NON-LINEAR DYNAMICS OF COLLECTIVE LEARNING IN UNCERTAIN ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Complex adaptive systems occur in all domains across all scales, from cells to societies. The question, however, of how the various forms of collective behavior can emerge from individual behavior and feedback to influence those individuals remains open. Complex systems theory focuses on emerging patterns from deliberately simple individuals. Fields such as machine learning and cognitive science emphasize individual capabilities without considering the collective level much. To date, however, little work went into modeling the effects of changing and uncertain environments on emergent collective behavior from individually self-learning agents. To this end, we derive and present deterministic reinforcement learning dynamics where the agents only partially observe the actual state of the environment. This paper aims to obtain an efficient mathematical description of the emergent behavior of biologically plausible and parsimonious learning agents for the typical case of environmental and perceptual uncertainty. We showcase the broad applicability of our dynamics across different classes of agent-environment systems, highlight emergent effects caused by partial observability and show how our method enables the application of dynamical systems theory to partially observable multi-agent learning. The presented dynamics have the potential to become a formal yet practical, lightweight, and robust tool for researchers in biology, social science, and machine learning to systematically investigate the effects of interacting partially observant agents.

1 INTRODUCTION

Motivation. Complex adaptive multi-agent systems are everywhere. The emergence of a collective level characterizes them through the interplay of individual entities, which in turn feeds back to influence those individuals. Advancing the understanding of complex adaptive systems is vital for explaining collective behavior in nature (Couzin, 2009), making technological multi-agent systems safe and efficient (Ferber & Weiss, 1999), and tackling societal collective action challenges (Bak-Coleman et al., 2021).

Complex systems research has made significant progress in this regard by employing a range of tools from stochastic processes, statistical physics, and non-linear dynamics. However, while focusing on the emergence of a collective level, individual entities are kept deliberately simple, often described as being in one of only a few states. Yet, in most complex systems, individual behavior is more sophisticated. Research fields such as machine learning and cognitive science focus on the intelligent behavior of one individual situated in an environment, whereas the collective level is not considered much, albeit notable exceptions (Rosa et al., 2019; Ha & Tang, 2021). We argue that a combined approach is required to describe the collective and individual levels' interplay adequately.

To this end, describing multi-agent reinforcement learning *as* a dynamical system has proven itself useful to gain improved, qualitative insights into the emerging collective learning dynamics (Bloembergen et al., 2015). However, existing learning dynamics are either applicable only to stateless environments, assume that agents do not tailor their response to the current environmental state, or, if they do, believe that agents observe the true states of the environment perfectly. Yet, in the real world, state observations are noisy and incomplete.

Overview. With this work, we advance the field of describing multi-agent reinforcement learning as a dynamical system and obtain an efficient mathematical description of the emergent behavior of biologically plausible and parsimonious learning agents for the typical case of environmental and perceptual uncertainty. With the derived dynamics, we can study the idealized reinforcement learning behavior in a wide range of environmental classes, from partially observable Markov decision processes to fully general, partially observable stochastic games.

We employ the widely-occurring principle of temporal-difference reinforcement learning (Sutton, 1988). Temporal-difference learning is not only a computational technique (Sutton & Barto, 2018), it also occurs in biological agents through the dopamine reward prediction error signal (Schultz et al., 1997; Dayan & Niv, 2008). We study boundedly rational agents who treat their observations (or a short history of those) as if they were the actual states of the environment. This has the advantage of being simple to act upon (Williams & Singh, 1998), and they are easy to realize at no or little additional computational cost.

We showcase the applicability of dynamics at two partially observable environment classes. We find instances where partial observability can lead to better learning outcomes faster by stabilizing a chaotic learning process in a multi-state zero-sum game and overcoming an uncertain social dilemma. Furthermore, our method allows applying dynamical systems theory to partially observable multi-agent learning. We find that partial observability can cause a critical slowing down of the learning processes between reward regimes and the separation of the learning dynamics into fast and slow eigendirections.

Related work. As in evolutionary game theory, learning dynamic agents are boundedly rational Simon (1997), taking into account *strategic uncertainty* by assuming that other agents are not perfectly rational either but instead by allowing agents to adapt to each other sequentially. Börgers & Sarin (1997) established the formal relationship between the learning behavior of one of the most basic reinforcement learning schemes, Cross learning (Cross, 1973), and the replicator dynamics of evolutionary game theory. Since then, the link between evolutionary game theory and reinforcement learning has been extended to stateless Q-learning (Tuyls et al., 2003; Sato & Crutchfield, 2003), regret-minimization (Klos et al., 2010) and temporal-difference learning (Barfuss et al., 2019), as well as discrete-time dynamics (Galla & Farmer, 2013), continuous strategy spaces (Galstyan, 2013) and extensive-form games (Panozzo et al., 2014). This learning dynamics approach offers a formal, lightweight, and deterministically reproducible way to gain improved, descriptive insights into the emerging multi-agent learning behavior.

Apart from strategic uncertainty, representing *stochastic uncertainty*, i.e., uncertainty about what will happen in the form of probabilistic events within the environment requires foremost the presence of an environment. Recent years have seen a growing interest in moving evolutionary and learning dynamics in stateless games to changing environments. Here, the term environment can mean external fluctuations (Assaf et al., 2013; Ashcroft et al., 2014), a varying population density (Hauert et al., 2006; Gokhale & Hauert, 2016), spatial network structure (Gracia-Lázaro et al., 2013; Szolnoki & Chen, 2018), or coupled systems out of evolutionary and environmental dynamics. Coupled systems may further be categorized into those with continuous environmental state spaces Weitz et al. (2016); Chen & Szolnoki (2018); Tilman et al. (2020); Wang & Fu (2020) or discrete ones (Hilbe et al., 2018; Barfuss et al., 2019; Hauert et al., 2019; Su et al., 2019). We'll be focusing on learning dynamics in stochastic games (Hilbe et al., 2018; Barfuss et al., 2019) which encode stochastic uncertainty via action-depended transition probabilities between environmental states.

While many works on partially observable decision domains are normative, ours is descriptive. For the normative agenda, agents are often enriched with, e.g., generative models and belief-state representations (Spaan, 2012; Oliehoek & Amato, 2016), abstractions (Sutton et al., 2006) or predictive state representations (Littman et al., 2001) in order to learn optimal policies in partially observable decision domains. Also, the economic value of signals is often studied by asking how fully rational agents optimally deal with a specific form of state uncertainty (Bagh & Kusunose, 2020). However, such techniques can become computationally extremely expensive (Loch & Singh, 1998). It is unlikely that biological agents perform those elaborate calculations (Gigerenzer & Gaissmaier, 2011) and the focus on unboundedly rational game equilibria lacks a dynamic perspective (Papadimitriou & Piliouras, 2019) making it unable to answer which equilibrium (of the often many) the agents select.

2 COLLECTIVE LEARNING DYNAMICS

We introduce the necessary background in Sec. 2.1 before we derive of our dynamics in Sec. 2.2.

2.1 BACKGROUND

Partially observable stochastic games. Formally, we employ the framework of stochastic games (Shapley, 1953; Levy & Solan, 2020). $N \in \mathbb{N}$ agents interact in a shared environment, which consists of $Z \in \mathbb{N}$ states $\mathcal{S} = (S_1, \dots, S_Z)$. In each state s, each agent i has $M \in \mathbb{N}$ available actions $\mathcal{A}^i = (A^i_1, \dots, A^i_M)$ to choose from. $\boldsymbol{\mathcal{A}} = \bigotimes_i \mathcal{A}^i$ is the joint-action set where $\bigotimes_i$ denotes the cartesian product over the sets indexed by i and agents choose their actions simultaneously. With $\boldsymbol{a}^{-i} = (a^1, \dots, a^{i-1}, a^{i+1}, \dots, a^N)$ we denote the joint action except agent i's. We chose an equal number of actions for all states and agents out of notational convenience.

The transition function $T : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to [0, 1]$ determines the probabilistic state changes. $T(s, \boldsymbol{a}, s')$ is the transition probability from current state s to next state s' under joint action $\boldsymbol{a}$. Throughout this work, we restrict ourselves to ergodic environments without absorbing states.

The reward function $\boldsymbol{R} : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to \mathbb{R}^N$ maps the triple of current state s, joint action $\boldsymbol{a}$ and next state s' to an immediate reward scalar for each agent. $R^i(s, \boldsymbol{a}, s')$ is the reward agent i receives. Agents are interested maximize their rewards over time, $\sum_t^\infty (\gamma^i)^t R^i_t$, discounted by their *discount factors* $\gamma^i \in [0, 1)$ and R^i_t being the reward agent i received at time step t. The discount factor regulates how much an agent cares for future rewards.

Instead of observing the states $s \in \mathcal{S}$ directly, each agent i observes one of $Q \in \mathbb{N}$ observations $\mathcal{O}^i = (O^i_1, \dots, O^i_Q)$ according to the observation functions $O^i : \mathcal{S} \times \mathcal{O}^i \to [0, 1]$. $O^i(s, o)$ is the probability that agent i observes observation $o \in \mathcal{O}^i$ given that the environment is in state $s \in \mathcal{S}$. $\boldsymbol{\mathcal{O}} = \bigotimes_i \mathcal{O}^i$ is the joint observation set and $\boldsymbol{O} = \bigotimes_i O^i : \mathcal{S} \times \boldsymbol{\mathcal{O}} \to [0, 1]^N$ is the joint observation function. We chose an equal number of observations for all agents out of notational convenience. By construction, this observation function can model both noisy state observations and hidden states.

We consider agents that choose their actions probabilistically according to their memoryless policy $\pi^i : \mathcal{O}^i \times \mathcal{A}^i \to [0, 1]$. $\pi^i(a^i|o^i)$ is the probability that agent i chooses action a^i given that it observed observation o^i. $\boldsymbol{\pi} = \bigotimes_i \pi^i$ is the joint policy.

Reinforcement Learning. Motivated by a maximum entropy approach Wolpert et al. (2012); Barfuss (2021), we parameterize the policy of agent i by a soft-max function

$$\pi^i_t(a^i|o^i) = \frac{e^{\beta^i Q^i_t(o^i, a^i)}}{\sum_{b \in \mathcal{A}^i} e^{\beta^i Q^i_t(o^i, b)}} \tag{1}$$

where the qualities $Q^i_t(o^i, a^i)$ express how agent i evaluates action a^i for observation o^i at time step t. The *intensity of choice* parameter β^i regulates the exploration-exploitation trade-off. Learning means a change of policy induced by an update of those state-action values $Q^i_t(o^i, a^i)$,

$$Q^i_{t+1}(o^i, a^i) = Q^i_t(o^i, a^i) + \alpha^i \cdot \delta^i_t(o^i, a^i), \tag{2}$$

where δ^i_t denotes the temporal-difference or reward-prediction error. Agents try to successively improve their qualities of the available actions at each observation. The *learning rate* parameter $\alpha^i \in (0, 1)$ regulates how much new information is used for an observation-action-value update. Different variants of temporal-difference learning exist in the literature. We focus on the famous Q-learning update (Watkins & Dayan, 1992) for which the temporal-difference error is obtained as,

$$\delta^i_t(o^i, a^i) := R^i_t + \gamma^i \max_b Q^i_t(o'^i, b) - Q^i_t(o^i, a^i), \tag{3}$$

where R^i_t is the reward agent i received at time step t and o'^i is the next observation agent i observed.

2.2 DERIVATION

In this section, we derive the deterministic reinforcement learning dynamics under partial observability in discrete time. As classic evolutionary dynamics operate in the theoretical limit of an infinite

population, the learning dynamics are derived by considering an infinite memory batch (Barfuss, 2020) or likewise a separation of time scales between the process of interaction and adaptation (Barfuss, 2021). In essence, deterministic learning dynamics consider policy averages instead of individual samples of rewards and observations. Thus, we need to construct the policy-average temporal-difference error $\bar{\delta}^i$ to be inserted in the update for the joint policy,

$$\pi^i_{t+1}(a^i|o^i) = \frac{\pi^i_t(a^i|o^i) \cdot \exp\left(\alpha \bar{\delta}^i(o^i, a^i)\right)}{\sum_b \pi^i_t(b|o^i) \cdot \exp\left(\alpha \bar{\delta}^i(o^i, b)\right)}. \tag{4}$$

Eq. 4 can be derived by combining Eqs. 2 and 1. The bar on top of δ^i indicates implicitly that $\bar{\delta}^i$ depends fully on the current joint policy $\boldsymbol{X}_t$. Computing $\bar{\delta}^i(o^i, a^i)$ involves averaging over policies, environmental transitions and observations for the first two terms of the temporal-difference error (Eq. 3), the immediate rewards and the qualities of the next observation. The quality of the current observation, $Q^i_t(o^i, a^i)$ becomes $\beta^{-1} \ln \pi^i(a^i|o^i)$ in the average temporal-difference error and serves as regularization term. This can be derived by inverting Eq. 1 and realizing that the dynamics induced by Eq. 4 are invariant under additive transformations which are constant in actions.

Beliefs. The challenge is that the rewards $R^i(s, \boldsymbol{a}, s')$ in the stochastic game model depend on the true states, not on the observations of the agents. Thus, in order to obtain the average observation-action rewards $\bar{R}^i(o^i, a^i)$, we need a mapping from observations to states. The observation function is a mapping from states to observations. With Bayes rule,

$$\bar{B}^i(o^i, s) = \frac{O^i(s, o^i)\bar{P}(s)}{\sum_s O^i(s, o^i)\bar{P}(s)} \tag{5}$$

we can transform the observation function into a belief function, following the rules of probability. $\bar{B}^i(o^i, s)$ is the belief of agent i (or simply the probability) that the environment is in state s when it observed observation o^i.

The only problem is how to obtain the policy-average stationary state distribution $\bar{P}(s)$. $\bar{P}(s)$ is the left-eigenvector of the average transition matrix $\bar{T}(\boldsymbol{s}, \boldsymbol{s})$ where the entry $\bar{T}(s, s')$ denotes the probability of transitioning from state s to state s'. This matrix could be obtained as $\bar{T}(s, s') = \prod_j \sum_{a^j} \bar{\rho}^j(a^j|s) T(s, \boldsymbol{a}, s')$ if we had the probability for each agent j to choose action a^j in state s, $\bar{\rho}^j(a^j|s)$. However, we assumed that agents condition their actions only on observations, $\pi^j(a^j|o^j)$. Yet, whenever the environment is in state s, agent j observes observation o^j with probability $O^j(s, o^j)$ and than chooses action a^j with probability $\pi^j(a^j|o^j)$. Thus, with

$$\bar{\rho}^j(a^j|s) := \sum_{o^j \in \mathcal{O}^j} O^j(s, o^j)\pi^j(a^j|o^j), \tag{6}$$

we can average out the observation and obtain the policy-average state-policies $\bar{\rho}^j(a^j|s)$. Note that $\bar{\rho}^j(a^j|s)$ are proper conditional probabilities, which can be seen by applying $\sum_{a^j}$ to both sides of Eq. 6. With $\bar{\rho}^j(a^j|s)$ we can then compute the policy-average transition matrix $\bar{T}(\boldsymbol{s}, \boldsymbol{s})$, its left-eigenvector, the stationary state distribution $\bar{P}(s)$, and thus, the policy-average belief of agent i that the environment is in state s when it observed observation o^i, $\bar{B}^i(o^i, s)$.

Rewards. Whenever agent i observes observation o^i, with probability $\bar{B}^i(o^i, s)$ the environment is in state s where all other agents $j \neq i$ behave according to $\bar{\rho}^j(a^j|s)$, the environment transitions to a next state s' with probability $T(s, \boldsymbol{a}, s')$, and agent i receives the reward $R^i(s, \boldsymbol{a}, s')$. Mathematically, the policy-average reward for action a^i under observation o^i reads

$$\bar{R}^i(o^i, a^i) := \sum_s \sum_{a^j} \sum_{s'} \prod_{j \neq i} \bar{B}^i(o^i, s)\bar{\rho}^j(a^j|s) T(s, \boldsymbol{a}, s') R^i(s, \boldsymbol{a}, s'). \tag{7}$$

Qualities. Second, the policy-average of the quality of the next observation ($\max_b Q^i_t(o^i_{t+1}, b)$ in Eq. 3) is computed by averaging over all states, all actions of the other agents, next states and next observations. Whenever agent i observes observation o^i, the environment is in state s with probability $\bar{B}^i(o^i, s)$. There, all other agents $j \neq i$ choose their action a^j with probability $\bar{\rho}^j(a^j|s)$. Consequently, the environment transitions to the next state s' with probability $T(s, \boldsymbol{a}, s')$. At s',

the agent observes observation o' with probability $O^i(s', o')$ and estimates the quality to be of value $\max_b \bar{Q}^i(o', b)$. Mathematically, we write

$$^{\max}\bar{Q}^i(o^i, a^i) := \sum_s \sum_{a^j} \sum_{s'} \sum_{o'} \prod_{j \neq i} \bar{B}^i(o^i, s) \bar{\rho}^j(a^j|s) T(s, \boldsymbol{a}, s') O^i(s', o') \max_b \bar{Q}^i(o', b). \tag{8}$$

Here, we replace the quality estimates $Q^i_t(o^i, a^i)$, which evolve in time t (Eq. 2), with the policy-average observation-action quality $\bar{Q}^i(o^i, a^i)$, which is the expected discounted sum of future rewards from executing action a^i at observation o^i and then following along the joint policy $\boldsymbol{\pi}$. It is obtained by a discount factor weighted average of the current policy-average reward $\bar{R}^i(o^i, a^i)$ and the policy-average observation quality of the following observation $\bar{V}^i(o')$,

$$\bar{Q}^i(o^i, a^i) = \bar{R}^i(o^i, a^i) + \gamma^i \sum_{o' \in \mathcal{O}^i} \bar{T}^i(o^i, a^i, o') \bar{V}^i(o'). \tag{9}$$

Here, $\bar{T}^i(o^i, a^i, o')$ is agent i's policy-average transition probability of observing observation o' at the next time step given it observed observation o^i at the current time step and chose action a^i. It is computed by averaging over all states, next states and all actions of the other agents. Whenever agent i observes observation o^i and selects action a^i, the environment is in state s with probability $\bar{B}^i(o^i, s)$, where all other agents $j \neq i$ select action a^j with probability $\bar{\rho}^j(a^j|s)$. Consequently, the environment will transition to the next state s' with probability $T(s, \boldsymbol{a}, s')$ which is observed by agent i as o' with probability $O^i(s', o')$. Mathematically, we write

$$\bar{T}^i(o^i, a^i, o') = \sum_s \sum_{a^j} \sum_{s'} \prod_{j \neq i} \bar{B}^i(o^i, s) \bar{\rho}^j(a^j|s) T(s, \boldsymbol{a}, s') O^i(s', o'). \tag{10}$$

Further at Eq. 9, $\bar{V}^i(o^i)$ is the policy-average observation quality, i.e., the expected discounted sum of future rewards from observation o^i and then following along the joint policy $\boldsymbol{\pi}$. They are computed via matrix inversion according to

$$\bar{V}^i(\underline{\boldsymbol{o}}) = [\mathbb{1}_Q - \gamma^i \bar{T}^i(\underline{\boldsymbol{o}}, \underline{\boldsymbol{o}})]^{-1} \bar{R}(\underline{\boldsymbol{o}}). \tag{11}$$

This equation is a direct conversion of the Bellman equation $\bar{V}^i(o^i) = \bar{R}(o^i) + \gamma^i \sum_{o'} \bar{T}^i(o^i, o') \bar{V}^i(o')$, which expresses that the value of the current observation is the discount factor weighted average of the current reward and the value of the next observation. Underlined observation variables indicate that the corresponding object is a vector or matrix and $\mathbb{1}_Q$ is a Q-by-Q identity matrix.

$\bar{T}^i(\underline{\boldsymbol{o}}, \underline{\boldsymbol{o}})$ denotes the policy-averaged transition matrix for agent i. The entry $\bar{T}^i(o^i, o')$ indicates the probability that agent i will observe observation o' after observing observation o^i at the previous time step, given all agents follow the joint policy $\boldsymbol{X}$. We compute them by averaging over all states, all actions from all agents and all next states,

$$\bar{T}^i(o^i, o') = \sum_s \sum_{a^j} \sum_{s'} \prod_j \bar{B}^i(o^i, s) \bar{\rho}^j(a^j|s) T(s, \boldsymbol{a}, s') O^i(s', o'). \tag{12}$$

For any observation o^i, $\bar{B}^i(o^i, s)$ is the probability to be in state s, where all agents j act according to $\bar{\rho}^j(a^j|s)$. Therefore, the environment transitions with probability $T(s, \boldsymbol{a}, s')$ from state s to the next state s', which is observed by agent i as observation o' with probability $O^i(s', o')$. Note that $\bar{T}^i(\underline{\boldsymbol{o}}, \underline{\boldsymbol{o}})$ is a proper probabilistic matrix. This can be seen by applying $\sum_{o'}$ to both sides of Eq. 12.

Further in Eq. 11, $\bar{R}^i(o^i)$ denotes the policy-average reward agent i obtains from observation o^i. We compute them by averaging over all states, all actions from all agents and all next states. Whenever agent i observes observation o^i, the environment is in state s with probability $\bar{B}^i(o^i, s)$. Here, all agents j choose action a^j with probability $\bar{\rho}^j(a^j|s)$. Hence, the environment transitions to the next state s' with probability $T(s, \boldsymbol{a}, s')$ and agent i receives the reward $R^i(s, \boldsymbol{a}, s')$,

$$\bar{R}^i(o^i) := \sum_s \sum_{a^j} \sum_{s'} \prod_j \bar{B}^i(o^i, s) \bar{\rho}^j(a^j|s) T(s, \boldsymbol{a}, s') R^i(s, \boldsymbol{a}, s'). \tag{13}$$

Note that the quality $^{\max}\bar{Q}^i(o^i, a^i)$ depends on o^i and a^i although it is the policy-averaged maximum observation-action value of the next observation.

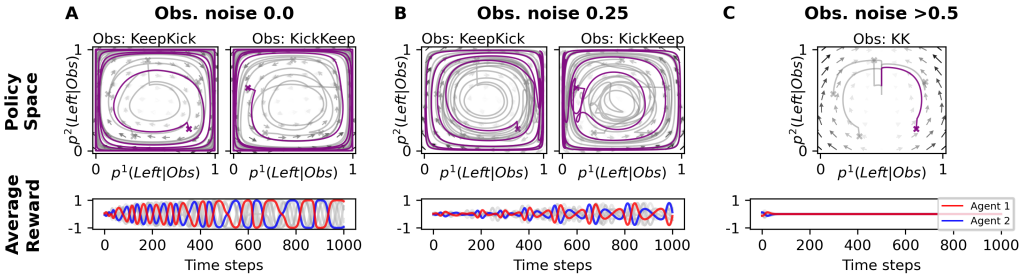

Figure 1: **Deterministic learning dynamics in an uncertain zero-sum competition.** Policy spaces and reward trajectories are shown for three different observational noise levels: (A) $\nu = 0.0$, i.e., perfect observation, (B) $\nu = 0.25$, and (C) $\nu > 0.5$, i.e., both states are observed inseparably as one. The probability of choosing action *left*, conditioned on the current observation, is plotted on the x-axis for agent 1 and on the y-axis for agent 2. Learning trajectories are shown from 5 initial policies around the center of the policy spaces. Only one of those trajectories is portrayed in color for better visual inspection. Arrows in gray indicate the flow of the dynamical learning system. Learning-parameters were $\alpha = 0.005$, $\beta = 200$, and $\gamma = 0.9$. Partial observability can stabilize the learning process and separate the learning dynamics into fast and slow eigendirections.

TD error. All together, the policy-average temporal-difference error, to be inserted into Eq. 4, reads

$$\bar{\delta}^i(o^i, a^i) = \bar{R}^i(o^i, a^i) + \gamma^{\max}\bar{Q}^i(o^i, a^i) - \beta^{-1} \ln \pi^i(a^i|o^i). \tag{14}$$

3 Experiments

In the following section, we will showcase our dynamics at two environment classes, an uncertain zero-sum competition 3.1 and an uncertain social dilemma 3.2.

3.1 Uncertain Zero-sum Competition

Environment description. The first environment we use to explore the derived learning dynamics is a two-agent, two-state, two-action zero-sum competition, also known as the two-state matching pennies game (Hennes et al., 2010). It roughly models the situation of penalty kicks between a kicker and a keeper. Both agents can choose between the *left* and the *right* side of the goal. The keeper agent scores one point if it catches the ball (when both agents have chosen the same action); otherwise, the kicker agent receives one point. The two states of the environment encode which agent is the keeper and which one is the kicker. In the state *KeepKick* agent 1 is the keeper, and agent 2 is the kicker. In the state *KickKeep* it is the other way around. Agents change roles under state transitions, which depend only on agent 1's actions. When agent 1 selects either *left* as keeper or *right* as kicker both agents will change roles. With symmetrical rewards but asymmetrical state transitions, this two-state zero-sum game presents the challenge of coordinating both agents on playing a mixed strategy with equiprobable actions. The agents' observations of the environmental states are obscured by a noise level ν.

Results. Fig. 1 shows how partial observability can stabilize the learning process. When both agents observe the environment perfectly, the learning dynamics are prone to be unstable, either unpredictably chaotic or on periodic orbits and limit cycles (Panel A, Barfuss et al., 2019). The rewards of agents 1 and 2 are circulating around zero. The learning dynamics are still unstable under a medium observational noise level of $\nu = 0.25$. Especially the transient dynamics in the policy space (Panel B, on the right) appear strange. The average reward trajectory is damped compared to the fully observant agents. Increasing the observational noise further such that the agents perceive the two environmental states (*KeepKick* and *KickKeep*) as a single observation (*KK*) stabilizes the learning process.

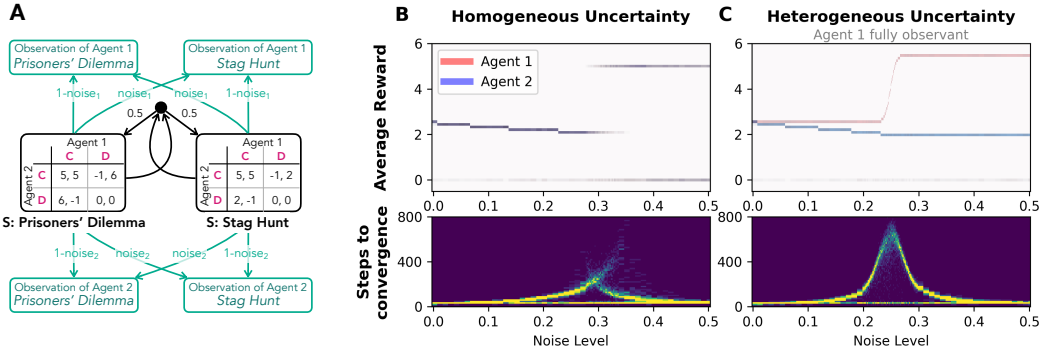

Figure 2: **Deterministic learning dynamics in an uncertain social dilemma.** Panel A illustrates the environment. Panels B and C show the average rewards at convergence for agent 1 in red and agent 2 in blue (top row) and the timesteps it takes the learners to convergence (bottom row) for various observational noise levels from 0 to 0.5. The plots show a histogram for each noise level via the color scale. Each histogram results from a Monte Carlo simulation from 100 random initial policies. Panel B shows the case of homogeneous uncertainty where both agents' observations are corrupted equally by noise. In Panel C, only agent 2 is increasingly unable to observe the environment correctly (Heterogeneous Uncertainty). The discount factor was set to $\gamma = 0.5$ since future states are independent of the agents' actions, making the discount factor irrelevant for the learning in this case. Remaining parameters were set to $\alpha = 0.01$ and $\beta = 50$. Homogeneous uncertainty can overcome the social dilemma through the emergence of a stable, mutually high rewarding fixed point above a critical level of observational noise. Heterogeneous uncertainty, however, leads to reward inequality. In both cases, the transition is accompanied by a critical slowing down of the convergence speed.

Interestingly, the flow of the learning dynamics is separated into fast and slow eigendirections. The fast directions have the form of two half circles directed at the upper half of the line at which agent 1 chooses both actions with equal probabilities. The slow direction is the movement down to the center of the policy space. At this downward movement, both agents play the different roles of kicker and keeper in equal amounts since only agent 1 is responsible for the state transitions. Any advantage agent 2 gains from deviating from the equiprobable policy as the kicker is balanced by the same amount of disadvantage agent 2 loses as the keeper. Thus, the rewards for both agents quickly stabilize at zero.

3.2 Uncertain Social Dilemma

Environment description. The emergence of cooperation in social dilemmas is a crucial research challenge for evolutionary biology, the social and sustainability sciences (Nowak, 2006; Kollock, 1998; Barfuss et al., 2020). We'll focus on the situation where two agents can either cooperate (C) or defect (D) and either face a Prisoner's Dilemma or a Stag Hunt game with equal probability (Fig. 2 A, cf., Levine & Ponssard, 1977; LiCalzi & Mühlenbernd, 2019). In the pure Prisoner's Dilemma, defection is the Nash equilibrium, which leads to a sub-optimal reward for both agents, also known as the tragedy of the commons (Hardin, 1968). In the pure Stag Hunt game, both mutual cooperation and mutual defection are Nash equilibria with the difference that mutual cooperation yields a higher reward than mutual defection for both agents. It is therefore also referred to as a coordination challenge (Barrett & Dannenberg, 2012). Here, we consider the situation when the agents are uncertain about the game they face at each decision point. Whether we are facing a tragedy or a coordination challenge is relevant for, e.g., the mitigation of human-caused climate change (Barrett & Dannenberg, 2017). We investigate two scenarios. Under homogenous uncertainty (Fig. 2 B), both agents' observations are blurred by an increasing level of observational noise. Under heterogeneous uncertainty (Fig. 2 C), only agent 2's observations become noisier. Since the environment is symmetric under exchanging the roles of the agents, it suffices to explore only one heterogeneous uncertainty scenario.

Results. Homogeneous uncertainty can overcome the social dilemma through the emergence of a stable, mutually high rewarding fixed point above a critical level of observational noise. Under perfect observation, both agents convergence to full defection when observing the Prisoner's Dilemma. When observing the Stag Hunt game, it depends on the initial joint policy whether the agents converge to mutual defection or cooperation. Reward values are as such that the defective basin of attraction is comparable small (see the light line at an average reward of 0 in Fig. 2 B). Increasing the observational noise level from zero under homogeneous uncertainty will first decrease the average reward at convergence. The agents still converge to the perfect observation policy, which leads them to defect when they observe the Prisoners' Dilemma, but the situation is actually the Stag Hunt. However, increasing observational noise further eventually leads to a bifurcation (Fig. 2 B). Mutual cooperation under both observations becomes a stable fixed point. Consequently, both agents obtain an average reward of 5 at convergence. Interestingly, there seems to be a small range of observational noise at which all three rewards 0, ~ 2, and 5 are supported by equilibria. Only the rewards at 0 and 5 are stable for large noise levels.

Thus, we find that the deterministic learning dynamics under homogeneous partial observability can converge to more rewarding policies than under perfect observation. The existence of those equilibria is long known in traditional static game theory (Levine & Ponssard, 1977). Here we show that our derived dynamics serve as dynamic micro-foundations for those static equilibria. They correspond not only to fixed points of the derived learning dynamics. The transitions between equilibria are also accompanied by the dynamical systems phenomenon of a critical slowing down of the convergence speed (Fig. 2 B, bottom).

However, the mutual benefit of uncertainty vanishes when not all agents' observations are uncertain (Fig. 2 C). Under slight uncertainty, only the reward of the ill-informed agent (Agent 2 in Fig. 2) decreases. After the bifurcation point under large uncertainty, the ill-informed agent converges to full cooperation under both observations. In contrast, the well-informed agent still defects in the Prisoner's Dilemma, earning an average reward of even more than 5. The knowledgable agent exploits the ill-informed and heterogeneous uncertainty leads to reward-inequality between the agents.

Interestingly, Fig. 2 suggests a difference in the type of phase transition between the policy of mediocre reward at low observational noise levels and the policies at high noise levels. The phase transition under homogeneous uncertainty seems to be discontinuous and shifted towards greater noise levels, whereas the transition under heterogeneous uncertainty appears to be continuous. Investigating the relationship between the learning dynamics and phase transitions is a promising direction of future work.

4 Conclusion

In this article, we introduced deterministic multi-agent reinforcement learning dynamics, in which the agents are only partially able to observe the actual states of the environment. These dynamics operate in the theoretical limit of an infinite memory batch and implicitly infer the actual states via Bayes rule. This limit allows us to systematically separate the stochasticity of reinforcement learning, resulting from probabilistic environmental dynamics, observations, and decisions, from the environmental uncertainty that originates in the agents' incomplete awareness of the actual state space.

Overall, we demonstrated how these dynamics serve as a practical, lightweight, deterministically reproducible, and robust tool to systematically study the combined effects of *strategic uncertainty*, *stochastic uncertainty* and *state uncertainty* in collectives of self-learning agents. For instance, we have shown that partial observability can lead to better learning outcomes faster by stabilizing a chaotic learning process in a multi-state zero-sum game and overcoming an uncertain social dilemma. Furthermore, our method allows applying dynamical systems theory to partially observable multi-agent learning. We find that partial observability can cause a critical slowing down of the learning processes between reward regimes and the separation of the learning dynamics into fast and slow eigendirections. It is an interesting direction for future work how such insights from dynamical systems theory can be used in technological applications of multi-agent reinforcement learning, e.g., with respect to training regimes, hyper-parameter tuning, and the development of novel algorithms.

REFERENCES

Peter Ashcroft, Philipp M Altrock, and Tobias Galla. Fixation in finite populations evolving in fluctuating environments. *Journal of The Royal Society Interface*, 11(100):20140663, 2014.

Michael Assaf, Mauro Mobilia, and Elijah Roberts. Cooperation dilemma in finite populations under fluctuating environments. *Physical Review Letters*, 111(23):238101, 2013.

Adib Bagh and Yoko Kusunose. On the economic value of signals. *The BE Journal of Theoretical Economics*, 20(1), 2020.

Joseph B Bak-Coleman, Mark Alfano, Wolfram Barfuss, Carl T Bergstrom, Miguel A Centeno, Iain D Couzin, Jonathan F Donges, Mirta Galesic, Andrew S Gersick, Jennifer Jacquet, et al. Stewardship of global collective behavior. *Proceedings of the National Academy of Sciences*, 118 (27), 2021.

Wolfram Barfuss. Reinforcement learning dynamics in the infinite memory limit. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pp. 1768–1770, 2020.

Wolfram Barfuss. Dynamical systems as a level of cognitive analysis of multi-agent learning. *Neural Computing and Applications*, pp. 1–19, 2021.

Wolfram Barfuss, Jonathan F. Donges, and Jürgen Kurths. Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E*, 99(4):043305, 2019. doi: 10.1103/PhysRevE.99.043305.

Wolfram Barfuss, Jonathan F Donges, Vítor V Vasconcelos, Jürgen Kurths, and Simon A Levin. Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. *Proceedings of the National Academy of Sciences*, 117(23):12915–12922, 2020.

S. Barrett and A. Dannenberg. Climate negotiations under scientific uncertainty. *Proceedings of the National Academy of Sciences*, 109(43):17372–17376, 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1208417109.

Scott Barrett and Astrid Dannenberg. Tipping versus cooperating to supply a public good. *Journal of the European Economic Association*, 15(4):910–941, 2017.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: a survey. *Journal of Artificial Intelligence Research*, 53:659–697, 2015.

Tilman Börgers and Rajiv Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.

Xiaojie Chen and Attila Szolnoki. Punishment and inspection for governing the commons in a feedback-evolving game. *PLoS Computational Biology*, 14(7):e1006347, 2018.

Iain D Couzin. Collective cognition in animal groups. *Trends in cognitive sciences*, 13(1):36–43, 2009.

John G. Cross. A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, 87(2):239, 1973.

Peter Dayan and Yael Niv. Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185–196, 2008.

Jacques Ferber and Gerhard Weiss. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-Wesley Reading, 1999.

Tobias Galla. Intrinsic noise in game dynamical learning. *Physical Review Letters*, 103(19):198702, 2009.

Tobias Galla and J. Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013.

Aram Galstyan. Continuous strategy replicator dynamics for multi-agent Q-learning. *Autonomous Agents and Multi-Agent Systems*, 26(1):37–53, 2013.

Gerd Gigerenzer and Wolfgang Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62:451–482, 2011.

Chaitanya S Gokhale and Christoph Hauert. Eco-evolutionary dynamics of social dilemmas. *Theoretical Population Biology*, 111:28–42, 2016.

Carlos Gracia-Lázaro, Luis M Floría, Jesús Gómez-Gardeñes, and Yamir Moreno. Cooperation in changing environments: Irreversibility in the transition to cooperation in complex networks. *Chaos, Solitons & Fractals*, 56:188–193, 2013.

David Ha and Yujin Tang. Collective Intelligence for Deep Learning: A Survey of Recent Developments. *arXiv:2111.14377 [cs]*, 2021.

Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968.

Christoph Hauert, Miranda Holmes, and Michael Doebeli. Evolutionary games and population dynamics: maintenance of cooperation in public goods games. *Proceedings of the Royal Society B: Biological Sciences*, 273(1600):2565–2571, 2006.

Christoph Hauert, Camille Saade, and Alex McAvoy. Asymmetric evolutionary games with environmental feedback. *Journal of Theoretical Biology*, 462:347–360, 2019.

Daniel Hennes, Michael Kaisers, and Karl Tuyls. RESQ-learning in stochastic games. In *Adaptive and Learning Agents Workshop at AAMAS (ALA)*, 2010.

Christian Hilbe, Štěpán Šimsa, Krishnendu Chatterjee, and Martin A Nowak. Evolution of cooperation in stochastic games. *Nature*, 559(7713):246–249, 2018.

Tomas Klos, Gerrit Jan Van Ahee, and Karl Tuyls. Evolutionary dynamics of regret minimization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 82–96. Springer, 2010.

Peter Kollock. Social dilemmas: The anatomy of cooperation. *Annual Review of Sociology*, 24(1): 183–214, 1998.

Pierre Levine and Jean-Pierre Ponssard. The values of information in some nonzero sum games. *International Journal of Game Theory*, 6(4):221–229, 1977.

Yehuda John Levy and Eilon Solan. Stochastic Games. In Marilda Sotomayor, David Pérez-Castrillo, and Filippo Castiglione (eds.), *Complex Social and Behavioral Systems*, pp. 229–250. Springer US, New York, NY, 2020. ISBN 978-1-07-160367-3 978-1-07-160368-0. doi: 10.1007/978-1-0716-0368-0_522.

Marco LiCalzi and Roland Mühlenbernd. Categorization and cooperation across games. *Games*, 10 (1):5, 2019.

Michael L Littman, Richard S Sutton, and Satinder P Singh. Predictive representations of state. In *International Conference on Neural Information Processing Systems (NeurIPS)*, volume 14, pp. 30, 2001.

John Loch and Satinder P. Singh. Using eligibility traces to find the best memoryless policy in partially observable markov decision processes. In *International Conference on Machine Learning (ICML)*, pp. 323–331, 1998.

Martin A Nowak. Five rules for the evolution of cooperation. *Science*, 314(5805):1560–1563, 2006.

Frans A Oliehoek and Christopher Amato. *A concise introduction to decentralized POMDPs*, volume 1. Springer, 2016.

Fabio Panozzo, Nicola Gatti, and Marcello Restelli. Evolutionary dynamics of Q-learning over the sequence form. In *Conference on Artificial Intelligence (AAAI)*, pp. 2034–2040, 2014.

Christos Papadimitriou and Georgios Piliouras. Game dynamics as the meaning of a game. *ACM SIGecom Exchanges*, 16(2):53–63, 2019.

Marek Rosa, Olga Afanasjeva, Simon Andersson, Joseph Davidson, Nicholas Guttenberg, Petr Hlubuček, Martin Poliak, Jaroslav Vítku, and Jan Feyereisl. BADGER: Learning to (Learn [Learning Algorithms] through Multi-Agent Communication). 2019.

Yuzuru Sato and James P. Crutchfield. Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E*, 67(1):015206, January 2003.

Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

L. S. Shapley. Stochastic Games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.39.10.1095.

Herbert Alexander Simon. *Models of bounded rationality: Empirically grounded economic reason*, volume 3. MIT press, 1997.

Matthijs TJ Spaan. Partially observable Markov decision processes. In *Reinforcement Learning: State-of-the-Art*, pp. 387–414. Springer, 2012.

Qi Su, Alex McAvoy, Long Wang, and Martin A Nowak. Evolutionary dynamics with game transitions. *Proceedings of the National Academy of Sciences*, 116(51):25398–25404, 2019.

R. S. Sutton, E. Rafols, and A. Koop. Temporal abstraction in temporal-difference networks. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1313–1320, 2006.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. MIT Press, 2018.

Attila Szolnoki and Xiaojie Chen. Environmental feedback drives cooperation in spatial social dilemmas. *EPL (Europhysics Letters)*, 120(5):58001, 2018.

Andrew R Tilman, Joshua B Plotkin, and Erol Akçay. Evolutionary games with environmental feedbacks. *Nature Communications*, 11(1):1–11, 2020.

Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for Q-learning in multi-agent systems. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 693–700, 2003.

Xin Wang and Feng Fu. Eco-evolutionary dynamics with environmental feedback: Cooperation in a changing world. *EPL (Europhysics Letters)*, 132(1):10001, 2020.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

Joshua S Weitz, Ceyhun Eksin, Keith Paarporn, Sam P Brown, and William C Ratcliff. An oscillating tragedy of the commons in replicator dynamics with game-environment feedback. *Proceedings of the National Academy of Sciences*, 113(47):E7518–E7525, 2016.

John K. Williams and Satinder P. Singh. Experimental results on learning stochastic memoryless policies for partially observable Markov decision processes. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1073–1079, 1998.

David H. Wolpert, Michael Harré, Eckehard Olbrich, Nils Bertschinger, and Jürgen Jost. Hysteresis effects of changing the parameters of noncooperative games. *Physical Review E*, 85(3):036102, 2012. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.85.036102.