

GAME-THEORETICAL CONTROL WITH CONTINUOUS ACTION SETS

STEVEN PERKINS, PANAYOTIS MERTIKOPOULOS, AND DAVID S. LESLIE

ABSTRACT. Motivated by the recent applications of game-theoretical learning techniques 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, and we propose an actor-critic reinforcement learning algorithm that provably converges to equilibrium in this class of problems. The method employed is to analyse the learning process under study through a mean-field dynamical system that evolves in an infinite-dimensional function space (the space of probability distributions over the players' continuous controls). To do so, we extend the theory of finite-dimensional two-timescale stochastic approximation to an infinite-dimensional, Banach space setting, and we 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 the other agents.

CONTENTS

1. Introduction	1
2. Actor-critic learning with continuous action spaces	3
3. Two-timescales stochastic approximation in Banach spaces	7
4. Stochastic approximation of the actor-critic algorithm	12
5. Convergence of the logit best response dynamics	14
6. Discussion	18
References	18

1. INTRODUCTION

There has been much recent activity in using techniques of learning in games to design distributed control systems. This research traverses from utility function design [1–3], through analysis of potential suboptimality due to the use of distributed

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 Univ. Grenoble Alpes, LIG, F-38000 Grenoble, France.

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

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 of [1, 2] 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”, and each player is assigned a utility function which depends on the actions of all players (as does the global system-level utility). As such, a player’s utility in a particular play of the game could be set to be the global utility of the joint action selected by all players. However, a more learnable choice is the so-called Wonderful Life Utility (WLU) [1, 2], in which the utility of any particular player is given by how much better the system is doing as a result of that player’s action (compared to the situation where no other player changes their action but the focal player uses a baseline action instead). A fundamental result in this domain is that setting the players’ utilities using WLUs results in a potential game [7] (see Section 2 below). There are alternative methods for converting a system-level utility function into individual utilities, such as Shapley value utility [8]; however, most of these also boil down to a potential game (possibly in the extended sense of [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 of the resulting game.

On the other hand, like much of the economic literature on learning in games [9, 10], the vast majority of this corpus of research has focused almost exclusively on situations where each player’s controls comprise a *finite* set. This allows results from the theory of learning in games to be applied directly, resulting in learning algorithms that converge to the set of equilibria – and hence system optima. However, the assumption of discrete action sets is frequently anomalous in control, engineering and economics: after all, prices are not discrete, and neither are the controls in a large number of engineering systems. For instance, in massively parallel grid computing networks (such as the Berkeley Open Infrastructure for Network Computing – BOINC) [11], the decision granularity of “bag-of-tasks” application scheduling gives rise to a potential game with continuous action sets [7]. A similar situation is encountered in the case of energy-efficient power control and power allocation in large wireless networks [12, 13]: mobile wireless users can transmit at different power levels (or split their power across different subcarriers [14]), and their throughput is a continuous function of their chosen transmit power profiles (which have to be optimized unilaterally and without recourse to user coordination or cooperation). Finally, decision-making in the emerging “smart grid” paradigm for power generation and management in electricity grids also revolves around continuous variables (such as the amount of power to generate, or when to power down during the day), leading again to game-theoretical model formulations with continuous action sets [15].

In this paper, we focus squarely on control problems (presented as potential games) with *continuous* action sets and we propose an actor-critic reinforcement learning algorithm that provably converges to equilibrium. To address this problem in an economic setting, very recent work by Perkins and Leslie [16] extended the theory of learning in games to zero-sum games with continuous action sets (see

also [17, 18]); however, from a control-theoretical point of view, zero-sum games are of limited practical relevance because they only capture adversarial interactions between two players. Owing to this fundamental difference between zero-sum and potential games, the two-player analysis of [16] no longer applies to our case, so a completely different approach is required to obtain convergence in the context of many-player potential games.

To accomplish this, our analysis relies on two theoretical contributions of independent interest. The first is the extension of stochastic approximation techniques for Banach spaces (otherwise known as “abstract stochastic approximation” [19–24]) to the so-called “two-timescales” framework originally introduced in standard (finite-dimensional space) stochastic approximation by [25]. This allows us to consider interdependent strategies and value functions evolving as a stochastic process in a Banach space (the space of signed measures over the players’ continuous action sets and the space of continuous functions from action space to \mathbb{R} respectively, both endowed with appropriate norms). Our second contribution is the asymptotic analysis of the mean field dynamics of this process on the space of probability measures on the action space; our analysis reveals that the dynamics’ rest points in potential games are globally attracting, so, combined with our stochastic approximation results, we obtain the convergence of our actor-critic reinforcement learning algorithm to equilibrium.

In Section 2 we introduce the framework and notation, and introduce our actor-critic learning algorithm. Following that, in Section 3 we introduce two-timescales stochastic approximation in Banach spaces, and prove our general result. Section 4 applies the stochastic approximation theory to the actor-critic algorithm to show that it can be studied via a mean field dynamical system. Section 5 then analyses the convergence of the mean field dynamical system in potential games, a result which allows us to prove the convergence of the actor-critic process in this context.

2. ACTOR-CRITIC LEARNING WITH CONTINUOUS ACTION SPACES

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 will frequently 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 the joint 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 $u^i : \underline{A} \rightarrow \mathbb{R}$. For the game to be a potential game, there must exist a potential function $\phi : \underline{A} \rightarrow \mathbb{R}$ such that

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

for all $i \in \{1, \dots, N\}$, for all a^{-i} and for all a^i, \tilde{a}^i . Thus if any player changes their action while the others do not, the change in utility for the player that changes their action is equal to the change in value of the potential function of the game. Methods for constructing potential games from system utility functions [1–3] usually ensure

¹We are only making this assumption for convenience; our analysis carries through to higher-dimensional convex bodies with minimal hassle.

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 . This is a simple concept when A^i is finite, but for the continuous action spaces A^i considered in this paper more care is required. 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 . Throughout this article we endow $\mathcal{P}(A^i, \mathcal{B}^i)$ with the weak topology, metrized by the bounded Lipschitz norm (see Section 4; also [16, 26, 27]). A mixed strategy is then an element $\pi^i \in \mathcal{P}(A^i, \mathcal{B}^i)$; for $B^i \in \mathcal{B}^i$ we have that $\pi^i(B^i)$ is the probability that Player i selects an action in the Borel set B^i . Note that a mixed strategy under this definition 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 definition of utilities to the space $\Delta = \prod_{i=1}^N \mathcal{P}(A^i, \mathcal{B}^i)$ of mixed strategy profiles. In particular, let $\underline{\pi} \in \Delta$ be a mixed strategy profile, and define

$$u^i(\underline{\pi}) = \int_{A^1} \cdots \int_{A^N} u^i(\underline{a}) \pi^1(da^1) \cdots \pi^N(da^N).$$

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 $\pi^{-i} = (\pi^1, \dots, \pi^{i-1}, \pi^{i+1}, \dots, \pi^N)$. In further abuse of notation, we write (a^i, π^{-i}) for the mixed strategy profile (δ_{a^i}, π^{-i}) , where δ_{a^i} is the Dirac measure at a^i (meaning that Player i selects action a^i with probability 1). 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 , i.e. the set of mixed strategies that maximise Player i 's utility given any particular opponent mixed strategy π^{-i} . A Nash equilibrium is a fixed point of this correspondence, in which all players are playing a best response to all other players. In a learning context however, the discontinuities that appear in best response correspondences can cause great difficulties [28]. 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 to be the mixed strategy $L_\eta^i(\pi^{-i}) \in \mathcal{P}(A^i, \mathcal{B}^i)$ such that

$$L_\eta^i(\pi^{-i})(B^i) = \frac{\int_{B^i} \exp\{\eta^{-1} u^i(a^i, \pi^{-i})\} da^i}{\int_{A^i} \exp\{\eta^{-1} u^i(b^i, \pi^{-i})\} db^i} \quad (1)$$

for each $B^i \in \mathcal{B}^i$. In [18] it is shown that $L_\eta^i(\underline{\pi}^{-i}) \in \mathcal{P}(A^i, \mathcal{B}^i)$ is absolutely continuous (with respect to Lebesgue measure), with density given by

$$l_\eta^i(\underline{\pi}^{-i})(a^i) = \frac{\exp\{\eta^{-1} u^i(a^i, \pi^{-i})\}}{\int_{A^i} \exp\{\eta^{-1} u^i(b^i, \pi^{-i})\} db^i}. \quad (2)$$

To ease notation in what follows, we let $L_\eta(\underline{\pi}) = (L_\eta^1(\pi^{-1}), \dots, L_\eta^N(\pi^{-N}))$.

The existence of fixed points of L_η is shown in [18] and [16]; such a fixed point is a joint strategy $\underline{\pi}$ such that $\pi^i = L_\eta^i(\pi^{-i})$ for each i , and so is a mixed strategy profile such that every player is playing a smooth best response to the strategies of the other players. Such profiles $\underline{\pi}$ are called *logit equilibria* and the set of all such fixed points will be denoted by \mathcal{LE}_η . A logit equilibrium is thus an approximation

of a local maximizer of the potential function of the game in the sense that for small η a logit equilibrium places most of the probability mass in areas where the joint action results in a high potential function value; in particular, logit equilibria approximate Nash equilibria when the noise level is sufficiently small.²

Smooth best responses also play an important part in discrete action games, particularly when learning is considered. In this domain they were introduced in stochastic fictitious play by [30], and later studied by, among others, [31–33] to ensure the played mixed strategies in a fictitious play process converge to logit equilibrium. This is in contrast to classical fictitious play in which the beliefs of players converge, but the played strategies are (almost) always pure. The technique was also required by [34–36] to allow simple reinforcement learners to converge to logit equilibria: as discussed in [34], players whose strategies are a function of the expected value of their actions cannot converge to a Nash equilibrium because, at equilibrium, all actions in the support of the equilibrium mixed strategies will receive the same expected reward.

Recently [18] developed the dynamical systems tools necessary to consider whether the smooth best response dynamics converge to logit equilibria in the infinite-dimensional setting. This was extended to learning systems in [16], where it was shown that stochastic fictitious play converges to logit equilibrium in two-player zero-sum games with compact continuous action sets.

One of the main requirements for efficient learning in a control setting is that the full utility functions of the game need not be known in advance, and players may not be able to observe the actions of all other players. Using fictitious play (or, indeed, many of the other standard game-theoretical tools) does not satisfy this requirement because they assume full knowledge and observability of payoff functions and opponent actions. This is what motivates the simple reinforcement learning approaches discussed previously [34–36], and also the actor-critic reinforcement learning approach of [37], which we extend in this article to the continuous action space setting. The idea is to learn both a value function $Q^i : A^i \rightarrow \mathbb{R}$ 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 .

In the continuous action space setting of this paper, we implement the actor-critic algorithm as the following iterative process (for a pseudo-code implementation, see Algorithm 1):

- (1) At the n -th stage of the process, each player $i = 1, \dots, N$ selects an action a_n^i by sampling from the distribution π_n^i and uses a_n^i to play the game.
- (2) Players update their critics using the update equation

$$Q_{n+1}^i = Q_n^i + \gamma_n \cdot (u^i(\cdot, a_n^{-i}) - Q_n^i) \tag{3a}$$

- (3) Each player samples $b_n^i \sim L_\eta^i(Q_n^i)$ and updates their actor using the update equation

$$\pi_{n+1}^i = \pi_n^i + \alpha_n \cdot (\delta_{b_n^i} - \pi_n^i). \tag{3b}$$

The algorithm above is the main focus of our paper, so some remarks are in order:

²We note here that the notion of a logit equilibrium is a special case of the more general concept of *quantal response equilibrium* introduced in [29].

Algorithm 1 Actor-critic Reinforcement Learning Based on Logit Best ResponsesParameters: step-size sequences α_n, γ_n .Initialize critics Q^i , actors π^i ; $n \leftarrow 0$.**Repeat**

```

   $n \leftarrow n + 1$ ;
  foreach player  $i = 1, \dots, N$  do
    select action  $a^i$  based on actor  $\pi^i$ ;                                #play the game
    update critic:  $Q^i \leftarrow Q^i + \gamma_n(u^i(a_1, \dots, a_N) - Q^i)$ ;    #update payoff estimates
    draw sample  $b^i \sim L_{\eta}^i(Q^i)$ ;                                       #sample logit best response
    update actor:  $\pi^i \leftarrow \pi^i + \alpha_n(\delta_{b^i} - \pi^i)$ ;           #update mixed strategies
  until termination criterion is reached.

```

Remark 1. In (3a), it is assumed that a player can access $u^i(\cdot, a_n^{-i})$, so they can calculate how much they would have received for each of their actions in response to the joint action that was selected by the other players. Even though this assumption restricts the applicability of our method somewhat, it is relatively harmless in many settings — for instance, in congestion games such estimates can be calculated simply by observing the utilization level of the system’s facilities. Note further that to implement this algorithm an individual need not actually observe the action profile a_n^{-i} , needing only the utility $u^i(\cdot, a_n^{-i})$. This means that a player need know nothing at all about the players who don’t directly affect her utility function, which allows a degree of separation and modularisation in large systems, as demonstrated in [38].

Remark 2. The logit response L_{η}^i used to sample the b_n^i used in (3b) is now parameterised by Q_n^i instead of π^{-i} . This is a trivial change in which we use $Q^i(\cdot)$ in place of $u^i(\cdot, \pi^{-i})$ in (1), which represents the fact that now players select smooth best responses to their critic Q^i instead of directly to the estimated mixed strategy of the other players.

Remark 3. Also in (3b), the players update towards a sampled b_n^i instead of toward the full function $L_{\eta}^i(Q_n^i)$. This is so that the critic π_n^i can be represented as a collection of weighted atoms, instead of as a complicated and continuous probability measure. Representing π_n^i as a collection of atoms means that sampling $a_n^i \sim \pi_n^i$ is particularly easy.

On the other hand, sampling $b_n^i \sim L_{\eta}^i(Q_n^i)$ could be extremely difficult for general Q_n^i . The gradual evolution of the Q_n^i however implies that a sequential Monte Carlo sampler [39] could be used to produce samples according to $L_{\eta}^i(Q_n^i)$. The representation of Q_n^i is also potentially troublesome and we do not address it fully here. However one could assume that each $u^i(a_n^i)$ can be represented as a finite linear combination of basis functions such as a spline, Fourier or wavelet basis. Another option 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.

Remark 4. Finally, we note that the updates (3a) and (3b) use different step size parameters α_n and γ_n . This separation is what allows the algorithm to be a two-timescales procedure, and is discussed at the start of Section 3.

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 (3) converges strongly to a component of the equilibrium set \mathcal{LE}_η (a.s.).*

Remark. We recall here that the notion of strong convergence of probability measures $\underline{\pi}_n \rightarrow \underline{\pi}^*$ is defined by asking that $\underline{\pi}_n(A) \rightarrow \underline{\pi}^*(A)$ for every measurable A . As such, this notion of convergence is even stronger than the notion of “convergence in probability” (vague convergence) used in the central limit theorem and other weak-convergence results.

3. TWO-TIMESCALES STOCHASTIC APPROXIMATION IN BANACH SPACES

The analysis of systems such as Algorithm 1 is enabled by the use of two-timescales stochastic approximation techniques [25]. By allowing $\alpha_n/\gamma_n \rightarrow 0$ as $n \rightarrow \infty$, the system can be analysed as if the ‘fast’ update (3a), with higher learning parameter γ_n , has fully converged to the current value of the ‘slow’ system (3b), with lower learning parameter α_n . Note that it is not the case that we have an outer and inner loop, in which (3a) is run to convergence for every update of (3b): both the actor Q_n and the critic π_n are updated on 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 [25] 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 [25] with the Banach space stochastic approximation framework of [16] to develop the tool necessary to analyse the recursion (3).

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

$$x_{n+1} = x_n + \alpha_{n+1} [F(x_n, y_n) + U_{n+1} + c_{n+1}], \quad (4a)$$

$$y_{n+1} = y_n + \gamma_{n+1} [G(x_n, y_n) + V_{n+1} + d_{n+1}], \quad (4b)$$

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 of the process.
- $F : X \times Y \rightarrow X$ and $G : X \times Y \rightarrow Y$ comprise the *mean field* of the process.
- $\{U_n\}$ and $\{V_n\}$ are stochastic processes in X and Y respectively. (For a detailed exposition of Banach-valued random variables, see [40].)
- $c_n \in X$ and $d_n \in Y$ are bias terms that converge almost surely to 0.

We will study this system using the asymptotic pseudotrajectory approach of [41], which is already cast in the language of metric spaces; since Banach spaces are metric, the framework of [41] still applies to our scenario. This modernises the approach of [22] while also introducing the two-timescales technique to ‘abstract stochastic approximation’.

To proceed, recall that a semiflow Φ on a metric space, M , is a continuous map $\Phi : \mathbb{R}^+ \times M \rightarrow M$, $(t, x) \mapsto \Phi_t(x)$, such that, $\Phi_0(x) = x$ and $\Phi_{t+s}(x) = \Phi_t(\Phi_s(x))$ for all $t, s \geq 0$. As in simple Euclidean spaces, well-posed differential equations on

Banach spaces induce a semiflow [42]. A continuous function $z : \mathbb{R}^+ \rightarrow M$ is an asymptotic pseudo-trajectory for Φ if for any $T > 0$,

$$\lim_{t \rightarrow \infty} \sup_{0 \leq s \leq T} d(z(t+s), \Phi_s(x(t))) = 0.$$

Properties of asymptotic pseudo-trajectories are discussed in detail in [41].

We will prove that interpolations of the stochastic approximation process (4) result in asymptotic pseudotrajectories to flows induced by dynamical systems on X and Y governed by F and G respectively. To do so, and to allow us to state necessary assumptions on the processes, we define timescales on which we will interpolate the stochastic approximation process. In particular, let $\tau_n^\alpha = \sum_{j=1}^n \alpha_j$ (with $\tau_0^\alpha = 0$), and for $t \in \mathbb{R}_+$ let $m^\alpha(t) = \sup\{k \geq 0; \tau_k^\alpha \leq t\}$. Similarly let $\tau_n^\gamma = \sum_{j=1}^n \gamma_j$ (with $\tau_0^\gamma = 0$), and for $t \in \mathbb{R}_+$ let $m^\gamma(t) = \sup\{k \geq 0; \tau_k^\gamma \leq t\}$.

With these timescales we define interpolations of the stochastic approximation processes (4). On the slow (α) timescale we define a continuous-time interpolation $\bar{x}^\alpha : \mathbb{R}_+ \rightarrow X$ of $\{x_n\}_{n \in \mathbb{N}}$ by letting

$$\bar{x}^\alpha(\tau_n^\alpha + s) = x_n + s \frac{x_{n+1} - x_n}{\alpha_{n+1}} \quad (5)$$

for $s \in [0, \alpha_{n+1})$. On the fast (γ) timescale we consider $z_n = (x_n, y_n) \in X \times Y$, and define the continuous time interpolation $\bar{z}^\gamma : \mathbb{R}_+ \rightarrow X \times Y$ of $\{z_n\}_{n \in \mathbb{N}}$ by letting

$$\bar{z}^\gamma(\tau_n^\gamma + s) = z_n + s \frac{z_{n+1} - z_n}{\gamma_{n+1}} \quad (6)$$

for $s \in [0, \gamma_{n+1})$.

Our assumptions, which are simple extensions to those of [25] and [41], can now be stated as follows:

A1) Noise control.

(a) For all $T > 0$,

$$\lim_{n \rightarrow \infty} \sup_{k \in \{n+1, \dots, m^\alpha(\tau_n^\alpha + T)\}} \left\{ \left\| \sum_{j=n}^{k-1} \alpha_{j+1} U_{j+1} \right\|_X \right\} = 0,$$

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

(b) $\{c_n\}_{n \in \mathbb{N}}$ and $\{d_n\}_{n \in \mathbb{N}}$ are bounded sequences such that $\|c_n\|_X \rightarrow 0$ and $\|d_n\|_Y \rightarrow 0$ as $n \rightarrow \infty$.

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 $\alpha_n \rightarrow 0$ and $\gamma_n \rightarrow 0$ as $n \rightarrow \infty$.

(b) $\alpha_n / \gamma_n \rightarrow 0$ as $n \rightarrow \infty$.

A4) Mean field behaviour.

(a) For any fixed $\tilde{x} \in C$ the differential equation

$$\frac{dy}{dt} = G(\tilde{x}, y) \quad (7)$$

has unique solution trajectories that remain in D for any initial value $y_0 \in D$. Furthermore the differential equation (7) has a unique globally attracting fixed point $y^*(\tilde{x})$, and the function $y^* : C \rightarrow D$ is Lipschitz continuous.

(b) The differential equation

$$\frac{dx}{dt} = F(x, y^*(x)) \tag{8}$$

has unique solution trajectories that remain in C for any initial value $x_0 \in C$.

Assumption A1 is the standard assumption for noise control in stochastic approximation. It has traditionally caused difficulty in abstract stochastic approximation, but recent solutions are discussed in the following paragraph. Assumption A2 is simply a boundedness and continuity assumption, but can cause difficulty with some norms in function spaces. Assumption A3 provides the two-timescales nature of the scheme, with both learning rate sequences converging to 0, but α_n becoming much smaller 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 [25]. Note that we do not make the stronger assumption that there exists a unique globally asymptotically stable fixed point in the slow timescale dynamics (8) [25, Assumption A2]; this assumption is not necessary for the theory presented here, and would unnecessarily restrict the applicability of the results.

Note that the noise assumption A1(a) has traditionally caused difficulty for stochastic approximation on Banach spaces: [23] considers the simple case where the stochastic terms are independent and identically distributed, whilst [22] prove a very weak convergence result for a particular process which again uses independent noise. However [16] provide criteria analogous to the martingale noise assumptions in \mathbb{R}^K which guarantee that the noise condition 1(a) holds in useful Banach spaces. In particular, if $\{U_n\}$ is a sequence of martingale differences in Banach space X , then

$$\lim_{n \rightarrow \infty} \sup_{k \in \{n+1, \dots, m^\alpha(\tau_n^\alpha + T)\}} \left\{ \left\| \sum_{j=n}^{k-1} \alpha_{j+1} U_{j+1} \right\|_X \right\} = 0$$

with probability 1 if X is:

- the space of L^p functions for $p \geq 2$, $\{\alpha_n\}_{n \in \mathbb{N}}$ is deterministic with $\sum_{n \in \mathbb{N}} \alpha_n^{1+q/2} < \infty$, $\{U_n\}_{n \in \mathbb{N}}$ is a martingale difference sequence with respect to some filtration $\{\mathcal{F}_n\}_{n \in \mathbb{N}}$, and $\sup_{n \in \mathbb{N}} \mathbb{E}[\|U_n\|_{L^p}^q] < \infty$ (cf. the remark following Proposition A.1 of [16]);
- the space of L^1 functions on bounded spaces (see [43]); or
- the space of finite signed measures on a compact interval of \mathbb{R} with the bounded Lipschitz norm (see [16, 26, 27] or Section 4 below) $\{\alpha_n\}_{n \in \mathbb{N}}$ is deterministic with $\sum_{n \in \mathbb{N}} \alpha_n^2 < \infty$, $U_n = \delta_{x_{n+1}} - P_n$ where there exists a filtration $\{\mathcal{F}_n\}_{n \in \mathbb{N}}$ such that U_n is measurable with respect to \mathcal{F}_n , P_n is a bounded absolutely continuous probability measure which is measurable with respect to \mathcal{F}_n and has density p_n , and x_{n+1} is sampled from the probability distribution P_n (Proposition 3.6 of [16]);

Clearly, if similar conditions also hold for Y then Assumption A1(a) holds.

Our first lemma demonstrates that we can analyse the system as if the fast system $\{y_n\}$ is fully calibrated to the slow system $\{x_n\}$. By this we mean that, for sufficiently large n , y_n is close to the value it would converge to if x_n were fixed and y_n allowed to fully converge.

Lemma 2. *Under Assumptions A1–A4,*

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

Proof. 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 (4) can be expressed as

$$z_{n+1} = z_n + \gamma_{n+1} \left[H(z_n) + W_{n+1} + \kappa_{n+1} \right], \quad (9)$$

where $H : Z \rightarrow Z$ is such that $H(z_n) = (0, G(z_n))$, for $0 \in X$, and

$$W_n = \begin{pmatrix} \frac{\alpha_n}{\gamma_n} U_n, V_n \end{pmatrix},$$

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

Assumptions A1–A4 imply the assumptions of Theorem 3.3 of [16]. Most are direct translations, but the noise must be carefully considered. For any $n \in \mathbb{N}$, any $T > 0$, and any $k \in \{n+1, \dots, m^\gamma(\tau_n^\gamma + T)\}$,

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

Since $\kappa_n \rightarrow 0$, the second term converges to 0 as $n \rightarrow \infty$. Hence, using assumption A1 to control the first term,

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

Therefore $\bar{z}^\gamma(\cdot) : \mathbb{R}_+ \rightarrow X \times Y$, defined in (6), is an asymptotic pseudotrajectory of the flow defined by

$$\frac{dz}{dt} = H(z(t)). \quad (10)$$

Assumption A4(a) implies that $\{(x, y^*(x)) : x \in C\}$ is globally attracting for (10). Hence Theorem 6.10 of [41] gives that $z_n \rightarrow \{(x, y^*(x)) : x \in C\}$. The result follows by the continuity of y^* assumed in A4(a). \square

We use this fact to consider the evolution of x_n on the slow timescale.

Theorem 3. *Under Assumptions A1–A4, the interpolation $\bar{x}^\alpha(\cdot) : \mathbb{R}_+ \rightarrow X$, defined in (5), is an asymptotic pseudo-trajectory to the flow induced by the differential equation (8).*

Proof. Rewrite (4a) as

$$x_{n+1} = x_n + \alpha_{n+1} \left[F(x_n, y^*(x_n)) + U_{n+1} + \tilde{c}_{n+1} \right], \quad (11)$$

where $\tilde{c}_{n+1} = F(x_n, y_n) - F(x_n, y^*(x_n)) + c_{n+1}$. We will show that this is a well-behaved stochastic approximation process. In particular, we need to show that \tilde{c}_n can be absorbed into U_n in such a way that the equivalent Assumption A1 of [16] can be applied to $U_n + \tilde{c}_n$.

By Lemma 2 we have that $\|y_n - y^*(x_n)\|_Y \rightarrow 0$. Hence we can define

$$\delta_n = \inf\{\delta > 0 : \forall m \geq n, \|y_m - y^*(x_m)\|_Y < \delta\}$$

with $\delta_n \rightarrow 0$ as $n \rightarrow \infty$. By the uniform continuity of F , it follows that we can define a sequence $\varepsilon_n \rightarrow 0$ such that for all $m \geq n$, $\|F(x_m, y_m) - F(x_m, y^*(x_m))\|_X < \varepsilon_n$.

From this construction, for any $n \geq 0$ and for any $k \in \{n+1, \dots, m^\alpha(\tau_n^\alpha + T)\}$,

$$\begin{aligned} & \left\| \sum_{j=n}^{k-1} \alpha_{j+1} \left[F(x_n, y_n) - F(x_n, y^*(x_n)) \right] \right\|_X \\ & \leq \left\| \sum_{j=n}^{k-1} \alpha_{j+1} \varepsilon_n \right\|_X \\ & \leq T \varepsilon_n. \end{aligned}$$

As in the proof of Lemma 2, similar arguments can be used for $\{c_n\}_{n \in \mathbb{N}}$ under assumption (A1)(b). Hence for all $T > 0$,

$$\lim_{n \rightarrow \infty} \sup_{k \in \{n+1, \dots, m^\alpha(\tau_n^\alpha + T)\}} \left\{ \left\| \sum_{j=n}^{k-1} \alpha_{j+1} \tilde{c}_{j+1} \right\|_X \right\} = 0.$$

Once again it is straightforward to show that, under (A1)–(A4), the slow timescale stochastic approximation (11) satisfies the assumptions of Theorem 3.3 of [16], and therefore $\bar{x}(\cdot) : \mathbb{R}^+ \rightarrow X$ is an asymptotic pseudo-trajectory to the flow induced by the differential equation (8). \square

While [41] provides several results that can be combined with Theorem 3, we summarise the result used in this paper with the following corollary:

Corollary 4. *Suppose that Assumptions A1–A4 hold. Then x_n converges to an internally chain transitive set of the flow induced by the mean field differential equation (8).*

Proof. This is an immediate consequence of Theorem 5 above and Theorem 5.7 of [41], where the definition of internally chain transitive sets can be found. \square

4. STOCHASTIC APPROXIMATION OF THE ACTOR–CRITIC ALGORITHM

In this section we demonstrate that the actor–critic algorithm (3) can be analysed using the two-timescales stochastic approximation framework of Section 3. 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.³ To define the BL norm, let

$$G^i = \{g : A^i \rightarrow \mathbb{R} : \sup_{a \in A^i} |g(a)| + \sup_{a, b \in A^i, a \neq b} \frac{|g(a) - g(b)|}{|a - b|} \leq 1\}.$$

Then, for $\mu \in M(A^i, \mathcal{B}^i)$ we define

$$\|\mu\|_{BL^i} = \sup_{g \in G^i} \left| \int_{A^i} g(d\mu) \right|.$$

$\mathcal{M}(A^i, \mathcal{B}^i)$ with norm $\|\cdot\|_{BL^i}$ is a Banach space [27], and convergence of a sequence of probability measures under $\|\cdot\|_{BL^i}$ corresponds to weak convergence of the measures [26]. Under the BL norm, $\mathcal{P}(A^i, \mathcal{B}^i)$ is a compact subset of $\mathcal{M}(A^i, \mathcal{B}^i)$ (see Proposition 4.6 of [16]), allowing Assumption A2 to be easily verified.

We consider mixed strategy profiles as existing in the subset Δ of the product space $\Sigma = \mathcal{M}(A^1, \mathcal{B}^1) \times \cdots \times \mathcal{M}(A^N, \mathcal{B}^N)$. We use the max norm to induce the product topology, so that if $\mu = (\mu^1, \dots, \mu^N) \in \Sigma$ we define

$$\|\mu\|_{BL} = \max_{i=1, \dots, N} \|\mu^i\|_{BL^i}. \quad (12)$$

Suppose also that utility functions u^i are bounded and Lipschitz continuous. Since their domain is a bounded interval of \mathbb{R} we can assume that the estimates Q_n^i are in the Banach space $L^2(A^i)$ of functions $A^i \rightarrow \mathbb{R}$ with a finite L^2 norm, under the 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 5. *Consider the actor–critic algorithm (3). 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*

$$\frac{d\underline{\pi}}{dt} = L_\eta(\underline{\pi}) - \underline{\pi}. \quad (13)$$

³For a discussion regarding the appropriateness of this norm for game-theoretical considerations, see [18, 26, 27], and, for stochastic approximation, especially [16].

Proof. We take $(X, \|\cdot\|_X) = (\Sigma, \|\cdot\|_{BL})$, and $(Y, \|\cdot\|_Y)$ as above. This allows a direct mapping of the actor-critic algorithm (3) to the stochastic approximation framework (4) by taking

$$\begin{aligned} x_n &= \underline{\pi}_n, \\ F(\underline{\pi}, \underline{Q}) &= L_\eta(\underline{Q}) - \underline{\pi}, \\ U_{n+1} &= (\delta_{b_n^1}, \dots, \delta_{b_n^N}) - L_\eta(\underline{Q}), \\ c_n &= 0 \end{aligned}$$

and

$$\begin{aligned} y_n &= \underline{Q}_n, \\ G(\underline{\pi}, \underline{Q}) &= (G^1(\underline{\pi}, \underline{Q}), \dots, G^N(\underline{\pi}, \underline{Q})), \\ G^i(\underline{\pi}, \underline{Q}) &= u^i(\cdot, \pi^{-i}) - Q^i, \\ V_{n+1} &= (V_{n+1}^1, \dots, V_{n+1}^N), \\ V_{n+1}^i &= u^i(\cdot, a_n^{-i}) - u^i(\cdot, \pi_n^{-i}), \\ d_n &= 0. \end{aligned}$$

By Corollary 4 we therefore only need to verify Assumptions A1–A4.

A1: U_n is of exactly the form studied by [16] 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 [16] 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, 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 [16].

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 exponentially quickly 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 (8) then becomes

$$\dot{\tilde{\pi}}^i = L_\eta^i(u^i(\cdot, \pi^{-i})) - \tilde{\pi}^i, \quad \text{for } i = 1, \dots, N.$$

Since we re-wrote L_η^i to depend on the utility functions instead of directly on π^{-i} , we find that we have recovered the logit best response dynamics of [18] and [16], which those authors show to have unique solution trajectories. \square

5. CONVERGENCE OF THE LOGIT BEST RESPONSE DYNAMICS

We have shown in Theorem 5 that the actor–critic algorithm (3) results in joint strategies $\{\pi_n\}_{n \in \mathbb{N}}$ that converge to an internally chain transitive set of the flow defined by the logit best response dynamics (13) under the bounded Lipschitz norm. It is demonstrated in [16] that in two-player zero-sum continuous action games the set \mathcal{LE}_η of logit equilibria (the fixed points of the logit best response L_η) is a global attractor of the flow. Hence, by Corollary 5.4 of [41] we instantly obtain the result that any internally chain transitive set is contained in \mathcal{LE}_η .

However two-player zero-sum games are not particularly relevant for control systems: multiplayer potential games are much more important. The logit best responses in a potential game are identical to the logit best responses in the identical interest game in which the potential function is the global utility function. Hence 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 (13) in N -player identical interest games with continuous action spaces. See [44] for related issues.

For the remainder of this section we work to prove the following theorem:

Theorem 6. *In a potential game with continuous bounded rewards, in which the connected components of the set \mathcal{LE}_η of logit equilibria of the game are isolated, any internally chain transitive set of the flow induced by the smooth best response dynamics (13) is contained in a connected component of \mathcal{LE}_η .*

Define

$$\Delta_D = \left\{ \underline{\pi} \in \Delta : \begin{array}{l} \forall i = 1, \dots, N, \pi^i \text{ is absolutely contin-} \\ \text{uous with density } p^i \text{ such that } D^{-1} \leq \\ p^i(x^i) \leq D \text{ for all } x^i \in A^i \text{ and } p^i \text{ is} \\ \text{Lipschitz continuous with constant } D \end{array} \right\}.$$

Appendix C of [16] shows that if the utility functions u^i are bounded and Lipschitz continuous then, for any $\eta > 0$, there exists a D such that $L_\eta(\underline{\pi}) \in \Delta_D$ for all $\underline{\pi} \in \Delta$, and that Δ_D is forward invariant under the logit best response dynamics. For the remainder of this article, D is taken to be sufficiently large for this to be the case.

Our method first demonstrates that the set Δ_D is globally attracting for the flow, so any internally chain transitive set of the flow is contained in Δ_D . The nice properties of Δ_D then allow the use of a Lyapunov function argument to show that any internally chain transitive set in Δ_D is a connected set of logit equilibria.

Lemma 7. *Let $\Lambda \subset \Delta$ be an internally chain-transitive set. Then $\Lambda \subset \Delta_D$.*

Proof. Consider the trajectory of (13) starting at an arbitrary $\underline{\pi}(0) \in \Delta$. We can write $\underline{\pi}(t)$ as

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

Defining

$$\underline{\sigma}(t) = \frac{\int_0^t e^{s-t} L_\eta(\underline{\pi}(s)) ds}{1 - e^{-t}}$$

it is immediate both that $\underline{\sigma}(t) \in \Delta_D$ and

$$\|\underline{\pi}(t) - \underline{\sigma}(t)\|_{BL} < 2e^{-t}. \quad (14)$$

Thus $\underline{\pi}(t)$ approaches Δ_D at an exponential rate, uniformly in $\underline{\pi}(0)$. Hence Δ_D is uniformly globally attracting.

We would like to invoke Corollary 5.4 of [41], but since Δ_D may not be invariant it is not an attractor in the terminology of [41] either. We therefore prove directly that $\Lambda \subset \Delta_D$. Suppose not, so there exists a point $p \in \Lambda \setminus \Delta_D$ and by the compactness of internally chain transitive sets there exists a $\delta > 0$ such that $\inf_{\pi \in \Delta_D} \|p - \pi\| = 2\delta$. There exists a $T > 0$ such that for the trajectory $p(t)$ with $p(0) = p$, $\inf_{\pi \in \Delta_D} \|p(T) - \pi\| < \delta$, and so $\|p(T) - p\| > \delta$. Hence, as in the proof of Proposition 5.3 of [41], p cannot be part of an internally chain recurrent set (see [41]). Since internally chain transitive sets are internally chain recurrent sets [41, Proposition 5.3] we have a contradiction. Hence $\Lambda \subset \Delta_D$. \square

We are now left to find the internally chain transitive sets of the flow restricted to Δ_D . Since all elements of Δ_D admit densities, we can define a Lyapunov function based on the densities of the mixed strategies. For an absolutely continuous mixed strategy π^i with density function p^i , we define the entropy

$$\nu^i(\pi^i) = - \int_{A^i} p(x^i) \log p(x^i) dx^i.$$

The Lyapunov function to be considered is

$$V_\eta(\underline{\pi}) = - \left[u(\underline{\pi}) + \eta \sum_{i=1}^N \nu^i(\pi^i) \right] \quad (15)$$

where $u^i(\underline{\pi}) = u(\underline{\pi})$ for all i . For V_η to be a useful Lyapunov function, it must be continuous with respect to the bounded Lipschitz norm that we use on strategy space.

Lemma 8. $V_\eta : \Delta_D \rightarrow \mathbb{R}$ is continuous with respect to the bounded Lipschitz norm.

Proof. Note that u is multilinear and therefore continuous. Therefore it suffices to show that the entropy $\nu(\pi^i)$ is continuous in π^i .

Consider two densities p and q corresponding to distributions P and Q on a finite interval $A \subset \mathbb{R}$, and assume that $p(x), q(x) \in [D^{-1}, D]$ for all $x \in A$, and both p and q are Lipschitz continuous with constant D . We calculate that

$$\begin{aligned} |\nu(P) - \nu(Q)| &= \left| \int_A p(x) \log(p(x)) - q(x) \log(q(x)) dx \right| \\ &\leq \int_A |p(x) - q(x)| |\log(p(x))| dx \\ &\quad + \int_A q(x) |\log(p(x)) - \log(q(x))| dx \\ &\leq \log(D) \int_A |p(x) - q(x)| dx \\ &\quad + D \int_A |\log(p(x)) - \log(q(x))| dx, \end{aligned}$$

since both $p(x)$ and $q(x)$ are uniformly bounded above by D . Furthermore, since \log is Lipschitz on $[D^{-1}, D]$ with constant D , $|\log(p(x)) - \log(q(x))| \leq D|p(x) - q(x)|$. We therefore see that

$$|\nu(P) - \nu(Q)| \leq (\log D + D^2) \int_A |p(x) - q(x)| dx.$$

It remains to show that this integral is arbitrarily small for sufficiently close P and Q under the bounded Lipschitz norm. Note that this is not the case for arbitrary P and Q , but the Lipschitz continuity of p and q ensure that we can complete the result. In particular, suppose that there exists an x^* such that $p(x^*) - q(x^*) > \epsilon$. To reduce the notational effort assume that $x^* \pm \epsilon/(4D) \in A$ to avoid boundary effects (which can be accommodated simply but with more notation). For $x \in [x^* - \epsilon/(4D), x^* + \epsilon/(4D)]$ we have that $p(x) > q(x) + \epsilon/2$. Define a test function $g(x) = \max(0, \epsilon/(8D) - |x - x^*|/2)$. We have that

$$\begin{aligned} \|P - Q\|_{BL} &\geq \left| \int_A (p(x) - q(x))g(x) dx \right| \\ &= \int_{x^* - \epsilon/(4D)}^{x^* + \epsilon/(4D)} (p(x) - q(x))g(x) dx \\ &\geq \int_{x^* - \epsilon/(4D)}^{x^* + \epsilon/(4D)} \frac{\epsilon}{2} g(x) dx \\ &= \frac{\epsilon^3}{64D}. \end{aligned}$$

So by taking $\|P - Q\|_{BL}$ small, we can force $p(x) - q(x)$ to be uniformly small, and hence $\int_A |p(x) - q(x)| dx$ to be small, giving the result. \square

Lemma 9. *The function V_η is strictly decreasing for any trajectory in Δ_D whenever $\underline{\pi} \notin \mathcal{LE}_\eta$.*

Proof. Using the Gateaux derivative,

$$\begin{aligned} \dot{V}_\eta(\underline{\pi}) &= dV_\eta(\underline{\pi}, \dot{\underline{\pi}}) \\ &= - \left[du(\underline{\pi}, \dot{\underline{\pi}}) + \eta \sum_{i=1}^N d\nu^i(\pi^i, \dot{\pi}^i) \right] \\ &= - \sum_{i=1}^N [du((\pi^i, \pi^{-i}), \dot{\pi}^i) + \eta d\nu^i(\pi^i, \dot{\pi}^i)]. \end{aligned}$$

It follows directly from the definition of the derivatives that $du((\pi^i, \pi^{-i}), \dot{\pi}^i) = \int_{A^i} u(a^i, \pi^{-i}) \dot{\pi}^i(da^i)$. Re-arranging the definition of $l_\eta^i(\pi^{-i})$ from (2) gives

$$\begin{aligned} u(a^i, \pi^{-i}) &= \eta \log(l_\eta^i(\pi^{-i})(a^i)) \\ &\quad + \eta \log \left[\int_{A^i} \exp\{\eta^{-1} u(\tilde{a}^i, \pi^{-i})\} d\tilde{a}^i \right]. \end{aligned}$$

So, noting that $\int_{A^i} \dot{\pi}^i(da^i) = 0$,

$$\int_{A^i} u(a^i, \pi^{-i}) \dot{\pi}^i(da^i) = \eta \int_{A^i} \log(l_\eta^i(\pi^{-i})(a^i)) \dot{\pi}^i(da^i).$$

It is shown in [16, equation (D.3)] that $d\nu^i(\pi^i, \hat{\pi}^i) = -\int_{A^i} \log(p^i(a^i)) \hat{\pi}^i(da^i)$. Hence

$$\begin{aligned} \dot{V}_\eta(\underline{\pi}) &= -\eta \sum_{i=1}^N \int_{A^i} [\log(l_\eta^i(\pi^{-i})(a^i)) - \log(p^i(a^i))] \hat{\pi}^i(da^i) \\ &= -\eta \sum_{i=1}^N \int_{A^i} [\log(l_\eta^i(\pi^{-i})(a^i)) - \log(p^i(a^i))] \times [l_\eta^i(\pi^{-i})(a^i) - p^i(a^i)] da^i \\ &= -\eta \sum_{i=1}^N \{KL(l_\eta^i(\pi^{-i}) \| p^i) + KL(p^i \| l_\eta^i(\pi^{-i}))\} \end{aligned}$$

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 $p^i = l_\eta^i(\pi^{-i})$ for all i , which is exactly the condition that $\underline{\pi} \in \mathcal{LE}_\eta$. \square

We thus have a continuous function which is decreasing whenever $\underline{\pi} \notin \mathcal{LE}_\eta$. However, as demonstrated by [41], this is insufficient to prove that all internally chain transitive sets are contained in \mathcal{LE}_η . We could use a further result, that the set of values V_η takes at points $\underline{\pi} \in \mathcal{LE}_\eta$ is a measure zero set. This is usually achieved by using Sard’s theorem (see [44] for example), but Smale’s generalisation of Sard’s theorem to Banach spaces does not apply in our case. We therefore prove a new result directly, using the provided condition that the connected components of the set of logit equilibria \mathcal{LE}_η are isolated.

Lemma 10. *Let $V : M \rightarrow \mathbb{R}$ 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 first that an internally chain transitive set Λ is a compact, connected, invariant and attractor-free set. Let $\Lambda_0 = \operatorname{argmin}\{V(x) : x \in \Lambda\}$, and $V_0 = \min\{V(x) : x \in \Lambda\}$. It then follows that Λ_0 only consists of equilibria of V : otherwise, if $x \in \Lambda_0$ is not an equilibrium, we would have $V(\Phi(x, t)) < V(x)$ for all $t > 0$, contradicting the fact that Λ is forward invariant and $V(x) \geq V_0$ for all $x \in \Lambda$.

Now, assume there exists some $x \in \Lambda$ with $V(x) > V_0$. Then, take $\epsilon > 0$ small enough so that the closed set $\Lambda_\epsilon = \{x \in \Lambda : V(x) \leq V_0 + \epsilon\}$ contains no other equilibria of Φ except those in Λ_0 (that this is possible follows from the fact that V is constant on equilibrium components and that these components are isolated). Since V is a strict Lyapunov function for Φ we will also have $\Phi(\Lambda_\epsilon, t) \subseteq \operatorname{int}(\Lambda_\epsilon)$ for all $t > 0$ (recall that Λ_0 is contained in the interior of Λ_ϵ and Λ_ϵ has no other equilibria), so Λ_ϵ contains an attractor of Φ for all $\epsilon > 0$ [41, Lemma 5.2]. This contradicts the fact that Λ is attractor-free, so we must have $V(x) = V_0$ for all $x \in \Lambda$, i.e. $\Lambda = \Lambda_0$. \square

We are now in a position to prove Theorem 6 and – finally – Theorem 1.

Proof of Theorem 6. V_η is necessarily constant on connected components of \mathcal{LE}_η , so the conditions of Lemma 10 are met. Therefore any internally chain transitive (under bounded Lipschitz norm) set of the flow defined by (13) is contained in a connected component of the set \mathcal{LE}_η . This is precisely Theorem 6. \square

Proof of Theorem 1. Theorem 5 shows that $\{\pi_n\}_{n \in \mathbb{N}}$ converges under the bounded Lipschitz norm to an internally chain transitive set of the flow defined by the logit best response dynamics. Theorem 6 shows that any internally chain transitive set of these dynamics is contained in \mathcal{LE}_η . It thus follows that π_n converges to \mathcal{LE}_η weakly.

To establish our strong convergence claim, recall first that every probability measure in \mathcal{LE}_η is nonatomic and absolutely continuous with respect to Lebesgue measure on \mathbb{R} . On the other hand, if π^* is a (weak) limit point of π_n , we will have $\pi_n(A) \rightarrow \pi^*(A)$ for every continuity set A of π^* (i.e. for every measurable set A such that $\pi(\partial A) = 0$). Since every weak limit point of π_n is contained in \mathcal{LE}_η and Borel sets are also continuity sets for absolutely continuous measures, our assertion follows. \square

6. DISCUSSION

In this paper, we introduced an actor-critic reinforcement learning algorithm for potential games with continuous action sets. By utilizing two different timescales for the actor and critic updates (fast and slow respectively), we showed that the algorithm converges strongly to the game’s set of logit equilibria with minimal information requirements – in particular, players are not assumed to observe their opponents’ actions or to have full knowledge of their individual payoff functions.

From a practical point of view, this provides an attractive algorithmic framework for distributed control and optimization in complex systems with sparse feedback – such as rate control and power allocation in large-scale, decentralized wireless networks. In addition, from a theoretical point of view, our approach provided a nontrivial extension of several finite-dimensional stochastic approximation techniques to infinite-dimensional Banach spaces. In this way, the proposed framework can be applied and extended to different scenarios of high practical relevance (especially in the context of wireless networks) such as the case of noisy/imperfect payoff observations, asynchronous and/or delayed player updates, etc. These research directions lie beyond the scope of the current work, but we intend to pursue them in a future paper.

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] T. Roughgarden and E. Tardos, “How bad is selfish routing?” *Jo. ACM*, vol. 49, pp. 236–259, 2002.
- [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] 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.

- [9] 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.
- [10] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [11] R. Bertin, A. Legrand, and C. Touati, "Toward a fully decentralized algorithm for multiple bag-of-tasks application scheduling on grids," in *GRID '08: Proceedings of the 3rd IEEE/ACM International Conference on Grid Computing*, 2008.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] S. Perkins and D. S. Leslie, "Stochastic fictitious play with continuous action sets," *J. Econ. Theory*, vol. 152, pp. 179–213, 2014.
- [17] M.-W. Cheung, "Pairwise comparison dynamics for games with continuous strategy space," *J. Econ. Theory*, vol. 153, pp. 344–375, 2014.
- [18] R. Lahkar and F. Riedel, "The Continuous Logit Dynamic and Price Dispersion," Tech. Rep., 2013.
- [19] H. Walk, "An invariance principle for the Robbins-Monro process in a Hilbert space," *Z. Wahrscheinlichkeit*, vol. 39, pp. 135–150, 1977.
- [20] E. Berger, "Asymptotic behaviour of a class of stochastic approximation procedures," *Probab. Theory Rel.*, vol. 71, pp. 517–552, 1986.
- [21] H. Walk and L. Zsidó, "Convergence of the Robbins-Monro method for linear problems in a Banach space," *J. Math. Anal. Appl.*, vol. 139, pp. 152–177, 1989.
- [22] A. Shwartz and N. Berman, "Abstract stochastic approximations and applications," *Stoch. Proc. Appl.*, vol. 31, pp. 133–149, 1989.
- [23] V. A. Koval, "Rate of convergence of stochastic approximation procedures in a Banach space," *Cybern. Syst. Anal.*, vol. 34, pp. 386–394, 1998.
- [24] J. Dippon and H. Walk, "The Averaged Robbins-Monro Method for Linear Problems in a Banach Space," *J. Theor. Probab.*, vol. 19, pp. 166–189, 2006.
- [25] V. S. Borkar, "Stochastic approximation with two time scales," *Syst. Control Lett.*, vol. 29, pp. 291–294, 1997.
- [26] J. Oechssler and F. Riedel, "On the Dynamic Foundation of Evolutionary Stability in Continuous Models," *J. Econ. Theory*, vol. 107, pp. 223–252, 2002.
- [27] 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.
- [28] M. Benaïm, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM J. Control Optim.*, vol. 44, pp. 328–348, 2006.
- [29] R. D. McKelvey and T. R. Palfrey, "Quantal Response Equilibria for Normal Form Games," *Game. Econ. Behav.*, vol. 10, pp. 6–38, 1995.
- [30] D. Fudenberg and D. M. Kreps, "Learning Mixed Equilibria," *Game. Econ. Behav.*, vol. 5, pp. 320–367, 1993.
- [31] 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.
- [32] J. Hofbauer and E. Hopkins, "Learning in Perturbed Asymmetric Games," *Game. Econ. Behav.*, vol. 52, pp. 133–152, 2005.
- [33] J. Hofbauer and W. H. Sandholm, "Evolution in games with randomly disturbed payoffs," *J. Econ. Theory*, vol. 132, pp. 47–69, 2007.
- [34] D. S. Leslie and E. J. Collins, "Individual Q-learning in normal form games," *SIAM J. Control Optim.*, vol. 44, pp. 495–514, 2005.

- [35] 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.
- [36] P. Coucheney, B. Gaujal, and P. Mertikopoulos, “Penalty-regulated dynamics and robust learning procedures in games,” *Math. Oper. Res.*, to appear.
- [37] 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.
- [38] A. C. Chapman, D. S. Leslie, A. Rogers, and N. R. Jennings, “Convergent Learning Algorithms for Unknown Reward Games,” *SIAM J. Control Optim.*, vol. 51, pp. 3154–3180, 2013.
- [39] P. Del Moral, A. Doucet, and A. Jasra, “Sequential Monte Carlo Samplers,” *J. Roy. Stat. Soc. B*, vol. 68, pp. 411–436, 2006.
- [40] M. Ledoux and M. Talagrand, *Probability in Banach spaces*. Springer-Verlag, 1991.
- [41] M. Benaïm, “Dynamics of stochastic approximation algorithms,” *Seminaire de probabilités XXXIII*, vol. 33, pp. 1–68, 1999.
- [42] D. Luenberger, *Optimization by vector space methods*. Prentice Hall, 1969.
- [43] S. Perkins, “Advanced Stochastic Approximation Frameworks and their Applications by,” Ph.D. dissertation, University of Bristol, 2013.
- [44] J. Hofbauer and W. H. Sandholm, “On the global convergence of stochastic fictitious play,” *Econometrica*, vol. 70, pp. 2265–2294, 2002.