The Evolutionary Dynamics of Soft-Max Policy Gradient in Games

Martino Bernasconi, [1, *] Federico Cacciamani, [1, *] Simone Fioravanti,[2, *] Nicola Gatti, [1] Francesco Trovò, [1]
[1] Politecnico di Milano, [2] Gran Sasso Science Institute
[1] {martino.bernasconideluca, federico.cacciamani, nicola.gatti, francesco1.trovo}@polimi.it
[2] simone.fioravanti@gssi.it
[*] Equal Contribuition

Abstract

In this paper, we study the mean dynamics of the *soft-max policy gradient* algorithm in multi-agent settings by resorting to *evolutionary game theory* and dynamical system tools. Such a study is crucial to understand the algorithm's weaknesses when employed in multi-agent settings. Unlike most multi-agent reinforcement learning algorithms, whose mean dynamics is a slight variant of the replicator dynamics not affecting the properties of the original dynamics, the soft-max policy gradient dynamics presents a structure significantly different from that of the replicator. Indeed the dynamics is equivalent to the replicator dynamics in a different game derived by a non-convex transformation of the payoffs of the original game. First we recover the properties—already known for the discrete-time soft-max policy gradient—for the continuous-time mean dynamics in the case of learning a best response. As it commonly happens, the continuous-time dynamics allow for a simpler analysis and deeper understanding of the algorithm that we use to characterize fully the dynamics and improve on its theoretical understanding. Then, we resort to models based on single- and multi-population games, showing that the dynamics preserve the volume as prove that, in arbitrary instances, it is not possible to obtain last-iterate convergence when the equilibrium of the game is fully mixed. Furthermore, we give empirical evidence that dynamics starting from close initial points may expand over time, thus showing that the behaviour of the dynamics in games with fully-mixed equilibrium is *chaotic*.

1 Introduction

In Multi-Agent Reinforcement Learning (MARL), every agent learns independently of the others how to play a strategic interaction situation (a.k.a. strategic game) in a shared environment. In particular, every agent acts in an unknown non-stationary Markov Decision Problem, where the non-stationarity is due to the evolution of the opponents' strategies over time. Most algorithm in Reinforcement Learning (RL) only provide theoretical guarantees in restricted settings, *i.e.*, every agent is guaranteed to converge to the optimal solution when facing non-learning opponents. Furthermore, some MARL algorithms also present convergence guarantees in self-play under very restrictive assumptions, *e.g.*, *Neural Fictitious Self Play* (Heinrich and Silver 2016) and *Deep-CFR* (Brown et al. 2019). Thanks to the dynamical systems formulation of evolutionary game theory (EGT), the algorithms can be studied in terms of properties—*e.g.*, the set of stationary strategies, the set of asymptotically stable strategies, and the convergence rate—in different settings—*e.g.*, best-response problem, single-population games, and multi-population games. Interestingly, most MARL algorithms, such as, *e.g.*, *Q-learning* (Kaisers and Tuyls 2010; Tuyls, Verbeeck, and Lenaerts 2003), and multiple no-regret algorithms (Klos, van Ahee, and Tuyls 2010) have mean dynamics that are slight variants of the replicator dynamics and having the same properties of the original dynamics.

One of the most interesting techniques developed in reinforcement learning is policy gradient (Peters and Schaal 2006; Sutton et al. 1999). It comes under various flavours such as, SAC (Haarnoja et al. 2018), DDPG (Lillicrap et al. 2016), MADDPG (Lowe et al. 2017), A3C (Mnih et al. 2016), and REINFORCE (Williams 1992). Policy-gradient methods work on a constrained space of policies, each of which is fully described by a parameters vector. Such an approach plays a crucial role whenever the space of an agent's (unparameterized) strategies is huge so that the learning of such strategies may result unaffordable in terms of samples complexity. Indeed, policy-gradient methods allow us to work on the policy parameters space that is generally smaller than the strategy space, so as to model a wide space of strategies with a compact number of parameters. Such an approach may result crucial in, *e.g.*, online settings where one cannot afford to simulate millions of samples to find an optimal strategy. On the other hand, this introduces an additional generalization error, instead, we focus on the unrestricted case, having a parameter per action, hence bringing to zero the further generalization error. Our paper focuses on the soft-max policy gradient algorithm, which is the most commonly adopted flavor of policy gradient. We find that exists a non-trivial connection between the replicator dynamics and the soft-max policy gradient dynamics. This feature was overlooked in the literature and both clarifies the underlying nature of the problem of the soft-max policy gradient algorithm and requires a new set of techniques, as the main bulk of the literature on the replicator is on games with linear payoffs. We first study the case in which an agent needs to learn the best response to a given opponents' joint strategy. As commonly happens when studying continuous-time

approximations, our analysis of the continuous-time dynamics both provides cleaner derivations of previously know results and a deeper theoretical understanding of the properties of the soft-max policy gradient algorithm. Namely, (i) we discover that the soft-max policy gradient algorithm corresponds to the replicator dynamics on a game with *non-linear* payoffs, and (ii) we are able to characterize exactly the set of points called in the literature as *bad initialization* points, *i.e.,* starting point for the dynamics s.t. the evolution moves initially toward sub-optimal strategies. This exact characterization is of paramount importance when shifting the attention from having to learn the best response to considering learning opponents. The conclusion of this analysis is that while the softmax policy gradient has sound theoretical guarantees when learning the best response, its properties make it less appealing in the presence of adversarial opponents.

In the second part of the work, we study the case in which all the agents learn simultaneously. To the best of our knowledge, this is the first work to study theoretically the behaviour of the soft-max policy gradient algorithm in multi-agent environments. This case is customarily tackled in the evolutionary game theory literature by investigating both the corresponding single and multiple-population games. In the former case, a single population of agents is playing against themselves. This model is customarily adopted as an abstraction of settings with many agents, *e.g.*, Hu et al. (2020) propose the use of single-population games as a tool for studying large populations of anonymous, independently learning agents and for studying the frequencies of competing biological traits such as genotypes. On the other hand, to analyze the multiple population case we resort to the above-mentioned, non-trivial, correspondence between the continuous-time dynamics of the soft-max policy gradient algorithm and the replicator dynamics. We rely on results based on the replicator dynamics on non-linear payoffs, while most of the current literature analyzes the case of linear payoffs. Indeed, the non-linearity of the payoffs of the dynamics makes the results currently available in the literature meaningless. Finally, we prove that the soft-max policy gradient algorithm demonstrates volume conservation when the game has an interior Nash equilibrium and, hence, it is Poincaré recurrent.

2 Preliminaries

Game Theory A *normal-form game* is defined as a tuple: $(\mathcal{N}, \{\mathcal{A}^{(i)}\}_{i\in\mathcal{N}}, \{r^{(i)}\}_{i\in\mathcal{N}})$, where $\mathcal{N}$ is a set of n agents, $\mathcal{A}^{(i)}$ is the action set of agent i, and $r^{(i)} : \mathcal{A}^{(1)} \times \ldots \times \mathcal{A}^{(n)} \to [0, 1]$ is the utility function that associates each agents' joint action with the payoff of agent i. From now on, for the sake of simplicity, we assume that all the agents have the same number of actions, or, formally, $|\mathcal{A}^{(i)}| = m$. The *strategy* $\mathbf{x}^{(i)} \in \Delta(\mathcal{A}^{(i)})$ of an agent i is defined as a probability distribution over her actions $\mathcal{A}^{(i)}$, where $\Delta(\mathcal{A}^{(i)})$ is the simplex over $\mathcal{A}^{(i)}$. We denote the j-th component of $\mathbf{x}^{(i)}$ with $x_j^{(i)}$, corresponding to the probability of playing action $a_j \in \mathcal{A}^{(i)}$. Furthermore, a *strategy profile* is defined as a tuple $\mathbf{x} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})$ specifying a strategy for each agent. In this paper, we focus on the central concept of Nash equilibrium, in which the strategy of every agent is a *best response* to the opponents' strategies. Formally in a NE $\bar{\mathbf{x}}$ it holds that, for all i $\bar{\mathbf{x}}^{(i)} = \arg\max_{\mathbf{x}^{(i)}} r^{(i)}(\mathbf{x}^{(i)}, \bar{\mathbf{x}}^{(-i)})$, where $(-i)$ denotes the set of indices different from i.

Evolutionary Game Theory and Replicator Dynamics Evolutionary game theory captures the situation in which the agents are not rational and adapt their strategies dynamically over time $t \in \mathbb{R}^+$. The central concept is that of *population*. A population $i \in \mathcal{N}$ is a potentially infinite collection of individuals with a common action set of actions $\mathcal{A}^{(i)}$, where each individual plays a fix action $a_j \in \mathcal{A}^{(i)}$. The aggregate behavior of population i is modeled by the frequency whereby an individual playing action a_j is met among all the possible individuals of that population. This leads to a direct connection between populations and agents, where every population i corresponds to an agent i and *vice versa*. Thus, at every time t, the state of the populations is described by a strategy profile $\mathbf{x}(t) := (\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t))$. The *fitness* of an individual of population i playing action a_j is provided by the function $\Pi_j^{(i)}(\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t)) \in \mathbb{R}$, while $\Pi^{(i)}(\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t))$ is the fitness vector for all actions of population i. Hence, the mean fitness of population i is $\mathbf{x}^{(i)}(t)^\top \Pi^{(i)}(\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t))$, where '$\top$' denotes the transpose operator. Notice that the mean fitness of population i is the expected (over agents' strategies) payoff $r^{(i)}$. The evolution of $\mathbf{x}(t)$ over time is determined by a continuous-time dynamical system. Replicator Dynamics are one of the most studied dynamics and have the property that the time derivative of each j-th component $\dot{x}_j^{(i)}(t)$ is proportional to the difference between the fitness associated with a_j and the average fitness of population i. Formally, we have:

$$\dot{x}_j^{(i)}(t) = x_j^{(i)}(t)\Big[\Pi_j^{(i)}\left(\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t)\right) - \mathbf{x}^{(i)}(t)^\top \Pi^{(i)}\left(\mathbf{x}^{(1)}(t), \ldots, \mathbf{x}^{(n)}(t)\right)\Big]. \quad (1)$$

Note that, the asymptotically-stable states of the above evolutionary model are a subset of the Nash equilibria (see for example (Hofbauer and Sigmund 1998; Cressman, Ansell, and Binmore 2003; Sandholm 2010b)). In particular, when $n = 1$, the model is called *single-population* or *symmetric* game, and the central concept is the one of Evolutionary Stable Strategies (ESS) to which the RD converge. Formally an ESS is defined as follows (Sandholm 2010b):

Definition 1. *A strategy $\mathbf{x} \in \Delta^{|\mathcal{A}|}$ is an ESS of a single population game defined by a fitness function $\Pi(\cdot)$ if there is a neighborhood $\mathcal{O}$ of $\mathbf{x}$ such that $(\mathbf{z} - \mathbf{x})^\top \Pi(\mathbf{z}) < 0$ for all the strategies $\mathbf{z} \in \mathcal{O} \setminus \{\mathbf{x}\}$.*[1]

Dynamical Systems Given an autonomous dynamical system $\dot{\mathbf{x}}(t) = f(\mathbf{x}(t))$ with $\mathbf{x} \in D \subseteq \mathbb{R}^d$ where D is an open domain and $f : \mathbb{R}^d \to \mathbb{R}^d$ a continuously differentiable vector field, its set of solutions is called *flow*, parametrized with the starting point of the dynamics. Formally, the flow is

[1] With Δ^m we denote a generic m-dimensional simplex.

defined as $\phi : \mathbb{R} \times D \to \mathbb{R}^d$, where $t \mapsto \phi(t, x)$ is the solution to the system such that $\phi(0, x) = x$. Given a set $S \subset D$, we call $S(t) = \{\phi(s, t) : \ s \in S\}$ the evolution of S under the flow ϕ at time t. Denoting with $\mathrm{vol}\, S(t)$ its volume, the *Liouville formula* offers a alternative method to compute its time derivative as:

$$\frac{d}{dt} \mathrm{vol}\, S(t) = \int_{S(t)} \mathrm{div}\, f \, d\mathbf{x}, \tag{2}$$

where $\mathrm{div}(f) = \sum_{i=1}^{d} \frac{\partial f_i}{\partial x_i}$ is the divergence of f *i.e.*, the sum of the diagonal elements of its Jacobian. The most immediate consequence is that, if the divergence is null, the flow preserves the volume. In this case, the flow $\phi(\cdot)$ is said to be *incompressible*, i.e., the set S is allowed to move or stretch under the flow of the dynamics, but cannot compress or enlarge. This affects the shape of the trajectories, which are not allowed to converge to a single point. This property will be analysed for the trajectories of two-population SPGD in Section 5.

Markov Decision Processes and Policy Gradient Algorithm A *Markov decision process* (MDP) is a tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu)$, where $\mathcal{S}$ and $\mathcal{A}$ are the sets of *states* and *actions*, respectively, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state-transition probability function returning the probability to transition to state $s(t+1)$ when performing action $a(t)$ in state $s(t)$, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the immediate reward function returning the reward associated with a given transition, $\gamma \in [0, 1]$ is the discount factor, and μ is the probability distribution over the initial states. The goal of an agent is to find a *policy* $\pi(a|s)$, *i.e.*, a mapping from states to actions, to maximize the expected sum of discounted future rewards. RL offers a set of algorithms and techniques to perform sequential decision-making in MDPs whose parameters are unknown. We focus on a particular class of widely used RL algorithms: the *Policy Gradient* (PG) ones. PG algorithms estimate the optimal policy by directly searching over a parameterized policy space $\pi(\cdot|s, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a real-valued vector of parameters. Formally, the problem consists of finding the $\arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$, where $J(\boldsymbol{\theta})$ is a performance surface (usually the expected reward) achieved by $\pi(\cdot|s, \boldsymbol{\theta})$. The search is performed by a stochastic gradient ascent procedure on $J(\boldsymbol{\theta})$, iteratively updating the parameters as follows:

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) + \eta \, \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}(t)), \tag{3}$$

where $\eta \in \mathbb{R}^+$ is a *learning rate*. We focus on the Soft-Max parameterization (SPG algorithm), which is widely adopted in practice with discrete action sets. In particular, the policy takes the form:

$$\pi(a|s, \boldsymbol{\theta}) = \frac{e^{\tau f^a(s, \boldsymbol{\theta})}}{\sum_{a' \in \mathcal{A}} e^{\tau f^{a'}(s, \boldsymbol{\theta})}},$$

where $\tau \in \mathbb{R}^+$ is an inverse temperature parameter, and $f^a(\cdot, \cdot)$ are function approximators, which are trained over parameters $\boldsymbol{\theta}$ to approximate the expected payoff of playing action a in state s.

3 Soft-Max Policy Gradient Mean Dynamics

In what follows, we adopt a single-agent i perspective, and derive the continuous-time mean dynamics of strategy $\mathbf{x}^{(i)}(t)$ evolving in a normal-form game against a generic set of $n-1$ opponents with joint strategy $\mathbf{y}(t)$.[2] This dynamics was already presented in (Srinivasan et al. 2018), but we report here the derivation for completeness. Normal-form games can be modeled as a direct extension of MDPs, namely *stochastic games*, in which the state-transitions and the rewards depend on the joint strategy of all agents, and a single state is present (for more details, we point the reader to the work by Shapley (1953)). Thus, we can safely drop the dependence of $\pi(\cdot|\cdot, \cdot)$ and $f^a(\cdot, \cdot)$ from the state s. In single-state environments, the SPG algorithm needs to estimate one value for each action, which is equivalent to $f^{a_j}(\boldsymbol{\theta}) = \theta_j$. Thus, the policy $\pi(a|\boldsymbol{\theta})$ is represented by a single strategy $\mathbf{x}(\boldsymbol{\theta}) \in \Delta^m$ which, through $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]$, defines a probability distribution over actions a_j, for every $j \in \{1, \ldots, m\}$, as follows:

$$x_j(t) = x_j(\boldsymbol{\theta}(t)) = \frac{e^{\tau \theta_j(t)}}{\sum_{k=1}^{m} e^{\tau \theta_k(t)}}. \tag{4}$$

Following the procedure proposed by Tuyls, Verbeeck, and Lenaerts (2003), the time derivative of $x_j(t)$ can be formulated in terms of time derivative of $\boldsymbol{\theta}(t)$, as follows:

$$\dot{x}_j(t) = \tau \, x_j(t) \left(\dot{\theta}_j(t) - \sum_{k=1}^{m} x_k(t) \, \dot{\theta}_k(t) \right). \tag{5}$$

Using the discrete-time variation of parameter vector $\boldsymbol{\theta}(t)$ provided in Equation (3), we obtain the corresponding continuous-time mean dynamics as follows:

$$\dot{\boldsymbol{\theta}}(t) := \lim_{\delta \to 0} \frac{\boldsymbol{\theta}(t+\delta) - \boldsymbol{\theta}(t)}{\delta} = \eta \, \nabla_{\boldsymbol{\theta}} \, J(\boldsymbol{\theta}(t)). \tag{6}$$

We denote the payoff n-dimensional tensor of agent i with A, and we obtain:

$$J(\boldsymbol{\theta}(t)) = \mathbf{x}^\top(t) \, A \, \mathbf{y}(t), \tag{7}$$

and, by applying the chain rule, we have:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{\partial J}{\partial \boldsymbol{\theta}} = \frac{\partial J}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} = \Psi(\mathbf{x}(\boldsymbol{\theta})) \, A \, \mathbf{y}, \tag{8}$$

where $\Psi(\mathbf{x})$ is the Jacobian of the Soft-Max function, which is a symmetric matrix defined as:

$$\Psi(\mathbf{x}) = \mathrm{diag}(\mathbf{x}) \, (I_m - X),$$

where $\mathrm{diag}(\mathbf{x})$ is the matrix of order m with diagonal entries equal to $\mathbf{x}$, I_m is the identity matrix of order m, and X is the matrix of order m where every row is $\mathbf{x}$. Finally, the mean dynamics of the SPG algorithm, called SPGD, are as follows:

$$\frac{\dot{x}_j(t)}{x_j(t)} = \tau \Big((\eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}(t)))_j - \mathbf{x}(t)^\top \eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}(t)) \Big) \tag{9}$$

[2] For the sake of simplicity, from now on, we omit the superscript '(i)' from $\mathbf{x}^{(i)}(t)$.

$$= \eta\tau \left(\left(\Psi(\mathbf{x}(t))A\,\mathbf{y}(t)\right)_j - \mathbf{x}(t)^\top \Psi(\mathbf{x}(t))A\,\mathbf{y}(t) \right), \tag{10}$$

where Equation (9) is derived by substituting Equation (8) into Equation (6), while Equation (10) follows from Equation (5). Notice that SPGD in Equation (10) resemble RD in Equation (1). Indeed, in both dynamics, the time derivative of $x_j(t)$ is proportional to the difference between the payoff provided by action a_j and the average payoff provided by strategy $\mathbf{x}(t)$. However, they differ as, in SPGD, the payoffs are weighted by the matrix $\Psi(\mathbf{x})$. In other words, SPGD and RD constitute the same set of differential equations except for an opportune, non-linear redefinition of the fitness function: $\Pi_{\mathsf{RD}}(\mathbf{x}(t), \mathbf{y}(t)) = A\,\mathbf{y}(t)$ in RD, while $\Pi_{\mathsf{SPGD}}(\mathbf{x}(t), \mathbf{y}(t)) = \Psi(\mathbf{x}(t))\,A\,\mathbf{y}(t)$ in SPGD. Moreover, note that the matrix $\Psi(\mathbf{x})$ is singular (see Proposition 2 by Gao and Pavel (2017)), and, therefore, no transformation applied to the payoff tensor A can lead to a new payoff tensor $\tilde{A}$ such that Π_{SPGD} on $\tilde{A}$ is equivalent to Π_{RD} on A, suggesting that the study of SPGD cannot be easily reduced to the study of RD. Finally, observe that $\Pi_{\mathsf{SPGD}}(\mathbf{x}(t), \mathbf{y}(t))_j = [\Psi(\mathbf{x}(t))A\,\mathbf{y}(t)]_j = x_j(t)(\mathbf{e}_j - \mathbf{x}(t))^\top A\,\mathbf{y}(t)$ which is the j-th component of the vector field associated with the RD.

4 Best-response Problem Analysis

The best-response problem is the central problem every agent i needs to face when converging to a Nash equilibrium. This corresponds to a setting in which an agent maximizes her utility, while the opponents' joint strategy $\mathbf{y}$ is fixed during time. The following analysis focuses on non-degenerate cases in which the best-response problem admits a unique optimal solution, *i.e.*, there is a single best response. The same analysis can be extended to the degenerate case in which there are multiple optimal solutions, and the convergence is required to a generic strategy of the (convex) set of the best responses. Initially, we state the following lemma, which is a variant of the Polyak-Łojasiewicz inequality (Karimi, Nutini, and Schmidt 2016) and is instrumental to our analysis.

Lemma 1. *Let* $\bar{\mathbf{e}}_j = \arg\max_k \left\{\mathbf{e}_k^\top\, A\,\mathbf{y}\right\}$*e the single (pure) best response, then it holds that:*

$$\left|\left|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right|\right|_2^2 \geq x_j(\boldsymbol{\theta})^2\,(J^* - J(\boldsymbol{\theta}))^2,\, \forall \boldsymbol{\theta} \in \mathbb{R}^m, \tag{11}$$

where $J^* = \bar{\mathbf{e}}_j^\top\, A\,\mathbf{y}$ *and* $||\cdot||_z$ *is the* z*-norm, and* $\bar{\mathbf{e}}_j \in \Delta^m$ *is the pure strategy in which action* a_j *is played with probability one.*[3]

Relying on the result provided by Lemma 1, we state that SPGD converge to the best response. Furthermore, we show that the function $V(t) = J^* - J(\boldsymbol{\theta}(t))$ is a Lyapunov function of those dynamics. Formally, we state:

Theorem 1. *If* $\mathbf{y}$ *is fixed,* $\mathbf{x}(0) \in \text{int}(\Delta^m)$ (i.e.*, it is fully mixed), and there is a single best response* $\bar{\mathbf{e}}_j$*, the SPGD asymptotically converge to the best response* $\bar{\mathbf{e}}_j$*.*

[3]All the proofs of the paper are deferred to the Appendix for space reasons.

Note that the Lyapunov function $V(t)$ is defined as the difference between the optimal value J^*, corresponding to the value provided by the best response, and the value of the current state $J(\boldsymbol{\theta}(t))$. Therefore, it directly follows that:

Corollary 1. *If* $\mathbf{y}$ *is fixed,* $\mathbf{x}(0) \in \text{int}\,\Delta^m$*, and there is a single best response* $\bar{\mathbf{e}}_j$*, SPGD are such that* $J(\boldsymbol{\theta}(t))$ *is strictly monotonically increasing in* t*.*

Finally, we derive the convergence rate of SPGD by a non-trivial adaptation of (Mei et al. 2020b, Theorem 2).

Theorem 2. *Given function* $V(t) := J^* - J(\boldsymbol{\theta}(t))$*, where* J^* *is the value of the best response and* $J(\boldsymbol{\theta}(t)) = \mathbf{x}(t)^\top A\,\mathbf{y}$*, then with SPGD it holds (for a suitable constant* $C_0 \in \mathbb{R}^+$*) that:*

$$V(t) \leq \frac{1}{\eta \left(\frac{m-\xi}{m+\xi}\right)^2 t + C_0}, \tag{12}$$

where ξ *is the optimality gap between the best response* $\bar{\mathbf{e}}_j$ *and the second best response,* i.e.*,* $\xi := \bar{\mathbf{e}}_j^\top A\,\mathbf{y} - \max_{k \neq j} \left\{\mathbf{e}_k^\top\, A\,\mathbf{y}\right\}$.

The idea behind the proof of Theorem 2 is to show that for a best-response problem, there is a *bad* set, that the dynamics leaves in finite time. Then, after the dynamics leaves the *bad* region, we have an asymptotic analysis that that gives linear convergence rate. Note that the results about the discrete version of the algorithm provided by Mei et al. (2020b, Theorem 2) is less general than what has been proposed here since it holds for stricter assumptions, *i.e.*, only for a learning rate $\eta = \frac{2}{5}$, but has the same asymptotic convergence rate of $\mathcal{O}(1/t)$.

Comparing the Behaviors of SPGD and RD in the Best-response Problem

As discussed in Section 3, the difference between RD and SPGD discussed that follows from the non-linear redefinition of the fitness function $\Pi(\cdot)$, results in dynamics that are dramatically different even if they both converge to the best response. In this section, we theoretically analyze this difference in the Best-Response problem. We also provide experimental results that highlight its relevance in the case of learning in games.

We start by recalling an interesting property of RD (Sandholm 2009). Let $\bar{\mathbf{e}}_j$ be the unique pure best response, for every action $a_k \neq a_j$, it holds that $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$, where the inequality is strict as the best response is unique. Therefore, in RD, $x_j(t)$ is strictly monotonically increasing in t, while the ratio $x_k(t)/x_j(t)$ is strictly monotonically decreasing in t for every $k \neq j$.[4] We show that such a monotonicity property does not generally hold in the case of SPGD, thus resulting in more inefficient dynamics than RD. In particular, we state the following exact characterization of the set of bad initialization.

[4]Note that this does not exclude that $x_k(t)$ with $k \neq j$ is monotonically increasing in $t \in [0, \bar{t}]$ for some $\bar{t} > 0$.

Theorem 3. *Let $\bar{\mathbf{e}}_j$ be the (unique) pure best response against the fixed opponents' joint strategy $\mathbf{y}$. Then, in SPGD, there is at least a $k \neq j$ such that $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$. Moreover, if $m > 2$, there exists a non-empty subspace $\mathcal{E} \subset \Delta^m$ such that if $\mathbf{x}(t) \in \mathcal{E}$ then $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) > 0$ for some $k \in \{1, \dots, m\}$, and the uniform initialization $\frac{\mathbf{1}}{m}$ is always outside $\mathcal{E}$. The set $\mathcal{E}$ is the set defined as:*

$$\mathcal{E} = \bigcup_{\boldsymbol{b} \in \mathcal{B}} \left\{ \mathbf{w} \in \Delta^m \,|\, \mathbf{w} = \alpha\, \boldsymbol{b} + (1-\alpha)\, \bar{\mathbf{e}}_j, 1 > \alpha > \mathfrak{B}(\boldsymbol{b}) \right\},$$

where the set $\mathcal{B} \subset \Delta^m$ is the set of $\mathbf{x}$ such that $x_j = 0$, and $\mathfrak{B}(\boldsymbol{b}) \in [0, 1]$ is a well defined quantity for each $\boldsymbol{b} \in \mathcal{B}$.

Theorem 3 shows that, in SPGD, there is always at least one non-optimal action k whose ratio with x_j is decreasing over time, but other actions might show an increasing rate. Trivially, it follows that when $m = 2$, the monotonicity property satisfied by RD also holds for SPGD. Furthermore, Theorem 3 states that the subspace $\mathcal{E}$ is an exact characterization of the *non-monotonic improvements*, which in the bandit setting the work by Mei et al. (2020a) call as *bad initialization*. In Figures 1 and 2, we provide an example of the bad initialization problem suffered by SPGD in the classical Rock-Paper-Scissors (RPS) game when the opponent's strategy is $\mathbf{y} = (0.05, 0.90, 0.05)^\top$. In particular, Figure 1 shows that a good initialization in $\Delta^3 \setminus \mathcal{E}$ leads to dynamics that approach the best response monotonically (green line), a bad initialization in $\mathcal{E}$ leads to dynamics that initially get far from the best response and subsequently approach the best response (red line). As a result, a bad initialization leads to a much slower convergence to the best response, as shown in Figure 2. It is worth remarking that Theorem 3 guarantees that an agent can always choose $\mathbf{1}/m$ as a good initialisation. This result is in line with the result stated by Mei et al. (2020b, Theorem 8) for the bandit setting, in which the convergence to the optimal arm is monotonic if the initial policy is $\mathbf{1}/m$. Theorem 3 also shows that the algorithm is particularly sensitive to slight variations to the opponents' joint strategy $\mathbf{y}$, as a slight modification in $\mathbf{y}$ results in moving the initialization of the algorithm from $\Delta^m \setminus \mathcal{E}$ to $\mathcal{E}$, thus leading to a dramatic stretching of the convergence time.

One key consequence of the continuous-time analysis of the Best-response problem is the following theorem.

Theorem 4. *Let A be a non-degenerate zero sum game with unique fully mixed equilibrium. Then for each $\mathbf{x}(t) \in \Delta^m$ s.t. $\mathbf{x}(t) \neq \mathbf{1}/m$ there exists a strategy $\mathbf{y} \in \Delta^m$ s.t. $\mathbf{x}(t) \in \mathcal{E}$. Moreover, the problem of finding such a $\mathbf{y}$ is a linear programming problem.*

The theorem shows that in normal-form games with a fully mixed equilibrium, every point $\mathbf{x} \in \Delta^m$ which is different from $\mathbf{1}/m$, can be made a bad initialization for a suitable choice of $\mathbf{y}$. Moreover, the problem of finding such $\mathbf{y}$, is a linear program. This means that an adversarial opponent can choose a fixed strategy $\mathbf{y}$ such that for any initial point $\mathbf{x}(t_0)$ is in a bad initialization region, with the already discussed consequences on the speed of convergence to the Best-response.

5 Multi-agent Problem Analysis

In this section, we focus on the properties of SPGD when multiple agents learn simultaneously. At first, we focus on the single-population setting, and, after that, we focus on the multiple-population games, restricting to the case of two populations. As mentioned in Section 1, the former case is instrumental for the subsequent study of the latter case, and crucialto give a complete analysis from an EGT perspective. In both cases, a central role is played by the connection between SPGD and RD we discussed in Section 3.

Single-Population Games

With a single population, SPGD is equivalent to RD, once the fitness function has been redefined as $\Pi_{\mathsf{SPGD}}(\mathbf{x}) = \Psi(\mathbf{x})\, A\, \mathbf{x}$. Therefore, SPGD satisfy the same properties (see, *e.g.*, (Sandholm 2009)) that RD would present when applied to the game with fitness function Π_{SPGD}. However, due to the non-linear correspondence between the two games, only a subset of these properties is preserved when considering the original game. In particular, we focus on the properties of the revision protocol of SPGD and on the asymptotic stability of NEs.

Initially, we derive the *revision protocol* $\rho^{(A)} : \mathbb{R}^m \to \mathbb{R}^{m \times m}$ of SPGD, which is crucial for the study of its properties. The revision protocol represents the switch rate of an individual of the population from strategy k to strategy j, formally for the SPGD we have:

$$\begin{aligned} \rho^{(A)}_{kj}(\mathbf{x}) &= x_j \left[\Pi_{\mathsf{SPGD}}(\mathbf{x})_j - \Pi_{\mathsf{SPGD}}(\mathbf{x})_k\right]_+ \\ &= x_j \left[x_j (A\,\mathbf{x})_j - x_k (A\,\mathbf{x})_k + (x_k - x_j)\, \mathbf{x}^\top A\, \mathbf{x}\right]_+ . \end{aligned} \tag{13}$$

In the work by Sandholm (2017), the author identifies four main properties related to the revision protocol of evolutionary dynamics in game theory: *continuity* (C), *scarcity of data* (SD), *Nash stationarity* (NS), and *positive correlation* (PC). The revision protocol of RD (or, more simply, RD) satisfies all these properties with the peculiarity that NS is satisfied only when restricting to $\mathrm{int}(\Delta^m)$. The revision protocol of SPGD (or, more simply, SPGD) satisfy the same properties of RD except for SD. More specifically, the SD property requires that the switch rate prescribed by the revision protocol from strategy k to strategy j depends only on x_j, $(A\,\mathbf{x})_k$, and $(A\,\mathbf{x})_j$. This property is related to the demand in terms of amount of information required by the evolutionary dynamics. In SPGD, this property does not hold as $\rho^{(A)}_{kj}(\mathbf{x})$ in Equation (13) also depends on x_k, and, therefore, SPGD is requiring stronger assumptions in terms of information available to the agents than those required by RD. The C property trivially holds in SPGD. The NS property requires that NEs are stationary points of the dynamics, whereas the PC property requires that in non-stationary points, strategies' growth rates are positively correlated with their payoffs. We show that these two properties hold in SPGD.

Lemma 2. *SPGD satisfy properties NS, when restricting to $\mathrm{int}(\Delta^m)$, and PC.*

Finally, we focus on the relationship between the asymptotically stable states of SPGD and those of RD. It is

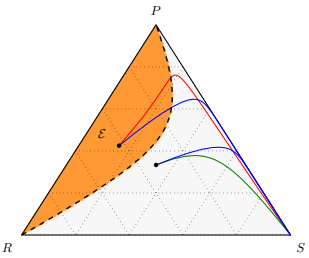

Figure 1: SPGD (red and green) and RD (in blue) trajectories in RPS game. The subspace in orange contains the bad initializations.

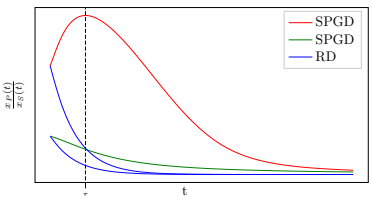

Figure 2: Ratio $x_P(t)/x_S(t)$ over time $t \geq 0$. At $t = \tau$ the dynamics in red leave subspace $\mathcal{E}$.

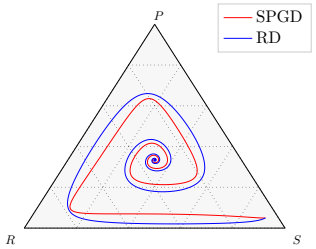

Figure 3: SPGD and RD trajectories with the good RPS population game.

well-known that RD converge to special points of interest such as Evolutionary Stable Strategies (ESS) when $\mathbf{x}(0) \in \text{int}(\Delta^m)$, see, *e.g.*, (Cressman and Tao 2014; Sandholm 2017). We show that SPGD converge to the same space of ESSs when restricting to $\text{int}(\Delta^m)$, and therefore, in the interior of the simplex, the spaces of asymptotically stable states of RD and SPGD coincide. To achieve this result, we use the concept of Regular ESS (RESS) (Sandholm 2010a), defined as follows:

Definition 2. *Strategy $\bar{\mathbf{x}} \in$ is a RESS for a population game with fitness function $\Pi(\cdot)$ if:*

(i) $\Pi_k(\bar{\mathbf{x}}) = \bar{\mathbf{x}}^\top \Pi(\bar{\mathbf{x}}) > \Pi_j(\bar{\mathbf{x}})$, *if* $\bar{x}_k > 0$ *and* $\bar{x}_j = 0$;

(ii) $\mathbf{z}^\top \mathcal{D}\Pi(\bar{\mathbf{x}})\mathbf{z} < 0$ *for all* $\mathbf{z} \neq 0, \mathbf{z} \in \mathfrak{T}$;

where $\mathfrak{T}$ is the tangent space to the m-simplex, and $\mathcal{D}\Pi(\mathbf{x})$ denotes the derivative of Π in $\mathbf{x}$.

The following lemma shows that the RESSs of the symmetric normal-form game defined by the payoff matrix A are RESS of the population game defined by the fitness $\Pi_{\mathsf{SPGD}}(\mathbf{x})$.

Lemma 3. *If $\bar{\mathbf{x}} \in \text{int}(\Delta^m)$ is a RESS for the symmetric normal-form game A, then $\bar{\mathbf{x}}$ is a RESS for the population game defined with fitness function $\Pi_{\mathsf{SPGD}}(\mathbf{x})$.*

Interestingly, Lemma 3 shows that the asymptotically stable states of SPGD and RD coincide whenever the space of RESS and the space of ESS coincide. This happens when we consider games such as the *good* RPS game (Sandholm 2017). Indeed, in Figure 3, we see, as prescribed by Lemma 3, that the trajectories of RD and SPGD converge to the center of the simplex, which in this case is a RESS. More in general, ESSs and RESSs coincide in symmetric normal-form games whenever they are fully mixed. By using this condition and Lemma 3 to show the following result.

Theorem 5. *Let $\bar{\mathbf{x}} \in \text{int}\,\Delta^m$ be an ESS for the symmetric normal-form game A. Then, it is asymptotically stable for SPGD.*

It is well-known that a ESS is an asymptotically stable rest point for the RD (Hofbauer and Sigmund 1998). Theorem 5 shows that the same behavior of RD also holds with SPGD, over the internal ESS.

Multiple-Population Games

We extend the results discussed above for the single-population case to multiple populations, investigating the convergence of the SPGD. We restrict to the case of two populations, each evolving according to the SPGD. Let $\mathbf{x}(t) \in \Delta^m$, and $\mathbf{y}(t) \in \Delta^m$ be the first and second populations, respectively, while $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{m\times m}$ are their payoff matrices, respectively. SPGD are described by the following coupled sets of differential equations for each $k \in \{1, \ldots, m\}$:

$$\begin{cases} \frac{\dot{x}_k}{x_k(t)} = \eta\tau \left(\left(\Psi(\mathbf{x}(t))A\,\mathbf{y}(t)\right)_k - \mathbf{x}(t)^\top \Psi(\mathbf{x}(t))A\,\mathbf{y}(t) \right) \\ \frac{\dot{y}_k}{y_k(t)} = \eta\tau \left(\left(\Psi(\mathbf{y}(t))B\,\mathbf{x}(t)\right)_k - \mathbf{y}(t)^\top \Psi(\mathbf{y}(t))B\,\mathbf{x}(t) \right) \end{cases} . \tag{14}$$

Let us define $\Pi^{(A)}_{\mathsf{SPGD}}(\mathbf{x},\mathbf{y}) = \Psi(\mathbf{x})\,A\,\mathbf{y}$, and $\Pi^{(B)}_{\mathsf{SPGD}}(\mathbf{x},\mathbf{y}) = \Psi(\mathbf{y})\,B\,\mathbf{x}$. To clarify further the relationship between SPGD and RD, we define the two following normal form games:

$$\mathcal{G} = (\{1,2\}, \{\mathcal{A}^1, \mathcal{A}^2\}, \{A, B\}),$$
$$\mathcal{P} = (\{1,2\}, \{\Delta^m, \Delta^m\}, \{\Pi^{(A)}_{\mathsf{SPGD}}, \Pi^{(B)}_{\mathsf{SPGD}}\}),$$

where, with abuse of notation, we use the payoffs matrices to identify the payoffs of $\mathcal{G}$. Similarly, the payoffs to the players in $\mathcal{P}$ are defined by $r^{(1)}(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\,\Pi^{(A)}_{\mathsf{SPGD}}(\mathbf{x},\mathbf{y})$, and $r^{(2)}(\mathbf{x},\mathbf{y}) = \mathbf{y}^\top\,\Pi^{(B)}_{\mathsf{SPGD}}(\mathbf{x},\mathbf{y})$. Once again, we observe that SPGD on $\mathcal{G}$ is equivalent to RD on the game $\mathcal{P}$.

Properties of $\mathcal{P}$. One of the main differences between the game $\mathcal{G}$ and $\mathcal{P}$ is that one cannot define the game $\mathcal{P}$ by redefining the payoff matrices A and B. Indeed, it also requires to change the correct action space in which to view the dynamics. Moreover, any pure strategy in the game $\mathcal{P}$ gives a zero payoff and the mixed extension of the game does not correspond to the expected value of pure strategies. Specifically, the game $\mathcal{P}$ is a differentiable game (using the definition by Letcher et al. (2019)), with payoffs $r^{(1)}$ and $r^{(2)}$. This shows that the study of SPGD is equivalent to the study of the properties of RD on the differentiable game $\mathcal{P}$ instead of the well studied normal form game $\mathcal{G}$. A complete study of the RD in a general differentiable game is left as a future work. Instead, we focus on the game theoretic

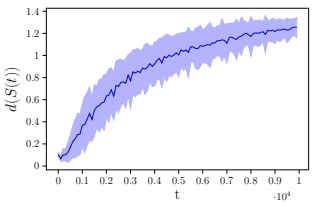

Figure 4: Evolution of the diameter $d(\cdot)$ of the set $S(t_0)$ over time.

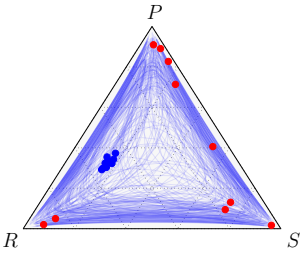

Figure 5: Trajectories of SPGD. The starting points of 10 trajectories are in blue, and the end points are in red.

properties that are preserved between the game $\mathcal{P}$ and $\mathcal{G}$, in particular concerning equilibria.

Let $\text{NE}(\mathcal{G})$ be the set of Nash equilibria of the normal-form game $\mathcal{G}$, and $\text{NE}(\mathcal{P})$ the set of Nash equilibria of the game $\mathcal{P}$. In the following theorem, we show that only over interior points $\text{NE}(\mathcal{G})$ and $\text{NE}(\mathcal{P})$ coincide:

Theorem 6. *In every normal-form game, it holds that:*

$$NE(\mathcal{G}) \cap \text{int}(\Delta^m)^2 = NE(\mathcal{P}) \cap \text{int}(\Delta^m)^2.$$

The behavior on to the border of the simplex is much more complex to study. Indeed, one can shows that the value of the payoffs at each NE of the non-linear game $\mathcal{P}$ is 0, and it is straightforward to observe that this value is attained for all couple of pure strategies. This suggests that game $\mathcal{P}$ has more NEs on the border with respect to $\mathcal{G}$, and some local stability properties could lead to unexpected behaviors when the dynamics are near pure strategies.

Volume and Convergence.

Even if the equivalence explored above points to which properties we can expect from SPGD, the non-linearity of $\mathcal{P}$ makes it impossible to apply the known results of RD to our case. In particular, two-population RD do not converge to interior points of $\Delta^m \times \Delta^m$ (Ritzberger and Weibull 1995, Proposition 6). This classical result is established by proving that RD in bimatrix games preserves a certain volume form: in particular, the dynamics of a suitable reparametrization of RD preserves the volume in the reparametrized space. Many recent papers (among others (Mertikopoulos, Papadimitriou, and Piliouras 2018; Cheung and Piliouras 2020)) exploit volume preservation properties of RD, to study the long-term behaviour of no-regret learning dynamics. The incompressibility results in these works are usually established for the RD in the cumulative payoff space as first done in (Hofbauer 1996). In what follows, we use the same argument as in the classical proof by Ritzberger and Weibull (1995) to show that SPGD preserves the volume in the interior of the product of simplices, allowing us to prove negative convergence results for each bimatrix game $\mathcal{G}$. As we already mentioned above, even with the equivalence that we have analyzed between RD and SPGD, we cannot apply known results or RD, since the multi-linearity of the payoffs is used as a key ingredient in the original proof. By the Liouville formula (2) described in Section 2, it is sufficient to show that the divergence of the reparametrized flow in the interior of the simplices is null, to obtain the invariance in time of the associated volume.

Lemma 4. *The flow of SPGD preserves a volume form in* $\text{int}(\Delta^m \times \Delta^m)$.

Thanks to Lemma 4 it is possible to obtain the non-convergence of the two-population SPGD in an interior point.

Theorem 7. *No closed set in* $\text{int}(\Delta^m \times \Delta^m)$ *is asymptotically stable for the SPGD.*

Since the volume considered blows up to infinity near the border of the simplex, its invariance does not prevent a priori a dynamics starting from the interior from converging asymptotically to the border, which may happen with pure NE. On the other hand, the theorem tells us that in games $\mathcal{G}$ like RPS, where the only NEs are in the interior of the strategies space, the dynamics will never converge to any NE. Theorem 7 also restricts the possible long-term behaviors to, either converge to the border, or being recurrent inside the interior of the action space.

Experimental Evaluation on Two Population Games

In this section, we analyze the behaviour of SPGD in two-population games. We analyse the same *rock-paper-scissor* game used in Section 4, in which both players strategies evolves according to the discrete SPG dynamics. We study the evolution of an initial set $S(t_0)$, and provide results on the evolution over time of its diameter $d(S(t))$, where the diameter $d(A)$ of a set A is formally defined as $d(A) = \sup_{\mathbf{x}_1, \mathbf{x}_2 \in A} ||\mathbf{x}_1 - \mathbf{x}_2||_2$. We run 50 independent experiments sampling uniformly (through rejection sampling) 10 points from the region $S(t_0)$ on the simplex with center in $(1/6, 1/3, 1/2)^\top$, and ℓ_1 diameter of $1/8$. We used $\eta = 0.1$ as learning rate of the SPG, time horizon $T = 10,000$, and $\mathbf{y}(t_0)$ has been initialized uniformly at random for each seed. Figure 4 shows the average approximate diameter $d(S(t))$, where the averaging is done over the 10 random points in the initial region $S(t_0)$ and light blue areas represents the standard deviation.

What emerges is that the diameter of an initial set $S(t_0)$ grows over time and converges to $\sqrt{2}$, which is the maximum ℓ_2 diameter in simplices. Intuitively, this means that any two strategies that are close at the beginning of the learning process, may end up in far points at the end of the learning process. The experimental evaluation points to the chaotic behaviour of the SPG algorithm, that also occurs for Multiplicative Weight Update (Arora, Hazan, and Kale 2012), *i.e.*, the discrete equivalent of RD in zero-sum games (Cheung and Piliouras 2020).

Figure 5 provides the dynamics of $\mathbf{x}(t)$ for one of the seeds. The starting strategies (depicted in blue) at t_0 are close together in the small region $S(t_0)$. Instead, at the end of the time horizon they are scattered throughout the entire strategy space (depicted in red). This suggest that there is indeed a chaotic behaviour, because small deviations in the starting initialization leads to large deviations in the final strategies.

References

Arora, S.; Hazan, E.; and Kale, S. 2012. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1): 121–164.

Brown, N.; Lerer, A.; Gross, S.; and Sandholm, T. 2019. Deep counterfactual regret minimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 793–802.

Cheung, Y. K.; and Piliouras, G. 2020. Chaos, Extremism and Optimism: Volume Analysis of Learning in Games. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Proceedings of the Neural Information Processing Systems Conference (NeurIPS)*.

Cressman, R.; Ansell, C.; and Binmore, K. 2003. *Evolutionary dynamics and extensive form games*, volume 5. MIT Press.

Cressman, R.; and Tao, Y. 2014. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences*, 111(Supplement 3): 10810–10817.

Gao, B.; and Pavel, L. 2017. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1861–1870.

Heinrich, J.; and Silver, D. 2016. Deep Reinforcement Learning from Self-Play in Imperfect-Information Games. *CoRR*, abs/1603.01121.

Hofbauer, J. 1996. Evolutionary dynamics for bimatrix games: a Hamiltonian system? *Journal of Mathematical Biology*, 34: 675–688.

Hofbauer, J.; and Sigmund, K. 1998. *Evolutionary Games and Population Dynamics*. Cambridge University Press.

Hu, S.; Leung, C.; Leung, H.; and Soh, H. 2020. The Evolutionary Dynamics of Independent Learning Agents in Population Games. *CoRR*, abs/2006.16068.

Kaisers, M.; and Tuyls, K. 2010. Frequency Adjusted Multi-Agent Q-Learning. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 309–316.

Karimi, H.; Nutini, J.; and Schmidt, M. 2016. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 795–811.

Klos, T.; van Ahee, G. J.; and Tuyls, K. 2010. Evolutionary Dynamics of Regret Minimization. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 82–96.

Letcher, A.; Balduzzi, D.; Racaniere, S.; Martens, J.; Foerster, J.; Tuyls, K.; and Graepel, T. 2019. Differentiable game mechanics. *The Journal of Machine Learning Research*, 20(1): 3032–3071.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2016. Continuous control with deep reinforcement learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 6382–6393.

Mei, J.; Xiao, C.; Dai, B.; Li, L.; Szepesvári, C.; and Schuurmans, D. 2020a. Escaping the gravitational pull of softmax. *Proceeding of the conference on Neural Information Processing Systems (NeurIPS)*, 33.

Mei, J.; Xiao, C.; Szepesvari, C.; and Schuurmans, D. 2020b. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the International Conference on Machine Learning (ICML)*, 6820–6829.

Mertikopoulos, P.; Papadimitriou, C. H.; and Piliouras, G. 2018. Cycles in Adversarial Regularized Learning. In Czumaj, A., ed., *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, (SODA)*, 2703–2717.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 1928–1937.

Peters, J.; and Schaal, S. 2006. Policy gradient methods for robotics. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2219–2225.

Ritzberger, K.; and Weibull, J. W. 1995. Evolutionary selection in normal-form games. *Econometrica: Journal of the Econometric Society*, 1371–1399.

Sandholm, W. H. 2009. *Evolutionary Game Theory*, 3176–3205. Springer New York.

Sandholm, W. H. 2010a. Local stability under evolutionary game dynamics. *Theoretical Economics*, 5(1): 27–50.

Sandholm, W. H. 2010b. *Population Games And Evolutionary Dynamics*. Economic learning and social evolution. MIT Press.

Sandholm, W. H. 2017. *Evolutionary Game Theory*, 1–38. Springer Berlin Heidelberg.

Shapley, L. S. 1953. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100.

Srinivasan, S.; Lanctot, M.; Zambaldi, V.; Perolat, J.; Tuyls, K.; Munos, R.; and Bowling, M. 2018. Actor-Critic Policy Optimization in Partially Observable Multiagent Environments. In *Proceedings of the Neural Information Processing Systems conference (NeurIPS)*, volume 31, 1–14.

Sutton, R. S.; McAllester, D. A.; Singh, S. P.; Mansour, Y.; et al. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the conference on Neural Information Processing Systems (NeurIPS)*, volume 99, 1057–1063.

Tuyls, K.; Verbeeck, K.; and Lenaerts, T. 2003. A Selection-Mutation Model for q-Learning in Multi-Agent Systems.

In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 693–700.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4): 229–256.