Perturbed Learning Automata in Potential Games

Chasparis, Georgios; Shamma, Jeff S.; Rantzer, Anders

2011

Perturbed Learning Automata in Potential Games

Georgios C. Chasparis Jeff S. Shamma Anders Rantzer

***Abstract*— This paper presents a reinforcement learning algorithm and provides conditions for global convergence to Nash equilibria. For several reinforcement learning schemes, including the ones proposed here, excluding convergence to action profiles which are not Nash equilibria may not be trivial, unless the step-size sequence is appropriately tailored to the specifics of the game. In this paper, we sidestep these issues by introducing a new class of reinforcement learning schemes where the strategy of each agent is perturbed by a state-dependent perturbation function. Contrary to prior work on equilibrium selection in games, where perturbation functions are globally state dependent, the perturbation function here is assumed to be local, i.e., it only depends on the strategy of each agent. We provide conditions under which the strategies of the agents will converge to an arbitrarily small neighborhood of the set of Nash equilibria almost surely. We further specialize the results to a class of potential games.**

I. Introduction

Lately, agent-based modeling has generated significant interest in various settings, such as engineering, social sciences and economics. In those formulations, agents make decisions independently and without knowledge of the actions or intentions of the other agents. Usually, the interactions among agents can be described in terms of a strategic-form game, and stability notions, such as the Nash equilibrium, can be utilized to describe desirable outcomes for all agents.

In this paper, we are interested in deriving conditions under which agents *learn* to play Nash equilibria. Assuming minimum information available to each agent, namely its *own* utilities and actions, we introduce a novel reinforcement learning scheme and derive conditions under which global convergence to Nash equilibria can be achieved.

Prior results in reinforcement learning has primarily focused on *common-payoff* games [1]. In reference [2], a reinforcement learning scheme is introduced and convergence to the set of Nash equilibria is shown when applied to a class of potential games. However, although the analysis is based on weak-convergence arguments (due to a constant step-size selection), an explicit characterization of the limiting invariant distribution is not provided, while the issue of non-convergence to unstable points on the boundary of the domain has been overlooked. In fact, as pointed out in [3], establishing non-convergence to the boundary of the probability simplex might not be trivial, since standard results of the ODE method for stochastic approximations (e.g., non-convergence to unstable equilibria [4]) are not applicable.

In this paper, we sidestep these issues by introducing a variation of reinforcement learning algorithms where the strategy of each agent is perturbed by a state-dependent perturbation function. Contrary to prior work on equilibrium selection, where perturbation functions are also state dependent [5], the perturbation function here is assumed to be *local*, i.e., it only depends on the strategy of each agent. Due to this perturbation function, the ODE method for stochastic approximations can be applied, since boundary points of the domain cease to be stationary points of the relevant ODE. This paper extends prior work [6] of the authors, where the perturbation function was assumed constant along the domain. In particular, we provide conditions under which the strategies of the agents will converge to an arbitrarily small neighborhood of the set of Nash equilibria almost surely. We further specialize the results to a class of games which belong to the family of *potential games* [7].

The remainder of the paper is organized as follows. Section II introduces the necessary terminology. Section III introduces the perturbed reinforcement learning scheme with a state-based perturbation function. Section IV states some standard results for analyzing stochastic approximations. Section V characterizes the set of stationary points for both the unperturbed and the perturbed learning scheme. Section VI discusses convergence properties of the unperturbed reinforcement learning scheme, while Section VII presents conditions under which the perturbed learning scheme converges to the set of Nash equilibria. Section VIII specializes the convergence analysis to a class of potential games. Finally, Section IX presents concluding remarks.

Notation:

- $|x|$ denotes the Euclidean norm of a vector $x \in \mathbb{R}^n$.
- $|x|_\infty$ denotes the ℓ_∞-norm of a vector $x \in \mathbb{R}^n$.
- $\mathcal{B}_\delta(x)$ denotes the δ-neighborhood of a vector $x \in \mathbb{R}^n$, i.e., $\mathcal{B}_\delta(x) \triangleq \{y \in \mathbb{R}^n : |x-y| < \delta\}$.
- $\mathrm{dist}(x, A)$ denotes the distance from a point x to a set A, i.e., $\mathrm{dist}(x, A) \triangleq \inf_{y \in A} |x-y|$.
- $\mathcal{B}_\delta(A)$ denotes the δ-neighborhood of the set $A \subset \mathbb{R}^n$, i.e., $\mathcal{B}_\delta(A) \triangleq \{x \in \mathbb{R}^n : \mathrm{dist}(x, A) < \delta\}$.
- $\Delta(m)$ denotes the probability simplex of dimension m, i.e., $\Delta(m) \triangleq \{x \in \mathbb{R}^m : x \geq 0, \mathbf{1}^{\mathrm{T}} x = 1\}$.
- $\Pi_\Delta : \mathbb{R}^m \to \Delta(m)$ is the projection to the probability simplex, i.e., $\Pi_\Delta[x] \triangleq \arg\min_{y \in \Delta(m)} |x-y|$.
- A^o is the interior of $A \subset \mathbb{R}^n$, and ∂A is its boundary.
- $\mathrm{col}\{\alpha_i\}_{i \in \mathcal{J}}$ is the column vector with entries $\{\alpha_i\}_{i \in \mathcal{J}}$

This work was supported by the Swedish Research Council through the Linnaeus Center LCCC and the AFOSR MURI project #FA9550-09-1-0538.

G. Chasparis is with the Department of Automatic Control, Lund University, 221 00-SE Lund, Sweden; E-mail: georgios.chasparis@control.lth.se; URL: http://www.control.lth.se/chasparis.

J. Shamma is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332. E-mail: shamma@gatech.edu. URL: http://www.prism.gatech.edu/˜jshamma3.

A. Rantzer is with the Department of Automatic Control, Lund University, 221 00-SE Lund, Sweden; E-mail: anders.rantzer@control.lth.se.

for some set of indexes $\mathcal{J}$.

II. Terminology

We consider the standard setup of finite strategic-form games.

1) Game: A finite strategic-form game involves a finite set of *agents* (or *players*), $\mathcal{I} \triangleq \{1, 2, ..., n\}$. Each agent $i \in \mathcal{I}$ has a *finite* set of available *actions*, $\mathcal{A}_i$. Let $\alpha_i \in \mathcal{A}_i$ be an action of agent i, and $\alpha = (\alpha_1, ..., \alpha_n)$ the *action profile* of all agents. The set $\mathcal{A}$ is the Cartesian product of the action spaces of all agents, i.e., $\mathcal{A} \triangleq \mathcal{A}_1 \times ... \times \mathcal{A}_n$.

The action profile $\alpha \in \mathcal{A}$ produces a *payoff* (or *utility*) for each agent. The utility of agent i, denoted by R_i, is a function which maps the action profile α to a payoff in $\mathbb{R}$. Denote $R : \mathcal{A} \to \mathbb{R}^n$ the combination of payoffs (or *payoff profile*) of all agents, i.e., $R(\cdot) \triangleq (R_1(\cdot), ..., R_n(\cdot))$.

2) Strategy: Let $\sigma_{ij} \in [0, 1]$ denote the probability that agent i selects action $\alpha_i = j$. Then, $\sigma_i \triangleq (\sigma_{i1}, ..., \sigma_{i|\mathcal{A}_i|})$ is a probability distribution over the set of actions $\mathcal{A}_i$ (or *strategy* of agent i), i.e., $\sigma_i \in \Delta(|\mathcal{A}_i|)$, where $|\mathcal{A}_i|$ denotes the cardinality of the set $\mathcal{A}_i$. We will also use the term *strategy profile* to denote the combination of strategies of all agents $\sigma = (\sigma_1, \sigma_2, ..., \sigma_n) \in \mathbf{\Delta}$ where $\mathbf{\Delta} \triangleq \Delta(|\mathcal{A}_1|) \times ... \times \Delta(|\mathcal{A}_n|)$ is the set of strategy profiles.

Note that if σ_i is a *unit vector* (or a vertex of $\Delta(|\mathcal{A}_i|)$), say e_j, then agent i selects an action j with probability one. Such a strategy will be called *pure strategy*. Likewise, a *pure strategy profile* is a profile of pure strategies. Denote $\mathbf{\Delta}^*$ the set of pure strategy profiles or *vertices* of $\mathbf{\Delta}$. We will use the term *mixed strategy* to define a strategy that is *not* pure.

3) Expected payoff and Nash equilibrium: Given a strategy profile $\sigma \in \mathbf{\Delta}$, the expected payoff *vector* of each agent i, $U_i : \mathbf{\Delta} \to \mathbb{R}^{|\mathcal{A}_i|}$, can be computed by[1]

$$U_i(\sigma) \triangleq \sum_{j \in \mathcal{A}_i} e_j \sum_{\alpha_{-i} \in \mathcal{A}_{-i}} \left(\prod_{s \in -i} \sigma_{s\alpha_s} \right) R_i(j, \alpha_{-i}). \quad (1)$$

We may think of the entry j of the expected payoff vector, denoted $U_{ij}(\sigma)$, as the payoff of agent i who is playing action j at the strategy profile σ. We denote the profile of expected payoff vectors as $U(\sigma) = (U_1(\sigma), ..., U_n(\sigma))$. Finally, let $u_i(\sigma)$ be the expected payoff of agent i at strategy profile $\sigma \in \mathbf{\Delta}$, defined as $u_i(\sigma) \triangleq \sigma_i^{\mathrm{T}} U_i(\sigma)$.

Definition 2.1 (Nash equilibrium): A strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*, ..., \sigma_n^*) \in \mathbf{\Delta}$ is a Nash equilibrium if, for each agent $i \in \mathcal{I}$,

$$u_i(\sigma_i^*, \sigma_{-i}^*) \geq u_i(\sigma_i, \sigma_{-i}^*) \quad (2)$$

for all $\sigma_i \in \Delta(|\mathcal{A}_i|)$ such that $\sigma_i \neq \sigma_i^*$.

In the special case where for all $i \in \mathcal{I}$, σ_i^* is a pure strategy, $\sigma^* \in \mathbf{\Delta}^*$ is called a *pure Nash equilibrium*. Any Nash equilibrium which is *not* pure is called a *mixed Nash equilibrium*. Also, in case the inequality in (2) is strict, the Nash equilibrium is called a *strict Nash equilibrium*.

[1]The notation $-i$ denotes the complementary set $\mathcal{I} \backslash \{i\}$. We will often split the argument of a function in this way, e.g., $F(\alpha) = F(\alpha_i, \alpha_{-i})$.

III. Perturbed Learning Automata

In this section, we introduce the basic form of the learning dynamics that we will consider in the remainder of the paper. They belong to the general class of *learning automata* [1].

For the remainder of the paper, we will assume:

Assumption 3.1 (Strictly positive payoffs): For every $i \in \mathcal{I}$, the utility function satisfies $R_i(\alpha) > 0$ for all $\alpha \in \mathcal{A}$.

Even in the case where utilities take on negative values, we can still analyze the game by considering an *equivalent* one with strictly positive payoffs (cf., [8]).

A. Modified Linear Reward-Inaction ($\tilde{\mathcal{L}}_{R-I}$) scheme

We consider a reinforcement scheme which is a small modification of the original *linear reward-inaction* ($\mathcal{L}_{R-I}$) scheme [9], [10]. This modified scheme, denoted by $\tilde{\mathcal{L}}_{R-I}$, was introduced in [6]. Contrary to $\mathcal{L}_{R-I}$, $R_i(\cdot)$ may take values other than $\{0, 1\}$, which increases the family of games this algorithm can be applied to.

Similarly to $\mathcal{L}_{R-I}$, the probability that agent i selects action j at time k is $\sigma_{ij}(k) = x_{ij}(k)$, for some probability vector $x_i(k)$ which is updated according to the recursion:

$$x_i(k+1) = \Pi_\Delta \left[x_i(k) + \epsilon(k) \cdot R_i(\alpha(k)) \cdot [\alpha_i(k) - x_i(k)] \right]. \quad (3)$$

Here we identify actions $\mathcal{A}_i$ with vertices of the simplex, $\{e_1, e_2, ..., e_{|\mathcal{A}_i|}\}$. For example, if agent i selects action j at time k, then $\alpha_i(k) = e_j$. Note that by letting the step-size sequence $\epsilon(k)$ to be sufficiently small and since the payoff function R_i is uniformly bounded in $\mathcal{A}$, $x_i(k) \in \Delta(|\mathcal{A}_i|)$ and the projection operator Π_Δ can be omitted.

We consider the following class of step-size sequences:

$$\epsilon(k) = \frac{1}{k^\nu + 1} \quad (4)$$

for some $\nu \in (1/2, 1]$. For these values of ν, the following two conditions can easily be verified:

$$\sum_{k=0}^{\infty} \epsilon(k) = \infty \quad \text{and} \quad \sum_{k=0}^{\infty} \epsilon(k)^2 < \infty. \quad (5)$$

The selection of ν is closely related to the desired rate of convergence. Compared with prior reinforcement learning schemes, both [11] and [3] consider comparable step-size sequences.

B. Pertubed Linear Reward-Inaction Scheme ($\tilde{\mathcal{L}}_{R-I}^{\lambda}$)

Here we consider a perturbed version of the scheme $\tilde{\mathcal{L}}_{R-I}$, in the same spirit with [6], where the decision probabilities of each agent are slightly perturbed. In particular, we assume that each agent i selects action $j \in \mathcal{A}_i$ with probability

$$\sigma_{ij} \triangleq (1 - \zeta_i(x_i, \lambda)) x_{ij} + \zeta_i(x_i, \lambda) / |\mathcal{A}_i|, \quad (6)$$

for some perturbation function $\zeta_i : \Delta(|\mathcal{A}_i|) \times [0, 1] \to [0, 1]$, where the probability vector x_i is updated according to (3).

We consider the following perturbation function:

$$\zeta_i(x_i, \lambda) = \begin{cases} 0 & |x_i|_\infty < \beta, \\ \frac{\lambda}{(1-\beta)^2} (|x_i|_\infty - \beta)^2 & |x_i|_\infty \geq \beta, \end{cases} \quad (7)$$

for some $\beta \in (0,1)$ which is close to one. In other words, an agent perturbs its strategy when the latter is close to a vertex of the probability simplex. Note that the perturbation function is continuously differentiable for some β sufficiently close to one. Furthermore, $\lim_{\lambda \downarrow 0} \zeta_i(x_i, \lambda) = 0$ uniformly in x, which establishes equivalence of the perturbed dynamics with the unperturbed dynamics as λ approaches zero.

The main difference with earlier work by the same authors [6] is that here we allow for the perturbation function to also depend on agent's *own* strategy. Similar ideas of state dependent perturbations have been utilized for equilibrium selection in adaptive learning by [5]. The difference here is that the perturbation function is *locally* state dependent, i.e., it only depends on the strategy of each agent and *not* on the strategy profile of all agents.

We will denote this scheme by $\tilde{\mathcal{L}}^{\lambda}_{R-I}$.

IV. Background Convergence Analysis

Let $\Omega \triangleq \mathbf{\Delta}^{\infty}$ denote the canonical path space with an element ω being a sequence $\{x(0), x(1), ...\}$, where $x(k) \triangleq (x_1(k), ..., x_n(k)) \in \mathbf{\Delta}$ is generated by the reinforcement learning process. An example of a random variable defined in Ω is the function $\psi_k : \Omega \to \mathbf{\Delta}$ such that $\psi_k(\omega) = x(k)$. In several cases, we will abuse notation by simply writing $x(k)$ or $\alpha(k)$ instead of $\psi_k(\omega)$. Let also $\mathcal{F}$ be a σ-algebra of subsets in Ω and $\mathbb{P}$, $\mathbb{E}$ be the probability and expectation operator on $(\Omega, \mathcal{F})$, respectively. In the following analysis, we implicitly assume that the σ-algebra $\mathcal{F}$ is generated appropriately to allow computation of the probabilities or expectations of interest.

A. Exit of a sample function from a domain

It is important to have conditions under which the process $\psi_k(\omega) = x(k)$, $k \geq 0$, with some initial distribution, will exit an open domain G in finite time.

Proposition 4.1 (Theorem 5.1 in [12]): Suppose there exists a nonnegative function, $V(k,x)$ in the domain $k \geq 0$, $x \in G$, such that

$$\Delta V(k,x) \triangleq \mathbb{E}[V(k+1, x(k+1)) - V(k, x(k)) | x(k) = x]$$

satisfies $\Delta V(k,x) \leq -a(k)$ in this domain, where $a(k)$ is a sequence such that

$$a(k) > 0, \quad \sum_{k=0}^{\infty} a(k) = \infty. \tag{8}$$

Then the process $x(k)$ leaves G in a finite time with probability one.

The following corollary is important in cases we would like to consider entrance of a stochastic process into the domain of attraction of an equilibrium. It is a direct consequence of Proposition 4.1. For details, see Exercise 5.1 in [12].

Corollary 4.1: Let $A \subset \mathbf{\Delta}$, $\mathcal{B}_{\delta}(A)$ its δ-neighborhood, and $\mathcal{D}_{\delta}(A) = \mathbf{\Delta} \backslash \mathcal{B}_{\delta}(A)$. Suppose there exists a nonnegative function $V(k,x)$ in the domain $k \geq 0$, $x \in \mathbf{\Delta}$ for which

$$\Delta V(k,x) \leq -a(k)\varphi(k,x), \quad k \geq 0, x \in \mathbf{\Delta}, \tag{9}$$

where the sequence $a(k)$ satisfies (8) and $\varphi(k,x)$ satisfies

$$\inf_{k \geq T, x \in \mathcal{D}_{\delta}(A)} \varphi(k,x) > 0$$

for all $\delta > 0$ and some $T = T(\delta)$. Then

$$\mathbb{P}[\liminf_{k \to \infty} \operatorname{dist}(x(k), A) = 0] = 1.$$

Corollary 4.1 implies that $x(k)$ enters an arbitrarily small neighborhood of a set A infinitely often with probability one.

B. Convergence to mean-field dynamics

The convergence properties of the reinforcement learning schemes can be described via the ODE method for stochastic approximations. The recursion of $\tilde{\mathcal{L}}^{\lambda}_{R-I}$, $\lambda \geq 0$, can be written in the following form:

$$x_i(k+1) = x_i(k) + \epsilon(k) \cdot [\overline{g}^{\lambda}_i(x(k)) + \xi^{\lambda}_i(k)], \tag{10}$$

where the observation sequence has been decomposed into a deterministic sequence, $\overline{g}^{\lambda}_i(x(k))$, (or *mean-field*) and a noise sequence $\xi^{\lambda}_i(k)$. The mean-field is defined as follows:

$$\overline{g}^{\lambda}_i(x) \triangleq \mathbb{E}[R_i(\alpha(k))[\alpha_i(k) - x_i(k)] | x(k) = x]$$

such that its s-th entry is

$$\overline{g}^{\lambda}_{is}(x) = U_{is}(x)\sigma_{is} - \sum_{q \in \mathcal{A}_i} U_{iq}(x)\sigma_{iq}x_{is}.$$

where σ_{iq}, $q \in \mathcal{A}_i$, is defined in (6). It is straightforward to verify that $\overline{g}^{\lambda}_i(\cdot)$ is continuously differentiable. The noise sequence is defined as

$$\xi^{\lambda}_i(k) \triangleq R_i(\alpha(k)) \cdot [\alpha_i(k) - x_i(k)] - \overline{g}^{\lambda}_i(x(k)),$$

where $\mathbb{E}[\xi^{\lambda}_i(k) | x(k) = x] = 0$ for all $x \in \mathbf{\Delta}$.

Note that for $\lambda = 0$, (10) coincides with $\tilde{\mathcal{L}}_{R-I}$. We will denote $\overline{g}(x)$ the corresponding vector field for $\lambda = 0$.

The more compact form of (10) will also be used:

$$x(k+1) = x(k) + \epsilon(k) \cdot [\overline{g}^{\lambda}(x(k)) + \xi^{\lambda}(k)], \tag{11}$$

where $\overline{g}^{\lambda}(\cdot) \triangleq \operatorname{col}\{\overline{g}^{\lambda}_i(\cdot)\}_{i \in \mathcal{I}}$ and $\xi^{\lambda}(\cdot) \triangleq \operatorname{col}\{\xi^{\lambda}_i(\cdot)\}_{i \in \mathcal{I}}$.

Proposition 4.2 (Theorem 6.6.1 in [13]): For the reinforcement scheme $\tilde{\mathcal{L}}^{\lambda}_{R-I}$, $\lambda \geq 0$, the stochastic iteration (11) is such that, for almost all $\omega \in \Omega$, $\{\psi_k(\omega) = x(k)\}$ converges to some invariant set of the ODE

$$\dot{x} = \overline{g}^{\lambda}(x). \tag{12}$$

Also, if $A \subset \mathbf{\Delta}$ is a locally asymptotically stable set in the sense of Lyapunov for (12),[2] and $x(k)$ is in some compact set in the domain of attraction of A infinitely often with probability $\geq \rho$, then $\mathbb{P}[\lim_{k \to \infty} x(k) \in A] \geq \rho$.

Proof: The proposition follows directly from Theorem 6.6.1 of [13], since the following conditions are satisfied:

– The function $\overline{g}^{\lambda}(\cdot)$ is continuous.

[2]If $\{x(t) : t \geq 0\}$ denotes the solution of the ODE (12), then a set $A \subset \mathbf{\Delta}$ is locally asymptotically stable set in the sense of Lyapunov for the ODE (12) if there exists $\delta > 0$ such that $\operatorname{dist}(x(0), A) < \delta$ implies $\lim_{t \to \infty} x(t) \in A$.

– The sequence $Y^\lambda(k) \triangleq \overline{g}^\lambda(x(k)) + \xi^\lambda(k)$ satisfies $\sup_k \mathbb{E}[|Y^\lambda(k)|^2] < \infty$ since, by Assumption 3.1, the utility functions are positive and bounded from above.
– The step-size sequence satisfies property (5).

■

V. Stationary Points

The stationary points of the mean-field dynamics are defined as the set of points $x \in \mathbf{\Delta}$ for which $\overline{g}^\lambda(x) = 0$. In this section, we characterize the set of stationary points for both the *unperturbed* ($\lambda = 0$) and the *perturbed* dynamics ($\lambda > 0$).

We will make the following distinction among stationary points of (12) for $\lambda > 0$, denoted $\mathcal{S}^\lambda$:

– $\mathcal{S}^\lambda_{\partial\mathbf{\Delta}}$: stationary points in $\partial\mathbf{\Delta}$;
– $\mathcal{S}^\lambda_{\mathbf{\Delta}^*}$: stationary points which are vertices of $\mathbf{\Delta}$;
– $\mathcal{S}^\lambda_{\mathbf{\Delta}^o}$: stationary points in $\mathbf{\Delta}^o$;
– $\mathcal{S}^\lambda_{\rm NE}$: stationary points which are Nash equilibria.

We will also use the notation $\mathcal{S}_{\partial\mathbf{\Delta}}$, $\mathcal{S}_{\mathbf{\Delta}^*}$ $\mathcal{S}_{\mathbf{\Delta}^o}$, and $\mathcal{S}_{\rm NE}$ to denote the corresponding sets when $\lambda = 0$.

A. Stationary points of unperturbed dynamics ($\lambda = 0$)

Proposition 5.1 (Stationary points for $\lambda = 0$): A strategy profile x^* is a stationary point of the ODE (12) if and only if, for every agent $i \in \mathcal{I}$, there exists a constant $c_i > 0$, such that for any action $j \in \mathcal{A}_i$, $x^*_{ij} > 0$ implies $U_{ij}(x^*) = c_i$.

Proof: See Proposition 3.3 in [6]. ■

The above result is quite well known for replicator learning dynamics. In fact, notice that the corresponding mean-field of the *share* of strategy s in agent i when $\lambda = 0$ is:

$$\overline{g}_{is}(x) = \left(U_{is}(x) - \sum_{q \in \mathcal{A}_i} U_{iq}(x) x_{iq} \right) x_{is} \qquad (13)$$

which coincides with the corresponding shares provided by the replicator dynamics (e.g., see equation (3.3) in [14]).

Two straightforward implications of Proposition 5.1 are:

Corollary 5.1 (Pure Strategies): For $\lambda = 0$, any pure strategy profile is a stationary point of the ODE (12).

Proof: According to Proposition 5.1 and for $\lambda = 0$, any strategy profile $x^* = (x^*_1, \cdots, x^*_n)$, such that x^*_i is a vertex of the probability simplex (pure strategy), is a stationary point of the ODE (12), since the support of a pure strategy is a single action. ■

Corollary 5.2 (Nash Equilibria): For $\lambda = 0$, any Nash equilibrium is a stationary point of the ODE (12).

Proof: Let σ^* be a (possibly mixed) Nash equilibrium. Then, for any $i \in \mathcal{I}$ and any $j \in \mathcal{A}_i$ such that $\sigma^*_{ij} > 0$, we should have

$$j \in \arg\max_{q \in \mathcal{A}_i} U_{iq}(\sigma^*).$$

Therefore, by Proposition 5.1, the conclusion follows. ■

Note that for some games not all stationary points of the ODE (12) are Nash equilibria. For example, if you consider the Typewriter Game of Table I, the pure strategy profiles which correspond to (A, B) or (B, A) are not Nash equilibria, although they are stationary points of (12).

	A	B
A	$4, 4$	$2, 2$
B	$2, 2$	$3, 3$

TABLE I
THE TYPEWRITER GAME.

On the other hand, any stationary point in the interior of $\mathbf{\Delta}$ will necessarily be a Nash equilibrium.

Corollary 5.3 (Mixed Nash equilibria): For $\lambda = 0$, any stationary point x^* of the ODE (12), such that $x^* \in \mathbf{\Delta}^o$, is a (mixed) Nash equilibrium of the game.

Proof: If $x^* \in \mathbf{\Delta}^o$ is a stationary point of the mean-field dynamics then, as Proposition 5.1 showed, for any agent i and for any pure strategy $j \in \mathcal{A}_i$, we have $U_{ij}(x^*) = c_i$, for some $c_i > 0$. Therefore, all pure strategies are best replies to the strategy x^*. Thus, x^* is also a Nash equilibrium. ■

Note that the above corollaries do not exclude the possibility that there exist stationary points in $\partial\mathbf{\Delta}$ without those necessarily being pure strategy profiles. ***For the remainder of the paper***, we will only consider games which satisfy:

Property 5.1: For the unperturbed dynamics, there are no stationary points in $\partial\mathbf{\Delta}$ other than the ones in $\mathbf{\Delta}^*$, i.e., $\mathcal{S}_{\partial\mathbf{\Delta}} \backslash \mathcal{S}_{\mathbf{\Delta}^*} = \varnothing$. Moreover, there exists $\delta > 0$ such that $\mathcal{B}_\delta(\mathcal{S}_{\mathbf{\Delta}^o}) \subset \mathbf{\Delta}^o$.

In other words, we only consider games for which, the stationary points of (12), $\lambda = 0$, in the boundary of $\mathbf{\Delta}$ are vertices of $\mathbf{\Delta}$, and the stationary points in $\mathbf{\Delta}^o$ are isolated from the boundary. Property 5.1 is not restrictive and is satisfied for most but trivial cases.

B. Stationary points of perturbed dynamics ($\lambda > 0$)

A straightforward implication of the properties of the perturbation function is the following:

Lemma 5.1 (Sensitivity of $S_{\mathbf{\Delta}^o}$): There exists $\beta_0 \in (0, 1)$ such that $\mathcal{S}_{\mathbf{\Delta}^o} \subseteq \mathcal{S}^\lambda_{\mathbf{\Delta}^o}$ for any $\beta_0 < \beta < 1$ and any $\lambda > 0$.

Proof: Due to Property 5.1, there exist $\beta_0 \in (0, 1)$ sufficiently close to one and $\delta > 0$, such that, for any $\beta_0 < \beta < 1$, we have $\zeta_i(x_i, \lambda) = 0$ for all $i \in \mathcal{I}$ and $x \in \mathcal{B}_\delta(\mathcal{S}_{\mathbf{\Delta}^o})$. Thus, the conclusion follows. ■

Vertices of $\mathbf{\Delta}$ cease to be equilibria for $\lambda > 0$. The following proposition provides the sensitivity of $\mathcal{S}_{\mathbf{\Delta}^*}$ to small values of λ.

Lemma 5.2 (Sensitivity of $S_{\mathbf{\Delta}^*}$): For any stationary point $x^* \in \mathcal{S}_{\mathbf{\Delta}^*}$, which corresponds to a strict Nash equilibrium and for sufficiently small $\lambda > 0$, there exists a unique continuously differentiable function $\nu^* : \mathbb{R}_+ \to \mathbb{R}^{|\mathcal{A}|}$, such that $\lim_{\lambda \downarrow 0} \nu^*(\lambda) = \nu^*(0) = 0$, and

$$\tilde{x} = x^* + \nu^*(\lambda) \in \mathbf{\Delta}^o \qquad (14)$$

is a stationary point of the ODE (12). If instead $x^* \in \mathcal{S}_{\mathbf{\Delta}^*}$ is not a Nash equilibrium, then for any sufficiently small $\delta > 0$ and $\lambda > 0$, the δ-neighborhood of x^* in $\mathbf{\Delta}$, $\mathcal{B}_\delta(x^*)$, does not contain any stationary point of the ODE (12).

Proof: The proof follows similar reasoning with the proof of Proposition 3.5 in [6]. ■

Note that the statements of Lemma 5.2 do not depend on the selection of β. Instead, they require λ to be sufficiently small. Also, note that Lemma 5.2 does not discuss the sensitivity of Nash equilibria which are *not* strict. However, it is straightforward to show that vertices *cannot* be stationary points for $\lambda > 0$.

Let also $\tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$ denote the set of stationary points in $\mathbf{\Delta}^o$ which are perturbations of the stationary points in $\mathcal{S}_{\mathbf{\Delta}^*} \cap \mathcal{S}_{\mathrm{NE}}$ (*strict* or *non-strict*) for some $\lambda > 0$.

Proposition 5.2 (Stationary points of perturbed dynamics): For any $\beta \in (0,1)$, let $\delta^* = \delta^*(\beta)$ be the smallest $\delta > 0$ such that, for all $x \in \mathbf{\Delta} \backslash \mathcal{B}_{\delta}(\mathbf{\Delta}^*)$, $\zeta_i(x_i, \lambda) = 0$ for some $i \in \mathcal{I}$. When β is sufficiently close to one and $\lambda > 0$ is sufficiently small, then: a) $\tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}} \subset \mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, and b) $\mathcal{S}^{\lambda} = \mathcal{S}_{\mathbf{\Delta}^o} \cup \tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$.

In other words, the stationary points of the perturbed dynamics are either the interior stationary points of the unperturbed dynamics or perturbations of pure Nash equilibria.

Proof: When we take $\beta > \beta_0$, where β_0 is defined in Lemma 5.1, then $\mathcal{S}_{\mathbf{\Delta}^o} \subseteq \mathcal{S}^{\lambda}_{\mathbf{\Delta}^o} \equiv \mathcal{S}^{\lambda}$. The rest of the stationary points are perturbations of the vertices characterized by Lemma 5.2. Due to the definition of $\delta^* = \delta^*(\beta)$, we have $\tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}} \subset \mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, since outside $\mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$ the dynamics coincide with the unperturbed dynamics for at least one agent. When we further take β to be sufficiently close to one (which implies that $\delta^* = \delta^*(\beta)$ approaches zero) and λ sufficiently small, then, according to Lemma 5.2, $\tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$ are the only stationary points in $\mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, and therefore $\mathcal{S}^{\lambda} = \mathcal{S}_{\mathbf{\Delta}^o} \cup \tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$. ∎

VI. Convergence to Boundary Points

Recall that, for the unperturbed dynamics, not all stationary points in $\mathbf{\Delta}^*$ are necessarily Nash equilibria. Convergence to non-desirable stationary points, such as the ones which are not Nash equilibria, cannot be excluded when agents employ the unperturbed reinforcement scheme $\tilde{\mathcal{L}}_{R-I}$.

Proposition 6.1 (Convergence to boundary points): If agents employ the reinforcement scheme $\tilde{\mathcal{L}}_{R-I}$, the probability that the same action profile will be played for all future times is uniformly bounded away from zero over all initial conditions if $R_i(\alpha) > 1$ for each $\alpha \in \mathcal{A}$, $i \in \mathcal{I}$.

Proof: Assume that agents play the action profile $\alpha = (\alpha_1, \alpha_2, ...\alpha_n) \in \mathcal{A}$ at time $k = 0$. Then $x_{i\alpha_i}(0) > 0$ for all $i \in \mathcal{I}$, since actions are selected according to the probability distribution $\sigma_i(0) = x_i(0)$. Define the following event:

$$A_{\tau} \triangleq \{\omega \in \Omega : \psi_k(\omega) = \alpha(k) = \alpha \text{ for all } k \leq \tau\}.$$

Thus, A_{τ} corresponds to the case where the same action profile has been performed for all times $k \leq \tau$. Note that the sequence of events $\{A_{\tau}\}$ is decreasing, since

$$A_{\tau} \supseteq A_{\tau+1}$$

for all $\tau = 1, 2, ...$ Define also the event

$$A_{\infty} \triangleq \bigcap_{\tau=1}^{\infty} A_{\tau} \equiv \{\alpha(\tau) = \alpha, \forall \tau > 0\}.$$

From continuity from above, we have:

$$\mathbb{P}[A_{\infty}] = \lim_{\tau \to \infty} \mathbb{P}[A_{\tau}] = \lim_{\tau \to \infty} \prod_{k=0}^{\tau} \prod_{i \in \mathcal{I}} x_{i\alpha_i}(k).$$

The above product is non-zero if and only if

$$\sum_{k=0}^{\infty} \log(x_{i\alpha_i}(k)) > -\infty \text{ for each } i \in \mathcal{I}. \tag{15}$$

Let us define the new variable

$$y_i(k) \triangleq 1 - x_{i\alpha_i}(k),$$

which corresponds to the probability of agent i selecting any action other than α_i. Condition (15) is equivalent to

$$-\sum_{k=0}^{\infty} \log(1 - y_i(k)) < \infty, \quad \text{for each } i \in \mathcal{I}. \tag{16}$$

We also have that

$$\lim_{k \to \infty} \frac{-\log(1 - y_i(k))}{y_i(k)} = \lim_{k \to \infty} \frac{1}{1 - y_i(k)} > \rho$$

for some finite $\rho > 0$, since $0 \leq y_i(k) \leq 1$. Thus, from the limit comparison test, we conclude that condition (16) holds if and only if

$$\sum_{k=0}^{\infty} y_i(k) < \infty, \quad \text{for each } i \in \mathcal{I}.$$

Since $\epsilon(k) = 1/(k^{\nu} + 1)$, for $1/2 < \nu \leq 1$, we also have:

$$\frac{y_i(k+1)}{y_i(k)} = 1 - \frac{R_i(\alpha)}{k^{\nu} + 1} \leq 1 - \frac{R_i(\alpha)}{k+1}.$$

By Raabe's criterion, the series $\sum_{k=0}^{\infty} y_i(k)$ is convergent if

$$\lim_{k \to \infty} k \left(\frac{y_i(k)}{y_i(k+1)} - 1 \right) > 1.$$

Since

$$k \left(\frac{y_i(k)}{y_i(k+1)} - 1 \right) \geq k \left(\frac{1}{1 - \frac{R_i(\alpha)}{k+1}} - 1 \right) = \frac{R_i(\alpha)}{1 + \frac{1 - R_i(\alpha)}{k}}$$

we conclude that the series $\sum_{k=0}^{\infty} y_i(k)$ is convergent if $R_i(\alpha) > 1$ for each $i \in \mathcal{I}$. In other words, the action profile α will be performed for all future times with positive probability if $R_i(\alpha) > 1$ for all $i \in \mathcal{I}$. Furthermore, if $R_i(\alpha) > 1$ for all $i \in \mathcal{I}$ and for all $\alpha \in \mathcal{A}$, then the probability that the same action profile will be played for all future times is uniformly bounded away from zero over all initial conditions. ∎

Proposition 6.1 reveals the main issue of applying reinforcement learning schemes, which is convergence with positive probability to boundary points which are not Nash equilibrium profiles.

Figure 1 shows a typical response of $\tilde{\mathcal{L}}_{R-I}$ in the Typewriter Game of Table I. We observe that it is possible for the process to converge to a non-Nash equilibrium profile since $R_i(\alpha) > 1$ for all $\alpha \in \mathcal{A}$ and $i \in \mathcal{I}$.

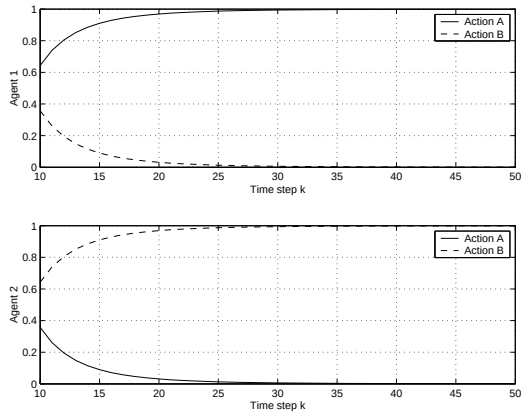

Fig. 1. Typical response of $\tilde{\mathcal{L}}_{R-I}$ on the Typewriter Game of Table I when $\nu = 0.78$.

These issues, which are also pointed out in [15], [3], will be resolved here due to the introduction of the perturbation function in $\tilde{\mathcal{L}}^{\lambda}_{R-I}$.

VII. Convergence of Perturbed Dynamics ($\tilde{\mathcal{L}}^{\lambda}_{R-I}$)

The convergence analysis of the perturbed dynamics $\tilde{\mathcal{L}}^{\lambda}_{R-I}$ will be subject to the following assumption:

Assumption 7.1: For the unperturbed dynamics, $\tilde{\mathcal{L}}_{R-I}$, there exists a twice continuously differentiable and nonnegative function $V : \mathbf{\Delta} \to \mathbb{R}_+$ such that a) $\nabla_x V(x)^{\mathrm{T}} \overline{g}(x) \leq 0$ for all $x \in \mathbf{\Delta}$, and b) $\nabla_x V(x)^{\mathrm{T}} \overline{g}(x) = 0$ if and only if $\overline{g}(x) = 0$.

For some $\delta > 0$, consider the δ-neighborhood of the set of stationary points $\mathcal{S}^{\lambda}$, $\mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$. Define also the closed set: $\mathcal{D}_{\delta}(\mathcal{S}^{\lambda}) \triangleq \mathbf{\Delta} \backslash \mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$.

Lemma 7.1: Under Assumption 7.1, for $\beta \in (0,1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, there exists $\delta = \delta(\beta, \lambda) > 0$ such that

$$\sup_{x \in \mathcal{D}_{\delta}(\mathcal{S}^{\lambda})} \nabla_x V(x)^{\mathrm{T}} \overline{g}^{\lambda}(x) < 0.$$

Proof: Pick $\delta^* = \delta^*(\beta)$ according to Proposition 5.2, such that, for all $x \in \mathbf{\Delta} \backslash \mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, $\zeta_i(x_i, \lambda) = 0$ for at least one i. Then, according to Proposition 5.2, when we take β sufficiently close to one (which implies that δ^* approaches zero) and λ sufficiently small, then a) $\tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}} \subset \mathcal{B}_{\delta^*}(\mathbf{\Delta}^*)$, and b) $\mathcal{S}^{\lambda} = \mathcal{S}_{\mathbf{\Delta}^o} \cup \tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$. Due to Assumption 7.1, there exists $\delta = \delta(\beta, \lambda) > \delta^*$ such that $\mathcal{B}_{\delta^*}(\mathbf{\Delta}^*) \subset \mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$ and

$$\sup_{x \in \mathcal{D}_{\delta}(\mathcal{S}^{\lambda})} \nabla_x V(x)^{\mathrm{T}} \overline{g}^{\lambda}(x) < 0.$$

■

Lemma 7.2 (LAS - $\tilde{\mathcal{L}}^{\lambda}_{R-I}$): For any $\lambda > 0$ sufficiently small, any stationary point $\tilde{x} \in \tilde{\mathcal{S}}^{\lambda}_{\mathrm{NE}}$, which is a perturbation of a strict Nash equilibrium according to (14), is a locally asymptotically stable point of the ODE (12).

Proof: The proof follows similar reasoning with the proof of Proposition 3.6 in [6]. ■

Theorem 7.1 (Convergence to Nash equilibria): Under Assumption 7.1, if agents employ the $\tilde{\mathcal{L}}^{\lambda}_{R-I}$ reinforcement scheme for some $\beta \in (0,1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, then there exists $\delta = \delta(\beta, \lambda)$ such that,

$$\mathbb{P}[\liminf_{k \to \infty} \mathrm{dist}(x(k), \mathcal{B}_{\delta}(\mathcal{S}^{\lambda})) = 0] = 1.$$

Also, for almost all ω, the process $\{\psi_k(\omega) = x(k)\}$ converges to some invariant set in $\mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$.

Proof: Consider the nonnegative function $V(x)$ of Assumption 7.1. We can approximate the expected incremental gain of $V(x)$ by applying a Taylor series expansion as follows:

$$\Delta V(k,x) = \nabla_x V(x)^{\mathrm{T}} \mathbb{E}[x(k+1) - x(k) | x(k) = x] + O(\epsilon(k)^2),$$

where $O(\epsilon(k)^2)$ denotes terms of order $\epsilon(k)^2$. Note that such an expansion is possible due to the fact that the second-order derivatives of $V(\cdot)$ are continuous in $\mathbf{\Delta}$. Equivalently,

$$\Delta V(k,x) = \epsilon(k) \nabla_x V(x)^{\mathrm{T}} \overline{g}^{\lambda}(x) + O(\epsilon(k)^2). \tag{17}$$

Due to Lemma 7.1, there exists $\delta = \delta(\beta, \lambda) > 0$ such that $-\bar{\rho} \triangleq \sup_{x \in \mathcal{D}_{\delta}(\mathcal{S}^{\lambda})} \nabla_x V(x)^{\mathrm{T}} \overline{g}^{\lambda}(x) < 0$. Thus,

$$\Delta V(k,x) \leq -\epsilon(k) \bar{\rho} + O(\epsilon(k)^2),$$

uniformly in $x \in \mathcal{D}_{\delta}(\mathcal{S}^{\lambda})$. The right hand side of the above inequality is strictly negative and can be formulated in the form of condition (9). Therefore, the conditions of Proposition 4.1 are satisfied and

$$\mathbb{P}[\liminf_{k \to \infty} \mathrm{dist}(x(k), \mathcal{B}_{\delta}(\mathcal{S}^{\lambda})) = 0] = 1.$$

From Proposition 4.2, we also have that the process $\{\psi_k(\omega) = x(k)\}$ will converge to some invariant set of the ODE in $\mathcal{B}_{\delta}(\mathcal{S}^{\lambda})$ almost surely. ■

VIII. Specialization to Potential Games

A. Potential games

In this section, we will specialize the convergence analysis to a class of games which belongs to the general family of potential games (cf., [7]). In particular, we will consider games which satisfy the following property:

Property 8.1: There exists a C^2 function $f : \mathbf{\Delta} \to \mathbb{R}$ such that $\nabla_{\sigma_i} f(\sigma) = U_i(\sigma)$ for all $\sigma \in \mathbf{\Delta}$ and $i \in \mathcal{I}$.

Example 1: (*Common-payoff games*) One class of games which satisfies Property 8.1 is *common-payoff games*, where the payoff function is the same for all players. An example of a common-payoff game is the Typewriter Game of Table I. It is straightforward to show that for this game the function

$$f(\sigma) = 4\sigma_{11}\sigma_{21} + 2\sigma_{11}\sigma_{22} + 2\sigma_{12}\sigma_{21} + 3\sigma_{12}\sigma_{22}$$

satisfies Property 8.1.

Example 2: (*Congestion games*) A typical congestion game consists of a set $\mathcal{I}$ of n players and a set $\mathcal{P}$ of m paths. For each player i, let the set of pure strategies $\mathcal{A}_i$ be the set of m paths. The cost to each player i of selecting the path p depends on the number of players that are using the same path. The expected number of players using path p is $\chi_p(\sigma) \triangleq \sum_{i \in \mathcal{I}} \sigma_{ip}$. Define $c_p = c_p(\chi_p)$ to be the cost of using path p when χ_p players are using path p and let $c_p(\chi_p)$

be linear on χ_p. The expected utility of player i is defined as: $u_i(\sigma) \triangleq -\sum_{p\in\mathcal{P}} c_p(\chi_p(\sigma))$. Note that the function

$$f(\sigma) \triangleq -\sum_{p\in\mathcal{P}} \int_0^{\chi_p(\sigma)} c_p(z)dz$$

satisfies Property 8.1.

B. Convergence to Nash equilibria

The following proposition establishes convergence to Nash equilibria for this class of potential games.

Proposition 8.1 (Convergence to Nash equilibria): In the class of games satisfying Property 8.1, the $\tilde{\mathcal{L}}_{R-I}^{\lambda}$ reinforcement scheme satisfies the conclusions of Theorem 7.1.

Proof: It suffices to show that the conditions of Assumption 7.1 are satisfied. In particular, define the nonnegative function

$$V(x) \triangleq f_{\max} - f(x) \geq 0, \quad x \in \mathbf{\Delta}, \tag{18}$$

where $f_{\max} \triangleq \sup_{x\in\mathbf{\Delta}} f(x)$. Note that $\nabla_{x_i} V(x) = -U_i(x)$, and

$$\begin{aligned} U_i(x)^{\mathrm{T}}\overline{g}_i(x) &= \sum_{s=1}^{|\mathcal{A}_i|}\sum_{j=1,j>s}^{|\mathcal{A}_i|} x_{is}x_{ij}(U_{is}(x)-U_{ij}(x))^2 \\ &= x_i^{\mathrm{T}}\tilde{D}_i(x)x_i/2 \end{aligned}$$

where $[\tilde{D}_i(x)]_{ss} = 0$ and $[\tilde{D}_i(x)]_{sj} = (U_{is}(x)-U_{ij}(x))^2$. Thus,

$$\nabla_x V(x)^{\mathrm{T}}\overline{g}(x) = -U(x)^{\mathrm{T}}\overline{g}(x) = -\sum_{i\in\mathcal{I}} U_i(x)^{\mathrm{T}}\overline{g}_i(x) \leq 0$$

for all $x \in \mathbf{\Delta}$.

We also observe that $\nabla_x V(x)^{\mathrm{T}}\overline{g}(x) = 0$ if and only if $U_{is}(x) = U_{ij}(x)$ for any $i \in \mathcal{I}$ and any $s, j \in \mathcal{A}_i$, $s \neq j$ such that $x_{is}, x_{ij} > 0$. By Proposition 5.1, these points correspond to the stationary points of $\overline{g}(x)$. Therefore, the conditions of Assumption 7.1 are satisfied. Thus, the conclusions of Theorem 7.1 hold for the class of games satisfying Property 8.1. ∎

C. Convergence to pure Nash equilibria

In several games, convergence to mixed Nash equilibria of the unperturbed dynamics $\mathcal{S}_{\mathbf{\Delta}^o}$ can be excluded. In this case, convergence to stationary points in $\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ which are perturbations of pure Nash equilibria can be established.

Let x_{-i} denote the distribution over action profiles of the group of agents $-i$. Let D_i be the matrix of payoffs of agent i and D_{-i} be the matrix of payoffs of $-i$. The vector of expected payoffs of agent i and $-i$ can be expressed as $U_i(x) = D_i x_{-i}$ and $U_{-i}(x) = D_{-i}x_i$, respectively.

To analyze the behavior around stationary points in $\mathbf{\Delta}^o$, we consider the nonnegative function $V(x) \triangleq f_{\max} - f(x) \geq 0$, $x \in \mathbf{\Delta}$, where $f_{\max} \triangleq \sup_{x\in\mathbf{\Delta}} f(x)$. It is straightforward to verify that the Jacobian matrix of $f(x)$ is:

$$\nabla_x^2 f(x) = \begin{pmatrix} O & D_i \\ D_{-i} & O \end{pmatrix}.$$

Higher-order derivatives of $f(x)$ will be zero, therefore from the extension of Taylor's Theorem (cf., Theorem 5.15 in [16]) to multivariable functions, we have:

$$\begin{aligned} \Delta V(k,x) = &-\nabla_x f(x)^{\mathrm{T}}\mathbb{E}[\delta x(k)|x(k)=x]- \\ &\mathbb{E}[\delta x_{-i}(k)^{\mathrm{T}}D_{-i}\delta x_i(k)|x(k)=x]- \\ &\mathbb{E}[\delta x_i(k)^{\mathrm{T}}D_i\delta x_{-i}(k)|x(k)=x], \end{aligned} \tag{19}$$

where $\delta x(k) \triangleq x(k+1) - x(k)$.

A direct consequence of the above formulation and Proposition 4.1 is the following:

Proposition 8.2 (Non-convergence to $\mathcal{S}_{\mathbf{\Delta}^o}$): If agents employ the $\tilde{\mathcal{L}}_{R-I}$ reinforcement scheme and $x^* \in \mathcal{S}_{\mathbf{\Delta}^o}$ satisfies

1) $\mathbb{E}[\delta x_{-i}(k)^{\mathrm{T}}D_{-i}\delta x_i(k)|x(k)=x] > 0$,
2) $\mathbb{E}[\delta x_i(k)^{\mathrm{T}}D_i\delta x_{-i}(k)|x(k)=x] > 0$

uniformly in $x \in \mathcal{B}_\delta(x^*)$, for some $\delta > 0$ sufficiently small, then $\mathbb{P}[\lim_{k\to\infty} x(k) = x^*] = 0$.

Proof: We consider the nonnegative function $V(x)$ defined above. Note that the expected incremental gain of $V(x)$ (19) can take the following form:

$$V(k,x) = -\epsilon(k)\phi(k,x)$$

where $\inf_{x\in\mathcal{B}_\delta(x^*)} \phi(k,x) > 0$ for some $\delta > 0$ sufficiently small and for all k. This is due to the fact that for any $x \in \mathcal{B}_\delta(x^*)$,

$$-\nabla_x f(x)^{\mathrm{T}}\mathbb{E}[\delta x(k)|x(k)=x] \leq 0$$

(due to Property 8.1), and the second-order terms of the incremental gain are strictly negative by assumption. Then, from Proposition 4.1, we conclude that the process will exit $\mathcal{B}_\delta(x^*)$ in finite time with probability one. Therefore, the conclusion follows. ∎

For several games testing the conditions of Proposition 8.2 may be difficult. For example, for two players and two actions, it is straightforward to show that:

$$\begin{aligned} &\mathbb{E}[\delta {x_i}^{\mathrm{T}} D_i \delta x_{-i} | x_i(k) = x_i, x_{-i}(k) = x_{-i}] = \\ &\quad \epsilon(k)^2 x_{i1}x_{i2}x_{(-i)1}x_{(-i)2}(d_{11}^i - d_{12}^i - d_{21}^i + d_{22}^i)\cdot \\ &\quad ((d_{11}^i)^2 - (d_{12}^i)^2 - (d_{21}^i)^2 + (d_{22}^i)^2), \end{aligned} \tag{20}$$

where $d_{s\ell}^i$ denotes the (s,ℓ) entry of D_i, $i = 1, 2$. Consider, for example, the Typewriter Game of Table I. Since the game is symmetric, and $d_{11}^i > d_{12}^i$, $d_{22}^i > d_{21}^i$, $i = 1, 2$, the second-order terms of the incremental gain will be positive. The above computation can be extended in a similar manner to the case of larger number of actions or players.

Proposition 8.3 (Convergence to pure Nash equilibria): In the framework of Proposition 8.1, let the conditions of Proposition 8.2 also hold. If the game exhibits pure Nash equilibria which are all strict, then, for some $\beta \in (0,1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, the process $\{\psi_k(\omega) = x(k)\}$ converges to the set $\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ for almost all ω, i.e., $\mathbb{P}[\lim_{k\to\infty} x(k) \in \tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}] = 1$.

Proof: Since the game exhibits pure Nash equilibria which are all strict, the set $\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ in non-empty for any $\lambda > 0$ sufficiently small.

Let x^* denote an action profile which is a strict pure Nash equilibrium, i.e., for every $i \in \mathcal{I}$ there exists $j^* = j^*(i)$ such that $x_{ij^*} = 1$ and $U_{is}(x^*) - U_{ij^*}(x^*) < 0$ for any $s \neq j^*$. Let also $\tilde{x} \in \tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ be the perturbed stationary point according to (14). Pick also $\delta^* = \delta^*(\beta) > 0$ similarly to the proof of

Lemma 7.1. Then, for any $x \in \mathcal{B}_{\delta^*}(\tilde{x})$, x_{is} is of order of δ^* and

$$\overline{g}_{is}^{\lambda}(x) \approx [U_{is}(x^*) - U_{ij^*}(x^*)]x_{is} \tag{21}$$

plus higher order terms of δ^* and λ, for all $s \neq j^*$. Since $U_{is}(x^*) - U_{ij^*}(x^*) < 0$ for all $s \neq j^*$, we conclude that the vector-field points towards the interior of $\mathcal{B}_{\delta^*}(\tilde{x})$ when evaluated at the boundary of $\mathcal{B}_{\delta^*}(\tilde{x})$. Thus, $\mathcal{B}_{\delta^*}(\tilde{x})$ is an invariant set of the ODE (12). Therefore, due to Proposition 8.2 and Theorem 7.1, if we take $\beta \in (0,1)$ sufficiently close to one and $\lambda > 0$ sufficiently small, then there exists $\delta = \delta(\beta, \lambda) > \delta^*$ such that the process $\{x(k)\}$ converges almost surely to some invariant set in $\mathcal{B}_{\delta}(\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda})$.

Furthermore, due to Lemma 7.2, we know that the points in $\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ are locally asymptotically stable, and therefore by (21), the set $\mathcal{B}_{\delta}(\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda})$ belongs to its region of attraction. Since the process visits $\mathcal{B}_{\delta}(\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda})$ infinitely often, by Proposition 4.2, we conclude that the process converges to $\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ with probability one. ■

D. Example

Consider the Typewriter Game of Table I. This game exhibits two pure Nash equilibria which are strict, (A, A) and (B, B). There is also a mixed Nash equilibrium, which satisfies the conditions of Proposition 8.2 as it can be verified from (20). Thus, the conditions of Proposition 8.3 are satisfied, and the process will converge to the stationary points in $\tilde{\mathcal{S}}_{\mathrm{NE}}^{\lambda}$ almost surely. Figure 2 shows the solution of the ODE (12) for an initial condition which corresponds to the non-Nash action profile (B, A). The solution converges

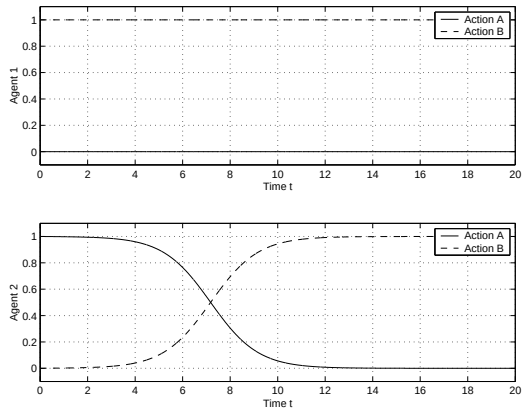

Fig. 2. ODE solution for $\tilde{\mathcal{L}}_{R-I}^{\lambda}$ and for the Typewriter Game of Table I when $\beta = 0.995$, $\lambda = 0.001$ and initial condition (B, A).

to the strict Nash equilibrium (B, B). Note that escaping from (B, A) would not be possible if $\lambda = 0$.

IX. Conclusions

This paper presented a new reinforcement learning scheme for distributed convergence to Nash equilibria. The main difference from prior schemes lies in the introduction of a perturbation function in the decision rule of each agent which depends only on its own strategy. The introduction of this perturbation function sidestepped issues regarding the behavior of the algorithm close to vertices of the simplex. In particular, we derived conditions under which the perturbed reinforcement learning scheme converges to an arbitrarily small neighborhood of the set of Nash equilibria almost surely. We further specialized the results to a class of games which belong to potential games.

References

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

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

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

[4] R. Pemantle, "Nonconvergence to unstable points in urn models and stochastic approximations," *The Annals of Probability*, vol. 18, no. 2, pp. 698–712, 1990.

[5] J. Bergin and B. L. Lipman, "Evolution with state-dependent mutations," *Econometrica*, vol. 64, no. 4, pp. 943–956, 1996.

[6] G. Chasparis and J. Shamma, "Distributed dynamic reinforcement of efficient outcomes in multiagent coordination and network formation," Georgia Institute of Technology, Atlanta, GA, Discussion Paper, 2009.

[7] D. Monderer and L. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, pp. 124–143, 1996.

[8] R. Myerson, *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press, 1991.

[9] M. F. Norman, "On linear models with two absorbing states," *Journal of Mathematical Psychology*, vol. 5, pp. 225–241, 1968.

[10] I. J. Shapiro and K. S. Narendra, "Use of stochastic automata for parameter self-organization with multi-modal performance criteria," *IEEE Transactions on Systems Science and Cybernetics*, vol. 5, pp. 352–360, 1969.

[11] W. B. Arthur, "On designing economic agents that behave like human agents," *Journal of Evolutionary Economics*, vol. 3, pp. 1–22, 1993.

[12] M. B. Nevelson and R. Z. Hasminskii, *Stochastic Approximation and Recursive Estimation*. Providence, RI: American Mathematical Society, 1976.

[13] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, 2nd ed. Springer-Verlag New York, Inc., 2003.

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

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

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