Version of Record: https://www.sciencedirect.com/science/article/pii/S0004370221002046
Manuscript_932ecdf3a277afc76cdb27915ece241a

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

Stefanos Leonardos[a,*], Georgios Piliouras[a]

[a]*Singapore University of Technology and Design, 8 Somapah Road, 487372 Singapore*

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 (1) that smooth Q-learning has bounded regret in arbitrary games for a cost model that explicitly balances game-rewards and exploration-costs, i.e., costs from testing potentially suboptimal actions, and (2) that it always converges to the set of quantal-response equilibria (QRE), the standard solution concept for games with bounded rationality, in arbitrary weighted potential games. In our main task, we then turn to measure the effect of exploration on 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.

Keywords: Exploration-Exploitation, Multi-Agent Learning, Game Theory, Catastrophe Theory
2010 MSC: 93A16, 91A26, 91A68, 58K35

1. 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 [1, 2]. From a behavioral perspective, it involves the design of realistic models to capture complex human behavior, such as the standard Experienced Weighted Attraction model [3, 4]. Learning (adapting) agents use time-varying parameters to explore potentially suboptimal, boundedly rational decisions, while, at the same time, they try to coordinate with other learning agents in order to maximize their long-term rewards [5, 6, 7].

From an Artificial Intelligence (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 [8]. Similar techniques have been applied to rank agents in tournaments according to performance for preferential evolvability [9, 10, 11] and to design *multi-agent learning (MAL)* algorithms that prevent the joint learning process from getting trapped in local optima [12, 13].

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 [14, 15]. While there is still no theory to formally explain the performance of MAL algorithms, and in particular, *the effects of exploration in*

*Corresponding author
Email addresses: stefanos_leonardos@sutd.edu.sg (Stefanos Leonardos), georgios@sutd.edu.sg (Georgios Piliouras)

MAS [16], 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 [17, 18, 19].

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 [20, 21, 22]. 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 [23, 24, 25].

Our approach and 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 being receiving increasing attention due to its connection with evolutionary game theory [13, 26, 23, 15].[1] Informally (see Section 2 for the rigorous definition), each agent k updates her choice distribution $x = (x_i)$ according to the rule

$$\dot{x}_i/x_i = \beta_k \underbrace{(u_i - \bar{u})}_{\text{exploitation}} - \alpha_k \underbrace{\Big(\ln x_i - \sum\nolimits_j x_j \ln x_j\Big)}_{\text{exploration}},$$

where $u_i, \bar{u}$ denote agent k's utility from action i and average utility, respectively, given all other agents' choice distributions (policies), and α_k/β_k is agent k's exploration rate.[2] Agents tune their exploration rate 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 3.1), 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 3.2). Second, we show that they converge to *Quantal Response Equilibria* (QRE), the prototypical extension of NE in games with bounded rationality [32], in weighted potential games with arbitrary numbers of heterogeneous agents (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 3.4).

While previously not formally known, these properties mirror results of strong regret guarantees for online algorithms, see e.g., [33, 34, 21], and convergence results for SQL in more restricted settings [35, 36].[3] 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 (bounded regret, convergence to QRE) and their unpredictable performance in practice (outcomes on agents' utilities after exploration).

In our effort to explain the latter, we are faced with two major unresolved complications. First, the outcome of the SQL algorithm in MAS is highly sensitive on the exploration parameters [38]. 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 [39, 40, 31].

[1] This variant of Q-learning has been also extensively studied in the economics and reinforcement learning literature under various names, see e.g., [27, 25] and [28, 29], respectively.

[2] Here, $\beta_k \in [0, +\infty)$ and $\alpha_k \in [0, 1)$ are positive constants that control the learning agent's rate of adaptation and memory loss, respectively [18]. Since β_k is only effective in rescaling the game's payoffs, these dynamics are frequently described via a single parameter having the same interpretation as α_k/β_k in the current formulation, see e.g., [30, 31].

[3] They are also of independent interest in the limited literature on the properties of the softmax function [37].

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 [41]. 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 4.3 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).

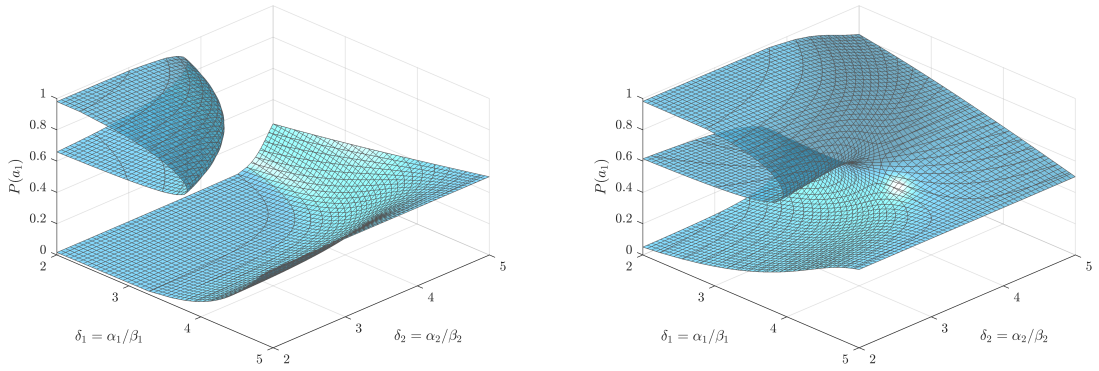

Figure 1: Geometry of QRE surfaces: single *saddle-node bifurcation* curve (left) vs two branches of saddle-node bifurcation curves meeting which is consistent with the emergence of a *co-dimension* 2 *cusp bifurcation* (right) on the QRE manifold of two-player, two-action games as function of the exploration rates, δ_x, δ_y (cf. Figures 3 and 4 for more details). The possible learning paths before, during and after exploration depend on the geometry of the QRE surface (Theorems 4.1,4.3).

In terms of performance, this is formalized in Theorem 4.1 which states that even in the simplest of MAS, *exploration can lead under different circumstances both to unbounded gain as well as unbounded loss.* Gains and losses refer here to agents' utilities in the original (unmodified) game that is played among them (of course, this implies unbounded gains and losses also in the modified game). While existential in nature, Theorem 4.1 does not merely say that anything goes when exploration is performed. When coupled with the characterization of the geometric locus of QRE in Theorem 4.3, 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.

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* (Section 5). 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 [42] for visualizing high dimensional loss functions in deep learning (Figure 10).

Related Literature

As mentioned in the previous part of the Introduction, our current work is related to several strands of the AI, game theoretic and multi-agent learning literature. The first involves the study of the same variant of Q-learning dynamics in multi-agent settings from a theoretical perspective. In this direction, our paper is most closely related to [13] who study SQL in the context of two-player, two-action games. Similar ideas concerning bifurcations and abrupt phase transitions can be found in [26] and [43] who study SQL in two-player, two-action games and in the evolution of metabolic tumor cells, respectively. In parallel with these works, [23] analyze the dependence of equilibria with bounded rationality on the parameters of

the underlying game and introduce the study of the resulting hysteresis effects. This connection is further elaborated in [31, 44] who use these effects to design new types of optimal control mechanisms in multi-agent systems. Motivated by these works, but moving in an orthogonal direction, our current paper formalizes the connections between exploration-exploitation in multi-agent learning and catastrophe theory and uses this framework to formally argue, for the first time, about the effects of exploration on collective system performance.

The second strand involves a line of works that study more general evolutionary dynamics (including SQL) in multi-agent settings. Benchmark papers in this direction include [30, 17, 18] and [36, 29]. These early works develop the formal connection between the exploration-exploitation scheme from reinforcement learning and the selection-mutation mechanisms from evolutionary game theory and document the possible variety of outcomes in collective learning dynamics (including quasi-periodicity, stable limit cycles, intermittency, and deterministic chaos). These works fall within a more general strand of literature that studies the connections of AI with evolutionary game theory such as [2, 38, 45, 15]. Along this line, our paper also contributes to the literature that studies the exploration-exploitation dilemma. Related models of multi-agent learning with variable exploration rates have been studied in [6, 38] and [7, 14, 12].

Our current contribution is also related to the study of behaviorally motivated dynamics in complex (either multi-action or multi-player) game-theoretic paradigms. Representative works from this line of literature include [46] who investigate two-person games with many possible actions in which the players learn based on experience-weighted attraction (EWA) (a behaviorally justified model that is closely related to Q-learning [3, 5, 4]) and [25] who find that complex non-equilibrium behavior, exemplified by chaos, is the norm for complicated games with many players. In a follow-up of the conference version of the current work [47], [48] leverage bifurcations of learning dynamics to find diverse solutions in optimization problems that involve multi-agent interactions [49]. Finally, concerning the game-theoretic setup, our work contributes to the growing literature on (weighted) potential games. Besides their natural emergence in distributed settings such as recommendation and congestion control systems ([50] and [51, 52, 53]), potential games are receiving increasing attention in the reinforcement learning literature as suitable frameworks to capture complex, multi-agent coordination [54, 55, 56, 57].

Paper Outline

Section 2 presents the smooth Q-learning (SQL) model and preliminary notation. Section 3 includes the theoretical results concerning regret and convergence of SQL in games. Section 4 focuses on performance of SQL in general and coordination games. Visualizations and experiments are presented in Section 5. Section 6 concludes the paper with final remarks and open questions. Omitted materials, all proofs, and systematic experiments are included in Appendices A to D.

2. Preliminaries: Game Theory and SQL

We consider a finite set $\mathcal{N}$ of interacting agents (players) 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 $k \in \mathcal{N}$.[4] 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 $t \geq 0$.

[4] 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.

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

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

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

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.[5] 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

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

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

3. Bounded Regret in All Games and Convergence in Weighted Potential Games

Our first observation is that the SQL dynamics (equation (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. This is formulated in Lemma 3.1. The superscript H refers to the regularizing term, $H(x_k) := -\langle x, \ln x_k \rangle = -\sum_{j \in A_k} x_{kj} \ln x_{kj}$ which denotes the *Shannon entropy* of choice distribution $x_k \in X_k$.

Lemma 3.1. *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*

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

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 $\Gamma^H = \left(\mathcal{N}, \left(A_k, u_k^H \right)_{k \in \mathcal{N}} \right)$.

[5] An explicit derivation due to [30, 18, 23, 26] (among others) of the connection between Q-learning [58] and the above dynamics (including their resting points) is given in Appendix A. Briefly, this connection rests on the assumptions that agents perform multiple Q-value updates by interacting with their environment for each choice distribution update which is then accomplished via a Boltzmann probability distribution. The updates are independent, and the agents' rewards remain constant over time for each action profile. A variation of this scheme, termed frequency adjusted Q-learning, has been proposed by [12].

Bounded Regret

To measure the performance of the SQL dynamics in (1), we will use the standard notion of (accumulated) *regret* [21]. The regret $R_k(T)$ at time $T > 0$ for agent k is

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

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 part is a constant upper bound on the regret of the SQL dynamics.

Theorem 3.2. *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 3.2 (see Appendix B), 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., [59]. 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 (see Section 4).

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 $k \in \mathcal{N}$, $u_k(i, a_{-k}) - u_k(j, a_{-k}) = w_k\left(\phi(i, a_{-k}) - \phi(j, a_{-k})\right)$, 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 3.3. *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.

The proof of Theorem 3.3 is based on the following intuitive argument. In the update rule of the SQL dynamics (equation (1)), the first term, $\beta_k\left(r_{ki}(x) - \langle x_k, r_k(x)\rangle\right)$, 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, $-\alpha_k\left(\ln x_{ki} - \langle x_k, \ln x_k\rangle\right)$, 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 3.3 can be established by extending the techniques of [60, 36] to the case of weighted potential games. This is the statement of Lemma 3.4 (which is also useful for the numerical experiments).

Lemma 3.4. *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*

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

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.

4. 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 4.1 and visualize in Section 5, exploration – exploitation may lead to (unbounded) improvement, but also to (unbounded) performance loss even in simple settings.

To compare outcomes between different exploration-exploitation policies, it will be convenient to use the following notation. If exploration remains *low* for all agents, i.e., if there exist thresholds $\bar{\delta}_k > 0$ (that may depend on the initial condition $x_k(0)$ of agent k) such that $\delta_k(t) < \bar{\delta}_k$ for all $k \in \mathcal{N}$, we will denote the sequence of each agent k's utilities by $u_k^{\text{exploit}}(t), t \geq 0$; otherwise, we will write $u_k^{\text{explore}}(t), t \geq 0$. Here, all utilities refer to the agents' utilities in the original (unmodified) game as intended. Using this notation, we can now formulate our main result.

Theorem 4.1 (**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*

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

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 4.1 is constructive and relies on Theorem 4.3 discussed next. In particular, in Theorem 4.3, we exploit the special structure of two-player, two-action coordination games (defined next) that form a subclass of weighted potential games, and provide a complete characterization of the possible geometries of the QRE surfaces in these games. As we show, this 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, and can be used to obtain the formal statement of Theorem 4.1.

Classification of 2×2 Coordination Games and Geometry of the QRE Surface

First, we introduce some minimal additional notation and terminology regarding *coordination games* (a subclass of weighted potential games). Concrete examples of such games are provided in Table 1 which will be relevant in the context of our numerical experiments. A two-player $\mathcal{N} = \{1, 2\}$, two-action $A_1 = A_2 = \{a_1, a_2\}$ game $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k\in\mathcal{N}}\right)$ can be described via two payoff matrices

$$u_1 = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix}, \; u_2 = \begin{pmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \end{pmatrix}, \tag{6a}$$

where u_{ij} (v_{ij}) denotes the payoff of agent 1 (2) when this agent uses action a_i and the other agent uses action a_j for $i, j = 1, 2$. Γ is a *coordination game*, if

$$u_{11} > u_{21}, u_{22} > u_{12} \;\text{ and }\; v_{11} > v_{21}, v_{22} > v_{12} \tag{6b}$$

hold for the payoffs of agents 1 and 2, respectively. Coordination games admit three NE that can be described in terms of the probabilities $x, y \in [0, 1]$ of agents 1 and 2 using action a_1: two pure NE $(x, y) = (1, 1)$ and $(x, y) = (0, 0)$ that correspond to the pure action profiles (a_1, a_1) and (a_2, a_2) and one fully mixed at $(x_{\text{mix}}, y_{\text{mix}}) = (\lambda_2/k_2, \lambda_1/k_1)$ where $\lambda_1 := u_{22} - u_{12}$, and $k_1 := u_{11} - u_{12} - u_{21} + u_{22}$, and, analogously, $\lambda_2 := v_{22} - v_{12}$, and $k_2 := v_{11} - v_{12} - v_{21} + v_{22}$, with $\lambda_i, k_i > 0$ for $i = 1, 2$. The equilibrium (a_2, a_2) is called *risk-dominant* if

$$(u_{22} - u_{12})(v_{22} - v_{12}) > (u_{11} - u_{21})(v_{11} - v_{21}). \tag{7}$$

In particular, a NE is risk dominant if it has the largest basin of attraction (is less risky) [61]. In symmetric games, i.e., when $u_2 = u_1^\top$, inequality (7) simplifies to $u_{22}+u_{21} > u_{11}+u_{12}$ and 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 the inequality in (7) is reversed, then (a_1, a_1) is risk dominant and if equality holds, then none of the pure equilibria risk-dominates the other. Finally, 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* or *Pareto-dominant*. Coordination games have the following properties.

Lemma 4.2 (Properties of 2×2 coordination games)**.** *Let $\Gamma = \big(\mathcal{N}, (A_k, u_k)_{k\in\mathcal{N}}\big)$ denote a two-player, $\mathcal{N} = \{1, 2\}$, two-action, $A_1 = A_2 = \{a_1, a_2\}$, coordination game with payoff functions (u_1, u_2) as in equations (6). Then, Γ is a weighted potential game with weights $(1, w)$ for some $w > 0$ and it holds that*

(i) $x_{mix} + y_{mix} > 1$ if and only if $(0, 0)$, i.e., the pure action profile (a_2, a_2), is the risk-dominant equilibrium.

If, in addition, Γ is symmetric, i.e., $u_2 = u_1^\top$, then $x_{mix} = y_{mix}$ and property (i) simplifies to

(i) $x_{mix} > 1/2$ if and only if $(0, 0)$ is the risk-dominant equilibrium.*

In this case, the (exact) potential is globally maximized at the pure action profile (a_2, a_2), i.e., at the risk-dominant equilibrium.

As a special case, it is immediate from Lemma 4.2 that a necessary and sufficient condition for Γ to be an exact potential game is that $k_1 = k_2$. Using Lemma 4.2, we can now reason about the possible locations of QRE in 2×2 coordination games. In particular, 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. This is established in Theorem 4.3 where we focus on the case $x_{\text{mix}} + y_{\text{mix}} > 1$ with $x_{\text{mix}} > 0.5$.[6] By Lemma 4.2, this is the case in which $(0, 0)$ is the risk-dominant equilibrium.

Theorem 4.3 (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 $\Gamma = \big(\mathcal{N}, (A_k, u_k)_{k\in\mathcal{N}}\big)$ with payoff functions (u_1, u_2) as in equations (6). Then, for any positive exploration rates $\delta_x, \delta_y > 0$, it holds that*

(i) If $x_{mix} > 0.5, y_{mix} \geq 0.5$, then any QRE, (x_Q, y_Q), satisfies either $(x_Q, y_Q) > (x_{mix}, y_{mix})$ or $(x_Q, y_Q) < (0.5, 0.5)$.

(ii) If $x_{mix} > 0.5, y_{mix} < 0.5$, then any QRE, (x_Q, y_Q), satisfies either of the following: (1) $(x_Q, y_Q) < (0.5, y_{mix})$, (2) $(x_Q, y_Q) = (0.5, y_{mix})$ or $(x_{mix}, 0.5)$, (3) $(0.5, y_{mix}) < (x_Q, y_Q) < (x_{mix}, 0.5)$, or (4) $(x_Q, y_Q) > (x_{mix}, 0.5)$.

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

Remark 4.4 (Intuition of Theorem 4.3). The statement of Theorem 4.3 is visualized in Figure 2. The main intuition is that the QRE surface may or may not be connected depending on the alignment of the interests of the two agents, i.e., on whether $x_{\text{mix}} > 1/2, y_{\text{mix}} \geq 1/2$ (left panel) or $x_{\text{mix}} > 1/2, y_{\text{mix}} < 1/2$ (right panel) of Figure 2. In turn, the connectedness of the QRE surface crucially affects the convergence of the learning process. 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 are consistent with the emergence of a *cusp bifurcation* point. In this case, after exploration, the dynamics may rest to either of the two boundary equilibria.[7].

Intuitively, if one agent sufficiently increases their exploration rate, then that agent's choice distribution will move towards the uniform distribution, i.e., $(1/2, 1/2)$, regardless of what the other agent is playing

[6] This is without loss of generality since the remaining cases on $x_{\text{mix}} + y_{\text{mix}}$ and x_{mix} yield symmetric results. The case

[7] Figure 2 should be further compared with Figures 3 to 5. In particular, the left panel of Figure 2 captures the games of Stag Hunt and Pareto Coordination and should be compared with Figure 3 and the left panel of Figure 5. The right panel of Figure 2 captures the game of Battle of the Sexes and should be compared with Figure 4 and the right panel of Figure 5 (the games of Stag Hunt, Pareto Coordination and Battle of the Sexes will be introduced in Table 1 in Section 5)

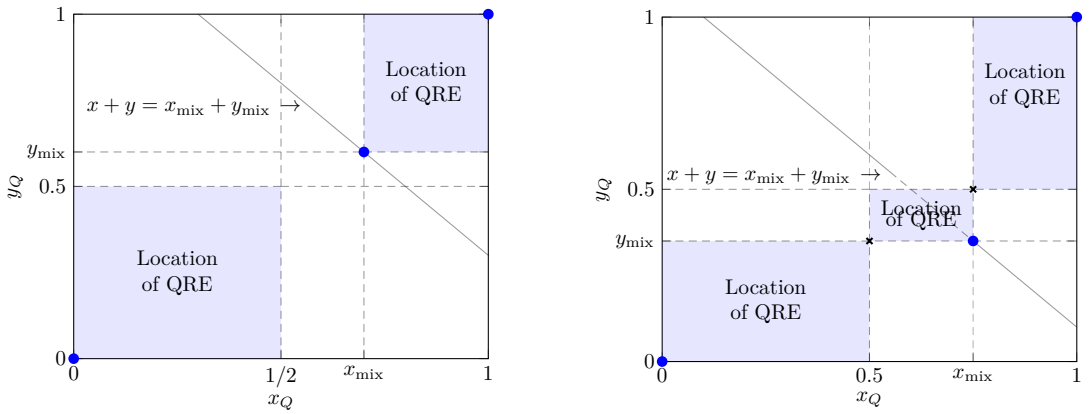

Figure 2: Geometric locus of QRE in 2×2 coordination games for all possible exploration rates in the two cases (i) $x_{\text{mix}} > 0.5, y_{\text{mix}} \geq 0.5$ (left panel) and (ii) $x_{\text{mix}} > 0.5, y_{\text{mix}} < 0.5$ (right panel) of Theorem 4.3. The NE of the underlying game Γ (when exploration is zero) are denoted with blue dots. In the right panel, the black cross marks denote the two connecting QREs at $(0.5, y_{\text{mix}})$ and $(x_{\text{mix}}, 0.5)$. In both panels, the risk-dominant equilibrium is $(0, 0)$.

(exploration dominates exploitation). In case that the payoff parameters of the underlying game are such that the game is described by the left panel of Figure 2, then the SQL dynamics will rest in some QRE in the bottom left (shaded) square. This lies entirely in the attracting region of the risk-dominant equilibrium, $(x, y) = (0, 0)$, at the bottom left corner. Hence, after reducing the exploration rate to 0, the SQL dynamics will converge to this equilibrium. Importantly, this outcome is independent of the starting point *and* the exploration-exploitation profile of the other agent and holds even if the other agent starts from a pure action and only exploits. The reason is that the changing choice distribution of the exploring player will prompt a change in that agent's optimal action as well.

In the special case of symmetric games, i.e., games in which $u_2 = u_1^\top$, it holds that $x_{\text{mix}} = y_{\text{mix}}$ by symmetry which implies that such games always fall under case (i) of Theorem 4.3. Thus, as the concluding remark in Theorem 4.3 suggests, such games cannot have a QRE with $x_Q \in (0.5, x_{\text{mix}})$ which provides a handy way to conclude that their QRE surface is disconnected. This is illustrated by the Pareto Coordination and Stag Hunt games (visualized in the next section).

By contrast, convergence to the risk-dominant equilibrium is not always the case if the payoff parameters of the underlying game are such that the game is described by the right panel of Figure 2. In this case, the QRE surface is connected and the effect of the exploration policy on the learning process depends on the exploration-exploitation profile of both agents, and importantly, also on their synchronicity. The critical observation is that the middle (shaded) rectangle in the right panel of Figure 2 is transcended by the $x + y = x_{\text{mix}} + y_{\text{mix}}$ line which is the threshold between the attracting regions of the two corner equilibria, $(x, y) = (0, 0)$ and $(x, y) = (1, 1)$. Hence, when one or both agents increase their exploration rates to reach the middle rectangle, they cannot say in advance in which attracting region the learning process will find itself. In this case, after the players reduce their exploration rates back to their initial zero levels, the process may well converge to the equilibrium where it started from. These observations are further elaborated in the context of concrete examples next.

5. 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 [62], which starts with (relatively) high exploration that reduces linearly to zero and a *Cyclical Learning Rate with one cycle* (CLR-1) policy [63], 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]

2×2 *Coordination Games*

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).[9]

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 Section 5.

In Pareto Coordination, (a_2, a_2) is both the risk- and payoff-dominant equilibrium whereas in Stag Hunt, the payoff-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 left panel in Figure 2. Accordingly, the QRE surface 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 and explained in Figure 3 (and in a similar fashion in Figure D.11 in Appendix D). In both these cases, the risk-dominant equilibrium is the global maximizer of the potential function, see Lemma 4.2 and [27, 64].

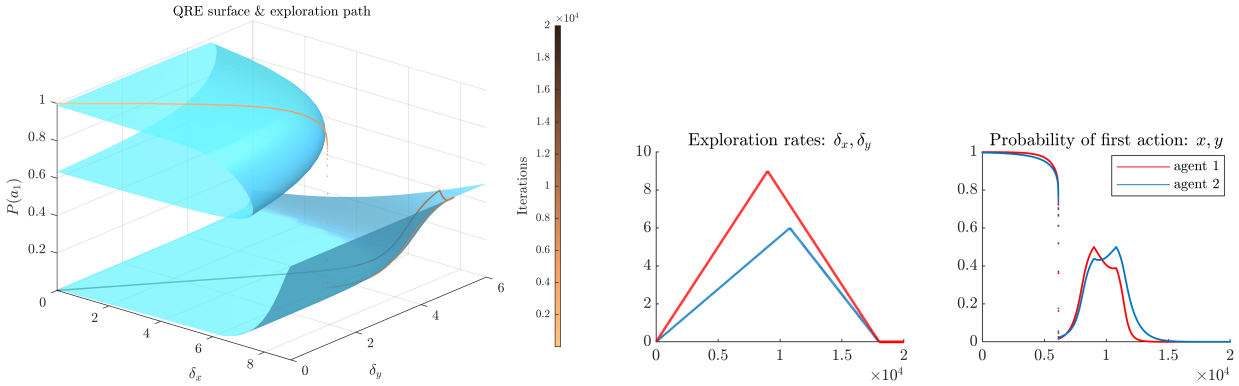

Figure 3: SQL in Stag Hunt. The left panel shows the QRE surface and the exploration path of agent 1 (light to dark line). The right panels show the CLR-1 exploration rates (left) and the probability of action 1 during the learning process for both agents. As agents increase exploration, their choice distributions undergo a *saddle-node bifurcation* (disconnected surface). This prompts a permanent transition from the vicinity of the payoff-dominant action profile, (a_1, a_1), in the upper component of the QRE surface to the (a_2, a_2) equilibrium when exploration reduces back to zero (right corner of the lower component). A top-down view of the QRE surface is provided in the left panel of Figure 5.

By contrast, if agents' interests are not perfectly aligned, then the outcome of the exploration process is ambiguous (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

[8]The findings are qualitatively equivalent for non-linear, e.g., quadratic, changes in the exploration rates in both policies and for more than one learning cycle in the CLR policy.

[9]In the numerical experiments, we have used the transformation in equation (A.4) in Lemma A.1 (see also Remark A.2 for a discussion) in Appendix A. This leads to a robust Euler discretization of the ODEs with batch updates in the space of Q-values that is used to implement the continuous time SQL dynamics of equation (1a).

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.[10] While not symmetric, this game satisfies equations (6) and is, thus, a coordination game and hence also a weighted potential game. This asymmetry generates an important difference in the geometry of the QRE surface. Namely, In this class of games, the location of the QRE is described by the right panel in Figure 2. The QRE surface is now 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 which denotes two different outcomes of the learning process under the same exploration policy for agent 1 but different exploration policies for agent 2. In Appendix D, we provide an exhaustive treatment of the possible outcomes under the ETE and CLR-1 exploration policies.

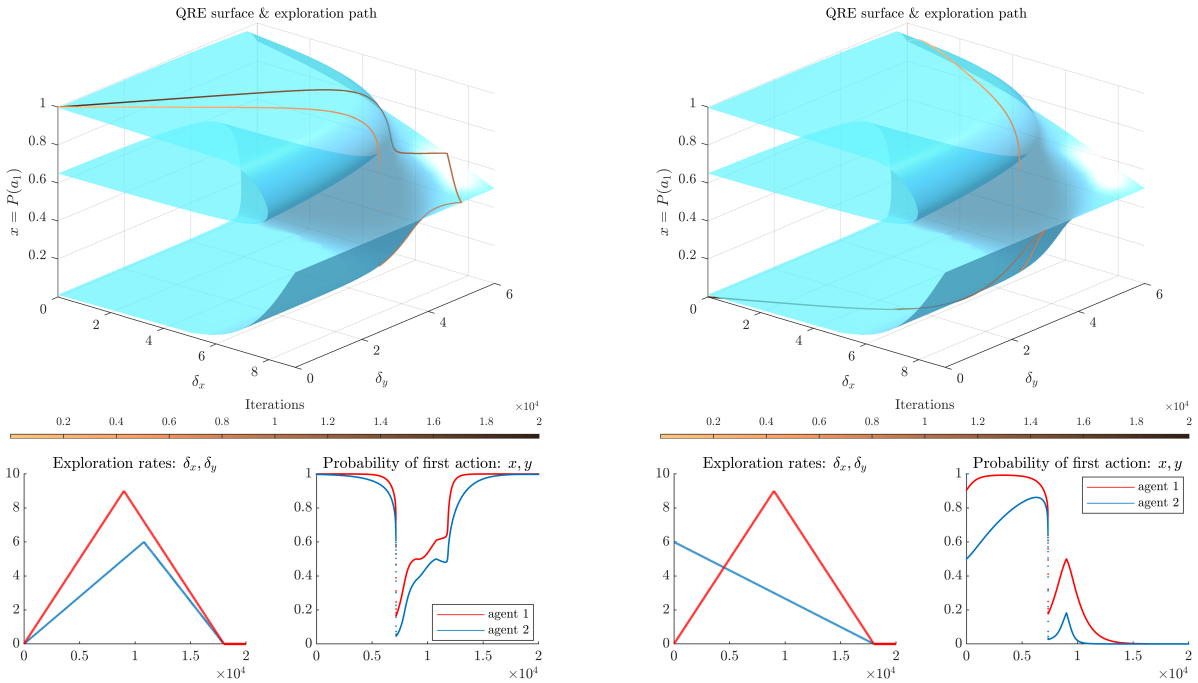

Figure 4: Exploration-Exploitation in Battle of the Sexes. In contrast to Stag Hunt, the QRE manifold has two branches of saddle-node bifurcation curves (consistent with the emergence of a co-dimension 2 cusp point) and the phase transition to the lower part of the QRE surface may not be permanent. These two cases are illustrated via the CLR-1 vs CLR-1 policies (left) and the CLR-1 vs ETE policies (right). A top-down view of the QRE surface is provided in the right panel of Figure 5.

Different types of bifurcation curves. Keeping Figure 2 in mind, we can now describe the bifurcation curves in the QRE surfaces of Figure 3 and Figure 4 as discussed in Remark 4.4. This is done in Figure 5 which shows the QRE surfaces (manifolds) from a bird's-eye perspective (top of the vertical axis) and highlights the bifurcation curves in the Stag Hunt and Battle of the Sexes games (Pareto Coordination is equivalent to Stag Hunt in this respect).

Games in Larger Dimensions

Pure coordination games. As a warm-up, we study the SQL dynamics in pure coordination games — coordination games with non-zero payoffs only on the diagonal — with action spaces of arbitrary size

[10] 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 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.

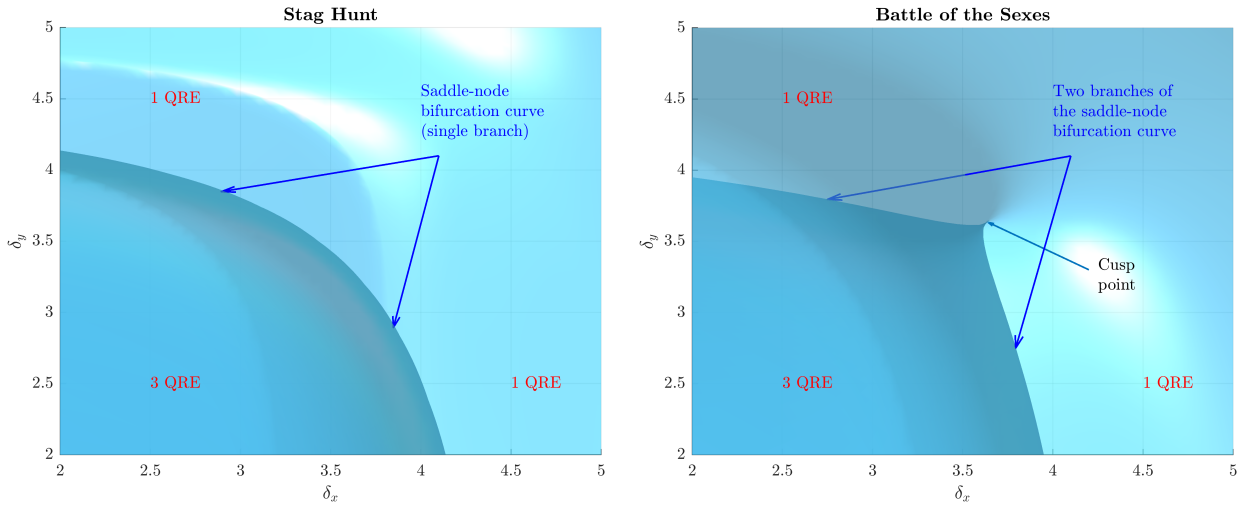

Figure 5: QRE surfaces in Stag Hunt (left panel) and Battle of the Sexes (right panel). The depicted surfaces are the same as in Figures 1, 3, 4 and D.11. The current perspective is from the top of the z-axis. The x-y axes have been set between 2 and 5 for a better focus on the bifurcation curve. 360° rotations of these surfaces are shown in the online supplementary material.

[65]. Specifically, we consider games $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k\in\mathcal{N}}\right)$ with two players $\mathcal{N} = \{1, 2\}$ and n actions $A_1 = A_2 = \{a_1, a_2, \ldots, a_n\}$ with diagonal payoff matrices

$$u_1 = \begin{array}{c} \\ a_1 \\ a_2 \\ \ldots \\ a_n \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & \ldots & a_n \end{array} \\ \begin{pmatrix} u_{11} & 0 & \ldots & 0 \\ 0 & u_{22} & \ldots & 0 \\ \ldots & \ldots & \ddots & \ldots \\ 0 & 0 & \ldots & u_{nn} \end{pmatrix} \end{array}, \quad u_2 = u_1^\top, \tag{8}$$

with $0 < u_{11} < u_{22} < \cdots < u_{nn}$. Any symmetric profile $(a_i, a_i), i = 1, 2, \ldots, n$ constitutes a pure NE of Γ. The equilibrium (a_n, a_n) is payoff-dominant and (by properly generalizing the notion of risk-dominance) also risk dominant.

For such games, Theorem 3.3 suggests that the SQL dynamics converge to a compact connected set of QRE of Γ. However, it is unclear whether this will be the payoff-dominant equilibrium or not and if not, how this outcome depends on the starting point and exploration policy of both agents. As we see in the following experiments, depending on the starting point and (especially) on the intensity and timing of the exploration performed by each agent, the SQL dynamics converge after the exploration phase to (typically) improved, yet possibly only locally optimal equilibria. Two instances are given in Figures 6 and 7.

Arbitrary potential games. We next turn to two-player potential games with arbitrary action spaces and payoffs. In Figure 8, we plot the SQL dynamics ($1e+20$ Q-value updates for each of $1e+03$ choice distribution updates) in a two-player potential game with $n = 10$ actions and potential, Φ, with randomly generated integer payoffs in $[1, 9]$ except for the payoff at the action profile (a_{10}, a_{10}) which has been deterministically set equal to 10 (absolute maximum). The exact matrix Φ is given in D. As in Figures 6 and 7, the first three panels show averages over a set of 10×10 different trajectories (starting points) one close to each pure action pair. Both agents use the same CLR-1 exploration policies in all cases (fourth panel).

As can be seen in Figure 8, the SQL dynamics rest at different local optima before exploration, converge to the uniform distribution when exploration rates reach their peak (first two panels) and then converge to the same (in this case, global) optimum when exploration is gradually reduced back to zero (horizontal line at the absolute maximum payoff of 10 and vanishing shaded region in the third panel). Figure 9 shows a second experiment in the same potential game, Φ, in which the SQL dynamics converge to a suboptimal outcome

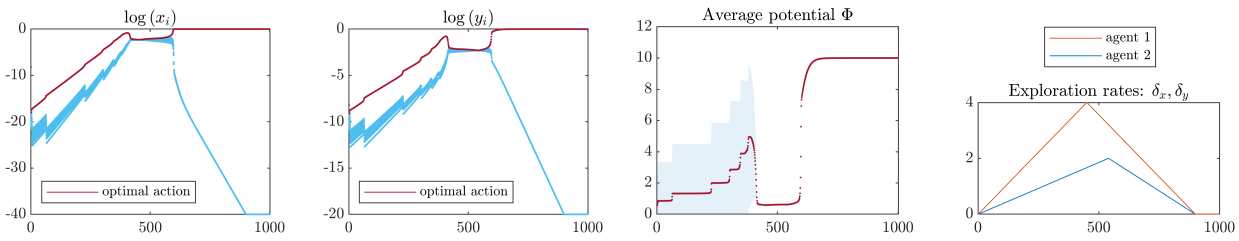

Figure 6: SQL dynamics ($1e+20$ Q-value updates for each of $1e+03$ choice distribution updates) in a pure coordination game with $n = 10$ actions and diagonal payoffs $u_{ii} = i$ for $i = 1, \ldots, 10$, cf. equation (8). The first two panels show the (log) choice distributions (with the optimal action in different color). The third 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 fourth panel shows the selected CLR-1 policies. Both agents perform CLR-1 exploration with different intensities and all trajectories of the SQL dynamics converge to the global optimum after the exploration phase regardless of the starting point (the standard deviation, depicted as the shaded region in the third panel, vanishes).

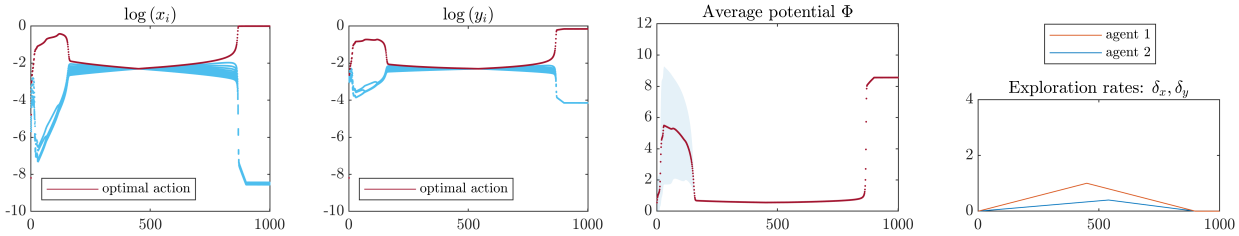

Figure 7: The SQL dynamics for the same experiment as in Figure 6 but with less intense exploration rate in the CRL-1 policies for both agents. In this case, the SQL dynamics converge to a suboptimal outcome after the exploration phase (the mean lies at 8.559 instead of the absolute maximum 10).

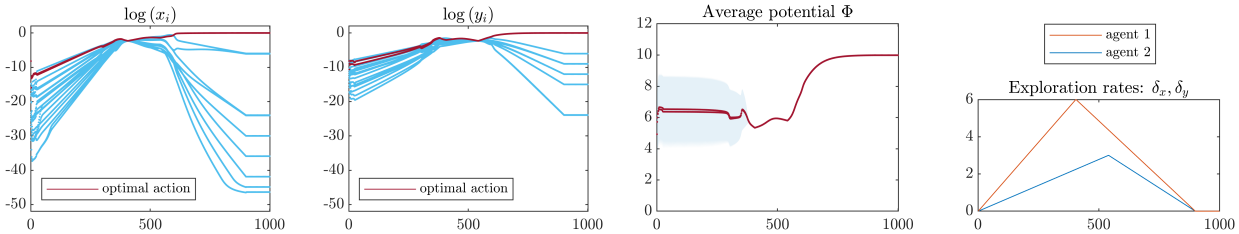

Figure 8: Exploration-Exploitation with the SQL dynamics in a potential game with $n = 10$ actions and a global maximum payoff of 10. The panels are as in Figure 6 and again show averages over a 10×10 grid of starting points, each of which is close to one pure action profile.

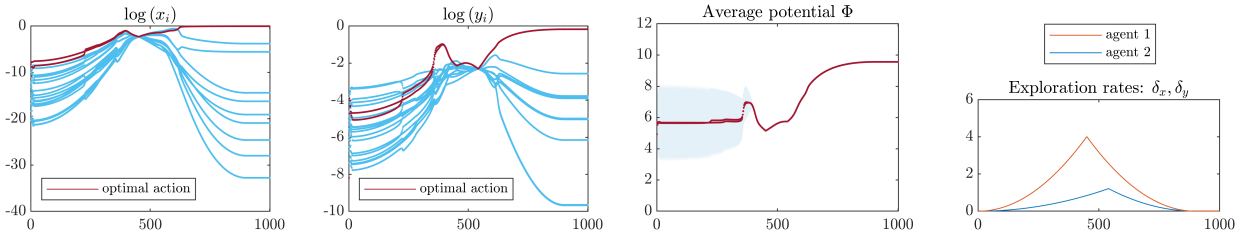

Figure 9: The game and panels are as in Figure 8. Agent 2 reduces their exploration rate in comparison to Figure 8 (fourth panel; also note the quadratic increase-decrease in the exploration rate). The SQL dynamics converge to a suboptimal state with potential value 9.561 (flat line in the right part of the third panel) for all trajectories (even for those initially close to the global optimum) as can be inferred by the vanishing shaded region (one standard deviation).

under different exploration policies (fourth panel). An interesting finding from these experiments is that the SQL dynamics converge to the same (local) optimum after exploration regardless of the starting point

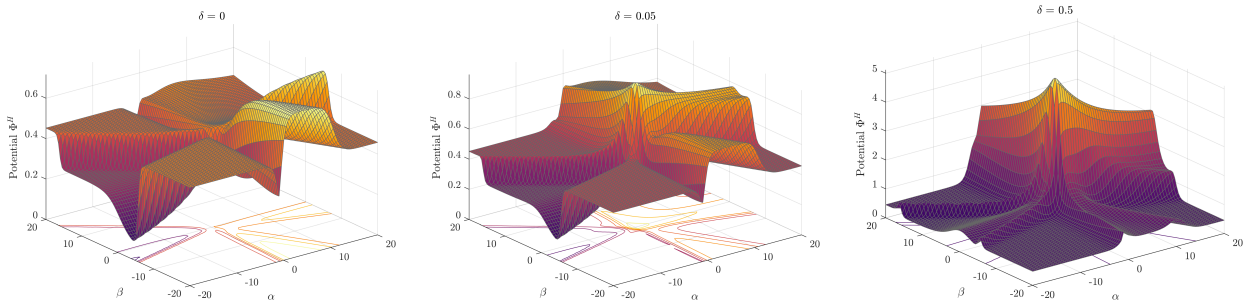

Figure 10: Snapshots of the modified potential Φ^H for different exploration rates in a symmetric 2-player potential game with random payoffs in $[0, 1]$. Unlike Figures 3 and 4, we now visualize the potential function instead of the QRE surface. Hence, we cannot reason about the bifurcation types. However, 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, i.e., the $(0, 0)$ point in the transformed coordinates.

(vanishing shaded regions of one standard deviation around the mean in the third panels). The reason is that the agents' choice distributions approximate the uniform distribution during *peak-exploration* in all cases. This means that *sufficient* exploration eliminates the initial condition effect. Subsequently, after exploration starts to decrease, the dynamics converge to the equilibrium that lies in the attraction region of the uniform distribution. Quantifying what *sufficient* means in different multi-agent setups and empirically testing the outcome of SQL in more complex (and potentially stateful) environments are very interesting directions for experimental work in this area.

Visualization of the modified potential. To better understand the above statement, it will be instructive to visualize the modified potential in equation (5) of Lemma 3.4. However, in arbitrary potential games (as the ones that we studied in the previous paragraph), the players' action spaces involve more than two actions and a direct visualization is not possible. To proceed, we instead adapt the two-dimensional projection technique of [42] which yields the insightful illustrations of Figure 10.

Given a potential game with potential Φ and $n, m \geq 2$ actions for agents 1 and 2, we first embed their choice distributions into $\mathbb{R}^{n+m-2}$ to remove the Simplex restrictions 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) and then 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). It is useful to observe that this transformation projects the uniform choice distribution to the origin. For simplicity, we keep the exploration ratio $\delta_k := \alpha_k/\beta_k$ equal to a common δ for both players.[11]

A visualization of the modified potential for different exploration rates in the case of a randomly generated 2-player potential game is given in Figure 10. As players increase their exploration rates, the SQL dynamics may converge to different QRE (local maxima) of these changing surfaces. However, similar to the low dimensional case, when exploration rates are sufficiently large, a single attracting QRE remains, visible as the global maximum in the vicinity of the (0,0) point which corresponds to the uniform choice distributions.

6. Conclusions

The interplay between exploration and exploitation is a fundamental concept in the learning algorithms that are currently used in multi-agent systems. However, its theoretical underpinnings are not well understood. In fact, numerous empirical papers report extremely opposing outcomes of exploration-exploitation algorithms that range from super-human performance in complex domains, e.g., mastering Starcraft, GO, or coordinating under communication constraints [66, 67, 68], to failure in the simplest possible settings, e.g., cycles and

[11] A detailed description of the routine is in Appendix D.

chaotic behavior in stateless games [17, 21]. In the current paper, we provide a theoretical framework that can be used to formally explain some of these phenomena. The main novelty of this framework is that it brings together three different theoretical areas: (evolutionary) game theory, online learning (and optimization) and, in particular, catastrophe theory. Catastrophe theory is a branch of mathematics that argues about abrupt phase transitions (bifurcations) in dynamical systems. Such transitions emerge in AI systems when learning agents explore their action spaces. As we show, the nature of these transitions can be explained by the underlying geometry of the game which in turn can be used to formally reason about the performance of these algorithms in practice. Among other results, this framework sheds light on the path-dependence of these learning algorithms, i.e., the dependence of their outcomes on the initial conditions and the history of play (hysteresis effects) and on the long-standing problem of equilibrium selection.

The techniques that are developed in this paper can be of interest to researchers that study the theoretical foundations of learning in multi-agent AI systems. Our results can also be used by empirically oriented researchers and reinforcement learning practitioners as benchmarks against experimental findings. Apart from establishing formal connections between multi-agent learning in game theoretic settings and catastrophe theory, our findings also provide a stepping stone to attack open questions that previously seemed intractable. In the current paper, we focus on multi-agent settings that involve an element of aligned incentives between agents (weighted potential games) and in which equilibrium selection has been thoroughly studied. This also provides a novel toolbox for the same question of equilibrium selection in competitive multi-agent settings. Such games have been predominantly studied in the abstraction of two-player zero-sum games (in which equilibrium selection is not an issue, as all equilibria have the same value) and not in full generality in which equilibrium selection is an important problem [69, 70]. The next natural challenges involve firstly, the application of these methods in arbitrary games that further close the gap between the two extremes of pure cooperation and pure competition, e.g., [71], and (ideally) also in stateful reinforcement learning setups [72, 73], and secondly, the analysis of alternative algorithms such as (stochastic) policy gradient with various exploration schemes, e.g., [74, 75, 56].

Acknowledgments

We are grateful to Sylvie Thiebaux, Editor-in-Chief, an anonymous associate editor and three anonymous referees for their helpful comments to improve this work. This research-project is supported in part by the National Research Foundation, Singapore under NRF 2018 Fellowship NRF-NRFF2018-07, AI Singapore Program (AISG Award No: AISG2-RP-2020-016), NRF2019-NRF-ANR095 ALIAS grant, AME Programmatic Fund (Grant No. A20H6b0151) from the Agency for Science, Technology and Research (A*STAR) and grant PIE-SGP-AI-2018-01.

References

[1] C. Claus, C. Boutilier, The Dynamics of Reinforcement Learning in Cooperative Multiagent Systems, in: Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence, AAAI '98/IAAI '98, American Association for Artificial Intelligence, USA, 1998, p. 746–752. doi:10.5555/295240.295800.

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

[3] C. F. Camerer, T.-H. Ho, Experience-weighted Attraction Learning in Normal Form Games, Econometrica 67 (4) (1999) 827–874. doi:10.1111/1468-0262.00054.

[4] T.-H. Ho, C. F. Camerer, J.-K. Chong, Self-tuning experience weighted attraction learning in games, Journal of Economic Theory 133 (1) (2007) 177–198. doi:10.1016/j.jet.2005.12.008.

[5] C. F. Camerer, T.-H. Ho, Experience-Weighted Attraction Learning in Coordination Games: Probability Rules, Heterogeneity, and Time-Variation, Journal of Mathematical Psychology 42 (2) (1998) 305–326. doi:10.1006/jmps.1998.1217.

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

[7] M. Kaisers, K. Tuyls, S. Parsons, F. Thuijsman, 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, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2009, p. 1255–1256. doi:10.5555/1558109.1558239.

[8] A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, Z. D. Guo, C. Blundell, Agent57: Outperforming the Atari Human Benchmark, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 507–517.
URL https://proceedings.mlr.press/v119/badia20a.html

[9] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Perolat, D. Silver, T. Graepel, A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning, in: I. G. et. al (Ed.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4190–4203.

[10] S. Omidshafiei, C. Papadimitriou, G. Piliouras, K. Tuyls, M. Rowland, J.-B. Lespiau, W. M. Czarnecki, M. Lanctot, J. Perolat, R. Munos, α-Rank: Multi-Agent Evaluation by Evolution, Scientific Reports 9 (1) (2019) 9937. doi:10.1038/s41598-019-45619-9.

[11] M. Rowland, S. Omidshafiei, K. Tuyls, J. Perolat, M. Valko, G. Piliouras, R. Munos, Multiagent Evaluation under Incomplete Information, in: H. W. et. al (Ed.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 12291–12303.

[12] M. Kaisers, K. Tuyls, Frequency adjusted multi-agent q-learning, in: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: Volume 1, AAMAS '10, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2010, p. 309–316. doi:10.5555/1838206.1838250.

[13] M. Kaisers, K. Tuyls, 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, AAAI Press, 2011, p. 36–42. doi:10.5555/2908738.2908744.

[14] Michael Wunder and Michael Littman and Monica Babes, Classes of multiagent Q-learning dynamics with epsilon-greedy exploration, in: J. Fürnkranz, T. Joachims (Eds.), Proceedings of the 27th International Conference on Machine Learning (ICML-10), Omnipress, Haifa, Israel, 2010, pp. 1167–1174.
URL http://www.icml2010.org/papers/191.pdf

[15] D. Bloembergen, K. Tuyls, D. Hennes, M. Kaisers, Evolutionary Dynamics of Multi-Agent Learning: A Survey, J. Artif. Int. Res. 53 (1) (2015) 659–697. doi:10.1613/jair.4818.

[16] T. Klos, G. J. Van Ahee, K. Tuyls, Evolutionary Dynamics of Regret Minimization, in: J. L. Balcázar, F. Bonchi, A. Gionis, M. Sebag (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010, pp. 82–96. doi:10.1007/978-3-642-15883-4_6.

[17] Y. Sato, J. P. Crutchfield, Coupled replicator equations for the dynamics of learning in multiagent systems, Phys. Rev. E 67 (2003) 015206. doi:10.1103/PhysRevE.67.015206.

[18] Y. Sato, E. Akiyama, J. P. Crutchfield, Stability and diversity in collective adaptation, Physica D: Nonlinear Phenomena 210 (1) (2005) 21–57. doi:10.1016/j.physd.2005.06.031.

[19] K. Tuyls, G. Weiss, Multiagent learning: Basics, challenges, and prospects, AI Magazine 33 (3) (2012) 41. doi:10.1609/aimag.v33i3.2426.

[20] D. Balduzzi, W. M. Czarnecki, T. Anthony, I. Gemp, E. Hughes, J. Leibo, G. Piliouras, T. Graepel, Smooth markets: A basic mechanism for organizing gradient-based learners, in: International Conference on Learning Representations, 2020, pp. 1–18.

[21] P. Mertikopoulos, C. Papadimitriou, G. Piliouras, Cycles in adversarial regularized learning, in: Proceedings of the 2018 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2018, pp. 2703–2717. doi:10.1137/1.9781611975031.172.

[22] E. Mazumdar, L. J. Ratliff, S. S. Sastry, On Gradient-Based Learning in Continuous Games, SIAM Journal on Mathematics of Data Science 2 (1) (2020) 103–131. doi:10.1137/18M1231298.

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

[24] G. Palaiopanos, I. Panageas, G. Piliouras, Multiplicative Weights Update with Constant Step-Size in Congestion Games: Convergence, Limit Cycles and Chaos, in: Advances in Neural Information Processing Systems, NIPS'17, 2017, p. 5874–5884. doi:10.5555/3295222.3295337.

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

[26] A. Kianercy, A. Galstyan, Dynamics of Boltzmann Q learning in two-player two-action games, Phys. Rev. E 85 (2012) 041145. doi:10.1103/PhysRevE.85.041145.

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

[28] L. P. Kaelbling, M. L. Littman, A. W. Moore, Reinforcement learning: A survey, Journal of Artificial Intelligence Research 4 (1) (1996) 237–285. doi:10.1613/jair.301.

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

[30] K. Tuyls, K. Verbeeck, T. Lenaerts, 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, 2003, pp. 693–700. doi:10.1145/860575.860687.

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

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

[33] N. Cesa-Bianchi, G. Lugosi, Prediction, learning, and games, Cambridge university press, 2006.

[34] J. Kwoon, P. Mertikopoulos, A continuous-time approach to online optimization, Journal of Dynamics & Games 4 (2) (2017) 125–148. doi:10.3934/jdg.2017008.
[35] Leslie, D. S. and Collins, E. J., Individual Q-Learning in Normal Form Games, SIAM Journal on Control and Optimization 44 (2) (2005) 495–514. doi:10.1137/S0363012903437976.
[36] P. Coucheney, B. Gaujal, P. Mertikopoulos, Penalty-Regulated Dynamics and Robust Learning Procedures in Games, Mathematics of Operations Research 40 (3) (2015) 611–633. doi:10.1287/moor.2014.0687.
[37] B. Gao, L. Pavel, On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning (2017). arXiv:1704.00805.
[38] K. Tuyls, P. J. T. Hoen, B. Vanschoenwinkel, An Evolutionary Dynamical Analysis of Multi-Agent Learning in Iterated Games, Autonomous Agents and Multi-Agent Systems 12 (1) (2006) 115–153. doi:10.1007/s10458-005-3783-9.
[39] M. Göcke, Various Concepts of Hysteresis Applied in Economics, Journal of Economic Surveys 16 (2) (2002) 167–188. doi:10.1111/1467-6419.00163.
[40] J. Romero, The effect of hysteresis on equilibrium selection in coordination games, Journal of Economic Behavior & Organization 111 (2015) 88–105. doi:10.1016/j.jebo.2014.12.029.
[41] S. H. Strogatz, Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering, 1st Edition, Westview Press (Studies in nonlinearity collection), Cambridge, MA, USA, 2000.
[42] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the Loss Landscape of Neural Nets, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018, pp. 6389–6399.
[43] A. Kianercy, R. Veltri, K. J. Pienta, Critical transitions in a game theoretic model of tumour metabolism, Interface Focus 4 (4) (2014) 20140014. doi:10.1098/rsfs.2014.0014.
[44] S. Leonardos, I. Sakos, C. Courcoubetis, G. Piliouras, Catastrophe by Design in Population Games: Destabilizing Wasteful Locked-In Technologies, in: Web and Internet Economics: 16th International Conference, WINE 2020, Beijing, China, December 7–11, 2020, Proceedings, Vol. 12495, Springer, 2020, p. 473, extended abstract. doi:10.1007/978-3-030-64946-3.
[45] K. Tuyls, S. Parsons, What evolutionary game theory tells us about multiagent learning, Artificial Intelligence 171 (7) (2007) 406–416, foundations of Multi-Agent Learning. doi:10.1016/j.artint.2007.01.004.
[46] T. Galla, J. D. Farmer, Complex dynamics in learning complicated games, Proceedings of the National Academy of Sciences 110 (4) (2013) 1232–1236. doi:10.1073/pnas.1109672110.
[47] S. Leonardos, G. Piliouras, Exploration-Exploitation in Multi-Agent Learning: Catastrophe Theory Meets Game Theory, Proceedings of the AAAI Conference on Artificial Intelligence 35 (13) (2021) 11263–11271.
URL https://ojs.aaai.org/index.php/AAAI/article/view/17343
[48] J. Lorraine, J. Parker-Holder, P. Vicol, A. Pacchiano, L. Metz, T. Kachman, J. Foerster, Using Bifurcations for Diversity in Differentiable Games, ICML Workshop: Beyond first-order methods in ML systems (2021).
[49] J. Parker-Holder, L. Metz, C. Resnick, H. Hu, A. Lerer, A. Letcher, A. Peysakhovich, A. Pacchiano, J. Foerster, Ridge Rider: Finding Diverse Solutions by Following Eigenvectors of the Hessian (2020). arXiv:2011.06505.
[50] O. Ben-Porat, M. Tennenholtz, A Game-Theoretic Approach to Recommendation Systems with Strategic Content Providers, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018.
[51] I. Panageas, G. Piliouras, 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, 2016, p. 703–720. doi:10.1145/2940716.2940784.
[52] B. Swenson, R. Murray, S. Kar, On Best-Response Dynamics in Potential Games, SIAM Journal on Control and Optimization 56 (4) (2018) 2734–2767. doi:10.1137/17M1139461.
[53] J. Perolat, R. Munos, J.-B. Lespiau, S. Omidshafiei, M. Rowland, P. Ortega, N. Burch, T. Anthony, D. Balduzzi, B. De Vylder, G. Piliouras, M. Lanctot, K. Tuyls, From Poincaré Recurrence to Convergence in Imperfect Information Games: Finding Equilibrium via Regularization, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, Vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 8525–8535.
URL https://proceedings.mlr.press/v139/perolat21a.html
[54] S. Valcarcel Macua, J. Zazo, S. Zazo, Learning Parametric Closed-Loop Policies for Markov Potential Games, in: International Conference on Learning Representations, 2020.
[55] D. H. Mguni, Y. Wu, Y. Du, Y. Yang, Z. Wang, M. Li, Y. Wen, J. Jennings, J. Wang, Learning in Nonzero-Sum Stochastic Games with Potentials, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, Vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 7688–7699.
URL https://proceedings.mlr.press/v139/mguni21a.html
[56] S. Leonardos, W. Overman, I. Panageas, G. Piliouras, Global Convergence of Multi-Agent Policy Gradient in Markov Potential Games (2021). arXiv:2106.01969.
[57] R. Zhang, Z. Ren, N. Li, Gradient play in stochastic games: stationary points, convergence, and sample complexity (2021). arXiv:2106.00198.
[58] C. J. Watkins, P. Dayan, Technical Note: Q-Learning, Machine Learning 8 (3) (1992) 279–292. doi:10.1023/A:1022676722315.
[59] T. Roughgarden, Intrinsic Robustness of the Price of Anarchy, J. ACM 62 (5). doi:10.1145/2806883.
[60] R. Kleinberg, G. Piliouras, E. Tardos, 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, 2009, pp. 533–542. doi:10.1145/1536414.1536487.
[61] J. Harsanyi, R. Selten, A General Theory of Equilibrium Selection in Games, The MIT Press, Massachusetts, USA, 1988.

[62] Y. Bai, C. Jin, Provable Self-Play Algorithms for Competitive Reinforcement Learning, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, Vol. 119 of Proceedings of Machine Learning Research, PMLR, 2020, pp. 551–560.
URL https://proceedings.mlr.press/v119/bai20a.html

[63] L. N. Smith, N. Topin, Super-convergence: very fast training of neural networks using large learning rates, in: T. Pham (Ed.), Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, Vol. 11006, International Society for Optics and Photonics, SPIE, 2019, pp. 369 – 386. doi:10.1117/12.2520589.

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

[65] Y. Kim, Equilibrium Selection in n-Person Coordination Games, Games and Economic Behavior 15 (2) (1996) 203–227. doi:10.1006/game.1996.0066.

[66] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (7587) (2016) 484–489. doi:10.1038/nature16961.

[67] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver, Grandmaster level in StarCraft II using multi-agent reinforcement learning, Nature 575 (7782) (2019) 350–354. doi:10.1038/s41586-019-1724-z.

[68] A. Mahajan, T. Rashid, M. Samvelyan, S. Whiteson, MAVEN: Multi-Agent Variational Exploration, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.

[69] Y. Cai, O. Candogan, C. Daskalakis, C. Papadimitriou, Zero-Sum Polymatrix Games: A Generalization of Minmax, Mathematics of Operations Research 41 (2) (2016) 648–655. doi:10.1287/moor.2015.0745.

[70] S. Leonardos, G. Piliouras, K. Spendlove, Exploration-Exploitation in Multi-Agent Competition: Convergence with Bounded Rationality, in: Advances in Neural Information Processing Systems, Vol. 35, Curran Associates, Inc., 2021.

[71] S.-H. Hwang, L. Rey-Bellet, Strategic decompositions of normal form games: Zero-sum games and potential games, Games and Economic Behavior 122 (2020) 370–390. doi:10.1016/j.geb.2020.05.003.

[72] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, T. Graepel, Open Problems in Cooperative AI, arXiv e-printsarXiv:2012.08630.

[73] A. Dafoe, Y. Bachrach, G. Hadfield, E. Horvitz, K. Larson, T. Graepel, Cooperative ai: machines must learn to find common ground., Nature 7857 (2021) 33–36. doi:10.1038/d41586-021-01170-0.

[74] A. Agarwal, S. M. Kakade, J. D. Lee, G. Mahajan, Optimality and Approximation with Policy Gradient Methods in Markov Decision Processes, in: J. Abernethy, S. Agarwal (Eds.), Proceedings of 33rd Conference on Learning Theory, Vol. 125 of PMLR, 2020, pp. 64–66.
URL http://proceedings.mlr.press/v125/agarwal20a.html

[75] C. Daskalakis, D. Foster, N. Golowich, Independent Policy Gradient Methods for Competitive Reinforcement Learning, in: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, Vol. 33, Curran Associates, Inc., 2020, pp. 5527–5540.

[76] D. Monderer, L. S. Shapley, Potential Games, Games and Economic Behavior 14 (1) (1996) 124–143. doi:10.1006/game.1996.0044.

A. Derivation of SQL Dynamics

The mathematical connection between Q-learning and the dynamics in (1) via a smoothening process (continuous time limit) at which the Q-values are interpreted as Boltzmann probabilities for the action selection can be found in [30, 17, 18, 23, 26] among others. To make the paper self-contained, we repeat here the main arguments. Each agent $k \in \mathcal{N}$ keeps track of the past performance of their actions $i \in A_k$ via the Q-learning update rule

$$Q_{ki}(n+1) = Q_{ki}(n) + \alpha_k \left[r_{ki}(n) - Q_{ki}(n) \right], \; i \in A_k \tag{A.1}$$

where $n \in \mathbb{N}$ denotes the time (in discrete steps) and $\alpha_k \in [0,1]$ the learning rate or memory decay of agent k, cf. [17, 26]. Here $Q_{ki}(n)$ is called the *memory* of agent k about the performance of action $i \in A_k$ up to time step $n \in \mathbb{N}$. By slightly overloading notation, we wrote $r_{ki}(n)$ to denote the reward of agent k from action i at time step n. In general, the $r_{ki}(n)$'s depend on the actions of other agents and may also change over time (i.e., time-varying rewards for each action profile). Here, we will restrict attention to rewards that remain constant over time, see e.g., [17]. After introducing the relevant notation, we will write $r_{ki}(x)$ (as

in the main text) to denote the reward of agent k for selecting action i when the choice distribution x is selected by all agents. Agent $k \in \mathcal{N}$ updates their actions (choice distributions) according to a Boltzmann type distribution, with

$$x_{ki}(n) = \frac{\exp(\beta_k Q_{ki}(n))}{\sum_{j \in A_k} \exp(\beta_k Q_{kj}(n))}, \quad \text{for each } i \in A_k, \tag{A.2}$$

where $\beta_k \in [0, +\infty)$ denotes agent k's learning sensitivity or adaptation, i.e., how much the choice distribution is affected by the past performance. Higher values of β_k indicate a higher exploitation rate, i.e., proclivity of the agent towards the best performing action, whereas values of β_k close to 0 lead to higher exploration or randomization among the agent's available choices in A_k. Combining the above equations, one obtains the recursive equation of the agent's choice distribution

$$x_{ki}(n+1) = \frac{x_{ki}(n) \exp(\beta_k (Q_{ki}(n+1) - Q_{ki}(n)))}{\sum_{j \in A_k} x_{kj}(n) \exp(\beta_k (Q_{kj}(n+1) - Q_{kj}(n)))},$$

for each $i \in A_k$. In practice, agents perform a large number of actions (updates of Q-values) for each choice-distribution (update of Boltzmann selection probabilities). This motivates to consider a continuous time version of the learning process for each agent $k \in \mathcal{N}$ which results in the following update rules for both the memories Q_{ki}

$$\dot{Q}_{ki} = \alpha_k [r_{ki}(x) - Q_{ki}],$$

where, as in the main text, $r_{ki}(x)$ denotes k's reward for selecting pure action $i \in A_k$ at state $x = (x_k, x_{-k}) \in X$, and the selection probabilities x_{ki} of each action $i \in A_k$

$$\dot{x}_{ki} = \beta_k x_{ki} \left(\dot{Q}_{ki} - \sum_{j \in A_k} \dot{Q}_{kj} \right).$$

Combining the last two equations under the assumptions that agents' rewards remain constant (for each action profile) over time, and that their choice distributions are independently distributed, yields the dynamics in equation (1), namely

$$\frac{\dot{x}_{ki}}{x_{ki}} = \beta_k \left(r_{ki}(x) - \sum_{j \in A_k} x_{kj} r_{kj}(x) \right) - \alpha_k \left(\ln x_{ki} - \sum_{j \in A_k} x_{kj} \ln x_{kj} \right)$$

for each $i \in A_k$ and for each agent $k \in \mathcal{N}$.

Timescale of updates.. In the above dynamics, there are three relevant time scales or rates from the perspective of agent $k \in \mathcal{N}$: (1) the rate of change of the environment, i.e., of $x_{-k} \in X_{-k}$, (2) the rate of adaptation, i.e., the rate of change of $x_k \in X_k$, which is captured by β_k and (3) the rate of memory dissipation or exploration of the action space which is captured by α_k. Typically, cf. [18], the rate of change of the environment — captured by the choices $x_{-k} \in Y$ of all other agents — is slower than the changes in the choices x_k of agent k — updates of equation (A.2) — which in turn, are slower than the rate of interaction — memory updates or equivalently updates of equation (A.1) — with the environment. In other words, the agent fixes a choice distribution $x_k \in X_k$ and interacts multiple times with the other agents (environment) before updating their choice distribution according to the above scheme.

To determine the interplay between updates of Q-values and choice distributions, let n and $n+1$ denote the time points of two successive updates of the choice distribution $x_k(n)$ and $x_k(n+1)$ of agent $k \in \mathcal{N}$. For any choice $i \in A_k$, let $n_i := x_{ki}(n) M$ denote the (on average or expectation) number of times that agent k selects action $i \in A_k$ given choice distribution $x_k(n)$, where $M \gg 0$ is the large number of interactions of the agent with its environment under the fixed choice distribution $x_k(n)$. Then, using the index $t \in [0, n_i]$ such that $n = 0 < 1 < \cdots < t < \cdots < n_i = n+1$ to denote the timing of the interactions (the actual timing is irrelevant; only the number of interactions matters), equation (A.1) yields the following recursion.

Lemma A.1. *Let $x_{-k}(n) \in X_{-k}$ and $x_k(n) \in X_k$ denote the choice distributions of agent $k \in \mathcal{N}$ and all other agents in $\mathcal{N}$ other than k at time point $n \geq 0$. Let $n+1$ denote the time point of the next update of the choice distribution of agent k. Then, if $n = 0 < 1 < \cdots < t < \cdots < n_i = n+1$ denote the time points of $M \gg 0$ interactions of agent k with its environment between time points n and $n+1$, the updates in the Q-values of agent k are governed in expectation by the equation*

$$Q_{ki}(n+1) = (1-\alpha_k)^{n_i} Q_{ki}(n) + \alpha_k \sum_{t=0}^{n_i} (1-\alpha_k)^t r_i(n_i - t), \tag{A.3}$$

with $n_i := M \cdot x_{ki}(n)$. If $x_{-k}(n)$ remains constant between two successive updates of the choice distribution of agent k at time points n and $n+1$, then $r_i(n_i - t) = r_i$ for any $t \in [0, n_i]$ and equation (A.3) *simplifies to*

$$Q_{ki}(n+1) = (1-\alpha_k)^{n_i} Q_{ki}(n) + \left(1 - (1-\alpha_k)^{n_i+1}\right) r_i. \tag{A.4}$$

Proof. Assuming that agent k interacts M times with their environment for each (fixed) choice distribution $x_k(n)$, where M is a large positive integer, the expected number of times that agent k selects choice $i \in A_k$ is equal to $n_i = M \cdot x_{ki}(n)$. Hence, equation (A.1) yields

$$\begin{aligned}
Q_{ki}(n+1) &= \alpha_k r_i(n_i) + (1-\alpha_k) Q_{ki}(n_i) \\
&= \alpha_k r_i(n_i) + (1-\alpha_k)\left[\alpha_k r_i(n_i - 1) + (1-\alpha_k) Q_{ki}(n_i - 1)\right] \\
&= \alpha_k r_i(n_i) + \alpha_k(1-\alpha_k) r_i(n_i - 1) + (1-\alpha_k)^2\left[\alpha_k r_i(n_i - 2) + (1-\alpha_k) Q_{ki}(n_i - 2)\right] \\
&= \ldots \\
&= (1-\alpha_k)^{n_i} Q_{ki}(n) + \alpha_k \sum_{t=0}^{n_i} (1-\alpha_k)^t r_i(n_i - t),
\end{aligned}$$

with t indexing the time points at which agent k interacts with their environment between the two successive updates of its choice distribution $x_k(n)$ and $x_k(n+1)$. Finally, assuming that the choice distribution of all other agents remains constant between time points n and $n+1$, it holds that $r_i(n_i - t) = r_i$ for any $t \in [0, n_i]$ and equation simplifies to

$$Q_{ki}(n+1) = (1-\alpha_k)^{n_i} Q_{ki}(n) + \left(1 - (1-\alpha_k)^{n_i+1}\right) r_i,$$

for $\alpha_k \in [0, 1)$, by the formula of the partial sums of the geometric series,

$$\sum_{t=0}^{n_i} (1-\alpha_k)^t = \frac{1}{\alpha_k}\left(1 - (1-\alpha_k)^{n_i+1}\right). \qquad \square$$

Remark A.2. Equations (A.4) and (A.3) can now be used to compute the memory updates of agent k between two successive updates of agent k's choice distribution that occur via equation (A.2). Both equations hold in expectation (or on average) since the acting agent chooses their choice according to the choice distribution $x_k(n)$ each of the M times that they interact with their environment. However, under the working assumption that $M \gg 0$, i.e., that M is a very large number, the law of large numbers suggests that the expected value $n_i = M \cdot x_{ki}(n)$ is close to the actual number of times that agent k uses action i while interacting with other agents under the (fixed) choice distribution $x_k(n)$. This approach has been used in all experiments resulting in a fast (scalable) implementation of the Q-learning process. Our code can be found in the online supplementary material or here: github repository.

Finally, equation (A.3) in Lemma A.1 suggests that for the extreme values of $\alpha \in [0, 1)$, the updates of the Q-values are

$$Q_i(n+1) = \begin{cases} r_i(n_i), & \text{for } \alpha \to 1, \\ Q_i(n) + \sum_{t=0}^{n_i} r_i(t), & \text{for } \alpha = 0. \end{cases}$$

This recovers the intuition that when $\alpha = 0$, the agent has *perfect memory*, whereas for $\alpha \to 1$, the agent is completely oblivious of past rewards.

B. Omitted Materials of Section 3

Proof of Lemma 3.1. Using the definition of $r_{ki}^H(x)$, we have that

$$r_{ki}^H(x) - \langle x_k, r_k^H(x)\rangle = \beta_k\left(r_{ki}(x) - \langle x_k, r_k(x)\rangle\right) - \alpha_k\left(\ln x_{ki} + 1 - \langle x_k, \ln x_k\rangle - \langle x_k, 1\rangle\right).$$

Since $\langle x_k, 1\rangle = \sum_{j\in A_k} x_{kj} = 1$, the right side reduces to the expression in (1b). □

To compare the performance of two different choice distributions $p, x \in X$, we will use the notion of *KL-divergence*, $D(p\,\|\,x)$, which is defined by $D(p\,\|\,x) := -\sum_{i=1}^n p_i \ln\left(\frac{x_i}{p_i}\right)$. While not a metric (due to its asymmetry), the KL-divergence will be sufficient for our purposes as measure of how one probability distribution (here x) is different from a second, reference probability distribution (here p), cf. [70]. To prove Theorem 3.2, we will need the following important property that follows immediately from the definition of the KL-divergence.

Lemma B.1. *Let $p = (p_1, p_2, \ldots, p_n) \in X$ be fixed and let $x = (x_1, x_2, \ldots, x_n) \in X$ denote an arbitrary probability distribution (mixed strategy or population state) in X. Also let $\langle\cdot,\cdot\rangle$ denote the inner product in $\mathbb{R}^n$. Then,*

$$\langle p, \ln p\rangle \geq \langle p, \ln x\rangle, \quad \text{for any } x \in X,$$

with equality if and only if $x = p$.

Proof. By linearity of the inner product in $\mathbb{R}^n$, and non-negativity of the KL-divergence (which follows from Gibbs' inequality and the fact that $\ln x \leq x - 1$ for all $x > 0$), we have that

$$\langle p, \ln p - \ln x\rangle = \sum_{i=1}^n p_i\left(\ln p_i - \ln x_i\right) = -\sum_{i=1}^n p_i \ln\left(\frac{x_i}{p_i}\right) = D(p\|x) \geq 0$$

with equality if and only if $x = p$. □

Proof of Theorem 3.2. Consider an agent $k \in \mathcal{N}$ and let $p_k \in X_k$ denote the agent's optimal strategy in hindsight, i.e., $p_k = \arg\max_{x_k'\in X_k} \int_0^T u_k^H(x_k'; x_{-k}(t))\,dt$. Let also $x_k(t)$ denote the sequence of play generated by the dynamics in (3) for an arbitrary initial condition $x_k(0) \in X_k$. Then, by taking the time derivative of the term $\sum_{i\in A_k} p_{ki}\ln(x_{ki}(t))$, we obtain

$$\begin{aligned}
\frac{d}{dt}\langle p_k, \ln x_k(t)\rangle &= \sum_{i\in A_k} p_{ki}\cdot\frac{\dot{x}_{ki}}{x_{ki}} = \sum_{i\in A_k} p_{ki}\left[r_{ki}^H(x) - \langle x_k, r_k^H(x)\rangle\right] \\
&= \sum_{i\in A_k} p_{ik}\left[\beta_k\left(r_{ki} - \langle x_k, r_k(x)\rangle\right) - \alpha_k\left(\ln x_{ki} - \langle x_k, \ln x_k\rangle\right)\right] \\
&= \beta_k\langle p_k, r_k(x)\rangle - \alpha_k\langle p_k, \ln x_k\rangle - \langle x_k, \beta_k r_k(x) - \alpha_k \ln x_k\rangle \\
&\geq \beta_k\langle p_k, r_k(x)\rangle - \alpha_k\langle p_k, \ln p_k\rangle - u_k^H(x_k; x_{-k}) \\
&= u_k^H(p_k; x_{-k}(t)) - u_k^H(x_k(t); x_{-k}(t))
\end{aligned}$$

where the inequality has been established in Lemma B.1. Integrating both sides of the previous inequality from timepoint 0 to $T > 0$, and using the definition of $R_k(T)$ in equation (4) we get

$$\sum_{i\in A_k} p_{ki}\left(\ln x_{ki}(T) - \ln x_{ki}(0)\right) \geq \int_0^T u_k^H(p_k; x_{-k}(t)) - u_k^H(x_k(t); x_{-k}(t))\,dt = R_k^H(T).$$

Since $x_{ki}(T) \in [0,1]$ for all $i = 1, 2, \ldots, n$, and since $\sum_{i\in A_k} p_{ki} = 1$, the left side is upper bounded by $-\sum_{i\in A_k} p_{ki}\ln(x_{ki}(0))$ which is a constant with respect to T. Hence,

$$\limsup_{T\to\infty} R_k^H(T) \leq -\sum_{i\in A_k} p_{ki}\ln x_{ki}(0),$$

which concludes the proof. □

Proof of Lemma 3.4. Before proceeding to the proof of the statement of Lemma 3.4, observe that the multiplicative constants $\beta_k, k \in \mathcal{N}$ in equation (1) are essentially equivalent to a rescaling of the payoffs of agent k in the underlying game. Thus, in a weighted potential game Γ with vector of positive weights $w = (w_k)_{k\in\mathcal{N}}$ satisfying

$$u_k(i, a_{-k}) - u_k(j, a_{-k}) = w_k(\phi(i, a_{-k}) - \phi(j, a_{-k})),$$

for all $i \neq j \in A_k, a_{-k} \in A_{-k}$, for all agents $k \in \mathcal{N}$, one may rescale the parameters β_k to $\beta'_k = \beta_k/w_k$ for all $k \in \mathcal{N}$ and consider the resulting *exact* potential game instead. This implies that we can adjust the techniques of [60, 36] to prove the statement of Lemma 3.4.

In particular, to see that $\Phi^H(x)$ defines a potential for Γ^H, consider the partial derivatives of $\Phi^H(x)$ at a point $x = (x_{ki})_{k\in N, i\in A_k} \in X$,

$$\begin{aligned}\frac{\partial}{\partial x_{ki}}\Phi^H(x) &= \frac{\partial}{\partial x_{ki}}\Phi(x) - \frac{\alpha_k}{\beta_k}(\ln x_{ki} + 1)\\ &= \frac{\partial}{\partial x_{ki}}u_k(x) - \frac{\alpha_k}{\beta_k}(\ln x_{ki} + 1)\\ &= r_{ki}(x) - \alpha_k(\ln x_{ki} + 1) = \frac{1}{\beta_k}r^H_{ki}(x)\end{aligned}$$

where $\frac{\partial}{\partial x_{ki}}u_k(x) = r_{ki}(x)$ by definition and $\frac{\partial}{\partial x_{ki}}\Phi(x) = \frac{\partial}{\partial x_{ki}}u_k(x)$ since $\Phi(x)$ is a potential function for the unmodified game with utilities $u_k(x), k \in \mathcal{N}$. Hence, Γ^H is a weighted potential game with potential function $\Phi^H(x)$ for $x \in X$ and vector of weights $\beta = (\beta_k)_{k\in\mathcal{N}}$. Given the above, taking the time derivative of the potential $\Phi^H(x)$ yields

$$\begin{aligned}\dot{\Phi}^H(x) &= \sum_{k\in\mathcal{N}}\sum_{i\in A_k}\left(\frac{\partial}{\partial x_{ki}}\Phi^H(x)\right)\dot{x}_{ki}\\ &= \sum_{k\in\mathcal{N}}\frac{1}{\beta_k}\sum_{i\in A_k}r^H_{ki}(x)\,\dot{x}_{ki}\\ &= \sum_{k\in\mathcal{N}}\frac{1}{\beta}_k\sum_{i\in A_k}r^H_{ki}(x)\,x_{ki}\left[r^H_{ki}(x) - \langle x_k, r^H_k(x)\rangle\right]\\ &= \sum_{k\in\mathcal{N}}\frac{1}{\beta}_k\left[\sum_{i\in A_k}x_{ki}\left(r^H_{ki}(x)\right)^2 - \left(\sum_{i\in A_k}x_{ki}r^H_{ki}(x)\right)^2\right] \geq 0,\end{aligned}$$

where the last inequality follows directly from the Cauchy-Schwartz inequality (equivalently by observing that the term in braces is the variance of the quantities $r^H_{ki}(x), i \in A_k$ under the distribution $x_k \in X_k$). Accordingly, equality holds if the dynamics are at a fixed point which concludes the proof. □

Proof of Theorem 3.3. Given Lemma 3.4, and the fact that the fixed points of (3) are precisely the QRE of Γ, it remains to show that any sequence of play $(x_k(t))_{k\in\mathcal{N}}, t \geq 0$ converges to a compact, connected set consisting entirely of equilibria, i.e., of points $x \in X$ for which $\Phi^H(x)$ is constant and equal to 0.

To see this, observe that for any sequence of play $(x(t))_{t\geq 0} \subseteq X$, the limit set Ω is defined as

$$\Omega = \bigcap_{s\in\mathbb{R}_+}\mathrm{cl}\{x(t) : t > s\}$$

where $\mathrm{cl}\{S\}$ denotes the closure of set S. Hence, Ω is compact and connected as the decreasing intersection of compact, connected sets. Moreover, since $\Phi^H(x(t))$ is increasing by Lemma 3.4 and bounded on X by definition, it follows that $\Phi^H(x(t))$ converges to a value $\Phi^* = \sup\Phi^H(x(t))$ for any sequence of play $x(t)_{t\geq 0} \subseteq X$. By continuity of Φ^H, this implies that $\Phi^* = \Phi^H(x^*)$ for all $x^* \in \Omega$. Hence, if $x^*(t)_{t\geq 0} \subseteq \Omega$, it must be that $\Phi^H(x^*) = \Phi^*$ for all $t \geq 0$. Since $\Phi^H(x(t))$ is strictly increasing in t unless $x(t)$ is a sequence of equilibria, it follows that Ω consists entirely of equilibria of the game Γ^H which are QRE in Γ. This concludes the proof. □

C. Omitted Materials of Section 4

Proof of Lemma 4.2. Since $k_1, k_2 > 0$, there exists $w > 0$ such that $k_1 = wk_2$. Then, it is immediate to check that

$$P = \begin{pmatrix} u_{11} - u_{21} & u_{11} - u_{21} - w\left(v_{11} - v_{21}\right) \\ 0 & k_1 - w\left(v_{11} - v_{21}\right) \end{pmatrix}$$

is a potential function for Γ with weights $(w_1, w_2) = (1, w)$. Hence, Γ is a $(1, w)$ weighted potential game with potential function P. For (i), using the coordination game assumption, i.e., that $u_{11} > u_{21}, u_{22} > u_{12}$ and $v_{11} > v_{21}, v_{22} > v_{12}$ and dividing both sides of inequality (7) with the product $k_1 k_2$, we have that (7) is equivalent to

$$\frac{\lambda_1}{k_1} \cdot \frac{\lambda_2}{k_2} > \frac{k_1 - \lambda_1}{k_1} \cdot \frac{k_2 - \lambda_2}{k_2} \iff y_{\text{mix}} x_{\text{mix}} > \left(1 - y_{\text{mix}}\right)\left(1 - x_{\text{mix}}\right) \iff x_{\text{mix}} + y_{\text{mix}} > 1$$

as claimed. Finally, if Γ is symmetric, i.e., if $u_2^\top = u_1$, then $k_1 = k_2$ and $\lambda_1 = \lambda_2$ which implies that $x_{\text{mix}} = y_{\text{mix}}$. In this case, inequality (7) yields the condition $x_{\text{mix}} > 1/2$ which proves (i*). To see that the global maximizer of the potential agrees with the risk dominant equilibrium, observe that $P = \begin{pmatrix} u_{11} - u_{21} & 0 \\ 0 & u_{22} - u_{12} \end{pmatrix}$ is a potential function for the game. Hence, the global maximum of P is at (a_2, a_2) whenever $u_{22} - u_{12} > u_{11} - u_{21}$, i.e., whenever $(\lambda_1/k_1) > 1 - (\lambda_1/k_1)$ or equivalently whenever $x_{\text{mix}} > 1/2$. Since any potential function must satisfy $P + c$ for some constant $c \in \mathbb{R}$ (see [76]), this concludes the proof. □

Proof of Theorem 4.1. The proof is constructive and utilizes Theorem 4.3. For $M > 0$, let $\Gamma_u^M = \{\mathcal{N} = \{1, 2\}, \left(\{a_1, a_2\}\right)_{k \in \mathcal{N}}, (u_1, u_2)\}$ with u_1, u_2 given by $u_1 = \begin{matrix} \\ a_1 \\ a_2 \end{matrix} \begin{matrix} \begin{matrix} a_1 & a_2 \end{matrix} \\ \begin{pmatrix} 2M & 0 \\ 2M - 1 & 2 \end{pmatrix} \end{matrix}$, $u_2 = u_1^\top$. Then, (i) $x_{\text{mix}} = y_{\text{mix}} = 2/3$ for any $M > 0$ and (ii) (a_2, a_2) is risk-dominant since $2M - 1 + 2 > 2M + 0$. Condition (i) implies that the basin of attraction of the (a_1, a_1) equilibrium is equal to $I_u = [2/3, 1]^2$ and hence that it has strictly positive measure (and in fact constant for any $M > 0$). By Theorem 4.3, (ii) implies that the QRE surface is disconnected — left panel of Figure 2. Hence, given a starting point in the interior of I_u, there exist exploration thresholds δ_k that depend on $(x_k(0))$, for $k = 1, 2$, so that if the exploration rates remain throughout low, i.e., $\delta_k(t) < \delta_k$ for all $t > 0$ and for both $k = 1, 2$, then the dynamics will never escape the basin of attraction of (a_1, a_1). Hence, since $\lim_{t \to \infty} \delta_k(t) = 0$ by assumption — i.e., at the end of the learning process, agents stop to explore the space — it holds that $\lim_{t \to \infty} u_k^{\text{exploit}}(t) = 2M$ for both agents, i.e., their choice distributions will approximate (to arbitrary precision), the (a_1, a_1) NE.

By contrast, if $\delta_k(t) > \delta_k$ for some agent $k \in \mathcal{N}$, then the coupled dynamics in equation (1) will reach a fixed point close to the uniform distribution which by Theorem 4.3 lies in the basin of attraction of the risk-dominant (a_2, a_2) equilibrium. When reducing the exploration rates back to zero, the agents' choice distribution will converge to the (a_2, a_2) NE, which implies that $\lim_{t \to \infty} u_k^{\text{explore}}(t) = 2$ for both agents. Since $M > 0$ was arbitrary, this concludes the case of unbounded loss.

A specific realization of the game Γ_u^M that is described above can be obtained by appropriately tuning the payoffs in the Stag Hunt game as described in Table 1.

To obtain the other direction, i.e., unbounded gain, consider (in a similar fashion) for $M > 0$ the game $\Gamma_v^M = \{\mathcal{N} = \{1, 2\}, \left(\{a_1, a_2\}\right)_{k \in \mathcal{N}}, (v_1, v_2)\}$ with u_1, u_2 given by $u_1 = \begin{matrix} \\ a_1 \\ a_2 \end{matrix} \begin{matrix} \begin{matrix} a_1 & a_2 \end{matrix} \\ \begin{pmatrix} 2M & 1.5 \\ 2M - 1 & 2 \end{pmatrix} \end{matrix}$, $u_2 = u_1^\top$. Then, (i) $x_{\text{mix}} = y_{\text{mix}} = 1/3$ for any $M > 0$ and (ii) (a_1, a_1) is risk-dominant since $2M - 1 + 2 < 2M + 1.5$. Proceeding as in the previous case, condition (i) implies that the basin of attraction of the (a_2, a_2) equilibrium is equal to $I_v = [0, 1/3]^2$ and hence that it has strictly positive measure (and in fact constant for any $M > 0$). By Theorem 4.3, (ii) implies that the QRE surface is disconnected — left panel of Figure 2. The difference is that now, the payoff-dominant equilibrium (a_1, a_1) is also the risk dominant equilibrium. Hence, starting by an arbitrary point in the interior of I_v and by a similar argument as above, we obtain that $\lim_{t \to \infty} v_k^{\text{explore}}(t) = 2M$ and $\lim_{t \to \infty} v_k^{\text{exploit}}(t) = 2$ for any agent $k \in \mathcal{N}$ which concludes the proof. □

Proof of Theorem 4.3. To prove Theorem 4.3, we will use that for an arbitrary 2-player, 2-action game $\Gamma = \left(\mathcal{N}, (A_k, u_k)_{k\in\mathcal{N}}\right)$, the coupled dynamic equations in (1) become

$$\frac{\dot{x}}{x\,(1-x)} = \beta_x\left[y\,(u_{11}-u_{21}) + (1-y)\,(u_{12}-u_{22})\right] + \alpha_x \ln\left(\frac{1}{x}-1\right) \tag{C.1a}$$

$$\frac{\dot{y}}{y\,(1-y)} = \beta_y\left[x\,(v_{11}-v_{21}) + (1-x)\,(v_{12}-v_{22})\right] + \alpha_y \ln\left(\frac{1}{y}-1\right) \tag{C.1b}$$

where $x, y \in [0,1]$ denote the probabilities that agent 1 and 2 respectively assign to pure action a_1.

At any QRE $(x_Q, y_Q) \in (0,1)\times(0,1)$, the right sides of the coupled equations in (C.1) are simultaneously equal to 0. Using the introduced notation $\lambda_1 := u_{22}-u_{12},\ k_1 := u_{11}-u_{12}-u_{21}+u_{22}$ and $\lambda_2 := v_{22}-v_{12},\ k_2 := v_{11}-v_{12}-v_{21}+v_{22}$, with $\lambda_i, k_i > 0$ for $i = 1,2$, we can rewrite the QRE conditions as

$$c_1\,(y_Q - y_{\text{mix}}) + \ln\left(\frac{1}{x_Q}-1\right) = 0 \tag{C.2a}$$

$$c_2\,(x_Q - x_{\text{mix}}) + \ln\left(\frac{1}{y_Q}-1\right) = 0 \tag{C.2b}$$

where, $c_1 := k_1\cdot\beta_x/\alpha_x$ and $c_2 := k_2\cdot\beta_y/\alpha_y$ are positive constants (with respect to x, y) by assumption. As above $(x_{\text{mix}}, y_{\text{mix}})$ denote the probabilities of pure action a_1 at the fully mixed equilibrium for agent 1 and 2, respectively. The cases in the statement of Theorem 4.3 now follow from an exhaustive sign analysis of the terms $\ln\left(\frac{1}{x_Q}-1\right)$ and $\ln\left(\frac{1}{y_Q}-1\right)$ which are negative for $x_Q, y_Q > 1/2$ and positive otherwise. Specifically, let $x_{\text{mix}} + y_{\text{mix}} > 1$ (the case $x_{\text{mix}} + y_{\text{mix}} < 1$ is analogous and the case $x_{\text{mix}} + y_{\text{mix}} = 1$ corresponds to coordination games at which none of the pure equilibria is risk dominant and is treated in [26]). Then, we have following cases

Case 1: $x_{\text{mix}}, y_{\text{mix}} > 1/2$. If $x_Q > 1/2$, then $\ln\left(\frac{1}{x_Q}-1\right) < 0$ which implies that $y_Q > y_{\text{mix}}$. In turn, since $y_{\text{mix}} > 1/2$, this implies in particular, that $y_Q > 1/2$ and hence, that $\ln\left(\frac{1}{y_Q}-1\right) < 0$ which imposes the condition $x_Q > x_{\text{mix}}$ for equation (C.2b) to be feasible. Hence, if (x_Q, y_Q) is a QRE with $x_Q > 1/2$, it must be the case that $x_Q > x_{\text{mix}}$ and $y_Q > y_{\text{mix}}$. This establishes the upper right region in the left panel of Figure 2. If $x_Q < 1/2$, then it holds that $x_Q - x_{\text{mix}} < 0$ since $x_{\text{mix}} > 1/2$ by assumption, and $\ln\left(\frac{1}{x_Q}-1\right) > 0$. This implies that $y_Q < y_{\text{mix}}$ for equation (C.2a) to hold and $y_Q < 1/2$ for equation (C.2b) to hold. Since $y_{\text{mix}} > 1/2$, this yields the feasible region $x_Q, y_Q < 1/2$ depicted in the bottom left shaded region in the left panel of Figure 2.
Finally, it is easy to see that there cannot be a QRE with either x_Q or y_Q equal to $1/2$. For a contradiction, assume without loss of generality that $x_Q = 1/2$. Then $\ln\left(\frac{1}{x_Q}-1\right) = 0$ and equation (C.2a) implies that $y_Q = y_{\text{mix}}$ which is larger than $1/2$ by assumption. However, this implies that both terms of equation (C.2b) must be negative which is a contradiction to (x_Q, y_Q) being a QRE.

Case 2: $x_{\text{mix}} > 1/2, y_{\text{mix}} < 1/2$. Proceeding as in Case 1, assume first that $x_Q > x_{\text{mix}}$. Then, y_Q must be larger than y_{mix} for equation (C.2a) to hold and larger than $1/2$ for equation (C.2b) to hold. This yields the upper right region of QREs in the right panel of Figure 2 which satisfy $x_Q > x_{\text{mix}}$ and $y_Q > 1/2$. Similarly, if $1/2 < x_Q < x_{\text{mix}}$, then it must be the case that $y_{\text{mix}} < y_Q < 1/2$ for equations (C.2a),(C.2b) to hold, which yields the middle region of QREs in the right panel of Figure 2. When $x_Q < 1/2$, it holds that $\ln\left(\frac{1}{x_Q}-1\right) > 0$ and hence, $y_Q < y_{\text{mix}}$ for equation (C.2a) to hold. In this case, (C.2b) is feasible which results in the bottom left region in the right panel of Figure 2.
In contrast to Case 1, it is easy to see (using a similar argument as in Case 1) that QREs of the form $(1/2, y_{\text{mix}})$ and $(x_{\text{mix}}, 1/2)$ are now feasible. This completes the right panel of Figure 2.

Case 3: $x_{\text{mix}} < 1/2, y_{\text{mix}} > 1/2$. This case is symmetric to Case 2. □

The above cases are illustrated in Figure 2 in the main body of the paper. The case $x_{\text{mix}}, y_{\text{mix}} > 1/2$ corresponds to a situation at which the interests are better aligned than in the case $x_{\text{mix}} > 1/2, y_{\text{mix}} < 1/2$. In particular, symmetric games are special instances of the situation depicted in the left panel of Figure 2. In this case, $x_{\text{mix}} = y_{\text{mix}}$ and there are no QRE in the segment $[1/2, x_{\text{mix}}]$.

D. Supplementary Experiments

To understand the effect of the exploration policy in the collective outcome of the SQL dynamics, we study the three coordination games in Table 1. For each game, we consider all possible combinations of the exploration policies CLR-1 and ETE for the two agents. The experiments for the three coordination games (Pareto Coordination, Battle of the Sexes and Stag Hunt) are presented in Figure D.11.

Arbitrary dimensions. For the plots of the SQL dynamics in Figures 8 and 9, the potential has been generated by the command `rnd('1',twister)` of Matlab — and the entry in $(10, 10)$ deterministically set to 10 — and is given by

$$\Phi = \begin{pmatrix} 4 & 4 & 8 & 1 & 9 & 1 & 1 & 9 & 8 & 2 \\ 7 & 7 & 9 & 4 & 7 & 7 & 4 & 2 & 6 & 9 \\ 1 & 2 & 3 & 9 & 3 & 2 & 7 & 2 & 7 & 5 \\ 3 & 8 & 7 & 5 & 8 & 3 & 4 & 8 & 4 & 6 \\ 2 & 1 & 8 & 7 & 1 & 5 & 1 & 4 & 3 & 4 \\ 1 & 7 & 9 & 3 & 5 & 1 & 5 & 2 & 9 & 3 \\ 2 & 4 & 1 & 7 & 9 & 6 & 6 & 9 & 4 & 9 \\ 4 & 6 & 1 & 8 & 3 & 2 & 5 & 4 & 9 & 6 \\ 4 & 2 & 2 & 1 & 3 & 6 & 9 & 7 & 6 & 1 \\ 5 & 2 & 8 & 7 & 2 & 7 & 6 & 7 & 6 & 10 \end{pmatrix}.$$

Visualization of the modified potential. To visualize the modified potential in games with action spaces of arbitrary dimension, we adapt the method of [42]. Specifically, for a two-player game with choice distribution spaces $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ consider the transformation $g : \mathbb{R}^n \to \mathbb{R}^{n-1}$ with $y_{1i} := \ln(x_{1i}/x_{1n})$, for each $i = 1, 2, \ldots, n$ and the same for player 2. Together with the constraint $\sum_{i=1}^{n} x_{1i} = 1$, the inverse $g^{-1} : \mathbb{R}^{n-1} \to \mathbb{R}^n$ of this transformation is given by $x_{1i} = x_{1n} e^{y_{1i}}$, for $i = 1, \ldots, n$. The benefit of working with the transformed variables is that they are not subject to the Simplex constraints. Hence, we may choose random choice distributions x_1, x_2, transform them to y_1, y_2 and then scale them with two real scalars α, β of arbitrary sign and magnitude to obtain a point on the hyperplane $\alpha \cdot y_1 + \beta \cdot y_2$. Then, we calculate the corresponding choice distributions (via the inverse transformation and the normalization equation) and evaluate the modified potential Φ^H at this point. This yields a tuple

$$\left(\alpha, \beta, \Phi^H\left(g^{-1}\left(\alpha \cdot y_1 + \beta \cdot y_2\right)\right)\right),$$

for each value of α, β which (if combined) yield the surface plots of Figures 10 and D.12. The process is summarized in Algorithm 1. The technique readily generalizes to $n > 2$ players with the same exploration rate. A specific instant of a potential game with $n = 20$ which shows the transformation of the potential manifold as δ increases is included in the multimedia appendix (online supplementary material).

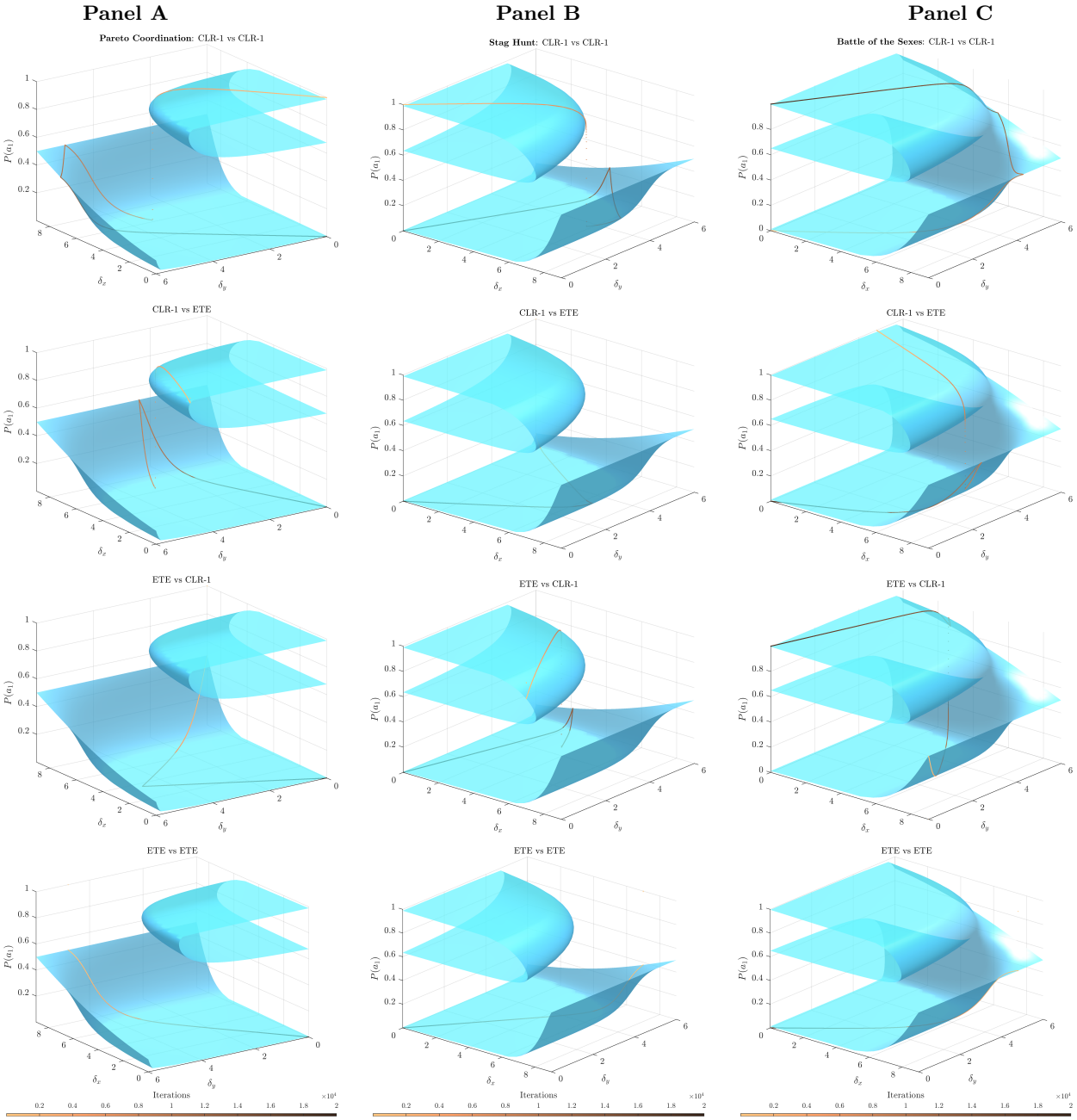

Figure D.11: SQL dynamics ($1e+03$ Q-value updates for each of $2e+04$ choice distribution updates) with the ETE and CLR-1 policies in coordination games. In Pareto Coordination (Panel A) and Stag Hunt (Panel B), the SQL dynamics converge to the risk dominant equilibrium regardless of the starting point (*saddle-node bifurcations*). By contrast, in the Battle of the Sexes (Panel C), the collective outcome of the exploration process is a priori ambiguous, and the SQL dynamics may converge to either of the pure action equilibria (consistent with *cusp bifurcations*). The decisive feature is the geometry of the QRE surface, i.e., whether it is connected or not, as formalized in Theorems 4.1 and 4.3.

Algorithm 1 3D Visualization of the Potential

procedure INPUT GAME(Φ, n, m)
Input: $\Phi \leftarrow$ Potential matrix
Input: $n \leftarrow$ # actions of player 1
Input: $m \leftarrow$ # actions of player 2

procedure GENERATE RANDOM DIRECTIONS(n, m)
    for $i \leftarrow u, v$ **do**
        u, v generate random vectors (not parallel)
        $u = [u_1, u_2], v = [v_1, v_2]$ with
        $u_1, v_1 \in \mathbb{R}^{n-1}, u_2, v_2 \in \mathbb{R}^{m-1}$

procedure TRANSFORM VARIABLES(u, v, α, β)
Input: $\alpha, \beta \leftarrow$ scalars in $\mathbb{R}$
    $x \leftarrow \alpha \cdot u_1 + \beta \cdot v_1$
    $x \leftarrow \exp(x)$
    $x \leftarrow$ normalized to sum up to 1
    Repeat to get y

procedure EVALUATE POTENTIAL($x, y, \delta, \alpha, \beta$)
Input: $\delta \leftarrow$ Common exploration rate
    for α, β **do**
        $\Phi^H(\alpha, \beta) = x'\Phi y - \delta \sum x_i \ln x_i - \delta \sum y_j \ln y_j$

return Plot tuples $\left(\alpha, \beta, \Phi^H(\alpha, \beta)\right)$

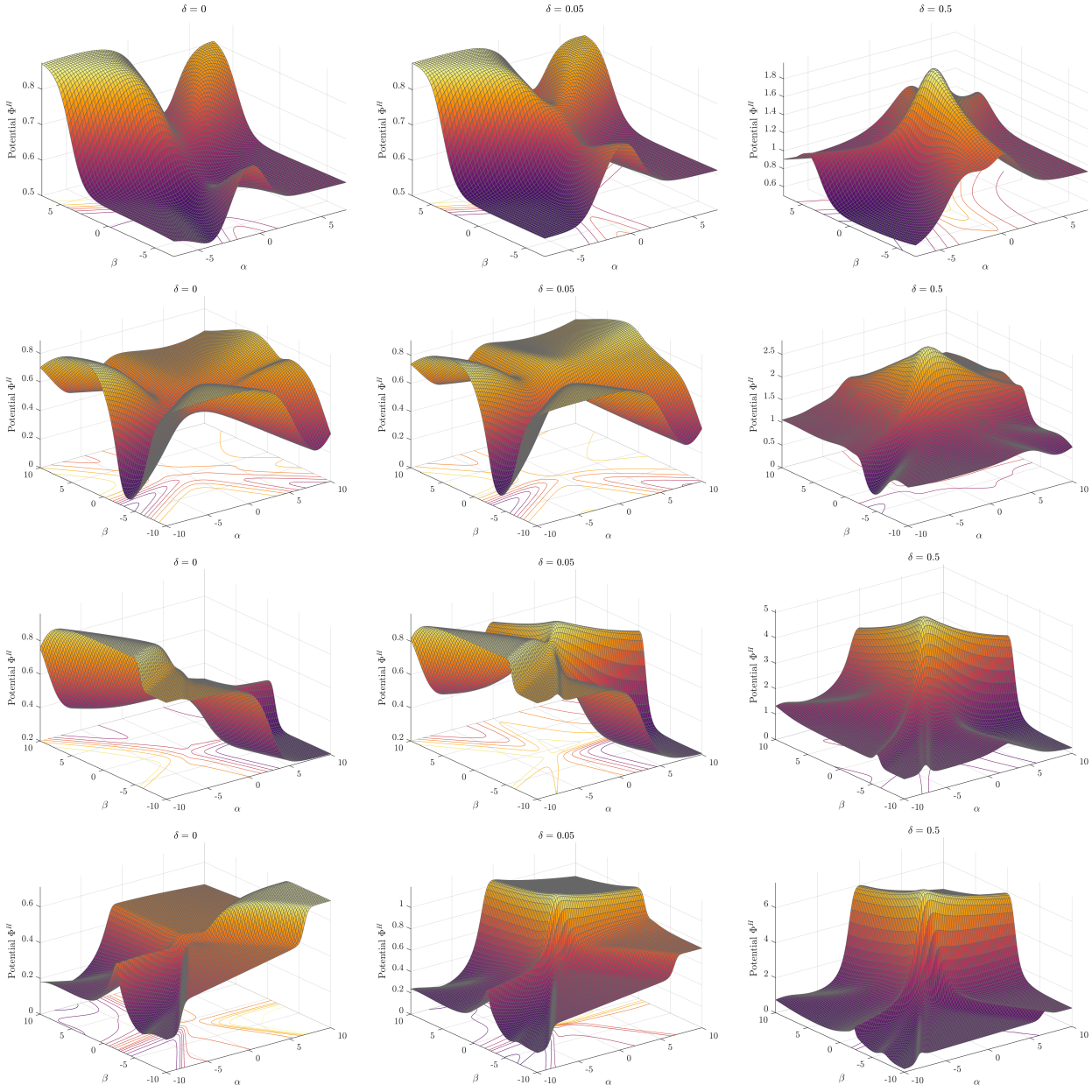

Figure D.12: From top to bottom: snapshots of the modified potential Φ^H surface in 2 player potential games with $n = 4, 10, 100$ and 1000 actions, respectively, and random payoffs in $[0, 1]$. The surfaces are plotted using Algorithm 1 (see [42]). From left to right: the exploration rate δ increases from $\delta = 0$ to $\delta = 0.05$ and $\delta = 0.5$. The surface has arbitrary maxima (resting points of the SQL dynamics) when δ is small (exploitation) but a single maximum at (or close to) $(0, 0)$ which corresponds to the uniform distribution when δ is large (exploration). Intuitively, this is what agents *see* as they increasingly incorporate exploration in their utilities.