Mixed-strategy learning with continuous action sets

Steven Perkins, Panayotis Mertikopoulos, *Member, IEEE*, and David S. Leslie

***Abstract*—Motivated by the recent applications of game-theoretical learning to the design of distributed control systems, we study a class of control problems that can be formulated as potential games with continuous action sets. We propose an actor-critic reinforcement learning algorithm that adapts mixed strategies over continuous action spaces. To analyse the algorithm we extend the theory of finite-dimensional two-timescale stochastic approximation to a Banach space setting, and prove that the continuous dynamics of the process converge to equilibrium in the case of potential games. These results combine to give a provably-convergent learning algorithm in which players do not need to keep track of the controls selected by other agents.**

I. Introduction

There has been much recent activity in using game-theoretical learning within distributed control systems. This research traverses from utility function design [1–3], through analysis of suboptimalities due to the use of distributed selfish controllers [4] to the design and analysis of game-theoretical learning algorithms with specific control-inspired objectives (reaching a global optimum, fast convergence, etc.) [5, 6].

In this context, considerable interest has arisen from the approach in which the independent controls available to a system are distributed among a set of agents, henceforth called "players". To complete the game-theoretical analogy, the controls available to a player are called "actions". Each player is assigned a utility function which depends on the actions of all players (as does the system-level utility). Each player's utility could be set to the global utility of the joint action selected by all players. However, as argued by [1], encoding the system utility into player-specific utility functions usually results in improved performance. Wonderful life utility [1, 2] and Shapley value utility [4] are two common approaches, and most proposed alternatives also result in a potential game [7] (possibly in an extended sense [3]) where the optimal system control is a Nash equilibrium of the game. Thus, by representing a control problem as a potential game, the controllers' main objective amounts to reaching a Nash equilibrium.

On the other hand, like much of the economic literature on learning in games [8], the vast majority of this corpus of research has focused on situations where each player's controls comprise a *finite* set. However, the assumption of discrete action sets is frequently problematic in control, engineering and economics: after all, prices are not discrete, and neither are the controls in a large number of engineering systems. For numerous examples see [9–13]. We therefore focus on control problems (presented as potential games) with *continuous* action sets and propose an actor-critic reinforcement learning algorithm that provably converges to equilibrium. Building on recent work on single population or two-player games in the economics literature [14, 15], our analysis relies on two novel contributions of independent interest. The first is the combination of two-timescales stochastic approximation techniques [16] with so-called "abstract stochastic approximation" on Banach spaces, e.g. [17]. Our second contribution is the convergence analysis of the mean field dynamics of this process on the space of probability measures on the action space. Combined with our stochastic approximation results, we thus obtain the convergence of our actor-critic reinforcement learning algorithm to equilibrium in potential games.

There are several other approaches to learning in continuous action games, from disconnected literatures. Deterministic differential equation approaches have focused on the replicator dynamics [18], the Brown–von Neumann–Nash dynamics [19], and the best-response dynamics [20, 21] in continuous time. In discrete time learning [22–24] consider a gradient-ascent-like scheme, while several approaches have been proposed to learn pure strategy equilibria based on variational inequality theory — see e.g. [25–28] and references therein. Our work initiates a new and different approach in which full mixed strategies are adapted in a discrete-time stochastic learning process which does not require full sharing of current strategies.

II. Actor–critic learning

Throughout this paper, we will focus on control problems presented as potential games with finitely many players and continuous action spaces. Such a game comprises a finite set of players labelled $i \in \{1,\dots,N\}$. For each i there exists an action set $A^i \subset \mathbb{R}$ which is a compact interval.[1] When each player selects an action $a^i \in A^i$ this results in a joint action $\underline{a} = (a^1,\dots,a^N) \in \underline{A} = \prod_{i=1}^N A^i$. We use the notation (a^i, a^{-i}) to refer to the joint action $\underline{a}$ in which Player i uses action a^i and all other players use action $a^{-i} = (a^1,\dots,a^{i-1},a^{i+1},\dots,a^N)$. Each player i is also associated with a bounded and continuous utility function $u^i : \underline{A} \to \mathbb{R}$. In a potential game [7], there exists a potential function $\phi : \underline{A} \to \mathbb{R}$ such that

$$u^i(a^i, a^{-i}) - u^i(\tilde{a}^i, a^{-i}) = \phi(a^i, a^{-i}) - \phi(\tilde{a}^i, a^{-i})$$

S. Perkins' Ph.D. research was funded by grant number EP/D063485/1 from the United Kingdom Engineering and Physical Sciences Research Council. P. Mertikopoulos' research was partially supported by the French National Research Agency under grant nos. ANR-GAGA-13-JS01-0004-01 and ANR-NETLEARN-13-INFR-004, and the CNRS grant PEPS-GATHERING-2014. D.S. Leslie's research was funded by grant number EP/I032622/1 from the United Kingdom Engineering and Physical Science Research Council.

S. Perkins carried out this research while a PhD candidate at the School of Mathematics, University of Bristol, United Kingdom.

P. Mertikopoulos is with the French National Center for Scientific Research (CNRS) and the Laboratoire d'Informatique de Grenoble, Grenoble, France.

David S. Leslie is with Department of Mathematics and Statistics, Lancaster University, United Kingdom.

[1] The analysis transfers easily to other convex, compact bodies.

for all $i \in \{1,\dots,N\}$, for all a^{-i} and for all a^i, $\tilde{a}^i$. Methods for constructing potential games from system utility functions [1–3] usually ensure that the potential corresponds to the system utility, so maximising the potential function corresponds to maximising the system utility.

Game-theoretical analyses usually focus on mixed strategies where a player selects an action to play randomly. A mixed strategy for Player i is defined to be a probability distribution over the action space A^i. Specifically, let $\mathcal{B}^i$ be the Borel sigma-algebra on A^i and let $\mathcal{P}(A^i,\mathcal{B}^i)$ denote the set of all probability measures on A^i. A mixed strategy is then an element $\pi^i \in \mathcal{P}(A^i,\mathcal{B}^i)$. Such a mixed strategy need not admit a density with respect to Lebesgue measure, and in particular may contain an atom at a particular action a^i.

Returning to our game-theoretical considerations, we extend the utilities u^i linearly to the space $\Delta = \prod_{i=1}^N \mathcal{P}(A^i,\mathcal{B}^i)$ of mixed strategy profiles. As before we use the notation (π^i,π^{-i}) to refer to the mixed strategy profile $\underline{\pi}$ in which Player i uses π^i and all other players use $\pi^{-i} = (\pi^1,\dots,\pi^{i-1},\pi^{i+1},\dots,\pi^N)$. We also write (a^i,π^{-i}) for the strategy profile (δ_{a^i},π^{-i}), where δ_{a^i} is the Dirac measure at a^i. Hence $u^i(a^i,\pi^{-i})$ is the utility to Player i for selecting a^i when all other players use strategy π^{-i}.

A central concept in game theory is the best response correspondence of Player i, the set of mixed strategies π^i that maximise $u^i(\pi^i,\pi^{-i})$ for any particular π^{-i}; a Nash equilibrium is a fixed point of this correspondence. In a learning context however, discontinuities in best response correspondences can cause difficulties [8]. We focus instead on a smoothing of the best response. For a fixed $\eta > 0$, the *logit best response with noise level* η of Player i to strategy π^{-i} is defined in [14, 15] and shown to be the absolutely continuous mixed strategy $L^i_\eta(\pi^{-i}) \in \mathcal{P}(A^i,\mathcal{B}^i)$ with density

$$l^i_\eta(\underline{\pi}^{-i})(a^i) = \frac{\exp\left\{\eta^{-1}u^i(a^i,\pi^{-i})\right\}}{\int_{A^i}\exp\left\{\eta^{-1}u^i(b^i,\pi^{-i})\right\}\,\mathrm{d}b^i}. \tag{1}$$

To ease notation, let $L_\eta(\underline{\pi}) = \left(L^1_\eta(\pi^{-1}),\dots,L^N_\eta(\pi^{-N})\right)$. Smooth best responses also play an important part in discrete action games, particularly when learning is considered. They were introduced in stochastic fictitious play by [29] (see also [30–32] for example) to ensure the played mixed strategies in a fictitious play process converge; in classical fictitious play the played strategies are (almost) always pure. The technique was also required by [33–35] to allow simple reinforcement learners to converge to logit equilibria.

The existence of fixed points of L_η is shown by [14, 15]; such a fixed point is a joint strategy $\underline{\pi}$ such that $\pi^i = L^i_\eta(\pi^{-i})$ for each i. Such profiles $\underline{\pi}$ are called *logit equilibria* and the set of all such fixed points will be denoted by $\mathcal{LE}_\eta$. Logit equilibria approximate Nash equilibria when η is sufficiently small, so for small η elements of $\mathcal{LE}_\eta$ concentrates most of its mass near to a critical point of the potential function ϕ.

One of the motivations for learning in a control setting is that full utility functions might not be known in advance, and players might not observe the actions of all other players. Using fictitious play (or, indeed, most standard game-theoretical tools) does not satisfy this requirement because they assume full knowledge of payoff functions and opponent actions. This is what motivates the simple reinforcement learning approaches [33–35]. The scheme in this article extends the actor–critic approach of [36] to the continuous action space setting. It learns both a value function $Q^i : A^i \to \mathbb{R}$ that estimates the function $u^i(a^i,\pi^{-i})$ for the current value of π^{-i}, while also maintaining a separate mixed strategy $\pi^i \in \mathcal{P}(A^i,\mathcal{B}^i)$. The critic, Q^i, informs the update of the actor, π^i. In turn the observed utilities received by the actor, π^i, inform the update of the critic Q^i. The procedure is given in Algorithm 1. It is the main focus of our paper, so some remarks are in order:

Algorithm 1 Actor-critic Reinforcement Learning

Parameters: step-size sequences α_n, γ_n.
Initialize critics Q^i_0, actors π^i_0; $n \leftarrow 0$.
Repeat
  $n \leftarrow n+1$;
  for each player $i = 1,\dots,N$ **do simultaneously**
    sample action a^i_n from distribution π^i;
    update critic:

$$Q^i_{n+1} = Q^i_n + \gamma_n\left(u^i(\cdot,a^{-i}_n) - Q^i_n\right) \tag{2a}$$

    sample $b^i_n \sim L^i_\eta(Q^i_n)$ and update actor:

$$\pi^i_{n+1} = \pi^i_n + \alpha_n\left(\delta_{b^i_n} - \pi^i_n\right). \tag{2b}$$

  until termination criterion is reached.

Remark 1. To implement this algorithm an individual need not actually observe the action profile a^{-i}_n, but only needs the utility $u^i(\cdot,a^{-i}_n)$. This means a player need know only about the players who directly affect her utility function, allowing a degree of modularisation in large systems. In (2a), it is assumed that a player can access the function $u^i(\cdot,a^{-i}_n)$ determining how much they would have received for each action against the selected joint action a^{-i}_n. Even though this assumption restricts the applicability of our method, it is harmless in many practical settings—e.g. in congestion games $u^i(\cdot,a^{-i}_n)$ can be calculated by observing the utilisation levels.

Remark 2. The logit response L^i_η used to sample b^i_n in (2b) is now parameterised by Q^i_n instead of π^{-i}. This is a trivial change in which $Q^i(\cdot)$ replaces $u^i(\cdot,\pi^{-i})$ in (1).

Remark 3. Also in (2b), the players update towards a sampled b^i_n instead of toward the full function $L^i_\eta(Q^i_n)$. This is so that the critic π^i_n can be represented as a collection of weighted atoms, instead of as a complicated and continuous probability measure. Representing π^i_n as a collection of atoms means that sampling $a^i_n \sim \pi^i_n$ is particularly easy.

On the other hand, sampling $b^i_n \sim L^i_\eta(Q^i_n)$ could be difficult for general Q^i_n. One solution would be to use a sequential Monte Carlo sampler [37] to produce samples from the slowly evolving distribution $L^i_\eta(Q^i_n)$. The representation of Q^i_n is also potentially troublesome and we do not address it here. A solution would be to assume that each $u^i(\underline{a}_n)$ can be represented as a finite linear combination of basis functions. Another would be to slowly increase the size of a

Fourier or wavelet basis as n gets large, resulting in vanishing bias terms which can be easily incorporated in the stochastic approximation framework introduced below.

Remark 4. Finally note that we assume all players use the same values for η, α_n and γ_n. Allowing player-dependent values of η^i causes only notational challenges later. Allowing player-dependent learning rates α_n^i and γ_n^i is easily accommodated so long as $\alpha_n^i/\alpha_n^j \to 1$ and $\gamma_n^i/\gamma_n^j \to 1$ as $n \to \infty$. See [36] for an analysis in which these conditions do not hold.

The remainder of this article works to prove the following theorem, while also providing several auxiliary results of independent interest along the way:

Theorem 1. *In a continuous-action-set potential game with bounded Lipschitz rewards and isolated equilibrium components, the actor–critic algorithm* (2)*, with step sizes satisfying (A3) below, converges strongly to a component of the equilibrium set $\mathcal{LE}_\eta$ (a.s.).*

III. Two-timescales stochastic approximation

The analysis of systems such as Algorithm 1 is enabled by the use of two-timescales stochastic approximation techniques [16]. By allowing $\alpha_n/\gamma_n \to 0$ as $n \to \infty$, the system can be analysed as if the 'fast' update (2a), with higher learning parameter γ_n, has fully converged to the current value of the 'slow' system (2b), with lower learning parameter α_n. Note that it is not the case that we have an outer and inner loop, in which (2a) is run to convergence for every update of (2b): both the actor Q_n and the critic π_n are updated at every iteration. It is simply that the two-timescales technique allows us to analyse the system *as if* there were an inner loop.

That being said, the results of [16] are only cast in the framework of finite-dimensional spaces. We have already observed that with continuous action spaces A^i, the mixed strategies π^i are probability measures in the space $\mathcal{P}(A^i, \mathcal{B}^i)$, and the critics Q^i are L^2 functions. Placing appropriate norms on these spaces results in Banach spaces, and in this section we combine the two-timescales results of [16] with the Banach space stochastic approximation framework of [14].

To that end, consider the general two-timescales stochastic approximation system

$$x_{n+1} = x_n + \alpha_{n+1}\left[F(x_n, y_n) + U_{n+1} + c_{n+1}\right], \tag{3a}$$
$$y_{n+1} = y_n + \gamma_{n+1}\left[G(x_n, y_n) + V_{n+1} + d_{n+1}\right], \tag{3b}$$

where

- x_n and y_n are sequences in the Banach spaces $(X, \|\cdot\|_X)$ and $(Y, \|\cdot\|_Y)$ respectively.
- $\{\alpha_n\}$ and $\{\gamma_n\}$ are the learning rate sequences.
- $F : X \times Y \to X$ and $G : X \times Y \to Y$ comprise the *mean field* of the process.
- $\{U_n\}$ and $\{V_n\}$ are stochastic processes in X and Y respectively.
- $c_n \in X$ and $d_n \in Y$ are transient bias terms.

We will study this system using the asymptotic pseudotrajectory approach of [38], which is already cast in the language of metric spaces; since Banach spaces are metric, the framework of [38] still applies to our scenario. This modernises the approach of [17] while also introducing the two-timescales technique to 'abstract stochastic approximation'.

Our assumptions, which are simple extensions to those of [16] and [38], can now be stated as follows:

A1) Noise control.
  a) For ξ being either α or γ, let $\tau_n^\xi = \sum_{j=1}^n \xi_j$ (with $\tau_0^\xi = 0$), and for $t \in \mathbb{R}_+$ let $m^\xi(t) = \sup\{k \geq 0; \tau_k^\xi \leq t\}$. For all $T > 0$, we assume that
$$\lim_{n\to\infty} \sup_{k\in\{n+1,\ldots,m^\alpha(\tau_n^\alpha+T)\}} \left\{ \left\| \sum_{j=n}^{k-1} \alpha_{j+1} U_{j+1} \right\|_X \right\} = 0,$$
$$\lim_{n\to\infty} \sup_{k\in\{n+1,\ldots,m^\gamma(\tau_n^\gamma+T)\}} \left\{ \left\| \sum_{j=n}^{k-1} \gamma_{j+1} V_{j+1} \right\|_Y \right\} = 0.$$
  b) $\{c_n\}_{n\in\mathbb{N}}$ and $\{d_n\}_{n\in\mathbb{N}}$ are bounded sequences such that $\|c_n\|_X \to 0$ and $\|d_n\|_Y \to 0$ as $n \to \infty$.

A2) Boundedness and continuity.
  a) There exist compact sets $C \subset X$ and $D \subset Y$ such that $x_n \in C$ and $y_n \in D$ for all $n \in \mathbb{N}$.
  b) F and G are bounded and uniformly continuous on $C \times D$.

A3) Learning rates.
  a) $\sum_{n=1}^\infty \alpha_n = \infty$ and $\sum_{n=1}^\infty \gamma_n = \infty$ with $\alpha_n \to 0$ and $\gamma_n \to 0$ as $n \to \infty$.
  b) $\alpha_n/\gamma_n \to 0$ as $n \to \infty$.

A4) Mean field behaviour.
  a) For any fixed $\tilde{x} \in C$ and initial value $y_0 \in D$,
$$\frac{\mathrm{d}y}{\mathrm{d}t} = G(\tilde{x}, y) \tag{4}$$
  has unique solution trajectories remaing in D. Furthermore (4) has a unique globally attracting fixed point $y^*(\tilde{x})$, and $y^* : C \to D$ is Lipschitz continuous.
  b) For any initial value $x_0 \in C$, the differential equation
$$\frac{\mathrm{d}x}{\mathrm{d}t} = F(x, y^*(x)) \tag{5}$$
  has unique solution trajectories that remain in C.

Assumption A1 is the standard assumption for noise control in stochastic approximation. It has caused difficulty in abstract stochastic approximation, but recent works [14, 39] provide martingale noise type criteria that guarantee A1(a) holds in useful Banach spaces. Assumption A2 is simply a boundedness and continuity assumption. Assumption A3 provides the two-timescales nature of the scheme, with both learning rate sequences converging to 0, but α_n doing so faster than γ_n. Finally Assumption A4 provides both the existence of unique solutions of the relevant mean field differential equations, and the useful separation of timescales in continuous time which is directly analogous to Assumption (A1) of [16].

Our first lemma demonstrates that we can analyse the system as if the fast system $\{y_n\}$ is calibrated to the slow system $\{x_n\}$.

Lemma 2. *Under Assumptions A1–A4,*

$$\|y_n - y^*(x_n)\|_Y \to 0 \quad \textit{as} \quad n \to \infty.$$

Proof: Let $z_n = (x_n, y_n)$ and let $Z = X \times Y$, with $\|\cdot\|_Z$ the induced product norm from the topologies of X and Y. Under this topology, Z is a Banach space, and $C \times D$ is compact. The updates (3) can be expressed as

$$z_{n+1} = z_n + \gamma_{n+1}\Big[H(z_n) + W_{n+1} + \kappa_{n+1}\Big], \qquad (6)$$

where $H : Z \to Z$ is such that $H(z_n) = (0, G(z_n))$ for $0 \in X$, $W_n = \left(\frac{\alpha_n}{\gamma_n} U_n, V_n\right)$, and

$$\kappa_{n+1} = \left(\frac{\alpha_{n+1}}{\gamma_{n+1}}\Big[F(z_n) + d_{n+1}\Big], e_{n+1}\right).$$

Assumptions A1–A4 imply the assumptions of Theorem 3.3 of [14]. Most are direct translations, but the noise must be carefully considered. For any $n \in \mathbb{N}$, any $T > 0$, and any $k \in \{n+1, \dots, m^\gamma(\tau_n^\gamma + T)\}$,

$$\begin{aligned}
&\left\|\sum_{j=n}^{k-1} \gamma_{j+1}(W_{n+1} + \kappa_{n+1})\right\|_Z \\
&\leq \left\|\sum_{j=n}^{k-1} \gamma_{j+1} W_{n+1}\right\|_Z + \left(\sup_{k' \geq n+1} \|\kappa_{k'}\|_Z\right) \sum_{j=n}^{m^\gamma(\tau_n^\gamma+T)-1} \gamma_{j+1}
\end{aligned}$$

Since $\kappa_n \to 0$ and $\sum_{j=n}^{m^\gamma(\tau_n^\gamma+T)-1} \gamma_{j+1} \approx T$, the second term converges to 0 as $n \to \infty$. Hence, using assumption A1 to control the first term, we get that

$$\lim_{n\to\infty} \sup_{k \in \{n+1,\dots,m^\gamma(\tau_n^\gamma+T)\}} \left\|\sum_{j=n}^{k-1} \gamma_{j+1}(W_{n+1} + \kappa_{n+1})\right\|_Z = 0.$$

Assumption A4(a) implies that $\{(x, y^*(x)) : x \in C\}$ is globally attracting for the flow defined by

$$\frac{\mathrm{d}z}{\mathrm{d}t} = H\big(z(t)\big). \qquad (7)$$

Hence Theorem 3.3 of [14] and Theorem 6.10 of [38] give that $z_n \to \{(x, y^*(x)) : x \in C\}$. The result follows by the continuity of y^* assumed in A4(a). ∎

This allows the consideration of the slow timescale.

Theorem 3. *Suppose that Assumptions A1–A4 hold. Then x_n converges to an internally chain transitive set [38] of the flow induced by the mean field differential equation* (5)*.*

Proof: Rewrite (3a) as

$$x_{n+1} = x_n + \alpha_{n+1}\Big[F\big(x_n, y^*(x_n)\big) + U_{n+1} + \tilde{c}_{n+1}\Big], \quad (8)$$

where $\tilde{c}_{n+1} = F(x_n, y_n) - F(x_n, y^*(x_n)) + c_{n+1}$. To show that this is a well-behaved stochastic approximation process we need to show that $\tilde{c}_n$ can be absorbed into U_n such that the equivalent Assumption A1 of [14] can be applied to $U_n + \tilde{c}_n$.

By Lemma 2 we have $\|y_n - y^*(x_n)\|_Y \to 0$. Uniform continuity of F implies that $\|F(x_m, y_m) - F(x_m, y^*(x_m))\|_X \to 0$. The rest of the proof uses identical arguments as that of Lemma 2 to show that Theorem 3.3 of [14] applies.

Combining with Theorem 5.7 of [38], where internally chain transitive sets are discussed, gives the result. ∎

IV. Convergence of the actor–critic algorithm

In this section we demonstrate that the actor–critic algorithm (2) can be analysed using the two-timescales stochastic approximation framework of Section III. Our first task is to define the Banach spaces in which the algorithm evolves.

Note that the set $\mathcal{P}(A^i, \mathcal{B}^i)$ of probability distributions on A^i is a subset of the space $\mathcal{M}(A^i, \mathcal{B}^i)$ of finite signed measures on $(A^i, \mathcal{B}^i)$. To turn this space into a Banach space, the most convenient norm for our purposes is the bounded Lipschitz (BL) norm, as in [14, 15] and references therein.

Suppose that utility functions u^i are bounded and Lipschitz continuous. Since their domain is a bounded interval of $\mathbb{R}$ we can assume that each Q_n^i lies in the Banach space $L^2(A^i)$ of functions $A^i \to \mathbb{R}$ with a finite L^2 norm. Hence we consider the vectors $\underline{Q}_n = (Q_n^1, \dots, Q_n^N)$ as elements of the Banach space $Y = \times_{i=1}^N L^2(A^i)$ with $\|\underline{Q}\|_Y = \max_{i=1,\dots,N} \|Q^i\|_{L^2}$.

Theorem 4. *Consider the actor–critic algorithm* (2)*. Suppose that for each i the action space A^i is a compact interval of $\mathbb{R}$, and the utility function u^i is bounded and uniformly Lipschitz continuous. Suppose also that $\{\alpha_n\}_{n\in\mathbb{N}}$ and $\{\gamma_n\}_{n\in\mathbb{N}}$ are chosen to satisfy Assumption A3 as well as $\sum_{n\in\mathbb{N}} \alpha_n^2 < \infty$ and $\sum_{n\in\mathbb{N}} \gamma_n^2 < \infty$. Then, under the bounded Lipschitz norm, $\{\underline{\pi}_n\}_{n\in\mathbb{N}}$ converges with probability 1 to an internally chain transitive set of the flow defined by the N-player logit best response dynamics*

$$\frac{\mathrm{d}\underline{\pi}}{\mathrm{d}t} = L_\eta(\underline{\pi}) - \underline{\pi}. \qquad (9)$$

Proof: We take $(X, \|\cdot\|_X) = (\Sigma, \|\cdot\|_{BL})$, and $(Y, \|\cdot\|_Y)$. This allows a direct mapping of the actor–critic algorithm (2) to the stochastic approximation framework (3) by taking

$$\begin{aligned}
&x_n = \underline{\pi}_n, \qquad y_n = \underline{Q}_n, \qquad c_n = d_n = 0, \\
&F(\underline{\pi}, \underline{Q}) = L_\eta(\underline{Q}) - \underline{\pi}, \quad U_{n+1} = (\delta_{b_n^1}, \dots, \delta_{b_n^N}) - L_\eta(\underline{Q}), \\
&G = (G^1, \dots, G^N), \qquad G^i(\underline{\pi}, \underline{Q}) = u^i(\cdot, \pi^{-i}) - Q^i, \\
&V_n = (V_n^1, \dots, V_n^N), \qquad V_{n+1}^i = u^i(\cdot, a_n^{-i}) - u^i(\cdot, \pi_n^{-i}).
\end{aligned}$$

Theorem 3 means we need only verify Assumptions A1–A4.

A1: U_n is of exactly the form studied by [14] and therefore Proposition 3.6 of that paper suffices to prove the condition on the tail behaviour of $\sum_j \alpha_{j+1} U_{j+1}$ holds with probability 1. The V_{n+1} are martingale difference sequences, since $\mathbb{E}(u^i(\cdot, a_n^{-i}) \,|\, \mathcal{F}_n) = u^i(\cdot, \pi_n^{-i})$, and the Q_{n+1} are L^2 functions. Hence Proposition A.1 of [14] suffices to prove the condition on the tail behaviour of $\sum_j \gamma_{j+1} V_{j+1}$ holds with probability 1 under the L^2 norm. Since c_n and d_n are identically zero, we have shown that A1 holds.

A2: Δ is a compact subset of Σ under the bounded Lipschitz norm (see Prop. 4.6 of [14]), so taking $C = \Delta$ suffices. Furthermore, with bounded continuous reward functions u^i it follows that the Q_n^i are uniformly bounded and equicontinuous and therefore remain in a compact set D. G is clearly uniformly continuous on the compact set $C \times D$. The continuity of L_η, and therefore F, is shown in Lemma C.2 of [14].

A3: The learning rates are chosen to satisfy this assumption.

A4: For fixed $\tilde{\pi}$, the differential equations $\dot{Q}^i = u^i(\cdot, \tilde{\pi}^{-i}) - Q^i$ converge to $Q^i = u^i(\cdot, \tilde{\pi}^{-i})$. Furthermore $u^i(\cdot, \pi^{-i})$ is Lipschitz continuous in π^{-i}, so part (a) is satisfied. Equation (5) then becomes the logit best response dynamics of [14, 15], which is shown to have unique solution trajectories. ∎

It is demonstrated in [14] that logit equilibria are global attractors in two-player zero-sum games with continuous action sets. Hence, by Corollary 5.4 of [38] we instantly obtain the result that any internally chain transitive set is contained in $\mathcal{LE}_\eta$. However two-player zero-sum games are not particularly relevant for control systems: an equivalent result is required for multiplayer potential games. Note that evolution of strategies under the logit best response dynamics in a potential game is identical to that in the identical interest game in which the potential acts as the global utility. We therefore carry out our convergence analysis for the logit best response dynamics (9) in N-player identical interest games with continuous action spaces. See [32] for related issues. For the remainder of this section we work to prove the following theorem:

Theorem 5. *In a potential game with continuous bounded rewards, in which the connected components of the set $\mathcal{LE}_\eta$ of logit equilibria of the game are isolated, any internally chain transitive set of the flow induced by the smooth best response dynamics* (9) *is contained in a connected component of $\mathcal{LE}_\eta$.*

Proof: The natural Lyapunov function for the logit best response dynamics (9) in an identical interest game is

$$V_\eta(\underline{\pi}) = -\left[u(\underline{\pi}) + \eta \sum_{i=1}^{N} \nu^i(\pi^i)\right] \tag{10}$$

where $u^i(\underline{\pi}) = u(\underline{\pi})$ for all i and $\nu^i(\pi^i) = -\int_{A^i} p^i(x^i) \log p^i(x^i)\, \mathrm{d}x^i$ is the entropy of of a distribution π^i with density p^i. However the entropy is only well-defined when the density exists, whereas we need to study the convergence of (9) over the whole space Δ. Therefore define Δ_D to be the subset of Δ consisting of $\underline{\pi}$ such that each π^i is absolutely continuous with a density p^i which is Lipschitz continuous with constant D and satisfying $p^i(a^i) \in [D^{-1}, D]$ for all $a^i \in A^i$. For the remainder of this article, assume that D is sufficiently large that $L_\eta(\underline{\pi}) \in \Delta_D$ for all $\underline{\pi} \in \Delta$ (see Appendix C of [14]). This in turn implies both that $\mathcal{LE}_\eta \subset \Delta_D$, and that Δ_D is forward invariant under the logit best response dynamics. We will first show that any internally chain transitive set of (9) is contained in Δ_D, then use Lyapunov function arguments on Δ_D.

Start by noting that

$$\underline{\pi}(t) = \mathrm{e}^{-t}\underline{\pi}(0) + \int_0^t \mathrm{e}^{s-t} L_\eta(\underline{\pi}(s))\, \mathrm{d}s.$$

Since $L_\eta(\underline{\pi}(s)) \in \Delta_D$ for all s it is immediate that $\underline{\pi}(t)$ approaches Δ_D at an exponential rate. It then follows from techniques in the proof of Proposition 5.3 of [38] that any internally chain transitive set of the flow is contained in Δ_D.

For V_η defined in (10) to be a useful Lyapunov function, it must be continuous on Δ_D with respect to the bounded Lipschitz norm that we use on strategy space Δ. Note that u is multilinear and therefore continuous. Therefore it suffices to show that the entropy $\nu^i(\pi^i)$ is continuous in π^i. Consider π^i and $\tilde{\pi}^i$ with densities p and q respectively such that $p(x), q(x) \in [D^{-1}, D]$ for all $x \in A^i$, and both p and q are Lipschitz continuous with constant D. We calculate that

$$\begin{aligned}
|\nu(P) - \nu(Q)| &\le \int_{A^i} |p(a^i) - q(a^i)|\, |\log(p(a^i))|\, \mathrm{d}a^i \\
&\quad + \int_{A^i} q(a^i) |\log(p(a^i)) - \log(q(a^i))|\, \mathrm{d}a^i \\
&\le (\log D + D^2) \int_{A^i} |p(a^i) - q(a^i)|\, \mathrm{d}a^i,
\end{aligned}$$

since $\log$ is Lipschitz on $[D^{-1}, D]$ with constant D. That the final integral is arbitrarily small for sufficiently close π^i and $\tilde{\pi}^i$ under the BL norm follows from trivial calculations, and relies on the fact that p and q are Lipschitz.

We now show that the function V_η is strictly decreasing for any trajectory in Δ_D whenever $\underline{\pi} \notin \mathcal{LE}_\eta$. Using the Gateaux derivative, we get that

$$\dot{V}_\eta(\underline{\pi}) = \mathrm{d}V_\eta(\underline{\pi}, \dot{\underline{\pi}}) = -\sum_{i=1}^{N} \left[\mathrm{d}u((\pi^i, \pi^{-i}), \dot{\pi}^i) + \eta \mathrm{d}\nu^i(\pi^i, \dot{\pi}^i)\right]$$

Note that $\mathrm{d}u((\pi^i, \pi^{-i}), \dot{\pi}^i) = \int_{A^i} u(a^i, \pi^{-i}))\dot{\pi}^i(\mathrm{d}a^i)$, and that $\mathrm{d}\nu^i(\pi^i, \dot{\pi}^i) = -\int_{A^i} \log(p^i(a^i))\dot{\pi}^i(\mathrm{d}a^i)$ [14, equation (D.3)]. Re-arranging the definition of $l^i_\eta(\pi^{-i})$ from (1), and noting that $\int_{A^i} \dot{\pi}^i(\mathrm{d}a^i) = 0$, then yields

$$\int_{A^i} u(a^i, \pi^{-i})\dot{\pi}^i(\mathrm{d}a^i) = \eta \int_{A^i} \log(l^i_\eta(\pi^{-i})(a^i))\dot{\pi}^i(\mathrm{d}a^i).$$

Putting these together and rearranging shows that

$$\dot{V}_\eta(\underline{\pi}) = -\eta \sum_{i=1}^{N} \left\{KL(l^i_\eta(\pi^{-i}) \,\|\, p^i) + KL(p^i \,\|\, l^i_\eta(\pi^{-i}))\right\}$$

where $KL(\cdot \,\|\, \cdot)$ is the Kullback–Leibler divergence, which is non-negative and zero only when the two arguments are equal. Therefore V_η is strictly decreasing unless $p^i = l^i_\eta(\pi^{-i})$ for all i, which is exactly the condition that $\underline{\pi} \in \mathcal{LE}_\eta$.

As demonstrated by [38], the existence of a Lyapunov function is insufficient to complete the result, and Sard's theorem (as used by [32] for example) does not apply in this case, even under Smale's generalisation to Banach spaces. However since V_η is necessarily constant on connected components of $\mathcal{LE}_\eta$, Lemma 6 suffices to prove the theorem. ∎

Proof of Theorem 1: Theorems 4 and 5 combine to show that $\underline{\pi}_n$ converges to $\mathcal{LE}_\eta$ under the bounded Lipschitz norm, which is weak convergence. To establish our strong convergence claim, note that every probability measure in $\mathcal{LE}_\eta$—and hence every (weak) limit point of $\underline{\pi}_n$—is nonatomic and absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$. The result follows immediately. ∎

APPENDIX

Lemma 6. *Let $V : M \to \mathbb{R}$ be a strict Lyapunov function for some flow Φ on a metric space M. If the connected equilibrium components of Φ are isolated, and V is constant on each component, every internally chain transitive set of Φ is contained in such a component.*

Proof: Recall that an internally chain transitive set Λ is a compact, connected, invariant and attractor-free set [38]. Let $V_0 = \min\{V(x) : x \in \Lambda\}$ and $\Lambda_0 = \{x \in \Lambda : V(x) = V_0\}$. It follows that Λ_0 only consists of equilibria of V: otherwise, a trajectory $x(t)$ with $x(0) \in \Lambda_0$ would have $V(x(t)) < V_0$ $\forall t > 0$, contradicting that Λ is forward invariant.

Suppose there exists some $x \in \Lambda$ with $V(x) > V_0$. V is continuous, and constant on equilibrium components (which are isolated). So we can take $\epsilon > 0$ small enough that the closed set $\Lambda_\epsilon = \{x \in \Lambda : V(x) \leq V_0 + \epsilon\}$ is strictly larger than Λ_0 and contains no other equilibria of Φ except those in Λ_0. Since V is continuous, and is a strict Lyapunov function, all forward trajectories with $x(0) \in \Lambda_\epsilon$ will be in the interior of Λ_ϵ for all $t > 0$. Hence by [38, Lemma 5.2] Λ_ϵ, and hence Λ, contains an attractor. This is a contradiction, so we must have $V(x) = V_0$ for all $x \in \Lambda$, i.e. $\Lambda = \Lambda_0$.

■

REFERENCES

[1] D. H. Wolpert and K. Tumer, "Optimal Payoff Functions for Members of Collectives ," *Adv. Complex Syst.*, vol. 4, pp. 265–279, 2001.

[2] G. Arslan, J. R. Marden, and J. S. Shamma, "Autonomous Vehicle-Target Assignment: A Game Theoretical Formulation," *J. Dyn. Syst.-T. ASME*, vol. 129, pp. 584–596, 2007.

[3] N. Li and J. R. Marden, "Designing Games for Distributed Optimization," *IEEE J. Sel. Top. Signa.*, vol. 7, no. 2, pp. 230–242, 2013.

[4] E. Anshelevich, A. Dasgupta, J. Kleinberg, E. Tardos, T. Wexler, and T. Roughgarden, "The Price of Stability for Network Design with Fair Cost Allocation," *SIAM J. Comput.*, vol. 38, pp. 1602–1623, 2008.

[5] 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, pp. 373–396, 2009.

[6] J. R. Marden, P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," in *Conference on Decision and Control*, 2012, pp. 7419—-7424.

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

[8] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, ser. MIT Press Series on Economic Learning and Solution Evolution. Cambridge, MA: MIT Press, 1998.

[9] R. Bertin, A. Legrand, and C. Touati, "Toward a fully decentralized algorithm for multiple bag-of-tasks application scheduling on grids," in *GRID '08: Proceddings of the 3rd IEEE/ACM International Conference on Grid Computing*, 2008.

[10] F. Meshkati, A. J. Goldsmith, H. V. Poor, and S. C. Schwartz, "A game-theoretic approach to energy-efficient modulation in CDMA networks with delay QoS constraints," *IEEE J. Sel. Area Comm.*, vol. 25, pp. 1069–1078, 2007.

[11] G. Scutari, D. P. Palomar, and S. Barbarossa, "Competitive design of multiuser MIMO systems based on game theory: a unified view," *IEEE J. Sel. Area Comm.*, vol. 26, pp. 1089–1103, 2008.

[12] P. Mertikopoulos, E. V. Belmega, A. L. Moustakas, and S. Lasaulce, "Distributed learning policies for power allocation in multiple access channels," *IEEE J. Sel. Area Comm.*, vol. 30, pp. 96–106, 2012.

[13] W. Saad, Z. Han, H. V. Poor, and T. Başar, "Game-theoretic methods for the smart grid: an overview of microgrid systems, demand-side management, and smart grid communications," *IEEE Signal Proc. Mag.*, vol. 29, pp. 86–105, 2012.

[14] S. Perkins and D. S. Leslie, "Stochastic fictitious play with continuous action sets," *J. Econ. Theory*, vol. 152, pp. 179–213, 2014.

[15] R. Lahkar and F. Riedel, "The Continuous Logit Dynamic and Price Dispersion," Tech. Rep., 2013.

[16] V. S. Borkar, "Stochastic approximation with two time scales," *Syst. Control Lett.*, vol. 29, pp. 291–294, 1997.

[17] A. Shwartz and N. Berman, "Abstract stochastic approximations and applications," *Stoch. Proc. Appl.*, vol. 31, pp. 133–149, 1989.

[18] R. Cressman, "Stability of the replicator equation with continuous strategy space," *Math. Soc. Sci.*, vol. 50, pp. 127–147, 2005.

[19] J. Hofbauer, J. Oechssler, and F. Riedel, "Brown von Neumann Nash dynamics : The continuous strategy case," *Game. Econ. Behav.*, vol. 65, pp. 406–429, 2009.

[20] O. Candogan, A. Ozdaglar, and P. A Parrilo, "Near-potential games: Geometry and dynamics," *ACM T. Econ. Comput.*, vol. 1, pp. 11:1–11:32, 2013.

[21] J. Hofbauer and S. Sorin, "Best response dynamics for continuous zero-sum games," *Discrete Contin. Dyn. Syst.*, vol. 6, pp. 215–224, (2006).

[22] M. S. Stanković, K. H. Johansson, and D. M. Stipanović, "Distributed seeking of Nash equilibria with applications to mobile sensor networks," *IEEE T. Automat. Contr.*, vol. 57, pp. 904–919, 2012.

[23] S.-J. Liu and M. Krstic, "Stochastic Nash equilibrium seeking for games with general nonlinear payoffs," *SIAM J. Control Optim.*, vol. 49, pp. 1659–1679, 2011.

[24] P. Frihauf, M. Krstic, and T. Basar, "Nash equilibrium seeking in non-cooperative games," *IEEE T. Automat. Contr.*, vol. 57, pp. 1192–1207, 2012.

[25] G. Scutari, D. P. Palomar, J.-S. Pang, and F. Facchinei, "Flexible design for cognitive wireless systems: from game theory to variational inequality theory," *IEEE Signal Proc. Mag.*, vol. 26, pp. 107–123, 2009.

[26] Y. Nesterov, "Primal-dual subgradient methods for convex problems," *Math. Prog.*, vol. 120, pp. 221–259, 2009.

[27] H. Jiang and H. Xu, "Stochastic approximation approaches to the stochastic variational inequality problem," *IEEE T. Automat. Contr.*, vol. 53, pp. 1462–1475, 2008.

[28] J. Koshal, A. Nedich, and U. V. Shanbhag, "Regularized iterative stochastic approximation methods for stochastic variational inequality problems," *IEEE T. Automat. Contr.*, vol. 58, pp. 594–609, 2013.

[29] D. Fudenberg and D. M. Kreps, "Learning Mixed Equilibria," *Game. Econ. Behav.*, vol. 5, pp. 320–367, 1993.

[30] M. Benaïm and M. W. Hirsch, "Mixed equilibria and dynamical systems arising from fictitious play in perturbed games," *Game. Econ. Behav.*, vol. 29, pp. 36–72, 1999.

[31] J. Hofbauer and E. Hopkins, "Learning in Perturbed Asymmetric Games," *Game. Econ. Behav.*, vol. 52, pp. 133–152, 2005.

[32] J. Hofbauer and W. H. Sandholm, "On the global convergence of stochastic fictitious play," *Econometrica*, vol. 70, pp. 2265–2294, 2002.

[33] D. S. Leslie and E. J. Collins, "Individual Q-learning in normal form games," *SIAM J. Control Optim.*, vol. 44, pp. 495–514, 2005.

[34] R. Cominetti, E. Melo, and S. Sorin, "A payoff-based learning procedure and its application to traffic games," *Game. Econ. Behav.*, vol. 70, pp. 71–83, 2010.

[35] P. Coucheney, B. Gaujal, and P. Mertikopoulos, "Penalty-regulated dynamics and robust learning procedures in games," *Math. Oper. Res.*, vol. 40, pp. 611–633, 2015.

[36] D. S. Leslie and E. J. Collins, "Convergent Multiple-timescales Reinforcement Learning Algorithms in Normal Form Games," *Ann. Appl. Probab.*, vol. 13, pp. 1231–1251, 2003.

[37] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo Samplers," *J. Roy. Stat. Soc. B*, vol. 68, pp. 411–436, 2006.

[38] M. Benaïm, "Dynamics of stochastic approximation algorithms," *Seminaire de probabilites XXXIII*, vol. 33, pp. 1–68, 1999.

[39] S. Perkins, "Advanced Stochastic Approximation Frameworks and their Applications by," Ph.D. dissertation, University of Bristol, 2013. http://www.openthesis.org/document/view/602202_0.pdf