# The Evolutionary Dynamics of Soft-Max Policy Gradient in Multi-Agent Settings

Extended Abstract

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

We investigate the mean dynamics of the *soft-max policy gradient* algorithm in multi-agent normal-form games by resorting to *evolutionary game theory* and dynamical system tools. First, we consider the best-response problem analysis, where a single learner plays against a fixed opponent in continuous time. For such dynamics, we provide a complete characterization of the set of *bad initializations* (points for which the dynamics initially move towards sub-optimal strategies). Then, we resort to models based on single- and multi-population games, showing that the dynamics preserve the volume and, in arbitrary instances, it is impossible 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 behavior of the dynamics in games with fully-mixed equilibrium is *chaotic*.

## KEYWORDS

Evolutionary Game Theory; Multiagent Reinforcement Learning

**ACM Reference Format:**
Martino Bernasconi, Federico Cacciamani, Simone Fioravanti, Nicola Gatti, and Francesco Trovò. 2022. The Evolutionary Dynamics of Soft-Max Policy Gradient in Multi-Agent Settings: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

## 1 LEARNING DYNAMICS AND POLICY GRADIENT

*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 [19] and Libratus [1]. In MARL, every agent learns 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 is available in the literature. These algorithms usually 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* [4] and *Deep-CFR* [2].

One of the mainstream approaches to study the learning dynamics of MARL algorithms, introduced by Börgers and Sarin [3], leverages the formalism and tools of *evolutionary game theory* [6, 11, 15, 16, 20] (EGT). This framework models the evolution of a population of agents as a dynamical system specified by a set of differential equations: the arguably most famous of such systems give rise to the so-called *Replicator Dynamics* (RD). In the same way, the continuous-time mean learning dynamics of an algorithm are modeled through a dynamical system. Such a system can, in turn, be studied in terms of specific properties, *i.e.*, the convergence rate, the set of stationary strategies, and asymptotic stability. These properties relate to different settings,*i.e.*, best-response problem, single-population games, and multi-population games, respectively. Interestingly, most MARL algorithms, such as *Q-learning* [7, 8, 12, 18], and a family of no-regret algorithms [10] have mean dynamics that are slight variants of the RD, sharing the same properties.

The algorithm we focus on in this paper belongs to the class of reinforcement learning techniques known as policy gradient methods [13, 17]. These methods search directly for the best policy in a constrained space where each policy is parameterized by a real-valued vector. Our paper focuses on the Soft-max Policy Gradient (SPG) algorithm, the most commonly adopted variant of policy gradient. In particular, we consider the evolutionary dynamics of SPG in the setting of normal-form games with linear reward functions. The SPG dynamics (shortly SPGD from here on) present a different structure not corresponding to any known evolutionary dynamics (see the work by Sandholm [16] for a detailed discussion of the main known dynamics), in contrast to the dynamics mentioned above of other classical RL algorithms. However, we draw an essential parallel between SPGD and RD: namely, SPGD correspond to RD on a game with *non-linear* payoffs. Even if this correspondence does not allow us to use already known results, it helps guide the analysis in many ways.

*Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.*

## 2 SINGLE-AGENT LEARNING ANALYSIS

This section considers the setting in which an agent maximizes her utility while the opponents' joint strategy is fixed over time.

Let $\mathbf{x}(t) \in \Delta^m$ be the mixed strategy of the learning agent, *i.e.*, a point in the simplex of probability distributions over a finite space of $m$ actions, and let the strategy of the opponent be $\mathbf{y}(t) = \mathbf{y}$. The expected reward of the first agent is $J(t) := \mathbf{x}(t)^\top A\,\mathbf{y}$, where $A$ is a real-valued matrix. We restrict ourselves to the non-degenerate case in which there is a unique, and thus pure, best response $\bar{\mathbf{e}}_j$. First, we show that SPGD converge to the best response, providing an upper bound to the convergence rate. Formally, we have:

Theorem 2.1. *Let $V(t)$ be defined as $V(t) := J^* - J(t)$, where $J^*$ is the value of the best response. If $\mathbf{x}(t)$ evolves according to SPGD, it holds (for a suitable constant $C_0 \in \mathbb{R}_+$) that:*

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

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

The idea behind the proof of Theorem 2.1 is to show that the dynamics leave in finite time a certain set of *bad* initialization points, such that the dynamics starting from them are initially more attracted towards a sub-optimal action rather than by the best response. We then show that, outside of this region, the dynamics monotonically converge to the best response. In the following theorem, we characterize precisely this set of points, which does not exist for RD and RD-like dynamics.

Theorem 2.2. *Let $\bar{\mathbf{e}}_j$ be the (unique) pure best response. If $\mathbf{x}(t)$ evolves according to SPGD, at every point in time, there is at least one $k \neq j$ such that the ratio between $x_j(t)$ and $x_k(t)$ increases over time. Moreover, if $m > 2$, there exists a non-empty subspace of the simplex of mixed strategies $\mathcal{E}$ such that, if $\mathbf{x}(t) \in \mathcal{E}$, the dynamics is attracted towards a sub-optimal action. Moreover the center of the simplex $\Delta^m$ is always outside $\mathcal{E}$. The set $\mathcal{E}$ is defined as:*

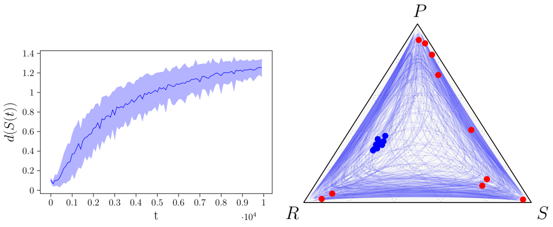
$$\mathcal{E} := \bigcup_{\boldsymbol{b} \in \mathcal{B}} \left\{ \mathbf{w} \in \Delta^m \mid \mathbf{w} = \alpha\,\boldsymbol{b} + (1-\alpha)\,\bar{\mathbf{e}}_j, 1 > \alpha > \mathfrak{B}(\boldsymbol{b}) \right\},$$

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

The theorem shows that, while SPG has sound theoretical guarantees when learning the best response, the non-monotonic behavior of the dynamics starting from $\mathcal{E}$ can make the convergence slow in practice. An adversarial opponent could exploit this fact in specific settings. On the other hand, the theorem shows that the uniform initialization is always outside the bad initialization region.

## 3 MULTI-AGENT LEARNING ANALYSIS

This section studies the case in which all the agents simultaneously learn. To the best of our knowledge, this is the first work to theoretically study the behavior of the SPG algorithm in multi-agent environments. For this study, we adopt an evolutionary game-theoretic perspective: the strategy of one learning agent is treated as a *population* of individuals evolving according to the equations of the dynamics. As commonly done in evolutionary game theory, we investigate separately the *single-population*, a.k.a. self-play, case and the *multiple-population* one, in which different independently learning agents co-evolve. In the single-population setting, we prove that, in terms of asymptotic convergence in the interior of the strategy space, the two dynamics present similar, desirable properties. In particular, we show that the spaces of asymptotically stable states in the interior of the simplex of RD and SPGD coincide. We tackle the multiple-population case in a two-population case. We cast the SPGD as RD on a game with non-linear payoffs and, in particular, a *differentiable* game. Let us denote with $\mathcal{G}$ the original game and $\mathcal{P}$ the derived differentiable one. We show:



(a) (b)

**Figure 1: (a) Diameter evolution. (b) SPGD trajectories, starting points $\mathbf{x}(t_0) \in S(t_0)$ in blue, end points $\mathbf{x}(T) \in S(T)$ in red.**

Theorem 3.1. *Let $NE(\mathcal{G})$ be the set of Nash equilibria of $\mathcal{G}$ and $NE(\mathcal{P})$ the set of Nash equilibria of $\mathcal{P}$. Then it holds:*

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

The theorem states that the equilibria in the interior of the simplex coincide. However, at the simplex borders, we may experience an increased instability of the dynamics near pure strategies. Finally, we give a negative result regarding the convergence of SPGD:

Theorem 3.2. *No closed set in* $\mathrm{int}(\Delta^m \times \Delta^m)$ *is asymptotically stable for the SPGD.*

Similar to what is proposed for RD [14, Proposition 6], the proof exploits the fact that SPGD preserve volume in a reparametrized space. Therefore, while convergence to an interior equilibrium is impossible, the dynamics are Poincaré recurrent or convergent to the border of the actions' simplex. Finally, we provide empirical evidence (Fig. 1) that the dynamics are chaotic. More precisely, we show that the diameter $d(S(t_0))$ of a set of close initial points $S(t_0)$ increases over time (Fig. 1a), and a small deviation from a starting point results in a large deviation in the ending point (Fig. 1b).

## 4 CONCLUSIONS

Our analysis paves the way to further evolutionary game theory studies of policy-gradient-based algorithms (including, *e.g.*, NeuRD [5]) when the policy space is restricted to assess the impact of special policy space structures on the evolutionary dynamics. A natural future direction is to analyze the dynamics of SPGD in multi-agent 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 [9]. Finally, it is an interesting line of research the study of the behavior of RD on general non-convex payoff functions, which may lead to even deeper insights on the behavior of SPGD by exploiting the same connection we drew between these two dynamics.

# REFERENCES

[1] 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.

[2] 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.

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

[4] 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

[5] 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.

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

[7] 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.

[8] ——, "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.

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

[10] 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.

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

[12] 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.

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

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

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

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

[17] 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.

[18] 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.

[19] 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.

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