Exploration-Exploitation in Multi-Agent Learning: Catastrophe Theory Meets Game Theory

Stefanos Leonardos, Georgios Piliouras
Singapore University of Technology and Design,
8 Somapah Road, 487372 Singapore, {stefanos_leonardos ; georgios}@sutd.edu.sg

Abstract

Exploration-exploitation is a powerful and practical tool in multi-agent learning (MAL), however, its effects are far from understood. To make progress in this direction, we study a smooth analogue of Q-learning. We start by showing that our learning model has strong theoretical justification as an optimal model for studying exploration-exploitation. Specifically, we prove that smooth Q-learning has bounded regret in arbitrary games for a cost model that explicitly captures the balance between game and exploration costs and that it always converges to the set of quantal-response equilibria (QRE), the standard solution concept for games under bounded rationality, in weighted potential games with heterogeneous learning agents. In our main task, we then turn to measure the effect of exploration in collective system performance. We characterize the geometry of the QRE surface in low-dimensional MAL systems and link our findings with catastrophe (bifurcation) theory. In particular, as the exploration hyperparameter evolves over-time, the system undergoes phase transitions where the number and stability of equilibria can change radically given an infinitesimal change to the exploration parameter. Based on this, we provide a formal theoretical treatment of how tuning the exploration parameter can provably lead to equilibrium selection with both positive as well as negative (and potentially unbounded) effects to system performance.

Introduction

The problem of optimally balancing exploration and exploitation in *multi-agent systems (MAS)* has been a fundamental motivating driver of online learning, optimization theory and evolutionary game theory (Claus and Boutilier 1998; Panait and Luke 2005). From a behavioral perspective, it involves the design of realistic models to capture complex human behavior, such as the standard Experienced Weighted Attraction model (Ho and Camerer 1999; Ho, Camerer, and Chong 2007). Learning agents use time varying parameters to explore suboptimal, boundedly rational decisions, while at the same time, they try to coordinate with other interacting agent and maximize their profits (Ho and Camerer 1998; Bowling and Veloso 2002; Kaisers et al. 2009).

From an AI perspective, the exploration-exploitation dilemma is related to the optimization of adaptive systems. For example, neural networks are trained to parameterize policies ranging from very exploratory to purely exploitative, whereas meta-controllers decide which policy to prioritize (Puigdomènech Badia et al. 2020). Similar techniques have been applied to rank agents in tournaments according to performance for preferential evolvability (Lanctot et al. 2017; Omidshafiei et al. 2019; Rowland et al. 2019) and to design *multi-agent learning (MAL)* algorithms that prevent collective learning from getting trapped in local optima (Kaisers and Tuyls 2010, 2011).

Despite these notable advances both on the behavioral modelling and AI fronts, the theoretical foundations of learning in MAS are still largely incomplete even in simple settings (Wunder, Littman, and Babes 2010; Bloembergen et al. 2015). While there is still no theory to formally explain the performance of MAL algorithms, and in particular, *the effects of exploration in MAS* (Klos, Van Ahee, and Tuyls 2010), existing research suggests that many pathologies of exploration already emerge at stateless matrix games at which naturally emerging collective learning dynamics exhibit a diverse set of outcomes (Sato and Crutchfield 2003; Sato, Akiyama, and Crutchfield 2005; Tuyls and Weiss 2012).

The reasons for the lack of a formal theory are manifold. First, even without exploration, MAL in games can result in complex behavior that is hard to analyze (Balduzzi et al. 2020; Mertikopoulos, Papadimitriou, and Piliouras 2018; Mazumdar, Ratliff, and Shankar Sastry 2018). Once explicit exploration is enforced, the behavior of online learning becomes even more intractable as Nash Equilibria (NE) are no longer fixed points of agents' behavior. Finally, if parameters are changed enough, then we get bifurcations and possibly chaos (Wolpert et al. 2012; Palaiopanos, Panageas, and Piliouras 2017; Sanders, Farmer, and Galla 2018).

Our approach & results. Motivated by the above, we study a smooth variant of stateless Q-learning, with softmax or Boltzmann exploration (one of the most fundamental models of exploration-exploitation in MAS), termed Boltzmann Q-learning or *smooth Q-learning (SQL)*, which has recently received a lot of attention due to its connection to evolutionary game theory (Tuyls, Verbeeck, and Lenaerts 2003;

Kianercy and Galstyan 2012). Informally (see the Preliminaries Section for the rigorous definition), each agent k updates her choice distribution $x = (x_i)$ according to the rule

$$\dot{x}_i/x_i = \beta_k \left(u_i - \bar{u}\right) - \alpha_k \left(\ln x_i - \textstyle\sum_j x_j \ln x_j\right)$$

where $u_i, \bar{u}$ denote agent k's utility from action i and average utility, respectively, given all other agents' actions and α_k/β_k is agent k's exploration rate.[1] Agents tune the exploration parameter to increase/decrease exploration during the learning process. We analyze the performance of SQL dynamics along the following axes.

Regret and Equilibration. First, we benchmark their performance against the optimal choice distribution in a cost model that internalizes agents' utilities from exploring the space (Lemma 1), and show that in this context, the SQL dynamics enjoy a *constant* total regret bound in arbitrary games that depends logarithmically in the number of actions (Theorem 2). Second, we show that they converge to *Quantal Response Equilibria* (QRE)[2] in weighted potential games with heterogeneous agents of arbitrary size (Theorem 3).[3] The underpinning intuition is that agents' deviations from pure exploitation are not a result of their bounded rationality but rather a perfectly rational action in the quest for more information about unexplored choices which creates value on its own. This is explicitly captured by a correspondingly modified Lyapunov function (potential) which combines the original potential with the entropy of each agent's choice distribution (Lemma 4).

While previously not formally known, these properties mirror results of strong regret guarantees for online algorithms (see e.g., Cesa-Bianchi and Lugosi (2006); Kwoon and Mertikopoulos (2017); Mertikopoulos, Papadimitriou, and Piliouras (2018); convergence results for SQL in more restricted settings (Leslie and Collins (2005); Coucheney, Gaujal, and Mertikopoulos (2015)).[4] However, whereas in previous works such results corresponded to main theorems, in our case they are only our starting point as they clearly not suffice to explain the disparity between the regularity of the SQL dynamics in theory and their unpredictable performance in practice.

We are faced with two major unresolved complications. First, the outcome of the SQL algorithm in MAS is highly sensitive on the exploration parameters (Tuyls, Hoen, and Vanschoenwinkel 2006). The set of QRE ranges from the NE of the underlying game when there is no exploration to uniform randomization when exploration is constantly high (agents never learn). Second, the collective system evolution is *path dependent*, i.e., in the case of time-evolving parameters, the equilibrium selection process cannot be understood by only examining the final exploration parameter but rather depends on its whole history of play (Göcke 2002; Romero 2015; Yang, Piliouras, and Basanta 2017).

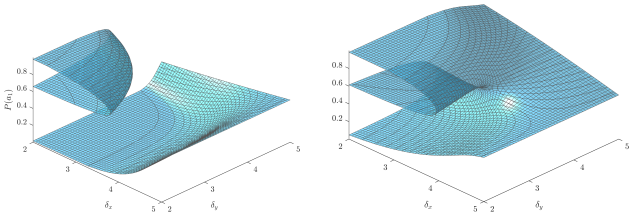

Figure 1: Single *saddle-node bifurcation* curve (left) vs two branches of saddle-node bifurcation curves meeting which is consistent with the emergence of a *co-dimension* 2 *cusp bifurcation* (right) on the QRE manifold of two player, two action games as function of the exploration rates, δ_x, δ_y (see also Figures 3,4). The possible learning paths before, during and after exploration depend on the geometry of the QRE surface (Theorems 5,6).

Catastrophe theory and equilibrium selection. We explain the fundamentally different outcomes of exploration-exploitation with SQL in games via catastrophe theory. The link between these two distinct fields lies on the properties of the underlying game which, in turn, shape the geometry of the QRE surface. As agents' exploration parameters change, the QRE surface also changes. This prompts dramatic phase transitions in the exploration path that ultimately dictate the outcome of the learning process. Catastrophe theory reveals that such transitions tend to occur as part of well-defined qualitative geometric structures.

In particular, the SQL dynamics may induce a specific type of catastrophes, known as *saddle-node* bifurcations (Strogatz 2000). At such bifurcation points, small changes in the exploration parameters change the stability of QRE and cause QRE to merge and/or disappear. However, as we prove, this is not always sufficient to stabilize desired states; the decisive feature is whether the QRE surface is connected or not (see Theorem 6 and Figure 2) which in turn, determines the possible types of bifurcations, i.e., whether there are one or two branches of saddle-node bifurcation curves, that may occur as exploration parameters change (Figure 1).

In terms of performance, this is formalized in Theorem 5 which states that even in the simplest of MAS, *exploration can lead under different circumstances both to unbounded gain as well as unbounded loss*. While existential in nature, Theorem 5 does not merely say that anything goes when exploration is performed. When coupled with the characterization of the geometric locus of QRE in Theorem 6, it suggests that we can identify cases where exploration can be provably beneficial or damaging. This provides a formal geometric argument why exploration is both extremely powerful but also intrinsically unpredictable.

[1] This variant of Q-learning has been also extensively studied in the economics and reinforcement learning literature under various names, see e.g., (Alós-Ferrer and Netzer 2010; Sanders, Farmer, and Galla 2018) and (Kaelbling, Littman, and Moore 1996; Mertikopoulos and Sandholm 2016), respectively.

[2] The prototypical extension of NE for games with bounded rationality (McKelvey and Palfrey 1995).

[3] Apart from their standard applications, see (Panageas and Piliouras 2016; Swenson, Murray, and Kar 2018; Perolat et al. 2020) and references therein, weighted potential games naturally emerge in distributed settings such as recommendation systems (Ben-Porat and Tennenholtz 2018).

[4] They are also of independent interest in the limited literature on the properties of the softmax function (Gao and Pavel 2017).

The above findings are visualized in systematic experiments in both low and large dimensional games along two representative exploration-exploitation policies, *explore-then-exploit and cyclical learning rates* (Experiments Section). We also visualize the modified potential (and how it changes during exploration) in weighted potential games of arbitrary size by properly adapting the technique of Li et al. (2018) for visualizing high dimensional loss functions in deep learning (Figure 5). Omitted materials, all proofs, and more experiments are included in Appendices A-D.

Preliminaries: Game Theory and SQL

We consider a finite set $\mathcal{N}$ of interacting agents indexed by $k = 1, 2, \ldots, N$. Each agent $k \in \mathcal{N}$ can take an action from a finite set $A_k = \{1, 2, \ldots, n_k\}$. Accordingly, let $A := \prod_{k=1}^{N} A_k$ denote the set of joint actions or pure action profiles, with generic element $a = (a_1, a_2, \ldots, a_N)$. To track the evolution of the agents' choices, let $X_k = \{x_k \in \mathbb{R}^{n_k} : \sum_{i=1}^{n_k} x_{ki} = 1, x_{ki} \geq 0\}$ denote the set of all possible choice distributions $x_k := (x_{ki})_{i \in A_k}$ of agent $k \in \mathcal{N}$.[5] We consider the dynamics in the collective state space $X := \prod_{k=1}^{N} X_k$, i.e., the space of all joint choice distributions $x = (x_k)_{k \in \mathcal{N}}$. Using conventional notation, we will write $(a_k; a_{-k})$ to denote the pure action profile at which agent $k \in \mathcal{N}$ chooses action $a_k \in A_k$ and all other agents in $\mathcal{N}$ choose actions $a_{-k} \in A_{-k} := \prod_{l \neq k} A_l$. Similarly, for choice distribution profiles, we will write (x_k, x_{-k}) with $x_{-k} \in X_{-k} := \prod_{l \neq k} X_l$. When time is relevant, we will use the index t for agent k's choice distribution $x_k(t) := (x_{ki}(t))_{i \in A_k}$ at time $t \geq 0$.

When selecting an action $i \in A_k$, agent $k \in \mathcal{N}$ receives a reward $u_k(i; a_{-k})$ which depends on the choices $a_{-k} \in A_{-k}$ of all other agents. Accordingly, the expected reward of agent $k \in \mathcal{N}$ for a choice distribution profile $x = (x_k, x_{-k}) \in X$ is equal to $u_k(x) = \sum_{a \in A} \left(x_{ki} u_k(i; a_{-k}) \prod_{l \neq k} x_{la_l} \right)$. We will also write $r_{ki}(x) := u_k(i; x_{-k})$ or equivalently $r_{ki}(x_{-k})$ for the reward of pure action $i \in A_k$ at the joint choice distribution profile $x = (x_k; x_{-k}) \in X$ and $r_k(x) := (r_{ki}(x))_{i \in A_k}$ for the resulting reward vector of all pure actions of agent k. Using this notation, we have that $u_k(x) = \langle x_k, r_k(x) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the usual inner product in $\mathbb{R}^{n_k}$, i.e., $\langle x_k, r_k(x) \rangle = \sum_{j \in A_k} x_{kj} r_{kj}(x)$. In particular, $\partial u_k(x) / \partial x_{ki} = r_{ki}(x)$. To sum up, the above setting can be represented in compact form with the notation $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k \in \mathcal{N}} \right)$.

We assume that the updates in the choice distribution x_k of agent $k \in N$ are governed by the dynamics

$$\dot{x}_{ki}/x_{ki} = \beta_k [r_{ki}(x) - \textstyle\sum_{j \in A_k} x_{kj} r_{kj}(x)] - \\ - \alpha_k [\ln x_{ki} - \textstyle\sum_{j \in A_k} x_{kj} \ln x_{kj}] \quad (1a)$$

$$= \beta_k [r_{ki}(x) - \langle x_k, r_k(x) \rangle] \\ - \alpha_k [\ln x_{ki} - \langle x_k, \ln x_k \rangle] \quad (1b)$$

where $\beta_k \in [0, +\infty)$ and $\alpha_k \in [0, 1)$ are positive constants that control the rate of choice adaptation and memory loss, respectively of the learning agent $k \in \mathcal{N}$ and $\ln x_k := (\ln x_{ki})_{i \in A_k}$ for $x_k \in X_k$. The first term, $r_{ki}(x) - \sum_{j \in A_k}^{n} x_{kj} r_{kj}(x)$, corresponds to the vector field of the replicator dynamics and captures the adaptation of the agents' choices towards the best performing strategy (exploitation). The second term, $\ln x_{ki} - \sum_{j \in A_k} x_{kj} \ln x_{kj}$, corresponds to the memory of the agent and the exploration of alternative choices. Due to their mathematical connection with Q-learning, we will refer to the dynamics in (1) as *smooth Q-learning* (SQL) dynamics.[6] The interior fixed points $x^Q \in X$ of the dynamics in equations (1) are the *Quantal Response Equilibria* (QRE) of Γ. In particular, each $x_k^Q \in X_k$ for $k = 1, 2, \ldots, N$ satisfies

$$x_{ki}^Q = \exp(r_{ki}(x_{-k}^Q)/\delta_k) / \textstyle\sum_{i \in A_k} \exp(r_{kj}(x_{-k}^Q)/\delta_k), \quad (2)$$

for $i \in A_k$, where $\delta_k := \alpha_k / \beta_k$ denotes the *exploration rate* for each agent $k \in \mathcal{N}$.

Bounded Regret in All Games and Convergence in Weighted Potential Games

Our first observation is that the SQL dynamics in (1) can be considered as replicator dynamics in a modified game with the same sets of agents and possible actions for each agent but with modified utilities.

Lemma 1. *Given* $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k \in \mathcal{N}} \right)$, *consider the modified utilities* $\left(u_k^H \right)_{k \in \mathcal{N}}$ *defined by* $u_k^H(x) := \beta_k \langle x_k, r_k(x) \rangle - \alpha_k \langle x_k, \ln x_k \rangle$, *for* $x \in X$. *Then, the dynamics described by the differential equation* $\dot{x}_{ik}/x_{ik}$ *in* (1) *can be written as*

$$\dot{x}_{ki}/x_{ki} = r_{ki}^H(x) - \langle x_k, r_k^H(x) \rangle \quad (3)$$

where $r_{ki}^H(x) := \frac{\partial}{\partial x_{ki}} u_k^H(x) = \beta_k r_{ki}(x) - \alpha_k (\ln x_{ki} + 1)$. *In particular, the dynamics in* (1) *describe the replicator dynamics in the modified setting* $\Gamma^H = \left(\mathcal{N}, \left(A_k, u_k^H \right)_{k \in \mathcal{N}} \right)$.

The superscript H refers to the regularizing term, $H(x_k) := -\langle x_k, \ln x_k \rangle = -\sum_{j \in A_k} x_{kj} \ln x_{kj}$ which denotes the *Shannon entropy* of choice distribution $x_k \in X_k$.

Bounded regret. To measure the performance of the SQL dynamics in (1), we will use the standard notion of (accumulated) *regret* (Mertikopoulos, Papadimitriou, and Piliouras 2018). The regret $R_k(T)$ at time $T > 0$ for agent k is

$$R_k(T) := \max_{x_k' \in X_k} \int_0^T [u_k(x_k'; x_{-k}(t)) - \\ - u_k(x_k(t), x_{-k}(t))] \, dt, \quad (4)$$

[5] Depending on the context, we will use either the indices $i, j \in A_k$ or the symbol $a_k \in A_k$ to denote the pure actions of agent k.

[6] An explicit derivation due to (Tuyls, Verbeeck, and Lenaerts 2003; Sato, Akiyama, and Crutchfield 2005; Wolpert et al. 2012; Kianercy and Galstyan 2012) (among others) of the connection between Q-learning (Watkins and Dayan 1992) and the above dynamics (including their resting points) is given in Appendix A.

i.e., $R_k(T)$ is the difference in agent k's rewards between the sequence of play $x_k(t)$ generated by the SQL dynamics and the best possible choice up to time T in hindsight. Agent k has *bounded regret* if for every initial condition $x_k(t)$ the generated sequence $x_k(t)$ satisfies $\limsup R_k(T) \leq 0$ as $T \to \infty$. Our main result in this respect is a constant upper bound on the regret of the SQL dynamics.

Theorem 2. *Consider the modified setting* $\Gamma^H = \left(\mathcal{N}, \left(A_k, u_k^H\right)_{k\in\mathcal{N}}\right)$. *Then, every agent* $k \in \mathcal{N}$ *who updates their choice distribution* $x_k \in X_k$ *according to the dynamics in equation* (3) *has bounded regret, i.e., there exists a constant* $C > 0$ *such that* $\limsup_{T\to\infty} R_k^H(T) \leq C$.

From the proof of Theorem 2, it follows that the constant C is logarithmic in the number of actions given a uniformly random initial condition as is the standard. This yields an optimal bound which is powerful in general MAL settings. In particular, regret minimization by the SQL dynamics at an optimal $O(1/T)$ rate implies that their time-average converges fast to *coarse correlated equilibria (CCE)*. These are CCE of the perturbed game, Γ^H, but if exploration parameter is low, they are approximate CCE of the original game as well. Even ϵ-CCE are $(\frac{\lambda(1+\epsilon)}{1-\mu(1+\epsilon)})$-optimal for $\lambda - \mu$ smooth games, see e.g., (Roughgarden 2015). However, for games that are not smooth (e.g., games with NE that have widely different performance and hence, a large Price of Anarchy), we need more specialized tools (Section on Performance).

Convergence to QRE in weighted potential games with heterogeneous agents. If $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k\in\mathcal{N}}\right)$ describes a potential game, then more can be said about the limiting behavior of the SQL dynamics. Formally, Γ is called a *weighted potential game* if there exists a function $\phi : A \to \mathbb{R}$ and a vector of positive weights $w = (w_k)_{k\in\mathcal{N}}$ such that for each player $k \in \mathcal{N}$, $u_k(i, a_{-k}) - u_k(j, a_{-k}) = w_k(\phi(i, a_{-k}) - \phi(j, a_{-k}))$, for all $i \neq j \in A_k$, and $a_{-k} \in A_{-k}$. If $w_k = 1$ for all $k \in \mathcal{N}$, then Γ is called an *exact potential game*. Let $\Phi : X \to \mathbb{R}$ denote the multilinear extension of ϕ defined by $\Phi(x) = \sum_{a\in A} \phi(a) \prod_{k\in N} x_{ka_k}$, for $x \in X$. We will refer to Φ as the *potential function* of Γ. Using this notation, we have the following.

Theorem 3. *If* $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k\in\mathcal{N}}\right)$ *admits a potential function,* $\Phi : X \to \mathbb{R}$*, then the sequence of play generated by the SQL dynamics in* (1) *converges to a compact connected set of QRE of* Γ.

Intuitively, the first term, $\beta_k(r_{ki}(x) - \langle x_k, r_k(x)\rangle)$, in equation (1) corresponds to agent k's replicator dynamics in the underlying game (with utilities rescaled by β_k that can also absorb agent k's weight) and thus, it is governed by the potential function. The second term, $-\alpha_k(\ln x_{ki} - \langle x_k, \ln x_k\rangle)$, is an idiosyncratic term which is independent from the environment, i.e., the other agents' choice distributions. Hence, the potential game structure is preserved — up to a multiplicative constant for each player which represents that players' exploration rate δ_k — and Theorem 3 can be established by extending the techniques of (Kleinberg, Piliouras, and Tardos 2009; Coucheney, Gaujal, and Mertikopoulos 2015) to the case of weighted potential games. This is the statement of Lemma 4 (which is also useful for the numerical experiments).

Lemma 4. *Let* $\Phi : X \to \mathbb{R}$ *denote a potential function for* $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k\in\mathcal{N}}\right)$*, and consider the modified utilities* $u_k^H(x) := \beta_k\langle x_k, r_k(x)\rangle - \alpha_k\langle x_k, \ln x_k\rangle$*, for* $x \in X$*. Then, the function* $\Phi^H(x)$ *defined by*

$$\Phi^H(x) := \Phi(x) + \textstyle\sum_{k\in\mathcal{N}} \delta_k H(x_k), \quad \text{for } x \in X, \quad (5)$$

is a potential function for the modified game $\Gamma^H = \left(\mathcal{N}, \left(A_k, u_k^H\right)_{k\in\mathcal{N}}\right)$. *The time derivative* $\dot{\Phi}^H(x)$ *of the potential function is positive along any sequence of choice distributions generated by the dynamics of equation* (3) *except for fixed points at which it is* 0.

From Topology to Performance

While the above establish some desirable topological properties of the SQL dynamics, the effects of exploration are still unclear in practice both in terms of equilibrium selection and agents' individual performance (utility). As we formalize in Theorem 5 and visualize in the Experiments Section, exploration – exploitation may lead to (unbounded) improvement, but also to (unbounded) performance loss even in simple settings.

To compare agents' utility for different exploration-exploitation policies, it will be convenient to denote the sequence of utilities of agent $k \in \mathcal{N}$ by $u_k^{\text{exploit}}(t), t \geq 0$ if there exist thresholds $\delta_k > 0$ (that may depend on the initial condition $x_k(0)$ of agent k) such that $\delta_k(t) < \delta_k$ for all $k \in \mathcal{N}$, i.e., if exploration remains *low* for all agents, and by $u_k^{\text{explore}}(t), t \geq 0$ otherwise. Then we have the following.

Theorem 5 (Catastrophes in Exploration-Exploitation). *For any number* $M > 0$*, there exist potential games* $\Gamma_u^M = \{\mathcal{N}, (X_k, u_k)_{k\in\mathcal{N}}\}$ *and* $\Gamma_v^M = \{\mathcal{N}, (X_k, v_k)_{k\in\mathcal{N}}\}$*, positive-measure sets of initial conditions* $I_u, I_v \subset X$*, and exploration rates* $\delta_k > 0$*, so that*

$$\lim_{t\to\infty} \left(u_k^{\text{exploit}}(t) / u_k^{\text{explore}}(t)\right) \geq M, \text{ and}$$
$$\lim_{t\to\infty} \left(v_k^{\text{exploit}}(t) / v_k^{\text{explore}}(t)\right) \leq 1/M$$

for all $k \in \mathcal{N}$*, whenever* $\limsup_{t\to\infty} \delta_k(t) = 0$ *for all* $k \in \mathcal{N}$*, i.e., whenever, after some point, exploration stops for all agents. In particular, for all agents* $k \in \mathcal{N}$*, the individual — and hence, also the aggregate — performance loss (gain) in terms of utility due to exploration can be unbounded, even if exploration is only performed by a single agent.*

The proof of Theorem 5 is constructive and relies on Theorem 6 discussed next. Theorem 6 characterizes the geometry of the QRE surface (connected or disconnected) which determines the bifurcation type that takes place during exploration. In turn, this dictates the possible outcomes — and hence, the individual and collective performance — after the exploration process, as formalized by Theorem 5.

Classification of 2×2 coordination games and geometry of the QRE surface. First, we introduce some minimal additional notation and terminology regarding *coordination games*.[7] Two player, $\mathcal{N} = \{1, 2\}$, two action, $A_k = (a_1, a_2), k = 1, 2$, *coordination* games are games in which the payoffs satisfy $u_{11} > u_{21}, u_{22} > u_{12}$ and $v_{11} > v_{21}, v_{22} > v_{12}$ where u_{ij} (v_{ij}) denotes the payoff of agent 1 (2) when that agent selects action i and the other agent action j. Such games admit three NE, two pure on the diagonal and one fully mixed $(x_{\text{mix}}, y_{\text{mix}})$, with $x_{\text{mix}}, y_{\text{mix}} \in (0, 1)$ (see Appendix C for details). The equilibrium (a_2, a_2) is called *risk-dominant* if

$$(u_{22} - u_{12})(v_{22} - v_{12}) > (u_{11} - u_{21})(v_{11} - v_{21}). \quad (6)$$

In particular, a NE is risk dominant if it has the largest basin of attraction (is less risky) (Harsanyi and Selten 1988). For symmetric games, inequality (6) has an intuitive interpretation: the choice at the risk dominant NE is the one that yields the highest expected payoff under complete ignorance, modelled by assigning $(1/2, 1/2)$ probabilities to the other agent's choices. If $u_{22} \geq u_{11}$ and $v_{22} \geq v_{11}$ with at least one inequality strict, then (a_2, a_2) is called *payoff-dominant*.

Depending on whether the interests of both agents are perfectly aligned — in the sense that $(u_{11} - u_{22})(v_{11} - v_{22}) > 0$ — or not, the QRE surface can be disconnected or connected. A formal characterization is provided in Theorem 6.

Theorem 6 (Geometric locus of the QRE equilibria in coordination games)**.** *Consider a two-player, $\mathcal{N} = \{1, 2\}$, two-action, $A_1 = A_2 = \{a_1, a_2\}$, coordination game $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k \in \mathcal{N}}\right)$ with payoff functions (u_1, u_2). If $x_{mix} + y_{mix} > 1$, then, for any exploration-exploitation rates $\alpha_x, \beta_x, \alpha_y, \beta_y > 0$ it holds that*

(i) If $x_{mix}, y_{mix} > 1/2$, then any QRE (x_Q, y_Q) satisfies either $x_Q > x_{mix}, y_Q > y_{mix}$ or $x_Q, y_Q < 1/2$.

(ii) If $x_{mix} > 1/2, y_{mix} \leq 1/2$, then any QRE (x_Q, y_Q) satisfies one of: $x_Q < 1/2, y_Q < y_{mix}$, $1/2 < x_Q < x_{mix}, y_{mix} < y_Q < 1/2$ and $x_Q > x_{mix}, y_Q > y_{mix}$.

In particular, if Γ is symmetric, i.e., if $u_2 = u_1^T$, then there exist no symmetric QRE, (x_Q, x_Q), with $1/2 < x_Q < x_{mix}$.

The statement of Theorem 6 is visualized in Figure 2. In the first case, disconnected QRE surface, the dynamics select the risk-dominant equilibrium after a *saddle-node bifurcation* in the exploration phase, regardless of whether it coincides with the payoff dominant equilibrium or not. In the second case, the QRE surface is connected via two branches of saddle-node bifurcations which is consistent with the emergence of a *cusp bifurcation* point. Hence, after exploration the learning process may rest to either of the two boundary equilibria. In short, the collective outcome of the SQL dynamics depends on the geometry of the QRE surface which is illustrated next.

Experiments: Phase Transitions in Games

To visualize the above, we start with 2×2 coordination games and then proceed to potential games with action spaces of arbitrary size. In all cases, we consider two representative exploration-exploitation policies: an *Explore-Then-Exploit* (ETE) policy (Bai and Jin 2020), which starts with (relatively) high exploration that reduces linearly to zero and a *Cyclical Learning Rate with one cycle* (CLR-1) policy (Smith and Topin 2017), which starts with low exploration, increases to high exploration around the middle of the cycle and decays to (ultimately) zero exploration (i.e., pure exploitation).[8]

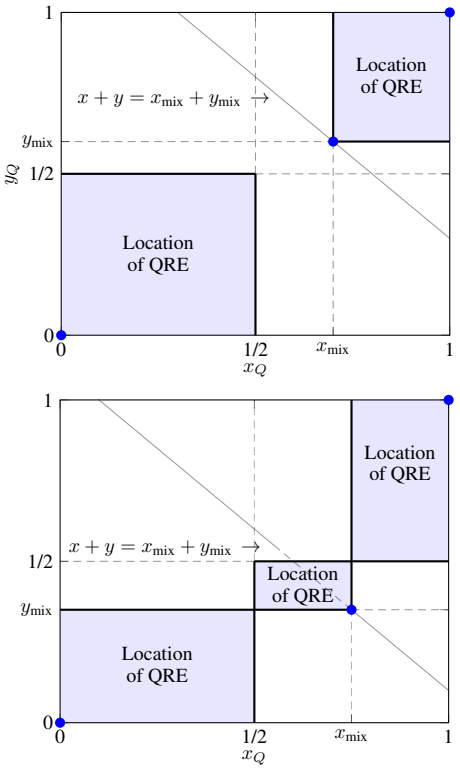

Figure 2: Geometric locus of QRE in 2×2 coordination games for all possible exploration rates in the two cases (i) $x_{\text{mix}}, y_{\text{mix}} \geq 1/2$ (upper panel) and (ii) $x_{\text{mix}} \geq 1/2, y_{\text{mix}} < 1/2$ (bottom panel) of Theorem 6. The blue dots are the NE of the underlying game Γ (when exploration is zero). In both panels, the risk-dominant equilibrium is $(0, 0)$.

Coordination Games 2×2. As long as agents' interests are aligned, sufficient exploration even by a single agent leads the learning process (after exploration is reduced back to zero) to the risk dominant equilibrium regardless of whether this equilibrium coincides with the payoff dominant equilibrium or not. Typical realizations of these cases are the Pareto Coordination and Stag Hunt games (Table 1).

In Pareto Coordination, (a_2, a_2) is both the risk- and payoff-dominant equilibrum whereas in Stag Hunt, the pay-

[7] A more general description is in Appendix C.

[8] The findings are qualitatively equivalent for non-linear, e.g., quadratic, changes in the exploration rates in both policies and for more that one learning cycle in the CLR policy. Moreover, in the numerical experiments, we have the used the transformation in Lemma A.1 in Appendix A which leads to a robust discretization of the ODEs in the theoretical analysis.

Pareto Coordination

	a_1	a_2
a_1	$1, 1$	$0, 0$
a_2	$0, 0$	$1.5, 1.8$

Battle of the Sexes

	a_1	a_2
a_1	$1.5, 1$	$0, 0$
a_2	$0, 0$	$1, 2$

Stag Hunt

	a_1	a_2
a_1	$3, 3$	$0, 2$
a_2	$2, 0$	$1.5, 1.5$

Table 1: Payoffs of the games in the Experiments Section.

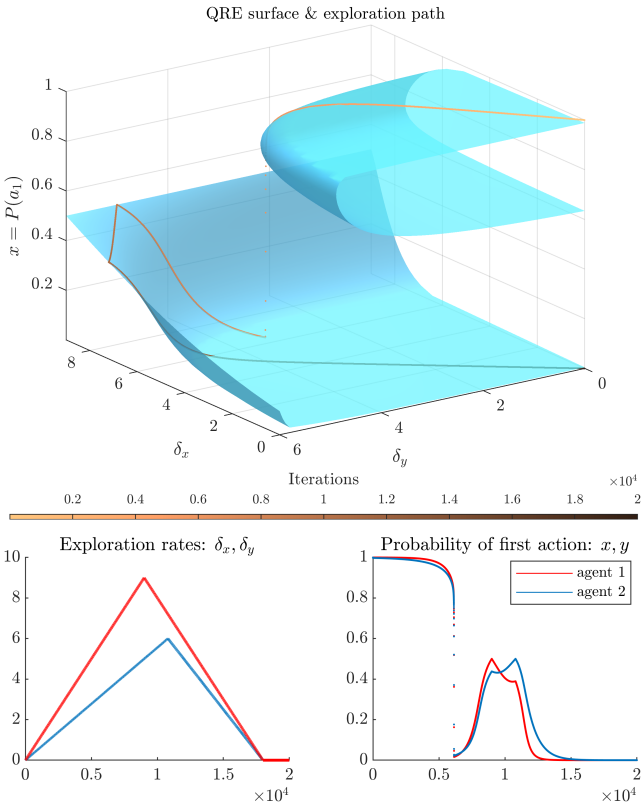

Figure 3: SQL in Stag Hunt. The upper panel shows the QRE surface and the exploration path of agent 1 (light to dark line). The bottom panels show the CLR-1 exploration rates (left) and the probability of action 1 during the learning process for both agents (right). As agents increase exploration, their choice distributions undergo a *saddle-node bifurcation* (disconnected surface). This prompts a permanent transition from the vicinity of the payoff dominant action profile, $(x, y) = (1, 1)$, in the upper component of the QRE surface to the (0,0) equilibrium when exploration reduces back to zero (right corner of the lower component).

off dominant equilibrium is (a_1, a_1). However, in both games $x_{\text{mix}}, y_{\text{mix}} > 1/2$ (due to the aligned interests of the players) which implies that the location of the QRE is described by the upper panel in Figure 2. Accordingly, the QRE surface

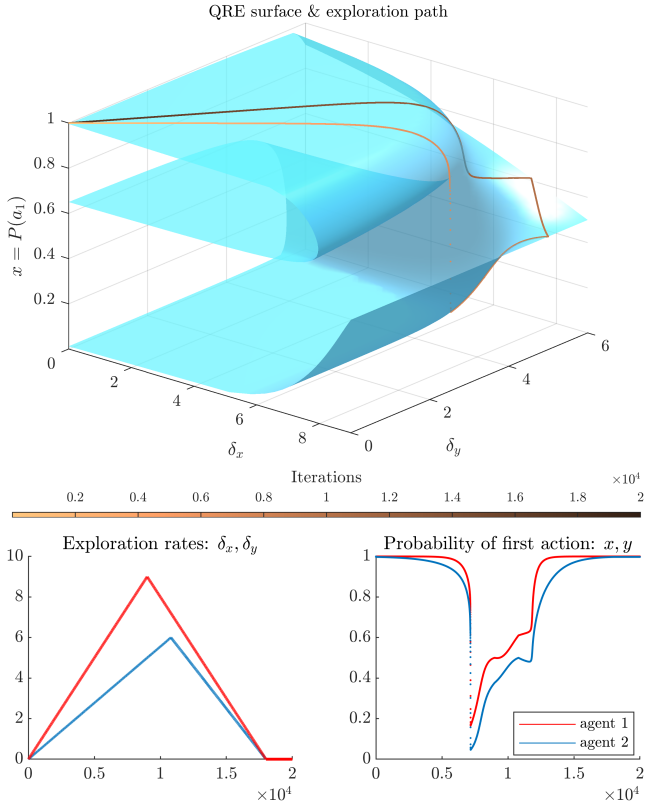

Figure 4: Exploration-Exploitation in Battle of the Sexes. In contrast to Stag Hunt, the QRE manifold has two branches of saddle-node bifurcation curves (consistent with the emergence of a co-dimension 2 cusp point) and the phase transition to the lower part of the QRE surface may not be permanent as illustrated here via the CLR-1 vs CLR-1 policies. (Examples with the CLR-1 vs ETE policies with permanent transitions are given in the full version).

is disconnected and if any agent sufficiently increases their exploration rate, the SQL dynamics converge to the risk-dominant equilibrium independently of the starting point and the exploration policy of the other agent. This is illustrated in Figure 3 (similarly, in Figure 8 in Appendix D). Note that in both theses cases, the risk-dominant equilibrium is the global maximizer of the potential function, see Lemma C.1 and (Alós-Ferrer and Netzer 2010; Schmidt et al. 2003).

By contrast, if agents' interests are not perfectly aligned, then the outcome of the exploration process is not unambiguous (even if the game remains a coordination game). A representative game of this class, in which no payoff dominant equilibrium exists, is the Battle of the Sexes in Table 1. The most preferable outcome is now different for the two agents which implies that there is no payoff dominant equilibrium. However, the pure joint profile (a_2, a_2) remains the risk-dominant equilibrium.[9] In this class of games, the loca-

[9] Although well defined, risk-dominance seems to be now less appealing: if agent 1 is completely ignorant about the equilibrium selection of agent 2 (and assigns a uniform distribution to agent 2's

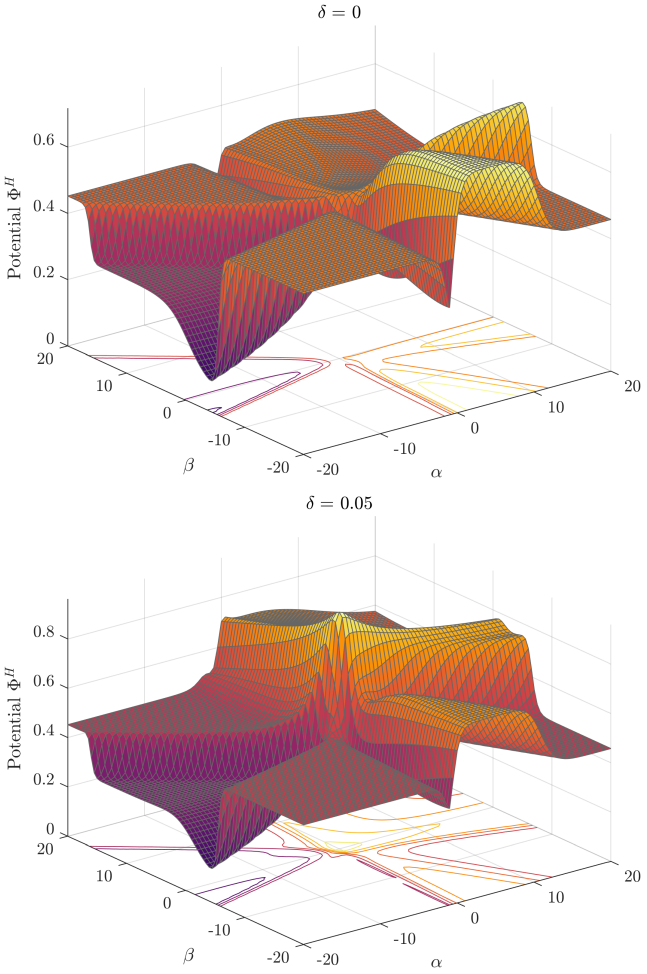

Figure 5: Snapshots of the modified potential Φ^H for different exploration rates in a symmetric 2-player potential game with random payoffs in $[0, 1]$. Unlike Figures 3 and 4, we now visualize the potential function instead of the QRE surface. Hence, we cannot reason about the bifurcation types. However, we see that without exploration, $\delta = 0$, the potential (equal to the potential, Φ, of the original game) has various local maxima, whereas as exploration increases, a unique remaining attractor (maximum) forms at the vicinity of the uniform distribution, $(0, 0)$ in the transformed coordinates.

tion of the QRE is described by the bottom panel in Figure 2. The QRE surface is connected and the collective output of the exploration process depends on the exploration policies (timing and intensity) of the two agents. This is illustrated in Figure 4. In Appendix D (see also full version), we provide an exhaustive treatment of the possible outcomes under combinations of the ETE and CLR-1 exploration policies.

Potential games in larger dimensions To visualize the modified potential in equation (5) of Lemma 4, we adapt

actions), then agent 1 is better off to select action a_1, despite the fact that (a_2, a_2) is the risk-dominant equilibrium.

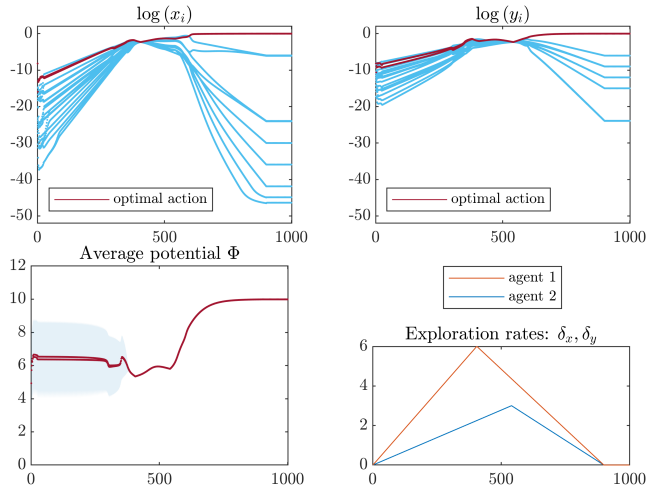

Figure 6: Exploration-Exploitation with the SQL dynamics in a potential game with $n = 10$ actions. The upper panels show the (log) choice distributions (with the optimal action in different color). The lower left panel shows the average potential over a set of 10×10 different trajectories (starting points) and one standard deviation (shaded region that disappears after all trajectories converge to the same choice distribution). The bottom right panel shows the selected CLR-1 policies.

the two-dimensional projection technique of Li et al. (2018). Given a potential game with potential Φ and n, m actions for agents 1 and 2, we first embed their choice distributions into $\mathbb{R}^{n+m-2}$ and remove the Simplex restrictions. This is done via the transformation $y_i := \log x_i / x_n$ from $\mathbb{R}^n \to \mathbb{R}^{n-1}$ (with $\sum_{i=1}^n x_i = 1$) for the first agent and similarly for the second agent. Then, we choose two arbitrary directions in $\mathbb{R}^{n+m-2}$ along which we plot the modified potential $\Phi^H(x) = \Phi(x) + \sum_{k \in \mathcal{N}} \delta_k H(x_k)$, for $x \in X$, cf. equation (5). For simplicity, we keep the exploration ratio $\delta_k := \beta_k / \alpha_k$ equal to a common δ for both players.[10]

A visualization of a randomly generated 2-player potential game is given in Figure 5. As players modify their exploration rates, the SQL dynamics converge to different QRE (local maxima) of these changing surfaces. However, when exploration is large, a single attracting QRE remains (similar to the low dimensional case).

In Figure 6, we plot the SQL dynamics ($1e-20$ Q-value updates for each of $1e-03$ choice distribution updates) in a 2-player potential game with $n = 10$ actions and potential, Φ, with random values in $[0, 10]$. Both agents use CLR-1 policies. Starting from a grid of initial conditions, one close to each pure action pair, the SQL dynamics rest at different local optima before the exploration, converge to the uniform distribution when exploration rates reach their peak and then converge to the same (in this case, global) optimum when exploration is gradually reduced back to zero (horizontal line and vanishing shaded region).

[10] A detailed description of the routine is in Appendix D. This method produces similar visualizations for any number of players.

Acknowledgements

Stefanos Leonardos gratefully acknowledges NRF 2018 Fellowship NRF-NRFF2018-07. Georgios Piliouras gratefully acknowledges grant PIE-SGP-AI-2020-01, NRF2019-NRF-ANR095 ALIAS grant and NRF 2018 Fellowship NRF-NRFF2018-07.

References

Alós-Ferrer, C.; and Netzer, N. 2010. The logit-response dynamics. *Games and Economic Behavior* 68(2): 413–427. doi:10.1016/j.geb.2009.08.004.

Bai, Y.; and Jin, C. 2020. Provable Self-Play Algorithms for Competitive Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. Madison, WI, USA: Omnipress.

Balduzzi, D.; Czarnecki, W. M.; Anthony, T.; Gemp, I.; Hughes, E.; Leibo, J.; Piliouras, G.; and Graepel, T. 2020. Smooth markets: A basic mechanism for organizing gradient-based learners. In *International Conference on Learning Representations*.

Ben-Porat, O.; and Tennenholtz, M. 2018. A Game-Theoretic Approach to Recommendation Systems with Strategic Content Providers. In et. al, S. B., ed., *Advances in Neural Information Processing Systems 31*, 1110–1120. Curran Associates, Inc.

Bloembergen, D.; Tuyls, K.; Hennes, D.; and Kaisers, M. 2015. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *J. Artif. Int. Res.* 53(1): 659–697.

Bowling, M.; and Veloso, M. 2002. Multiagent learning using a variable learning rate. *Artificial Intelligence* 136(2): 215–250. doi:10.1016/S0004-3702(02)00121-2.

Cesa-Bianchi, N.; and Lugosi, G. 2006. *Prediction, learning, and games*. Cambridge university press.

Claus, C.; and Boutilier, C. 1998. The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems. AAAI '98/IAAI '98, 746–752.

Coucheney, P.; Gaujal, B.; and Mertikopoulos, P. 2015. Penalty-Regulated Dynamics and Robust Learning Procedures in Games. *Mathematics of Operations Research* 40(3): 611–633. doi:10.1287/moor.2014.0687.

Gao, B.; and Pavel, L. 2017. On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning. *arXiv e-prints* arXiv:1704.00805.

Göcke, M. 2002. Various Concepts of Hysteresis Applied in Economics. *Journal of Economic Surveys* 16(2): 167–188. doi:10.1111/1467-6419.00163.

Harsanyi, J.; and Selten, R. 1988. *A General Theory of Equilibrium Selection in Games*. Massachusetts, USA: The MIT Press.

Ho, T.-H.; and Camerer, C. 1998. Experience-weighted attraction learning in coordination games: Probability rules, heterogeneity, and time-variation. *Journal of mathematical psychology* 42: 305–326.

Ho, T.-H.; and Camerer, C. 1999. Experience-weighted attraction learning in normal form games. *Econometrica* 67: 827–874.

Ho, T.-H.; Camerer, C. F.; and Chong, J.-K. 2007. Self-tuning experience weighted attraction learning in games. *Journal of economic theory* 133: 177–198.

Kaelbling, L. P.; Littman, M. L.; and Moore, A. W. 1996. Reinforcement Learning: A Survey. *Journal of Artificial Intelligence Research* 4(1): 237–285. doi:10.1613/jair.301.

Kaisers, M.; and Tuyls, K. 2010. Frequency Adjusted Multi-Agent Q-Learning. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1 - Volume 1*, AAMAS '10, 309–316.

Kaisers, M.; and Tuyls, K. 2011. FAQ-Learning in Matrix Games: Demonstrating Convergence near Nash Equilibria, and Bifurcation of Attractors in the Battle of Sexes. In *Proceedings of the 13th AAAI Conference on Interactive Decision Theory and Game Theory*, AAAIWS'11-13, 36–42. AAAI Press.

Kaisers, M.; Tuyls, K.; Parsons, S.; and Thuijsman, F. 2009. An Evolutionary Model of Multi-Agent Learning with a Varying Exploration Rate. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, 1255–1256.

Kianercy, A.; and Galstyan, A. 2012. Dynamics of Boltzmann *Q* learning in two-player two-action games. *Phys. Rev. E* 85: 041145. doi:10.1103/PhysRevE.85.041145.

Kleinberg, R.; Piliouras, G.; and Tardos, E. 2009. Multiplicative Updates Outperform Generic No-Regret Learning in Congestion Games. In *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*, STOC '09, 533–542. doi:10.1145/1536414.1536487.

Klos, T.; Van Ahee, G. J.; and Tuyls, K. 2010. Evolutionary Dynamics of Regret Minimization. In Balcázar, J. L.; Bonchi, F.; Gionis, A.; and Sebag, M., eds., *Machine Learning and Knowledge Discovery in Databases*, 82–96. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kwoon, J.; and Mertikopoulos, P. 2017. A continuous-time approach to online optimization. *Journal of Dynamics & Games* 4(2): 125–148. doi:10.3934/jdg.2017008.

Lanctot, M.; Zambaldi, V.; Gruslys, A.; Lazaridou, A.; Tuyls, K.; Perolat, J.; Silver, D.; and Graepel, T. 2017. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. In et. al, I. G., ed., *Advances in Neural Information Processing Systems 30*, 4190–4203. Curran Associates, Inc.

Leslie, D. S.; and Collins, E. J. 2005. Individual Q-Learning in Normal Form Games. *SIAM Journal on Control and Optimization* 44(2): 495–514. doi:10.1137/S0363012903437976.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the Loss Landscape of Neural Nets. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31*, 6389–6399. Curran Associates, Inc.

Mazumdar, E.; Ratliff, L. J.; and Shankar Sastry, S. 2018. On Gradient-Based Learning in Continuous Games. *arXiv e-prints* arXiv:1804.05464.

McKelvey, R. D.; and Palfrey, T. R. 1995. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior* 10(1): 6–38. doi:10.1006/game.1995.1023.

Mertikopoulos, P.; Papadimitriou, C.; and Piliouras, G. 2018. Cycles in Adversarial Regularized Learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, 2703–2717. USA: Society for Industrial and Applied Mathematics.

Mertikopoulos, P.; and Sandholm, W. H. 2016. Learning in Games via Reinforcement and Regularization. *Mathematics of Operations Research* 41(4): 1297–1324. doi:10.1287/moor.2016.0778.

Omidshafiei, S.; Papadimitriou, C.; Piliouras, G.; Tuyls, K.; Rowland, M.; Lespiau, J.-B.; Czarnecki, W. M.; Lanctot, M.; Perolat, J.; and Munos, R. 2019. a-Rank: Multi-Agent Evaluation by Evolution. *arXiv preprint arXiv:1903.01373* .

Palaiopanos, G.; Panageas, I.; and Piliouras, G. 2017. Multiplicative Weights Update with Constant Step-Size in Congestion Games: Convergence, Limit Cycles and Chaos. NIPS'17, 5874–5884. doi:10.5555/3295222.3295337.

Panageas, I.; and Piliouras, G. 2016. Average Case Performance of Replicator Dynamics in Potential Games via Computing Regions of Attraction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, 703–720. doi:10.1145/2940716.2940784.

Panait, L.; and Luke, S. 2005. Cooperative Multi-Agent Learning: The State of the Art. *Autonomous Agents and Multi-Agent Systems* 11(3): 387–434. doi:10.1007/s10458-005-2631-2.

Perolat, J.; Munos, R.; Lespiau, J.-B.; Omidshafiei, S.; Rowland, M.; Ortega, P.; Burch, N.; Anthony, T.; Balduzzi, D.; De Vylder, B.; Piliouras, G.; Lanctot, M.; and Tuyls, K. 2020. From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization. *arXiv e-prints* arXiv:2002.08456.

Puigdomènech Badia, A.; Piot, B.; Kapturowski, S.; Sprechmann, P.; Vitvitskyi, A.; Guo, D.; and Blundell, C. 2020. Agent57: Outperforming the Atari Human Benchmark. *arXiv e-prints* arXiv:2003.13350.

Romero, J. 2015. The effect of hysteresis on equilibrium selection in coordination games. *Journal of Economic Behavior & Organization* 111: 88–105.

Roughgarden, T. 2015. Intrinsic Robustness of the Price of Anarchy. *J. ACM* 62(5). doi:10.1145/2806883.

Rowland, M.; Omidshafiei, S.; Tuyls, K.; Perolat, J.; Valko, M.; Piliouras, G.; and Munos, R. 2019. Multiagent Evaluation under Incomplete Information. In et. al, H. W., ed., *Advances in Neural Information Processing Systems 32*, 12291–12303.

Sanders, J. B. T.; Farmer, J. D.; and Galla, T. 2018. The prevalence of chaotic dynamics in games with many players. *Scientific Reports* 8(1): 4902. doi:10.1038/s41598-018-22013-5.

Sato, Y.; Akiyama, E.; and Crutchfield, J. P. 2005. Stability and diversity in collective adaptation. *Physica D:Nonlinear Phenomena* 210(1): 21–57.

Sato, Y.; and Crutchfield, J. P. 2003. Coupled replicator equations for the dynamics of learning in multiagent systems. *Phys. Rev. E* 67: 015206. doi:10.1103/PhysRevE.67.015206.

Schmidt, D.; Shupp, R.; Walker, J. M.; and Ostrom, E. 2003. Playing safe in coordination games:: the roles of risk dominance, payoff dominance, and history of play. *Games and Economic Behavior* 42(2): 281–299. doi:10.1016/S0899-8256(02)00552-3.

Smith, L. N.; and Topin, N. 2017. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. *arXiv e-prints* arXiv:1708.07120.

Strogatz, S. H. 2000. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. Cambridge, MA, USA: Westview Press (Studies in nonlinearity collection), first edition.

Swenson, B.; Murray, R.; and Kar, S. 2018. On Best-Response Dynamics in Potential Games. *SIAM Journal on Control and Optimization* 56(4): 2734–2767. doi:10.1137/17M1139461.

Tuyls, K.; Hoen, P. J. T.; and Vanschoenwinkel, B. 2006. An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games. *Autonomous Agents and Multi-Agent Systems* 12(1): 115–153. doi:10.1007/s10458-005-3783-9.

Tuyls, K.; Verbeeck, K.; and Lenaerts, T. 2003. A Selection-mutation Model for Q-learning in Multi-agent Systems. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, AAMAS '03, 693–700. doi:10.1145/860575.860687.

Tuyls, K.; and Weiss, G. 2012. Multiagent Learning: Basics, Challenges, and Prospects. *AI Magazine* 33(3): 41. doi:10.1609/aimag.v33i3.2426.

Watkins, C. J.; and Dayan, P. 1992. Technical Note: Q-Learning. *Machine Learning* 8(3): 279–292. doi:10.1023/A:1022676722315.

Wolpert, D. H.; Harré, M.; Olbrich, E.; Bertschinger, N.; and Jost, J. 2012. Hysteresis effects of changing the parameters of noncooperative games. *Phys. Rev. E* 85: 036102. doi:10.1103/PhysRevE.85.036102.

Wunder, M.; Littman, M.; and Babes, M. 2010. Classes of Multiagent Q-Learning Dynamics with ϵ-Greedy Exploration. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, 1167–1174. Madison, WI, USA: Omnipress.

Yang, G.; Piliouras, G.; and Basanta, D. 2017. Bifurcation Mechanism Design - from Optimal Flat Taxes to Improved Cancer Treatments. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, EC '17, 587–587. doi:10.1145/3033274.3085144.