On the Evolutionary Dynamics of Soft-Max Policy Gradient in Multi-Agent Games

Martino Bernasconi
Politecnico di Milano
Milan, Italy
martino.bernasconideluca@polimi.it

Federico Cacciamani
Politecnico di Milano
Milan, Italy
federico.cacciamani@polimi.it

Simone Fioravanti
GSSI
L'Aquila, Italy
simone.fioravanti@gssi.it

Nicola Gatti
Politecnico di Milano
Milan, Italy
nicola.gatti@polimi.it

Francesco Trovò
Politecnico di Milano
Milan, Italy
francesco1.trovo@polimi.it

ABSTRACT

Policy gradient is one of the most famous algorithms in reinforcement learning. 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. Nevertheless, we show that the soft-max policy gradient dynamics in a given game is equivalent to the replicator dynamics in a different game derived by a non-linear, non-convex transformation of the payoffs of the original game. In this work, we first 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. Indeed, using such an approach, we can provide a complete characterization of the set of the *bad initializations* (points for which the dynamics initially moves towards sub-optimal strategies) for such dynamics, while this result was previously known only empirically. In the context of multi-agent environments, we also give a method that an agent can use to choose an initial strategy making the opponent to start always in a bad initialization region, thus slowing its learning process. 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].

[1]An extended version of this paper has been published at as an extended abstract at AAMAS 2022 [2].

Appears at the 1st Workshop on Learning with Strategic Agents (LSA 2022). Held as part of the Workshops at the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), N. Bishop, M. Han, L. Tran-Thanh, H. Xu, H. Zhang (chairs), May 9–10, 2022, Online.

KEYWORDS

Game Theory; Evolutionary Game Theory; Reinforcement Learning; Multiagent Learning

1 INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) recently demonstrated to be one of the most effective research fields in tackling complex multi-agent settings and leading to major Artificial Intelligence (AI) achievements, such as AlphaStar [47] and Libratus [5]. In 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. A plethora of MARL algorithms are available in the literature. We mention just a few of them: *Cross Learning* [7, 11], *Q-learning* [46] and its variations *Frequency-Adjusted Q-Learning* [20, 21] and *Lenient Frequency-Adjusted Q-Learning* [34], and *Polynomial Weights* [3, 24]. These algorithms provide theoretical guarantees only 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* [15] and *Deep-CFR* [6].

One of the mainstream approaches to study the learning dynamics of MARL algorithms is resorting to evolutionary game theory tools [18, 28, 39, 41, 49]. Introduced by Börgers and Sarin [7], such an approach models the continuous-time mean dynamics of an algorithm by resorting to evolutionary models based on dynamical systems. Thanks to that, these dynamical systems 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* [20, 21, 34, 46], and a family of no-regret algorithms [24] 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 [35, 45]. It comes under various flavours such as, *e.g.*, SAC [13], DDPG [26], MADDPG [27], A3C [33], REINFORCE [50]. 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. In the work by Hennes et al. [16], the authors provide experimental evidence that in some settings its performance may be more inefficient than other MARL algorithms and suggest a modification of the update rule. Such a modification leads to an algorithm named NeuRD, whose mean dynamics are equivalent to the replicator dynamics. Conversely, our work follows a different approach, maintaining the original definition of the soft-max policy gradient algorithm and providing its evolutionary game theory analysis. In doing that, we also find 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 simultaneously learn. 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. [19] 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. Furthermore, the study of single-population games is a crucial preliminary step for the study of multi-population games. This latter setting captures the case in which multiple agents learn their best strategy, each using a learning algorithm independently from the others. 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. Our analysis paves the way to further evolutionary game theory studies of policy-gradient-based algorithms (including, *e.g.*, NeuRD [16]) when the policy space is restricted to assess the impact of special policy space structures on the evolutionary dynamics.

Original Contributions. Differently from most MARL algorithms which have mean dynamics that are slight variants of the Replicator Dynamics (shortly RD from here on), the Soft-max Policy Gradient Dynamics (shortly SPGD from here on) present a different structure not corresponding to any known evolutionary dynamics (see the work by Sandholm [41] for a detailed discussion of the main known dynamics), and, thus, their study is an open problem. However, SPGD preserves a close connection with RD, corresponding to RD but applied to a non-linear, non-convex fitness function, in which the correct action space is no longer the discrete action space, but the entire Cartesian product of simplices. We separately analyze the dynamics when learning the best response from the cases of single and multi-population games. In particular, we show that SPGD always converges to the best response, and provide an upper bound to the convergence rate. Moreover, we show that, differently from RD, SPGD suffers from a non-empty simplex subspace of bad initializations with the property that, when starting from these points, the dynamics are initially more attracted towards a sub-optimal action rather than by the best response. Such an attraction holds until the dynamics do not leave that subspace, and, after that, the dynamics monotonically converge to the best response. Such a non-monotonic behavior makes the convergence slow in practice. Interestingly, we show that such a subspace is always non-empty, we characterize it exactly, and we show that there are always good initializations—*e.g.*, the center of the simplex—from which the dynamics monotonically converge to the best response. Moreover, in the class of games in which there exists a fully mixed equilibrium, we show that the opponent can always make the agent strategy to converge slowly. We also analyze the case of multiple agents employing the SPG dynamic, both in the case in which we have single and multiple-populations. In this setting, we can prove that, in terms of asymptotic convergence in the interior of the strategy space, SPGD and RD present similar properties. In particular, in single-population games, the spaces of asymptotically stable states in the interior of the simplex of RD and SPGD coincide. Moreover, we show that in multi-population games, the volume is conserved

in a reparametrized space, implying SPGD cannot converge to a fully-mixed Nash equilibrium. Following this direction, we provide experimental evaluation showing that the dynamics are chaotic. More precisely, we show that the diameter of a small set of initial points increases over time and that a small deviation from an initial starting point results in a large deviation in the ending point.

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. A *solution* to a normal-form game is a strategy profile that is in equilibrium according to some equilibrium concept. 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:

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

Note that, the asymptotically-stable states of the above evolutionary model are a subset of the Nash equilibria (see for example [10, 18, 42]). 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 [42]:

Definition 2.1. 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 O of $\mathbf{x}$ such that $(\mathbf{z} - \mathbf{x})^\top \Pi(\mathbf{z}) < 0$ for all the strategies $\mathbf{z} \in O \setminus \{\mathbf{x}\}$.[2]

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 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 $\mathrm{div}(f) = \sum_i^d \frac{\partial f_i}{\partial x_i}$ 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 convergence to a single point. This property will be analysed for the trajectories of two-population SPGD in Section 5.2.

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 define as:

$$\mathop{\mathbb{E}}_{\substack{a(t)\sim\pi(\cdot|s(t))\\ s(t+1)\sim P(\cdot|s(t),a(t))\\ s(0)\sim\mu}} \left[\sum_{t\geq 0} \gamma^t\, r\Big(s(t), a(t), s(t+1)\Big)\,\middle|\, s(0)\right].$$

Reinforcement Learning (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

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

policy space $\pi(\cdot|s,\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a real-valued vector of parameters. Formally, the problem consists in 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:

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

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 $\mathbf{y}(t)$.[3] This dynamics was already presented in [44], 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 [43]). 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:

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

Following the procedure proposed by Tuyls et al. [46], 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:

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

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

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:

$$\Psi(\mathbf{x}) = \begin{bmatrix} x_1\,(1-x_1) & -x_1\,x_2 & \cdots & -x_1\,x_m \\ -x_1\,x_2 & & & -x_2\,x_m \\ \vdots & & \ddots & \vdots \\ -x_1\,x_m & -x_2\,x_m & \cdots & x_m\,(1-x_m) \end{bmatrix}.$$

Note that it can also be expressed 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:

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

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

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: $\Pi_{\mathrm{RD}}(\mathbf{x}(t),\mathbf{y}(t)) = A\,\mathbf{y}(t)$ in RD, while $\Pi_{\mathrm{SPGD}}(\mathbf{x}(t),\mathbf{y}(t)) = \Psi(\mathbf{x}(t))\,A\,\mathbf{y}(t)$ in SPGD. Moreover, note that the matrix $\Psi(\mathbf{x})$ is singular (see Proposition 2 by Gao and Pavel [12]), 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 $\Pi_{\mathrm{SPGD}}(\mathbf{x}(t),\mathbf{y}(t))_j = [\Psi(\mathbf{x}(t))A\,\mathbf{y}(t)]_j = x_j(t)(\mathbf{e}_j - \mathbf{x}(t))^\top A\,\mathbf{y}(t)$ 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 [23] and is instrumental to our analysis.

Lemma 4.1. *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:*

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

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.*[4]

Relying on the result provided by Lemma 4.1, 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 4.2. *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* $\bar{\mathbf{e}}_j$*.*

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 4.3. *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 [30, Theorem 2].

Theorem 4.4. *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 4.4 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. [30, 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 $O(1/t)$.

4.1 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 [39].

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

Let $\bar{\mathbf{e}}_j$ be the unique pure best response, for every action $a_k \neq a_j$, it holds that:

$$\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) = \frac{\dot{x}_k(t)\,x_j(t) - \dot{x}_j(t)\,x_k(t)}{x_j^2(t)}$$
$$= \frac{x_k(t)}{x_j(t)}\left[(A\ \mathbf{y})_k - (A\ \mathbf{y})_j\right] < 0,$$

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 $k \neq j$.[5] 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.

Theorem 4.5. *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:*

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

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 4.5 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 4.5 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. [29] 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 4.5 guarantees that an agent always can choose $\mathbf{1}/m$ as a good initialisation. This result is in line with the result stated by Mei et al. [30, 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 4.5 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.

[5]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 $\bar{t} > 0$.

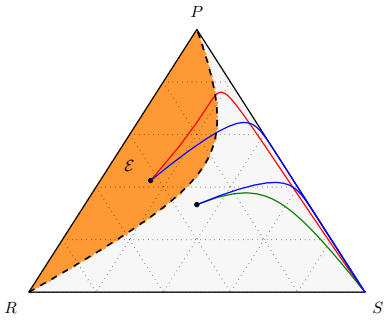

Figure 1: SPGD (red and green) and RD (in blue) trajectories in RPS game. The subspace in orange contains the bad initializations.

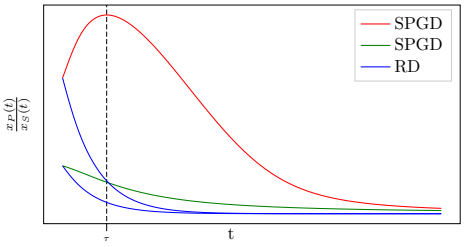

Figure 2: Ratio $x_P(t)/x_S(t)$ over time $t \geq 0$. At $t = \tau$ the dynamics in red leave subspace $\mathcal{E}$.

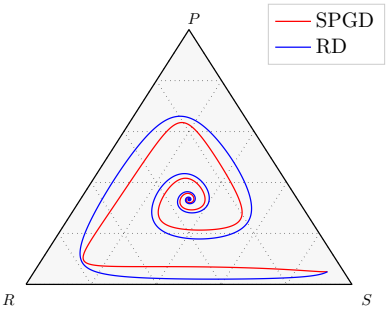

Figure 3: SPGD and RD trajectories with the good RPS population game.

Theorem 4.6. *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.

5.1 Single-Population Games

With a single population, SPGD is equivalent to RD, once the fitness function has been redefined as $\Pi_{\text{SPGD}}(\mathbf{x}) = \Psi(\mathbf{x})\, A\, \mathbf{x}$. Therefore, SPGD satisfy the same properties (see, *e.g.*, [39]) 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:

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

In the work by Sandholm [41], 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 $\text{int}(\Delta^m)$. 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 5.1. *SPGD satisfy properties NS, when restricting to* $\text{int}(\Delta^m)$, *and PC.*

Finally, we focus on the relationship between the asymptotically stable states of SPGD and those of RD. It is 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.*, [9, 41]. 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) [40], defined as follows:

Definition 5.2. Strategy $\bar{\mathbf{x}} \in$ is a RESS for a population game with fitness function $\Pi(\cdot)$ if:

(i) $\Pi_k(\bar{\mathbf{x}}) = \bar{\mathbf{x}}^\top \Pi(\bar{\mathbf{x}}) > \Pi_j(\bar{\mathbf{x}})$, whenever $\bar{x}_k > 0$ 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 $\Pi_{\text{SPGD}}(\mathbf{x})$.

Lemma 5.3. *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 $\Pi_{\text{SPGD}}(\mathbf{x})$.*

Interestingly, Lemma 5.3 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 [41]. Indeed, in Figure 3, we see, as prescribed by Lemma 5.3, 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 together with Lemma 5.3 we show the following result.

Theorem 5.4. *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 (for a proof we point at the book by Hofbauer and Sigmund [18]). Theorem 5.4 shows that the same behavior of RD also holds with SPGD, over the internal ESS.

5.2 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 $k \in \{1, \dots, m\}$:

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

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

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

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 $r^{(1)}(\mathbf{x},\mathbf{y}) = \mathbf{x}^\top\,\Pi^{(A)}_{\text{SPGD}}(\mathbf{x},\mathbf{y})$, and $r^{(2)}(\mathbf{x},\mathbf{y}) = \mathbf{y}^\top\,\Pi^{(B)}_{\text{SPGD}}(\mathbf{x},\mathbf{y})$. 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 (see Lemma C.1 below) 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. [25]), 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 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 5.5. *In every normal-form game, it holds that:*

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

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 (see for instance Lemma C.1 in the Appendix). 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 $\Delta^m \times \Delta^m$ [38, 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 [4, 8, 32, 37]) 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 [17].In what follows, we use the same argument as in the classical proof by Ritzberger and Weibull [38] 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 5.6. *The flow of SPGD preserves a volume form in* $\text{int}(\Delta^m \times \Delta^m)$.

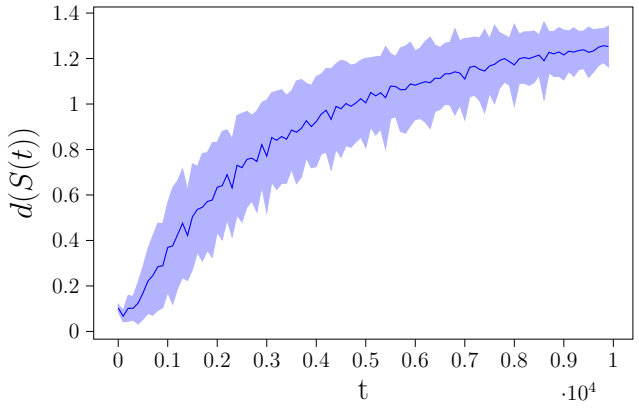

Figure 4: Evolution of the diameter $d(\cdot)$ of the set $S(t_0)$ over time.

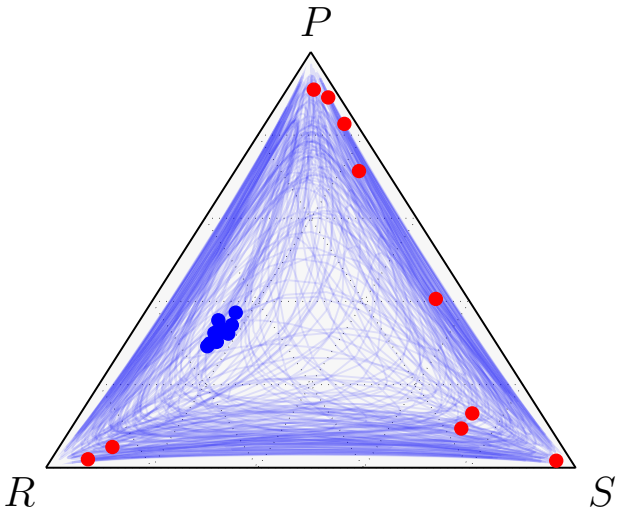

Figure 5: Trajectories of SPGD. The starting points $\mathbf{x}(t_0) \in S(t_0)$ of 10 trajectories are depicted in blue, while the end points $\mathbf{x}(T) \in S(T)$ are depicted in red.

Thanks to Lemma 5.6 it is possible to obtain the non-convergence of the two-population SPGD in an interior point.

Theorem 5.7. *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 5.7 also restricts the possible long-term behaviors to, either converge to the border, or being recurrent inside the interior of the action space.

5.3 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 et al. [1], *i.e.*, the discrete equivalent of RD in zero-sum games [8].

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.

6 CONCLUSIONS AND FUTURE WORKS

In this work, we derive and study the continuous-time mean dynamics of soft-max policy gradient, namely SPGD, in different scenarios. First, we focus on the scenario in which an agent learns the best response in a normal-form game against a fixed-strategy opponent. We show that SPGD converge to the best response strategy in non-degenerate games and derive a convergence rate for the dynamics. In the same setting, we show that SPGD are less efficient than RD since the ratio between the probability of using a non-optimal action and the one of using the best response is not always decreasing over time. Moreover, we show that an opponent can always exploit such features in RPS-like games that have a unique fully-mixed equilibrium. In the self-play scenario, we discuss the classical EGT properties for SPGD, and show that an internal ESS for the normal-form game is asymptotically stable for SPGD. Finally, we study the case of two populations jointly evolving according to SPGD. By leveraging the connection with RD, we show that SPGD do not converge to interior NEs.

A natural future direction is to analyze the dynamics of SPGD in multi-agents games with multiple states, such as Extensive Form Games. Another interesting direction is to analyze, under the EGT lenses, other flavors of policy-gradients, such as Natural Policy

Gradient [22]. Finally, it is an interesting line of research the study of the behaviour of RD on general non-convex payoff functions, which may lead to even deeper insights on the behaviour of SPGD, by exploiting the exact connection we drew between these two dynamics.

REFERENCES

[1] S. Arora, E. Hazan, and S. Kale, "The multiplicative weights update method: a meta-algorithm and applications," *Theory of Computing*, vol. 8, no. 1, pp. 121–164, 2012.

[2] M. Bernasconi, F. Cacciamani, S. Fioravanti, N. Gatti, and F. Trovò, "The evolutionary dynamics of soft-max policy gradient in multi-agent settings," in *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2022.

[3] A. Blum and Y. Mansour, *Learning, Regret Minimization, and Equilibria.* Cambridge University Press, 2007, p. 79–102.

[4] V. Boone and G. Piliouras, "From darwin to poincaré and von neumann: Recurrence and cycles in evolutionary and algorithmic game theory," in *Web and Internet Economics - 15th International Conference, WINE 2019, New York, NY, USA, December 10-12, 2019, Proceedings*, ser. Lecture Notes in Computer Science, I. Caragiannis, V. S. Mirrokni, and E. Nikolova, Eds., vol. 11920. Springer, 2019, pp. 85–99.

[5] N. Brown and T. Sandholm, "Superhuman ai for heads-up no-limit poker: Libratus beats top professionals," *Science*, vol. 359, no. 6374, pp. 418–424, 2018.

[6] N. Brown, A. Lerer, S. Gross, and T. Sandholm, "Deep counterfactual regret minimization," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 793–802.

[7] T. Börgers and R. Sarin, "Learning through reinforcement and replicator dynamics," *Journal of Economic Theory*, vol. 77, no. 1, pp. 1–14, 1997.

[8] Y. K. Cheung and G. Piliouras, "Chaos, extremism and optimism: Volume analysis of learning in games," in *Proceedings of the Neural Information Processing Systems Conference (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[9] R. Cressman and Y. Tao, "The replicator equation and other game dynamics," *Proceedings of the National Academy of Sciences*, vol. 111, no. Supplement 3, pp. 10 810–10 817, 2014.

[10] R. Cressman, C. Ansell, and K. Binmore, *Evolutionary dynamics and extensive form games.* MIT Press, 2003, vol. 5.

[11] J. G. Cross, "A stochastic learning model of economic behavior," *The Quarterly Journal of Economics*, vol. 87, no. 2, pp. 239–266, 1973.

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

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

[14] W. Hahn, H. H. Hosenthien, and H. Lehnigk, *Theory and application of Liapunov's direct method.* Prentice-Hall Englewood Cliffs, NJ, 1963, vol. 3(4).

[15] J. Heinrich and D. Silver, "Deep reinforcement learning from self-play in imperfect-information games," *CoRR*, vol. abs/1603.01121, 2016. [Online]. Available: http://arxiv.org/abs/1603.01121

[16] D. Hennes, D. Morrill, S. Omidshafiei, R. Munos, J. Perolat, M. Lanctot, A. Gruslys, J.-B. Lespiau, P. Parmas, E. Duenez-Guzman, and K. Tuyls, "Neural replicator dynamics," 2020.

[17] J. Hofbauer, "Evolutionary dynamics for bimatrix games: a hamiltonian system?" *Journal of Mathematical Biology*, vol. 34, pp. 675–688, 1996.

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

[19] S. Hu, C. Leung, H. Leung, and H. Soh, "The evolutionary dynamics of independent learning agents in population games," *CoRR*, vol. abs/2006.16068, 2020.

[20] M. Kaisers and K. Tuyls, "Frequency adjusted multi-agent q-learning," in *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010, p. 309–316.

[21] ——, "Faq-learning in matrix games: Demonstrating convergence near nash equilibria, and bifurcation of attractors in the battle of sexes," in *Proceedings of the Workshop on Interactive Decision Theory and Game Theory*, 2011, p. 36–42.

[22] S. M. Kakade, "A natural policy gradient," in *Proceedings of the Neural Information Processing Systems conference (NeurIPS)*, vol. 14, 2001, pp. 1–8.

[23] H. Karimi, J. Nutini, and M. Schmidt, "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)*, 2016, pp. 795–811.

[24] T. Klos, G. J. van Ahee, and K. Tuyls, "Evolutionary dynamics of regret minimization," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2010, pp. 82–96.

[25] A. Letcher, D. Balduzzi, S. Racaniere, J. Martens, J. Foerster, K. Tuyls, and T. Graepel, "Differentiable game mechanics," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 3032–3071, 2019.

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

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

[28] J. Maynard Smith and G. R. Price, "The logic of animal conflict," *Nature*, vol. 246, no. 5427, pp. 15–18, 1973.

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

[30] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 6820–6829.

[31] J. D. Meiss, *Differential dynamical systems.* SIAM, 2007.

[32] P. Mertikopoulos, C. H. Papadimitriou, and G. Piliouras, "Cycles in adversarial regularized learning," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, (SODA)*, A. Czumaj, Ed., 2018, pp. 2703–2717.

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

[34] L. Panait and K. Tuyls, "Theoretical advantages of lenient q-learners: An evolutionary game theoretic perspective," in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2007.

[35] J. Peters and S. Schaal, "Policy gradient methods for robotics," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2006, pp. 2219–2225.

[36] M. M. Petrovitch, "Sur une manière d'étendre le théorème de la moyenne aux équations différentielles du premier ordre," *Mathematische Annalen*, vol. 54, no. 3, pp. 417–436, 1901.

[37] G. Piliouras and J. S. Shamma, "Optimization despite chaos: Convex relaxations to complex limit sets via poincaré recurrence," in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, (SODA)*, 2014, pp. 861–873.

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

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

[40] ——, "Local stability under evolutionary game dynamics," *Theoretical Economics*, vol. 5, no. 1, pp. 27–50, 2010.

[41] ——, *Evolutionary Game Theory.* Springer Berlin Heidelberg, 2017, pp. 1–38.

[42] ——, *Population Games And Evolutionary Dynamics*, ser. Economic learning and social evolution. MIT Press, 2010.

[43] L. S. Shapley, "Stochastic games," *Proceedings of the national academy of sciences*, vol. 39, no. 10, pp. 1095–1100, 1953.

[44] S. Srinivasan, M. Lanctot, V. Zambaldi, J. Perolat, K. Tuyls, R. Munos, and M. Bowling, "Actor-critic policy optimization in partially observable multiagent environments," in *Proceedings of the Neural Information Processing Systems conference (NeurIPS)*, vol. 31, 2018, pp. 1–14.

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

[46] K. Tuyls, K. Verbeeck, and T. Lenaerts, "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)*, 2003, p. 693–700.

[47] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[48] B. Von Stengel, "Computing equilibria for two-person games," *Handbook of game theory with economic applications*, vol. 3, pp. 1723–1759, 2002.

[49] J. Weibull, *Evolutionary game theory.* MIT Press, 1995.

[50] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

A APPENDIX: OMITTED PROOFS FROM SECTION 4

A.1 Proof of Theorem 4.2

In this section we provide the proof of Theorem 4.2, and we will also prove the required intermediate results.

Lemma 4.1. *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:*

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

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.*[6]

Proof. It follows by the relation between 2-norm and ∞-norm, and by definition of best response that:

$$||\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})||_2^2 \geq ||\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})||_\infty^2 \tag{15}$$

$$= \max_k \left(x_j(\boldsymbol{\theta})\,(\mathbf{e}_k - \mathbf{x}(\boldsymbol{\theta}))^\top A\ \mathbf{y}\right)^2 \tag{16}$$

$$\geq x_j\left(\boldsymbol{\theta}\right)^2 (J^* - J(\boldsymbol{\theta}))^2\Big)^2. \tag{17}$$

This concludes the proof. □

Theorem 4.2. *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* $\bar{\mathbf{e}}_j$*.*

Proof. Let us recall the definition of the function $V(t)$:

$$V(t) := J^* - J(\boldsymbol{\theta}(t)) = \bar{\mathbf{e}}_j^\top A\ \mathbf{y} - \mathbf{x}(\boldsymbol{\theta}(t))^\top A\ \mathbf{y}\,. \tag{18}$$

Since J^* is the optimum of the function $J(t)$, we have $V(t) \geq 0$, and, in particular, $V(t) > 0$ if $\mathbf{x}(t) \neq \bar{\mathbf{e}}_j$ being $\bar{\mathbf{e}}_j$ the single best response. The time derivative of $V(t)$ is:

$$\dot{V}(t) = -\frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\,\frac{\partial \boldsymbol{\theta}}{\partial t} = -\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}(t))^\top\,\dot{\boldsymbol{\theta}}(t). \tag{19}$$

Substituting the definition of $\dot{\boldsymbol{\theta}}(t)$ provided in Equation (6), we have:

$$\dot{V}(t) = -\eta\,||\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}(t))||_2^2, \tag{20}$$

and, using Lemma 4.1 and the positivity of $V(r)$, we have:

$$\dot{V}(t) \leq -\eta\,x_j(\boldsymbol{\theta}(t))^2\,V(t)^2 < 0, \tag{21}$$

where the inequality is strict as, by the nature of Soft-Max, using an initial strategy s.t. $x_j(0) > 0$ implies that $x_j(t) > 0$ for each $t > 0$. As a consequence, we have that the unique zero of the $V(t)$ function is $\bar{\mathbf{e}}_j$, formally:

$$V(t) = 0 \quad \Longleftrightarrow \quad \mathbf{x}(t) = \bar{\mathbf{e}}_j.$$

Hence, $V(t)$ is a Lyapunov function for the system as the following conditions hold:

$$\begin{array}{ll} V(t) > 0 & \forall\, \mathbf{x}(t) \neq \bar{\mathbf{e}}_j, \\ V(t) = 0 & \mathbf{x}(t) = \bar{\mathbf{e}}_j, \\ \dot{V}(t) < 0 & \forall\, \mathbf{x}(t) \neq \bar{\mathbf{e}}_j, \\ \dot{V}(t) = 0 & \mathbf{x}(t) = \bar{\mathbf{e}}_j. \end{array}$$

Thus, SPGD satisfies the assumptions required by Lyapunov's Lemma [14], and, consequently, they asymptotically converge to $\bar{\mathbf{e}}_j$. □

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

A.2 Proof of Theorem 4.4

In what follows we give the proof of Theorem 4.4. We will first introduce some definitions. Let us define two subspaces $\mathcal{S}_1, \mathcal{S}_2 \in \Delta^m$ as follows (subscript j corresponds to action a_j played with probability one in the best response $\bar{\mathbf{e}}_j$):

$$\mathcal{S}_1 = \left\{ \mathbf{x} \in \Delta^m : \frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0, \quad \forall k \neq j \right\}, \tag{22}$$

$$\mathcal{S}_2 = \left\{ \mathbf{x} \in \Delta^m : x_j \geq \frac{m-\xi}{m+\xi} \right\}, \tag{23}$$

where ξ is the optimality gap between the best response $\bar{\mathbf{e}}_j$ and the second best response, *i.e.*, $\xi := \bar{\mathbf{e}}_j^\top A\,\mathbf{y} - \max_{k \neq j}\left\{\mathbf{e}_k^\top A\,\mathbf{y}\right\}$. Notice that the definition of $\mathcal{S}_1$ does not depend on time. Indeed, the flow (see [31] for details) of SPGD does not depend explicitly on time, which means that the derivative $\bar{\mathbf{x}}(t)$ depends only on the current state $\mathbf{x}(t)$ in the simplex. Therefore, subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ are invariant in time. We introduce the following three lemmas stating some interesting properties satisfied by $\mathcal{S}_1$, and $\mathcal{S}_2$, which are instrumental in deriving the upper bound on the convergence rate for SPGD. The following lemma shows that, if SPGD enter the subspace $\mathcal{S}_1$ at a certain point in time, then it will never leave this subspace.

Lemma A.1. *If it exists a time $t_0 > 0$ such that $\mathbf{x}(t_0) \in \mathrm{int}(\Delta^m) \cap \mathcal{S}_1$, then SPGD is such that $\mathbf{x}(t) \in \mathcal{S}_1$ for every $t > t_0$.*

Proof. Let us remark that subscript j is associated with the best response action a_j and that $\mathbf{x}(t) \in \mathcal{S}_1$. Since $\mathbf{x}(t) \in \mathcal{S}_1$ if and only if $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$ for each $k \neq j$, in what follows, we show that this inequality is satisfied by each component separately for every $t \geq t_0$. Initially, we rewrite the above derivative as:

$$\begin{aligned}
\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) &= \frac{\dot{x}_k(t)\,x_j(t) - \dot{x}_j(t)\,x_k}{x_j(t)^2} \\
&= \eta\,\tau\,\frac{x_k(t)}{x_j(t)}\left[x_k(t)\left((A\,\mathbf{y})_k - \mathbf{x}(t)^\top A\,\mathbf{y}\right) - x_j(t)\left((A\,\mathbf{y})_j - \mathbf{x}(t)^\top A\,\mathbf{y}\right)\right] \\
&= \eta\,\tau\,\frac{x_k(t)}{x_j(t)}\left[x_j(t)\,(A\,\mathbf{y})_k - x_j(t)\,(A\,\mathbf{y})_j + (x_j(t) - x_k(t))\,\mathbf{x}^\top(t)\,A\,\mathbf{y}\right] \\
&= \eta\,\tau\,\frac{x_k(t)}{x_j(t)}\left[x_k(t)\,\left((A\,\mathbf{y})_k - (A\,\mathbf{y})_j\right) + (x_j(t) - x_k(t))\left(\mathbf{x}^\top(t)\,A\,\mathbf{y} - (A\,\mathbf{y})_j\right)\right].
\end{aligned} \tag{24}$$

Notice that, being j the index associated to the best response, $x_k(t)\,\left((A\,\mathbf{y})_k - (A\,\mathbf{y})_j\right) < 0$ and $\left(\mathbf{x}^\top(t)\,A\,\mathbf{y} - (A\,\mathbf{y})_j\right) < 0$. Let us study separately two cases, according to the sign of the term $x_j(t_0) - x_k(t_0)$.

Case 1: $x_j(t_0) \geq x_k(t_0)$. From Equation (24) we have that $\frac{d}{dt}\left(\frac{x_k(t_0)}{x_j(t_0)}\right) < 0$, since the summands in the square brackets are both negative. Therefore, it suffices to show that for each $\delta > 0$ holds that $x_j(t_0+\delta) \geq x_k(t_0+\delta)$. Assume, by contradiction, that exists $\delta := \inf_{\delta>0}\{x_j(t_0+\delta) - x_k(t_0+\delta)\} \leq 0$. By continuity, we have that $x_j(t_0+\delta) = x_k(t_0+\delta)$. Using the fact that for all $t \in [t_0,\ t_0+\delta)$ it holds that $x_j(t) - x_k(t) > 0$, we have that $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$ for every $t \in [t_0,\ t_0+\delta)$. The fact that the derivative is negative implies that $\frac{x_k(t_0+\delta)}{x_j(t_0+\delta)} < \frac{x_k(t_0)}{x_j(t_0)}$. Summarizing, we have the following implications:

$$1 = \frac{x_k(t_0+\delta)}{x_j(t_0+\delta)} < \frac{x_k(t_0)}{x_j(t_0)} < 1,$$

where the last inequality holds by hypothesis of **Case 1**, which is a contradiction. This implies that in **Case 1** for each $t > t_0$ we have $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$.

Case 2: $x_j(t_0) < x_k(t_0)$. Let us define:

$$K_k(t) := \frac{1}{\eta\,\tau}\,\frac{1}{x_k(t)}\,\frac{x_j(t)}{x_k(t)}\,\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right). \tag{25}$$

Since $\eta\,\tau\,x_k^2(t)/x_j(t) > 0$, Equation (24) implies that:

$$\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0 \Leftrightarrow K_k(t) < 0.$$

Since $(A\ \mathbf{y})_j = \sum_{h=1}^{m} x_h(t)(A\ \mathbf{y})_j$ as $\sum_{h=1}^{m} x_h(t) = 1$, we can write:

$$\mathbf{x}(t)^\top A\ \mathbf{y} - (A\ \mathbf{y})_j = -\sum_{h=1}^{m} x_h(t)\left[(A\ \mathbf{y})_j - (A\ \mathbf{y})_h\right] = -\sum_{h=1}^{m} x_h(t)\, d_h, \tag{26}$$

where $d_h := (A\ \mathbf{y})_j - (A\ \mathbf{y})_h$. Writing $K_k(t)$ in terms of d_h, we have:

$$K_k(t) = \left(1 - \frac{x_j(t)}{x_k(t)}\right) \sum_{h=1}^{m} x_h(t)\, d_h - d_k. \tag{27}$$

where we used the definition of d_h and Equation (26) in Equation (24).

The time derivative of $K_k(t)$ is:

$$\dot{K}_k(t) = -\frac{d}{dt}\left(\frac{x_j(t)}{x_k(t)}\right) \sum_{h=1}^{m} x_h(t)\, d_h + \left(1 - \frac{x_j(t)}{x_k(t)}\right) \sum_{h=1}^{m} \dot{x}_h(t)\, d_h. \tag{28}$$

In the following, we show that $\dot{K}_k(t)\big|_{t=t_0} = 0$. Indeed, we have:

- $\frac{d}{dt}\left(\frac{x_j(t)}{x_k(t)}\right)\Big|_{t=t_0} > 0$ by hypothesis as $\mathbf{x}(t_0) \in \mathcal{S}_1$;
- $\sum_{h=1}^{m} x_h(t)\, d_h > 0$ as, by definition of d_h, every $d_h > 0$ for every $h \neq j, t \geq 0$;
- $\left(1 - \frac{x_j(t)}{x_k(t)}\right)\Big|_{t=t_0} > 0$ as we are in **Case 2**.

Since $\sum_{h=1}^{m} \dot{x}_h(t) = 0$ as $\sum_{h=1}^{m} x_h(t) = 1$ for every t, we have that:

$$\begin{aligned}
\sum_{h=1}^{m} \dot{x}_h(t)\, d_h &= \sum_{h} \dot{x}_h(t)(A\,\mathbf{y})_j - \sum_{h=1}^{m} \dot{x}_h(t)\ (A\ \mathbf{y})_h \\
&= -\sum_{h=1}^{m} \dot{x}_h(t)\ (A\ \mathbf{y})_h \\
&= -\dot{\mathbf{x}}(t)^\top A\ \mathbf{y} \\
&= -\dot{J}(\boldsymbol{\theta}(t)),
\end{aligned}$$

where we used the definitions of d_h and $\dot{J}(\cdot)$. A consequence of Theorem 4.2 is that $\dot{J}(\boldsymbol{\theta}(t)) \geq 0$, which implies that $\sum_{h=1}^{m} \dot{x}_h(t)\, d_h \leq 0$. Hence, Equation (28) consists in the the sum of two negative summands, and, thus, $\dot{K}_k(t) < 0$ in $t = t_0$. Most importantly, the same holds for every t is such that $\mathbf{x}(t) \in \mathcal{S}_1$ and $x_k(t)$ satisfies the assumption of **Case 2**. Assume by contradiction that there is $\delta > 0$ such that at $t = t_0 + \delta$ the trajectory of the dynamics leaves $\mathcal{S}_1$, *i.e.*, $\delta = \inf_{\delta > 0}\{K_k(t_0 + \delta) \geq 0\}$. By continuity, we have that $K_k(t_0 + \delta) = 0$. Moreover, it holds that $\dot{K}_k(t) < 0$ for all $t \in [t_0, t_0 + \delta)$ which means that $K_k(t_0 + \delta) < K_k(t_0) < 0$, thus contradicting the above assumption. Overall, we have that in **Case 2** we have that for each $t > t_0$ we have $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$.

The theorem statement follows from the fact that for each $t > t_0$ we have $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$ and the definition of $\mathcal{S}_1$.

□

The following lemma, instead, shows the inclusion relationship between $\mathcal{S}_2$ and $\mathcal{S}_1$.

Lemma A.2. *It holds $\mathcal{S}_2 \subset \mathcal{S}_1$.*

Proof. Let us assume w.l.o.g. that the payoffs are in $[0, 1]$. This can be always achieved by an opportune affine transformation of the original game. Recall that j is the index corresponding to the coordinate associated with the best response a_j. Let now $\mathbf{x}(t) \in \mathcal{S}_2$. We divide again the study into two cases, based on whether $x_k(t) \leq x_j(t)$ for all k, or $x_k(t) > x_j(t)$ for least one k.

Case 1: $x_k(t) \leq x_j(t)$ for all k. As showed in the proof of Lemma A.1[**Case 1**], if $\mathbf{x}(t)$ is such that $x_k(t) \leq x_j(t)$, it holds $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$. As a consequence, if $\mathbf{x}(t)$ is such that $x_k(t) \leq x_j(t)$ for every $k \neq j$, then $\mathbf{x}(t) \in \mathcal{S}_1$.

Case 2: $x_k(t) > x_j(t)$ for at least one index k. From the proof of Lemma A.1 we have that:

$$\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) = \eta\,\tau\,\frac{x_k(t)}{x_j(t)}\left[\underbrace{x_k(t)\left((A\ \mathbf{y})_k - \mathbf{x}(t)^\top A\ \mathbf{y}\right) - x_j(t)\left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y}\right)}_{=:P}\right].$$

The sign of $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right)$ has the same sign of P, since the factor multiplying the square brackets is positive. In what follows, we show that $\mathbf{x} \in \mathcal{S}_2$ implies that $-P \geq 0$, for all $k \neq j$. We have:

$$-P = x_j(t)\left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y}\right) - x_k(t)\ \left((A\ \mathbf{y})_k - \mathbf{x}(t)^\top A\ \mathbf{y}\right) \tag{29}$$

$$= x_j(t)\left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y}\right) - x_k(t)\ \left((A\ \mathbf{y})_k - \mathbf{x}(t)^\top A\ \mathbf{y}\right) + \sum_{h=1}^{m} x_h(t)\ \left((A\ \mathbf{y})_h - \mathbf{x}(t)^\top A\ \mathbf{y}\right) \tag{30}$$

$$= 2x_j(t)\left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\,\mathbf{y}\right) + \sum_{h=1,h\neq k,j}^{m} x_h\left((A\ \mathbf{y})_h - \mathbf{x}(t)^\top A\ \mathbf{y}\right) \tag{31}$$

$$= 2x_j(t)\left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y}\right) + \sum_{h=1,h\neq k,j}^{m} x_h\left((A\ \mathbf{y})_h - \mathbf{x}(t)^\top A\ \mathbf{y} + (A\ \mathbf{y})_j - (A\ \mathbf{y})_j\right) \tag{32}$$

$$= 2x_j(t)\left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y}\right) - \sum_{h=1,h\neq k,j}^{m} x_h(t)\,d_h + \sum_{h=1,h\neq k,j}^{m} x_h(t)\left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y}\right) \tag{33}$$

$$= \left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y}\right)\left(2x_j(t) + \sum_{h=1,h\neq i,j}^{m} x_h(t)\right) - \sum_{h=1,h\neq k,j}^{m} x_h(t)\,d_h \tag{34}$$

$$\geq \left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y}\right)\left(2x_j(t) + \sum_{h=1,h\neq k,j}^{m} x_h(t)\right) - \sum_{h=1,h\neq k,j}^{m} x_h(t), \tag{35}$$

where the inequality in Equation (35) holds because we have assumed positive and normalized payoffs.

Let us focus on the first multiplicative factor of Equation (35), we have:

$$\begin{aligned}
\left((A\ \mathbf{y})_j - \mathbf{x}^\top(t)\,A\ \mathbf{y}\right) &= \sum_{h=1,h\neq j}^{m} x_h(t)\,d_h \\
&\geq \xi \sum_{h=1,h\neq j}^{m} x_h(t) \\
&\geq \xi \max_{h\in\{1,\ldots,m\},h\neq j}\{x_h(t)\} \\
&= \xi \max_h\{x_h(t)\} \qquad (\text{since } x_k > x_j,\ \forall k \neq j) \\
&\geq \left(\frac{\xi}{m}\right) \qquad (\max_h\{x_h(t)\} \geq 1/m \quad \forall\,\mathbf{x} \in \Delta^m)
\end{aligned}$$

Substituting in Equation (35), we have:

$$-P \geq \frac{\xi}{m}\left(2x_j(t) + \sum_{h=1,h\neq k,j}^{m} x_h(t)\right) - \sum_{h=1,\neq k,j}^{m} x_h(t).$$

Using the fact that $\sum_{h=1,h\neq k,j}^{m} x_h(t) = 1 - x_k(t) - x_j(t)$ in the above equation, we obtain:

$$-P \geq x_j(t)\left(1 + \frac{\xi}{m}\right) - \left(1 - \frac{\xi}{m}\right) + x_k(t)\left(1 - \frac{\xi}{m}\right) \tag{36}$$

$$\geq x_k(t)\left(1 - \frac{\xi}{m}\right) \geq 0, \tag{37}$$

where the inequality in Equation (37) uses the definition of $\mathcal{S}_2$. Summarizing, we showed that the sign of $-P$ is positive, which implies that the sign of $\frac{d}{dt}\left(\frac{x_k}{x_j}\right)$ is negative, which concludes the proof. □

Finally, we show that the strategies in $\mathcal{S}_1$ share the desirable property that the component corresponding to the best response always has positive time-derivative *i.e.*, is monotonically increasing in time.

Lemma A.3. *It holds that $\dot{x}_j > 0$ for all $\mathbf{x} \in \mathcal{S}_1$.*

Proof. Since $x_j(t) = 1 - \sum_{k \neq j} x_k(t) = 1 - x_j(t) \sum_{k \neq j} \frac{x_k(t)}{x_j(t)}$, we can write $\dot{x}_j(t)$ as follows:

$$\begin{aligned}\dot{x}_j(t) &= \frac{d}{dt}\left[1 - x_j(t) \sum_{k \neq j} \frac{x_k(t)}{x_j(t)}\right] \\ &= -\dot{x}_j(t) \sum_{k \neq j} \frac{x_k(t)}{x_j(t)} - x_j(t) \sum_{k \neq j} \frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right),\end{aligned}$$

Solving for $\dot{x}_j(t)$ gives:

$$\dot{x}_j(t) = -\frac{x_j(t) \sum_{k \neq j} \frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right)}{1 + \sum_{k \neq j} \frac{x_k(t)}{x_j(t)}}$$

From the fact that the numerator of the r.h.s. of the above equation is always negative in $\mathcal{S}_1$ and the denominator is positive, we get the theorem statement. □

Lemmas A.1, A.2, and A.3 allow us to prove Theorem 4.4 which states that the convergence rate of SPGD to the unique best response $\bar{\mathbf{e}}_j$ is linear. We briefly sketch the proof of the theorem. We already know that SPGD converge to the best response by Theorem 4.2. Convergence implies that the dynamics eventually enter subspace $\mathcal{S}_2$ as $\mathbf{e}_j \in \mathcal{S}_2$. (In particular, $\mathcal{S}_2$ includes the portion of the simplex such that $x_j \geq 1/2$ as $m \geq$ and $0 \leq \xi \leq 1$.) By Lemma A.2, we know that $\mathcal{S}_2$ is contained by $\mathcal{S}_1$. Lemma A.3 leads to a direct extension of Lemma A.1 to $\mathcal{S}_2$: if the dynamics enter $\mathcal{S}_2$, they do not leave this subspace. The inequality in the definition of $\mathcal{S}_2$ coupled with calculations borrowed from the proof of Theorem 4.2 lead to the result.

Theorem 4.4. *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., $\xi := \bar{\mathbf{e}}_j^\top A\, \mathbf{y} - \max_{k \neq j} \left\{\mathbf{e}_k^\top A\, \mathbf{y}\right\}$.

Proof. From Theorem 4.2 and the continuity of the dynamics, we have that $t_0 := \inf_{t \geq 0}\{\mathbf{x}(t) \in \mathcal{S}_2\}$, exists and is finite. We want to show that for all $t \geq t_0$ it holds $\mathbf{x}(t) \in \mathcal{S}_2$, and, therefore, that $x_j(t) > \left(\frac{m-\xi}{m+\xi}\right)$ for all $t \geq t_0$.

Using Lemma A.2, we have that $\mathbf{x}(t_0) \in \mathcal{S}_1$ and by Lemma A.1 that for all $t \geq t_0$ $\mathbf{x}(t) \in \mathcal{S}_1$. Finally, by Lemma A.3, we have that $x_j(t)$ is increasing over t, and, consequently, $x_j(t) \geq \frac{m-\xi}{m+\xi}$ for all $t \geq t_0$. From the proof of Theorem 4.2 (Equation (21)), we have that:

$$\dot{V}(t) \leq -\eta\, x_j(t)^2\, V(t)^2 \tag{38}$$

$$\dot{V}(t) \leq -\eta \left(\frac{m-\xi}{m+\xi}\right)^2 V(t)^2, \tag{39}$$

where we used that $x_j(t) \geq \frac{m-\xi}{m+\xi}$ for all$t \geq t_0$. By Petrovitsch Theorem [36], we know that $V(t) \leq F(t)$, where $F(t)$ is the solution of the following differential equation:

$$\dot{F}(t) = -\eta \left(\frac{m-\xi}{m+\xi}\right)^2 F(t)^2. \tag{40}$$

Integrating Equation (40) from t_0 to t, we obtain that:

$$V(t) \leq \frac{1}{\eta \left(\frac{m-\xi}{m+\xi}\right)^2 t + C_0} \tag{41}$$

where $C_0 = \frac{1}{V(t_0)} - \eta \left(\frac{m-\xi}{m+\xi}\right)^2 t_0$. This concludes the proof. □

Theorem 4.5. *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:*

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

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 $\boldsymbol{b} \in \mathcal{B}$.

Proof. Let us focus on the first statement of the theorem. From Equation (24), it follows that, if $x_j(t) \geq x_k(t)$, then $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) < 0$ as $(A\,\mathbf{y})_k < (A\,\mathbf{y})_j$ and $\mathbf{x}(t)^\top A\ \mathbf{y} < (A\,\mathbf{y})_j$. However, such a conclusion cannot be drawn when $x_j(t) < x_k(t)$. We recall that, by defining $d_k := (A\,\mathbf{y})_j - (A\,\mathbf{y})_k$, and $D := \sum_{h=1}^m x_h(t)\, d_h$, Equation (24) can be rewritten as:

$$\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) = \frac{x_k(t)}{x_j(t)}\Big[(x_k(t) - x_j(t))\, D - x_k(t)\, d_k\Big]. \tag{42}$$

Assume, by contradiction, that $(x_k(t) - x_j(t))\, D - x_k(t)\, d_k > 0, \forall k \neq j$. By summing up all the inequalities for every k, we have:

$$0 < \sum_{k=1}^m (x_k(t) - x_j(t))\, D - \sum_{k=1}^m x_k(t)\, d_k = -x_j(t)\, m\, D < 0, \tag{43}$$

which contradicts our above assumption. This concludes the proof of the first statement of the theorem.

Let us focus on the second statement of the theorem and discuss separately the cases $m > 2$ and $m = 2$.

Case 1 $m > 2$**:** Consider the simplex subspace:

$$\mathcal{B} := \left\{ \mathbf{x} \in \Delta^m | x_j = 0, x_k > 0,\ \forall k \neq j \right\}.$$

We define $\mathbf{w} \in \Delta^m$ as a convex combination with parameter $\alpha \in (0,1)$ between $\boldsymbol{b} \in \mathcal{B}$ and $\bar{\mathbf{e}}_j$, formally, $\mathbf{w}(\alpha) = \alpha\, \boldsymbol{b} + (1-\alpha)\, \bar{\mathbf{e}}_j$. We claim that for every $\boldsymbol{b} \in \mathcal{B}$ there is a scalar $\mathfrak{B}(\boldsymbol{b}) \in (0,1)$ such that, for every $\alpha > \mathfrak{B}(\boldsymbol{b})$, it holds $\mathbf{w}(\alpha) \in \mathcal{E}$ and therefore, when $\mathbf{x}(t) = \mathbf{w}(\alpha)$, it holds $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) > 0$ for some $k \in \{1, \dots, m\}$. We observe that:

$$w_k(\alpha) = \begin{cases} \alpha\, b_k & if k \neq j \\ (1-\alpha) & if k = j \end{cases}. \tag{44}$$

where w_k and b_k are the k-th components of $\mathbf{w}$ and $\boldsymbol{b}$, respectively. By using Equation (44) in Equation (42) with $\mathbf{x}(t) = \mathbf{w}(\alpha)$, we obtain $\forall k \neq j$:

$$\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) > 0 \iff \frac{\alpha\, b_k}{(1-\alpha)}\left[\alpha\,(\alpha\, b_k - (1-\alpha)) \sum_{h=1}^m b_h\, d_h - \alpha\, b_k\, d_k\right] > 0. \tag{45}$$

We can thus solve this inequality for α for each $k \neq j$ and find that $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) > 0$ holds if and only if $\alpha > B_k(\boldsymbol{b})$, where:

$$B_k(\boldsymbol{b}) := \left(\frac{b_k\, d_k}{\sum_{h=1}^m b_h\, d_h} + 1\right)\frac{1}{b_k + 1}. \tag{46}$$

To conclude, we can define $\mathfrak{B}(\boldsymbol{b})$ as:

$$\mathfrak{B}(\boldsymbol{b}) := \min_{k \neq j}\left\{ B_k(\boldsymbol{b}) \right\}, \tag{47}$$

since we need that $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) > 0$ for at least one k. Finally, to prove that $\mathcal{E}$ is non-empty for $m > 2$, we show that $\mathfrak{B}(\boldsymbol{b}) < 1$. To prove that, we assume by contradiction that $\mathfrak{B}(\boldsymbol{b}) \geq 1$, this is equivalent to $B_k(\boldsymbol{b}) \geq 1$, $\forall k \neq j$, which can also be written as:

$$\frac{b_k\, d_k}{\sum_{h=1}^m b_h\, d_h} \geq b_k,\ \forall k \neq j. \tag{48}$$

Since by definition $1 \geq b_k \geq 0$, $\forall k \neq j$, we have that:

$$d_k \geq \sum_{h=1}^m b_h\, d_h > \sum_{h=1}^m d_h > d_k,\ \forall k \neq j, \tag{49}$$

which is a contradiction and, therefore, $B_k(\boldsymbol{b}) < 1$. Hence, $\mathcal{E}$ is non-empty and is defined as:

$$\mathcal{E} = \bigcup_{\boldsymbol{b}\in\mathcal{B}} \left\{ \mathbf{w} \in \Delta^m \mid \mathbf{w} = \alpha\, \boldsymbol{b} + (1-\alpha)\, \bar{\mathbf{e}}_j, 1 > \alpha > \mathfrak{B}(\boldsymbol{b}) \right\}. \tag{50}$$

Case 2 $m = 2$**:** We have that $\mathfrak{B}(\boldsymbol{b}) > \frac{2}{1+b_k} > 1$ for $k \neq j$ and, therefore, $\mathcal{E} = \emptyset$.

To show that the uniform initialization $\mathbf{x}(t) = \mathbf{1}/m \notin \mathcal{E}$, we plug $x_k(t) = 1/m\ \forall k$ in Equation (42), which provides:

$$\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) = -x_k(t) d_k,$$

which is negative for all k, as $x_k(t) > 0$, and $d_k > 0$. □

B PROOF OF THEOREM 4.6

Theorem 4.6. *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.*

Proof. Form the proof of Lemma A.2 we know that:

$$\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) = \eta\, \tau\, \frac{x_k(t)}{x_j(t)} \left[x_k(t) \left((A\ \mathbf{y})_k - \mathbf{x}(t)^\top A\ \mathbf{y} \right) - x_j(t) \left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y} \right) \right], \tag{51}$$

and we want to find $\mathbf{y}$ so that $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) > 0$ for some k. If $\mathbf{x}(t) \neq \frac{\mathbf{1}}{m}$ then there is one index in which $\mathbf{x}(t)$ attains lowest value and an index in which $\mathbf{x}(t)$ attains highest value. Formally, define $j = \arg\inf_i x_i(t)$ and k any index different from j, so that $x_j(t) < x_k(t)$. Clearly, $\frac{d}{dt}(x_k(t)/x_j(t)) > 0$ if and only if:

$$x_k(t) \left((A\ \mathbf{y})_k - \mathbf{x}(t)^\top A\ \mathbf{y} \right) > x_j(t) \left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y} \right). \tag{52}$$

Suppose that one can find $\mathbf{y}$ such that $\Omega := (A\,\mathbf{y})_j - (A\,\mathbf{y})_k > 0$ and $(A\,\mathbf{y})_j > (A\,\mathbf{y})_i\ \forall i \neq j$ (this assumption will be justified later in the proof). Define $f := \sum_{i\neq j,k} x_i(t)(A\,\mathbf{y})_i$, then the condition of Equation (52) can be rewritten as:

$$x_k(t) \left((A\ \mathbf{y})_k - x_k(t)(A\,\mathbf{y})_k - x_j(t)((A\,\mathbf{y})_k + \Omega) - f \right) > \tag{53}$$

$$x_j(t) \left((A\ \mathbf{y})_k + \Omega - x_k(t)(A\,\mathbf{y})_k - x_j(t)((A\,\mathbf{y})_k + \Omega) - f \right), \tag{54}$$

which, by algebraic manipulations, is in turn equivalent to:

$$\Omega x_k(t)(x_k(t) - x_j(t) - 1) > \tag{55}$$

$$(x_k(t) - x_j(t)) \left[(A\,\mathbf{y})_j (x_k(t) + x_j(t) - 1) + f \right]. \tag{56}$$

which is verified for any Ω such that:

$$\Omega < \frac{x_k(t) - x_j(t)}{x_k(t)(x_k(t) - x_j(t) - 1)} \left[(A\,\mathbf{y})_j (x_k(t) + x_j(t) - 1) + f \right].$$

Note that the denominator and the term in squared brackets are negative. The denominator is negative simply because $x_k(t), x_j(t)$ are two component of a probability vector and hence their sum is less than one. On the other hand, the term in squared brackets can be rewritten as follows:

$$(A\,\mathbf{y})_j (x_k(t) + x_j(t) - 1) + f = (A\,\mathbf{y})_j \left(- \sum_{i\neq j,k} x_i(t) \right) + \sum_{i\neq j,k} x_i(t)(A\,\mathbf{y})_i \tag{57}$$

$$= \sum_{i\neq j,k} x_i(t) \left((A\,\mathbf{y})_i - (A\,\mathbf{y})_j \right), \tag{58}$$

which is negative as by assumption $(A\,\mathbf{y})_i < (A\,\mathbf{y})_j$ for all $i \neq j$. To conclude we have that if $(A\,\mathbf{y})_i < (A\,\mathbf{y})_j$ for all $i \neq j$ then for all Ω positive and small enough one has that $\frac{d}{dt}\left(\frac{x_k(t)}{x_j(t)}\right) > 0$. The assumption of being able to find a strategy $\mathbf{y} \in \Delta^m$ such that $(A\,\mathbf{y})_i < (A\,\mathbf{y})_j$ for all $i \neq j$ (i.e., $\bar{\boldsymbol{e}}_j$ is the best response) is guaranteed by the fact that there is a unique fully mixed equilibrium, by arguments based on linear complementarity and the Lemke–Howson algorithm [48].

The problem of finding a $\mathbf{y}$ that induces the strategy $\mathbf{x}(t)$ to be a bad initialization is a linear problem. Indeed one could find such a $\mathbf{y}$ by solving the following problem for all k:

$$\max_{\mathbf{y}} \frac{x_k(t)}{x_j(t)} \left[x_k(t) \left((A\ \mathbf{y})_k - \mathbf{x}(t)^\top A\ \mathbf{y} \right) - x_j(t) \left((A\ \mathbf{y})_j - \mathbf{x}(t)^\top A\ \mathbf{y} \right) \right]$$
$$\text{s.t.} \sum_{i=1}^{m} y_i = 1$$
$$y_i \geq 0$$
$$(A\,\mathbf{y})_j > (A\,\mathbf{y})_i \; \forall\, i \neq j,$$

which is a linear problem in $\mathbf{y}$ and always feasible for the assumption of fully mixed equilibrium.

□

C OMITTED PROOFS FROM SECTION 5

In this section we provide all the results regarding the behaviour of SPGD in multi-agent environments.

C.1 Omitted Proofs from Section 5.1

In this section we will provide all the proofs for the behaviour of SPGD in the case of a single population.

Lemma 5.1. *SPGD satisfy properties NS, when restricting to* $\text{int}(\Delta^m)$*, and PC.*

Proof. Let us focus on the NS property. Initially, we show that a stationary strategy for SPGD is a Nash equilibrium of the population game. Let $\bar{\mathbf{x}} \in \text{int}(\Delta^m)$ be a stationary strategy of SPGD. For every k, we have $\Pi_{\text{SPGD}}(\bar{\mathbf{x}})_k = \bar{\mathbf{x}}^\top \Pi_{\text{SPGD}}(\bar{\mathbf{x}})$. Notice that, for every $\mathbf{x} \in \Delta^m$, SPGD satisfy the following property:

$$\sum_k \Pi_{\text{SPGD}}(\mathbf{x})_k = \sum_k (\Psi(\mathbf{x}) A\,\mathbf{x})_k = \sum_k x_k((A\,\mathbf{x})_k - \mathbf{x}^\top A\,\mathbf{x})$$
$$= \sum_k x_k (A\,\mathbf{x})_k - \mathbf{x}^\top A\,\mathbf{x} \sum_k x_k = 0,$$

for every $\mathbf{x} \in \Delta^m$. Therefore, it holds that:

$$m\,\bar{\mathbf{x}}^\top \Pi_{\text{SPGD}}(\bar{\mathbf{x}}) = \sum_k \bar{\mathbf{x}}^\top \Pi_{\text{SPGD}}(\bar{\mathbf{x}}) = \sum_k \Pi_{\text{SPGD}}(\bar{\mathbf{x}})_k = 0,$$

and, thus, $\bar{\mathbf{x}}^\top \Pi(\bar{\mathbf{x}}) = 0$. Recalling that $\bar{\mathbf{x}}$ is stationary for SPGD, we obtain $\Pi_{\text{SPGD}}(\bar{\mathbf{x}})_k = \bar{\mathbf{x}}^\top \Pi_{\text{SPGD}}(\bar{\mathbf{x}}) = 0$ for every k. Using the definition of Π_{SPGD} we have:

$$\Pi_{\text{SPGD}}(\mathbf{x})_k = x_k((A\ \mathbf{x})_k - \mathbf{x}^\top\ A\ \mathbf{x}) = 0,$$

for every $k \in \{1, \ldots, m\}$. Therefore, $(A\,\bar{\mathbf{x}})_k - \bar{\mathbf{x}}^\top A\,\bar{\mathbf{x}} = 0$ as $\bar{\mathbf{x}} \in \text{int}(\Delta^m)$, and, thus, $\bar{\mathbf{x}}$ is a Nash equilibrium.

Now, we show that a Nash equilibrium of the population game is stationary for SPGD. We recall that a NE $\bar{\mathbf{x}} \in \text{int}(\Delta^m)$ is such that $(A\bar{\mathbf{x}})_k = \bar{\mathbf{x}}^\top A\bar{\mathbf{x}}$ for every $k \in \{1, \ldots, m\}$. By definition of Π_{SPGD}, we have:

$$\Pi_{\text{SPGD}}(\bar{\mathbf{x}})_k = \bar{x}_k \underbrace{\left((A\,\bar{\mathbf{x}})_k - \bar{\mathbf{x}}^\top A\,\bar{\mathbf{x}}\right)}_{=0} = 0,$$

for every $k \in \{1, \ldots, m\}$. Therefore, $\Pi_{\text{SPGD}}(\bar{\mathbf{x}})_k = 0$ and thus $\bar{\mathbf{x}}$ is also a stationary strategy of SPGD.

Let us focus on the PC property. We need to show that $\dot{\mathbf{x}}^\top A\,\mathbf{x} > 0$ for all strategies $\mathbf{x}$ that are not stationary. In particular, we have:

$$\dot{\mathbf{x}}^\top A\,\mathbf{x} = \eta\tau \sum_i x_i \left(\Pi_{\text{SPGD}}(\mathbf{x})_i - \mathbf{x}^\top \Pi_{\text{SPGD}}(\mathbf{x}) \right) (A\,\mathbf{x})_i \tag{59}$$
$$= \eta\tau \sum_i x_i \Pi_{\text{SPGD}}(\mathbf{x})_i (A\,\mathbf{x})_i - \mathbf{x}^\top \Pi_{\text{SPGD}}(\mathbf{x}) \sum_i x_i (A\,\mathbf{x})_i \tag{60}$$
$$= \eta\tau \sum_i \Pi_{\text{SPGD}}(\mathbf{x}) x_i ((A\,\mathbf{x})_i - \mathbf{x}^\top A\,\mathbf{x}) \tag{61}$$
$$= \eta\tau ||\Pi_{\text{SPGD}}(\mathbf{x})||_2^2 > 0, \tag{62}$$

where the strict inequality in Equation (62) holds since we are in a non-stationary point $\mathbf{x}$. □

Lemma 5.3. *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* $\Pi_{\text{SPGD}}(\mathbf{x})$.

Proof. Since $\bar{\mathbf{x}} \in \text{int}(\Delta^m)$, Condition (i) of Definition 5.2 is trivially satisfied. Thus, it is sufficient to check Condition (ii). By the definition of derivative, we have:

$$\frac{\partial \Pi_{\text{SPGD}}(\mathbf{x})_k}{\partial x_j} = \delta_{k,j}(\mathbf{e}_k - \mathbf{x})^\top A\,\mathbf{x} + x_k\left(A_{k,j} - \mathbf{e}_j^\top\left(A + A^\top\right)\mathbf{x}\right).$$

Since $\bar{\mathbf{x}}$ is a RESS, it is also a NE, and, therefore, it holds $\mathbf{e}_i\, A\, \mathbf{x} = \mathbf{x}^\top\, A\, \mathbf{x}$. Thus, we have the following:

$$\left.\frac{\partial \Pi_{\text{SPGD}}(\mathbf{x})_k}{\partial x_j}\right|_{\mathbf{x}=\bar{\mathbf{x}}} = \bar{x}_k\left(A_{k,j} - \mathbf{e}_j^\top\left(A + A^\top\right)\bar{\mathbf{x}}\right) = X \odot P,$$

where $\odot$ is the element-wise (Hadamard) product, $X \in \mathbb{R}^{m\times m}$ is a matrix whose rows are all equal to $\bar{\mathbf{x}}$ (i.e. $X = \mathbf{1}\,\bar{\mathbf{x}}^\top$) and $P \in \mathbb{R}^{m\times m}$ is a matrix whose elements are defined by $[P]_{k,j} := (A_{k,j} - \mathbf{e}_j^\top(A + A^\top)\bar{\mathbf{x}})$. We observe that matrix P is negative definite on $\mathfrak{T}$ as the following holds:

$$\sum_{k,j} z_k\, z_j\, P_{k,j} = \sum_{k,j}\left[z_k\, z_j\, A_{k,j} - \mathbf{e}_j^\top\,(A + A^\top)\,\mathbf{x}\right] \tag{63}$$

$$< -\sum_{k,j} z_k\, z_j\, \mathbf{e}_j^\top\,(A + A^\top)\,\mathbf{x} \tag{64}$$

$$= -\sum_k z_k \sum_j \mathbf{e}_j^\top\,(A + A^\top)\,\mathbf{x} = 0. \tag{65}$$

By resorting to the proof of the Schur Product Theorem, we can also prove that $X \odot P$ is negative definite. Let us define $B = -P$ which is positive definite. In particular also $B^\top$ is positive definite and thus admits a square root $\sqrt{B^\top}$. Notice that $X^2 = \mathbf{1}\,\bar{\mathbf{x}}^\top\,\mathbf{1}\,\bar{\mathbf{x}}^\top = \mathbf{1}\,\bar{\mathbf{x}}^\top = X$, since $\bar{\mathbf{x}}^\top\mathbf{1} = 1$, and hence we can write $\sqrt{X} = X$, and, hence, $\sqrt{X}$ is well defined. Therefore, the matrix $X \odot B$ is definite positive since:

$$\mathbf{z}^\top\,(X \odot B)\,\mathbf{z} = tr(B^\top\ \text{diag}(\mathbf{z})\,X\ \text{diag}(\mathbf{z})) \tag{66}$$

$$= tr\left(\underbrace{\sqrt{B^\top}\ \text{diag}(\mathbf{z})\ \sqrt{X}}_{C^\top}\ \underbrace{\sqrt{X}\ \text{diag}(\mathbf{z})\ \sqrt{B^\top}}_{C}\right) \tag{67}$$

$$= tr(C^\top C) > 0. \tag{68}$$

This proves Condition (ii) and therefore it concludes the proof. □

Theorem 5.4. *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.*

Proof. Let $\bar{\mathbf{x}} \in \text{int}\,\Delta^m$ be a fully mixed ESS for a symmetric normal-form game. By [39, Section VII.4], we know that for a symmetric normal-form game an internal ESS if and only if $\mathbf{z}^\top A\mathbf{z} < 0$, $\forall \mathbf{z} \in \mathfrak{T}$. This condition is equivalent to condition (ii) of the Definition 5.2 of RESSs. Hence, every fully mixed ESS is also a RESS in symmetric normal form games. To conclude, we use Lemma 5.3 which states that every RESS of the symmetric normal-form game with matrix A is a RESS of the population game defined by fitness function $\Pi_{\text{SPGD}}(\mathbf{x})$. Hence, they are asymptotically stable for the SPGD. □

C.2 Proofs Omitted from Section 5.2

In this section we will provide all the proofs for the behaviour of SPGD in the case of two agents co-learning using the SPGD dynamics.

Lemma C.1. *Let $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ be a NE for $\mathcal{P}$. Then $\Pi^{(A)}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})_k = 0$ for all actions a_k in the support of $\bar{\mathbf{x}}$ and $\Pi^{B}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})_j = 0$ for all action a_j in the support of $\bar{\mathbf{y}}$.*

Proof. Let us start by considering $\Pi^{(A)}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})_k$, for an action a_k in the support of $\bar{\mathbf{x}}$. Since $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a NE, $\Pi^{(A)}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})_k = \bar{\mathbf{x}}^\top \Pi^{(A)}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$. Summing over actions a_k in the support of $\mathbf{x}$, we get:

$$\sum_{k:x_k>0}\left[\bar{\mathbf{x}}^\top \Pi^{(A)}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})\right] = \sum_{k:x_k>0} \Pi^{(A)}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})_k = 0,$$

which means that $\mathbf{x}^\top\,\Pi^{(A)}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = 0$, and, therefore, we have that $\Pi^{(A)}_{\text{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})_k = 0$ for very a_k in the support. The same arguments can be used to prove the statement of the lemma concerning Π^{B}_{SPGD}. □

Theorem 5.5. *In every normal-form game, it holds that:*

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

Proof. Let $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \mathrm{NE}(\mathcal{G}) \cap \{\mathrm{int}(\Delta^m) \times \mathrm{int}(\Delta^m)\}$. We prove the first inclusion by showing that $\Pi^A_{\mathrm{SPGD}}(\mathbf{x}, \bar{\mathbf{y}}) = \mathbf{0}^m$ for all $\mathbf{x} \in \Delta^m$ and $\Pi^B_{\mathrm{SPGD}}(\bar{\mathbf{x}}, \mathbf{y}) = \mathbf{0}^m$ for all $\mathbf{y} \in \Delta^m$. Observe that, by Lemma C.1, these two conditions together are equivalent to satisfying the NE conditions for $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ in $\mathcal{P}$. Indeed, since $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is fully mixed, it holds that $(A\,\bar{\mathbf{y}}) = c\,\mathbf{1}$ for some constant c. Plugging this into the definition of Π^A_{SPGD}, we get that:

$$\Pi^A_{\mathrm{SPGD}}(\mathbf{x}, \bar{\mathbf{y}})_k = x_k((A\,\bar{\mathbf{y}})_k - \mathbf{x}^\top A\,\bar{\mathbf{y}}) = x_k(c - c\ \mathbf{x}^\top\ \mathbf{1}) = 0, \tag{69}$$

for all k and $\mathbf{x} \in \Delta^m$. The same reasoning can be used to show that $\Pi^B_{\mathrm{SPGD}}(\bar{\mathbf{x}}, \mathbf{y}) = \mathbf{0}^m$ for all $\mathbf{y} \in \Delta^m$, thus showing the inclusion.

Let now $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \mathrm{NE}(\mathcal{P}) \cap \{\mathrm{int}(\Delta^m) \times \mathrm{int}(\Delta^m)\}$. By Lemma C.1, since $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are interior points, it holds that $\Pi^{(A)}_{\mathrm{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})_k = 0$ for all $0 \leq k \leq m$ and $\Pi^B_{\mathrm{SPGD}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})_j = 0$ for all for all $0 \leq j \leq m$. Thus the only way in which this necessary condition can be satisfied is that $(A\,\mathbf{y})_k - \mathbf{x}^\top A\,\mathbf{y} = 0$ and $(B\,\mathbf{x})_j - \mathbf{y}^\top B\,\mathbf{x} = 0$ for all $0 \leq k \leq m$ and $0 \leq j \leq m$. This in turns shows that $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a NE for $\mathcal{G}$, which concludes the proof. □

Lemma 5.6. *The flow of SPGD preserves a volume form in* $\mathrm{int}(\Delta^m \times \Delta^m)$.

Proof. As in the proof for RD made in [38, Proposition 6] we will show that a reparametrization of the RD w.r.t. time is divergence free in $\mathrm{int}\,\Delta^m \times \mathrm{int}\,\Delta^m$. In particular the reparametrization used in [38] is defined on $\mathrm{int}\,\Delta^m \times \mathrm{int}\,\Delta^m$ as follows:

$$\zeta^1(\mathbf{x}, \mathbf{y}) = \frac{1}{P(\mathbf{x}, \mathbf{y})}\dot{\mathbf{x}}$$
$$\zeta^2(\mathbf{x}, \mathbf{y}) = \frac{1}{P(\mathbf{x}, \mathbf{y})}\dot{\mathbf{y}}$$

where $P(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^m x_i \cdot \prod_{j=1}^m y_j$. Observe that the reparametrization is well-defined, since we are in $\mathrm{int}\,\Delta^m \times \mathrm{int}\,\Delta^m$. The divergence is computed on the simplex, *i.e.* substituting the usual i -th partial derivative, with the directional derivatives in the directions $\mathbf{x} - \bar{\mathbf{e}}_i$ or $\mathbf{y} - \bar{\mathbf{e}}_i$. In particular, if we call the directional derivative '∂^i' we have that, for each real function $g(\mathbf{z})$:

$$\partial^i g = \frac{\partial g}{\partial z_i} - \mathbf{z} \cdot \nabla(g(\mathbf{z})). \tag{70}$$

From now on, we will do all the calculations for ζ^1. We omit the ones for ζ^2, since they are exactly the same. In particular, the i-th directional derivative of ζ^1_i is the following:

$$\begin{aligned}\partial^i \zeta^1_i(\mathbf{x}, \mathbf{y}) &= \frac{1}{P(\mathbf{x}, \mathbf{y})}\left[\partial^i \dot{x}_i - \frac{\dot{x}_i}{P(\mathbf{x}, \mathbf{y})}\partial^i P(\mathbf{x}, \mathbf{y})\right] \\ &= \frac{1}{P(\mathbf{x}, \mathbf{y})}\left[\partial^i \dot{x}_i - \eta\tau(1 - m x_i)\left(\Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i - \mathbf{x}^\top \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})\right)\right].\end{aligned}$$

In the following, we will compute $\partial^i \dot{x}_i$. In particular:

$$\begin{aligned}\frac{\partial^i \dot{x}_i}{\eta\tau} &= \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i - \mathbf{x}^\top \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y}) + x_i\left(\frac{\partial \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i}{\partial x_i} - \frac{\partial\, \mathbf{x}^\top \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})}{\partial x_i}\right) \\ &\quad - x_i \sum_k x_k\left(\frac{\partial \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i}{\partial x_k} - \frac{\partial\, \mathbf{x}^\top \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})}{\partial x_k}\right) - x_i\left(\Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i - \mathbf{x}^\top \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})\right).\end{aligned}$$

Observe that the following equalities hold (by direct calculation):

$$\begin{aligned}\frac{\partial \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i}{\partial x_i} &= \left((A\,\mathbf{y})_i - \mathbf{x}^\top A\,\mathbf{y}\right) - x_i(A\,\mathbf{y})_i \\ \frac{\partial \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i}{\partial x_k} &= -x_i(A\,\mathbf{y})_k \\ \frac{\partial\, \mathbf{x}^\top \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})}{\partial x_k} &= 2\Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_k - (A\,\mathbf{y})_k \sum_j x_j^2.\end{aligned}$$

Thus we can rewrite $\partial^i \dot{x}_i / \eta\tau$ as:

$$\frac{\partial^i \dot{x}_i}{\eta\tau} = (1 - x_i)\left(\Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i - \mathbf{x}^\top \Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})\right) + (1 - 2x_i)\Pi^{(A)}_{\mathrm{SPGD}}(\mathbf{x}, \mathbf{y})_i + x_i(A\,\mathbf{y})_i\left(\sum_j x_j^2 - x_i\right)$$

$$- x_i \sum_k x_k \left[(A\,\mathbf{y})_k \left(\sum_j x_j^2 - x_i \right) - 2\Pi^{(A)}_{\text{SPGD}}(\mathbf{x},\mathbf{y})_k \right].$$

By summing up on all indexes i we obtain:

$$\begin{aligned}
\frac{P(\mathbf{x},\mathbf{y})}{\eta\tau} \sum_i \zeta^1_i(\mathbf{x},\mathbf{y}) &= \sum_i \left[\partial \dot{x}_i - (\Pi^{(A)}_{\text{SPGD}}(\mathbf{x},\mathbf{y})_i - \mathbf{x}^\top \Pi^{(A)}_{\text{SPGD}}(\mathbf{x},\mathbf{y}))(1 - m x_i) \right] \\
&= (m-1) \sum_i \dot{x}_i + \sum_i (1 - 2x_i) \Pi^{(A)}_{\text{SPGD}}(\mathbf{x},\mathbf{y})_i + 2 \sum_i x_i\, \mathbf{x}^\top \Pi^{(A)}_{\text{SPGD}}(\mathbf{x},\mathbf{y}) + \\
&+ \sum_i x_i (A\,\mathbf{y})_i \left(\sum_j x_j^2 - x_i \right) - \sum_i x_i \sum_k x_k (A\,\mathbf{y})_k \left(\sum_j x_j^2 - x_i \right) = 0.
\end{aligned}$$

where we have used the fact that:

$$\sum_i \Pi^{(A)}_i(\mathbf{x},\mathbf{y}) = \sum_i \dot{x}_i = 0.$$

The calculation for ζ^2 are exactly the same and thus we obtain that the vector field has null divergence. □

Theorem 5.7. *No closed set in* $\text{int}(\Delta^m \times \Delta^m)$ *is asymptotically stable for the SPGD.*

Proof. This theorem has been proved for the RD (and for more general kinds of dynamics) in [38, Proposition 6]. The key passage of their proof is exactly the equivalent of Lemma 5.6, and it is also the only passage in the proof in which the multi-linearity of the payoffs has been used. Thus the rest of the proof holds exactly the same, given Lemma 5.6. We point to [38] for the rest of the proof. □