal{Z}$. We use the same notation for the elements (α, x, ρ) of the space \mathcal{Z} and for the coordinates of the process $Z_t = (\alpha(t), x(t), \rho(t))$. 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}^∞ , initialized at $z = (\alpha, x, \rho)$.

B. Stochastic stability

First, note that both P and P_λ ($\lambda, h > 0$) satisfy the *weak-Feller property* (cf., [17, Definition 4.4.2]).

Proposition 4.1: Both the unperturbed process P ($\lambda = 0$) and the perturbed process P_λ ($\lambda > 0$) satisfy the weak-Feller property.

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

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

Since \mathcal{Z} defines a locally compact separable metric space and P, P_λ satisfy the weak-Feller property, they both admit an i.p.m., denoted μ and μ_λ , respectively [17, 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 $\liminf_{\lambda \rightarrow 0} \mu_\lambda(z) > 0$. 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 4.1 (Pure Strategy State): A pure strategy state is a state $s = (\alpha, x, \rho) \in \mathcal{Z}$ such that for all $i \in \mathcal{I}$, $x_i = e_{\alpha_i}$ and $\rho_i = u_i(\alpha)$, i.e., x_i coincides with the vertex of the probability simplex $\Delta(|\mathcal{A}_i|)$ which assigns probability 1 to action α_i , and ρ_i coincides with the utility of agent i under action profile α .

We will denote the set of pure strategy states by \mathcal{S} . Pure strategy states that correspond to pure Nash equilibria, will be referred to as *pure Nash equilibrium states* and will be denoted by \mathcal{S}^* . For any pure strategy state $s^* = (\alpha^*, x^*, \rho^*)$, define the δ -neighborhood of s^* as follows

$$\mathcal{N}_\delta(s^*) \doteq \{z = (\alpha, x, \rho) \in \mathcal{Z} : \alpha = \alpha^*, |x - x^*| < \delta, |\rho - \rho^*| < \delta\}.$$

Theorem 4.1 (Stochastic Stability): In any coordination game (Definition 2.2) and under the aspiration-based perturbed learning automata of Table II (APLA), there exists a unique probability vector $\pi = (\pi_1, \dots, \pi_{|\mathcal{S}^*|})$ such that, for any collection of i.p.m.'s $\{\mu_\lambda \in \mathfrak{P}(\mathcal{Z}) : \mu_\lambda P_\lambda = \mu_\lambda, \lambda > 0\}$,

- $\lim_{\lambda \downarrow 0} \mu_\lambda(\cdot) = \hat{\mu}(\cdot) \doteq \sum_{s \in \mathcal{S}^*} \pi_s \delta_s(\cdot)$, where convergence is in the weak sense.
- The probability vector π is an invariant distribution of the (finite-state) Markov process \hat{P} , such that, for any $s, s' \in \mathcal{S}^*$,

$$\hat{P}_{ss'} \doteq \lim_{t \rightarrow \infty} Q P^t(s, \mathcal{N}_\delta(s')), \quad (9)$$

for any $\delta > 0$ sufficiently small, where Q is the t.p.f. corresponding to *at least two players trembling* (i.e., following the uniform distribution of (4)).

Theorem 4.1 establishes weak convergence of the i.p.m. of P_λ with the invariant distribution of a finite Markov chain \hat{P} , whose support is on the set of pure Nash equilibrium states. Thus, from the ergodicity of μ_λ , we have that the expected percentage of time that the process spends in any $O \in \mathfrak{B}(\mathcal{Z})$ such that $\partial O \cap \mathcal{S}^* \neq \emptyset$ is given by $\hat{\mu}(O)$ as $h, \lambda \downarrow 0$ and time increases, i.e.,

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

The methodology for assessing which Nash equilibria are stochastically stable will follow in the forthcoming Section VI.

V. TECHNICAL DERIVATION

In this section, we provide the main steps for the proof of Theorem 4.1. 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 pure Nash equilibrium states \mathcal{S}^* .

A. Unperturbed Process

Recall that the unperturbed process with t.p.f. P has been defined such that *at most one agent may tremble*. We first present two technical lemmas that will help us identify the behavior of the unperturbed process.

Let $\tau(D)$ denote the first hitting time of the unperturbed process to a set $D \subset \mathfrak{B}(\mathcal{Z})$.

For some action profile $\alpha \in \mathcal{A}$ and $\delta > 0$, define the set:

$$D_\delta(\alpha) \doteq \{z \in \mathcal{Z} : \rho_i \leq u_i(\alpha) + \delta, x_{i\alpha_i} > 0, \forall i \in \mathcal{I}\}.$$

The set $D_\delta(\alpha) \subset \mathfrak{B}(\mathcal{Z})$, corresponds to any state $z \in \mathcal{Z}$ at which the aspiration level is below $u_i(\alpha) + \delta$ for some given $\alpha \in \mathcal{A}$ and $\delta > 0$. Define also the event

$$\Gamma_{\delta, t} \doteq \{\exists \alpha \in \mathcal{A} : \tau(D_\delta(\alpha)) < t\},$$

i.e., the event Γ_t corresponds to the case that the process first reaches $D_\delta(\alpha)$ for some action profile α before time t .

The first lemma states that, for any initial state z , at least one $\Gamma_{\delta, t}$ occurs for some $t > 0$.

Lemma 5.1: For any $\delta > 0$, and any $z = (\alpha, x, \rho) \in \mathcal{Z}$, $\mathbb{P}_z[\Gamma_{\delta, \infty}] = 1$.

The second lemma states that the unperturbed process reaches (infinitely often) a state at which the aspiration level is below the utility level of a Nash equilibrium and its strategy assigns positive probability to it. Define the event:

$$\Gamma_t^* \doteq \{\exists \alpha^* \in \mathcal{A}^* : \tau(D_0(\alpha^*)) < t\},$$

which corresponds to the case that the aspiration level is below the utility level of a Nash equilibrium.

Lemma 5.2: For any $z = (\alpha, x, \rho) \in \mathcal{Z}$, $\mathbb{P}_z[\Gamma_\infty^*] = 1$.

Next, we will use Lemma 5.2 to show that the process will reach a δ -neighborhood of a pure Nash equilibrium infinitely often with probability one.

For any $\delta > 0$, define the event:

$$B_{\delta, t} \doteq \{\exists s^* \in \mathcal{S}^* : \tau(\mathcal{N}_\delta(s^*)) < t\}.$$

In other words, $B_{\delta,t}$ corresponds to the event that the unperturbed process has reached a δ -neighborhood of a pure Nash equilibrium state before time instance t .

Proposition 5.1: For any $\delta > 0$ and any initial state $z = (\alpha, x, \rho) \in \mathcal{Z}$, $\mathbb{P}_z[B_{\delta,\infty}] = 1$.

Proposition 5.1 states that the unperturbed process will reach a δ -neighborhood of a Nash equilibrium infinitely often with probability one. Note that this derivation is independent of the size of h and λ .

Proposition 5.2 (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 $f \in C_b(\mathcal{Z})$, $\lim_{t \rightarrow \infty} \|P^t f - \Pi f\|_\infty = 0$;
- (c) μ is an i.p.m. for Π ;
- (d) the support¹ of Π is on \mathcal{S}^* for all $z \in \mathcal{Z}$.

Proposition 5.2 states that the limiting unperturbed t.p.f. converges weakly to a t.p.f. Π which accepts the same invariant probability measure as P . Furthermore, *the support of Π is the set of pure Nash equilibrium states in \mathcal{S}^** . This is a rather handy property, 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

Note that the t.p.f. of the perturbed process can be decomposed as follows:

$$P_\lambda = (1 - \varphi(\lambda))P + \varphi(\lambda)Q \quad (10)$$

where Q is the t.p.f. of the one-step process where *at least two agents tremble simultaneously*, i.e., they play an action uniformly at random. Note that

$$1 - \varphi(\lambda) = (1 - \lambda)^n + n\lambda(1 - \lambda)^{n-1}$$

is the probability that *at most one agent trembles*. It is straightforward to check that $\varphi(\lambda) \rightarrow 0$ as $\lambda \downarrow 0$.

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

$$P_\lambda^L \doteq \varphi(\lambda) \sum_{t=0}^{\infty} (1 - \varphi(\lambda))^t Q P^t = Q R_\lambda \quad (11)$$

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

- (a) For $f \in C_b(\mathcal{Z})$, $\lim_{\lambda \rightarrow 0} \|R_\lambda f - \Pi f\|_\infty = 0$.
- (b) For $f \in C_b(\mathcal{Z})$, $\lim_{\lambda \rightarrow 0} \|P_\lambda^L f - Q\Pi f\|_\infty = 0$.

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

- (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 $\lambda \rightarrow 0$, is an i.p.m. of $Q\Pi$.

Proposition 5.3 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 (9).

Proposition 5.4 (Unique i.p.m. of $Q\Pi$): There exists a unique i.p.m. $\hat{\mu}$ of $Q\Pi$. It satisfies

$$\hat{\mu}(\cdot) = \sum_{s \in \mathcal{S}^*} \pi_s \delta_s(\cdot) \quad (12)$$

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

D. Proof of Theorem 4.1

Theorem 4.1(a)–(b) is a direct implication of Propositions 5.3–5.4.

VI. CONVERGENCE TO EFFICIENT OUTCOMES

A. Payoff-dominant states

According to Theorem 4.1, the stochastically stable states is a subset of the set of pure Nash equilibrium states. We may further refine the set of stochastically stable states if we introduce additional structural constraints. One such constraint is the existence of a payoff-dominant set of action profiles.

Definition 6.1 (Payoff-dominant states): The set of payoff-dominant states, denoted by $\bar{\mathcal{S}} \subseteq \mathcal{S}^*$, is such that, for any $s^* = (\alpha^*, x^*, \rho^*) \in \bar{\mathcal{S}}$ and any $s = (\alpha, x, \rho) \in \mathcal{S}^* \setminus \bar{\mathcal{S}}$, $u_i(\alpha^*) > u_i(\alpha)$, for all $i \in \mathcal{I}$.

Note that the set of payoff-dominant states may not be empty in a coordination game of Definition 2.2, since the set of pure Nash equilibrium states is not empty. However, it might be the case that $\bar{\mathcal{S}} \equiv \mathcal{S}^*$.

Proposition 6.1: Consider a coordination game of Definition 2.2, and let every agent implement the learning dynamics of Table II with sufficiently small $\epsilon > 0$ such that $\epsilon u_i(\alpha) < 1$ for all i and $\alpha \in \mathcal{A}$. Then,

$$\lim_{h, \lambda \downarrow 0} \mu_\lambda(\bar{\mathcal{S}}) = 1. \quad (13)$$

B. Simulation study

We consider the Stag-Hunt coordination game of Table I where agents implement the APLA dynamics of Table II with $\epsilon = 0.0001$, $\nu = 0.001$, $\lambda = 0.01$ and $h = 0.01$. The response of the aspiration-based reinforcement learning is shown in Figure 2, while the corresponding response under standard perturbed learning automata (i.e., for $h \equiv 1$), is shown in Figure 1. It is evident that under standard perturbed learning automata, the risk-dominant equilibrium, which corresponds to the action profile (B, B) is the stochastically-stable state. On the other hand, Figure 2 demonstrates that

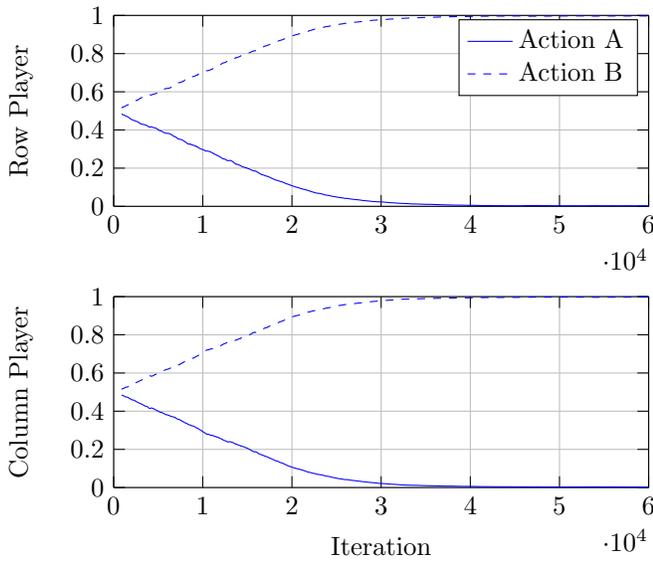


Fig. 1. Response of standard perturbed learning automata (PLA) ($h \equiv 1$).

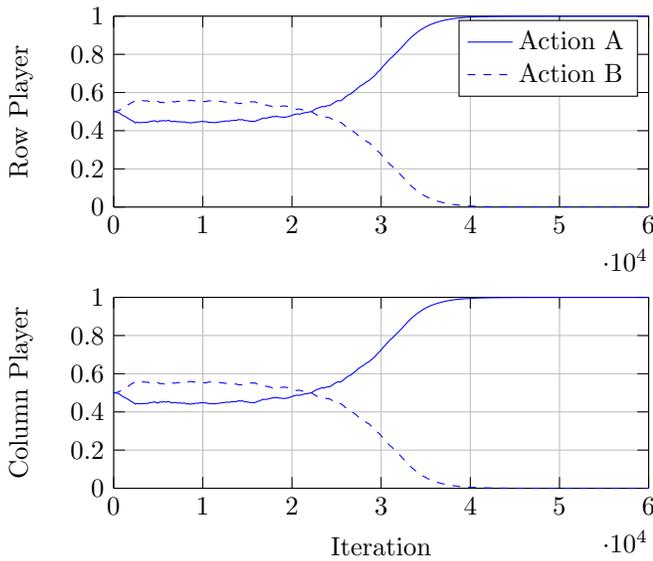


Fig. 2. Response of aspiration-based perturbed learning automata (APLA) (small $h > 0$).

the payoff-dominant equilibrium (which corresponds to the action profile (A, A)) is the stochastically-stable state under APLA, which verifies the conclusion of Proposition 6.1.

VII. CONCLUSIONS

This paper introduced a novel learning automata dynamics, namely *aspiration-based perturbed learning automata* (APLA). Contrary to standard perturbed learning automata, actions are reinforced based on both repeated selection and their satisfaction levels. This modification, that is solely based on local information available to each agent (i.e., its own utility function), is able to alter the stochastic stability properties of standard learning automata in strategic-form games. In particular, the payoff-dominant (pure) Nash equilibrium states become the only stochastically stable states. This overcomes several limitations observed in reinforcement-based learning dynamics (e.g., discrete-time replicator dynamics or learning-automata) where risk-dominant pure Nash equilibria can be the only stochastically stable outcomes.

REFERENCES

- [1] G. C. Chasparis, "Aspiration-based Perturbed Learning Automata," *arXiv:1803.02751 [cs]*, Mar. 2018.
- [2] W. B. Arthur, "On designing economic agents that behave like human agents," *J. Evolutionary Econ.*, vol. 3, pp. 1–22, 1993.
- [3] M. Tsetlin, *Automaton Theory and Modeling of Biological Systems*. Academic Press, 1973.
- [4] K. Narendra and M. Thathachar, *Learning Automata: An introduction*. Prentice-Hall, 1989.
- [5] 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.
- [6] M. Thathachar and P. Sastry, *Networks of Learning Automata: Techniques for Online Stochastic Optimization*. Kluwer Academic Publishers, 2004.
- [7] G. C. 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.
- [8] G. C. Chasparis, A. Arapostathis, and J. Shamma, "Aspiration learning in coordination games," *SIAM J. Control and Optim.*, vol. 51, no. 1, 2013.
- [9] G. C. Chasparis, "Stochastic Stability of Perturbed Learning Automata in Positive-Utility Games," *arXiv:1709.05859 [cs]*, Sept. 2017.
- [10] 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.
- [11] J. C. Harsanyi and R. Selten, *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: MIT Press, 1988.
- [12] 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.
- [13] 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.
- [14] H. P. Young, "Learning by trial and error," *Games and Economic Behavior*, vol. 65, no. 2, pp. 626–643, Mar. 2009.
- [15] E. Hopkins and M. Posch, "Attainability of boundary points under reinforcement learning," *Games Econ. Behav.*, vol. 53, pp. 110–125, 2005.
- [16] H. P. Young, *Strategic Learning and Its Limits*. New York, NY: Oxford University Press, 2004.
- [17] O. Hernandez-Lerma and J. B. Lasserre, *Markov Chains and Invariant Probabilities*. Birkhauser Verlag, 2003.