Non-linear dynamics of multi-agent reinforcement learning in partially observable environments
(Extended Abstract)

Wolfram Barfuss
University of Tübingen
Tübingen, Germany
wolfram.barfuss@uni-tuebingen.de

Richard P. Mann
University of Leeds
Leeds, United Kingdom
r.p.mann@leeds.ac.uk

ABSTRACT

We derive and present deterministic reinforcement learning dynamics where the agents only partially observe the actual state of the environment. Our aim with this work is to obtain an efficient mathematical description of the emergent behavior of biologically plausible and parsimonious learning agents for the common 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 allows 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 MOTIVATION

We do not observe the world as it is but instead as our limited sensory and cognitive apparatus perceives it. There are always elements of the world hidden from us, such as the detailed physical state of our environment and the internal states of other agents. As such, uncertainty is a fundamental feature of life [24, 31, 37]. Thus, among other forms of uncertainty, we might not know what will happen (*stochastic uncertainty*), what currently is (*state uncertainty*) and what others are going to do (*strategic uncertainty*).

Given the cognitive demands of fully integrating all sources of uncertainty when learning from experience and making decisions, natural agents must employ methods of bounded rationality [44] that use cognitive resources efficiently to obtain acceptable solutions in a timely manner [23]. As such, evolutionary game theory [29] takes into account *strategic uncertainty* by assuming that other agents are not perfectly rational but instead by allowing agents to adapt to each other sequentially. Tools and methods from evolutionary game theory have also been used successfully to formally study the dynamics of multi-agent reinforcement learning [5, 12]. Börgers and Sarin [13] established the formal relationship between the learning behavior of one of the most basic reinforcement learning schemes, Cross learning [15], and the replicator dynamics of evolutionary game theory. Since then, this approach of evolutionary reinforcement learning dynamics has been extended to stateless Q-learning [42, 54], regret-minimization [30] and temporal-difference learning [7], as well as discrete-time dynamics [17], continuous strategy spaces [18] and extensive-form games [40]. This learning dynamic 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 [1, 2], a varying population density [21, 26], spatial network structure [22, 52], or coupled systems out of evolutionary and environmental dynamics. Coupled systems may further be categorized into those with continuous environmental state spaces [14, 53, 55, 56] or discrete ones [7, 27, 28, 48]. We'll be focusing on learning dynamics in stochastic games [4, 6, 7, 28] which encode stochastic uncertainty via action-depended transition probabilities between environmental states.

However, all dynamics discussed so far 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, often in real-world settings, state observations are noisy and incomplete. Thus, they lack a systematic way to describe interacting agents under *state uncertainty*.

This work relaxes the assumption of perfect observations and introduces deterministic reinforcement learning dynamics for partially observable environments. 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 [POMDPs, 46], decentralized POMDPs [39], and fully general partially observable stochastic games [25].

Note, 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 [39, 46], abstractions [51] or predictive state representations [35] 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 [3]. However, such techniques can become computationally extremely expensive [36]. It is unlikely that biological agents perform those elaborate calculations [20] and the focus on unboundedly rational game equilibria lacks a dynamic perspective [41] making it unable to answer which equilibrium (of the often many) the agents select.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2022), Cruz, Hayes, da Silva, Santos (eds.), May 9-10, 2022, Online, https://ala2022.github.io/. 2022.

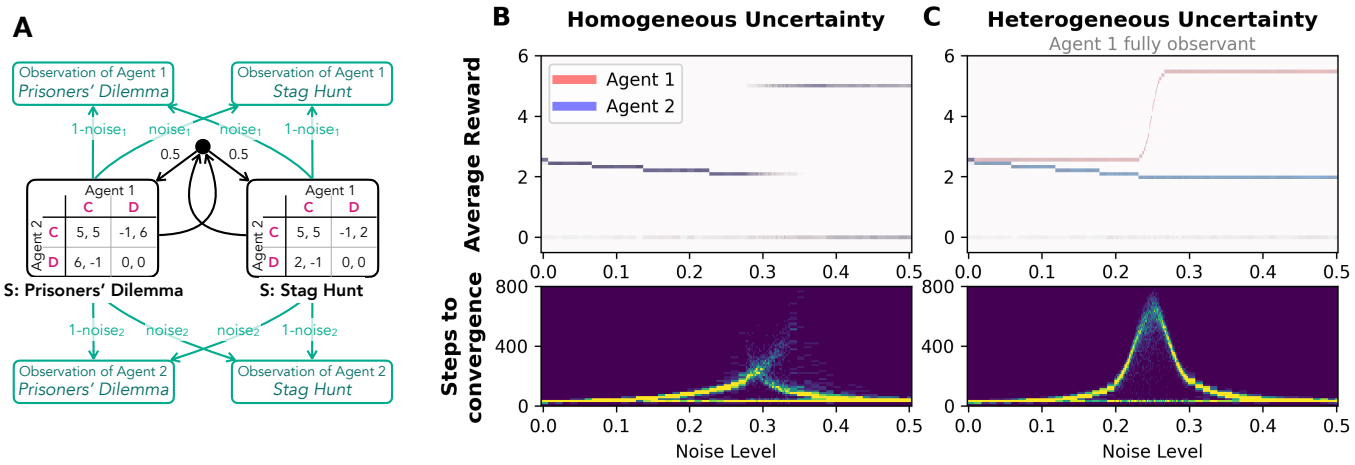

Figure 1: Deterministic learning dynamics in an uncertain social dilemma. Panel A illustrates the environment. Panels B and C show the average rewards at convergence for agent 1 in red and agent 2 in blue (top row) and the time steps it takes the learners to convergence (bottom row) for various observational 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).

2 OVERVIEW OF RESULTS

This work takes a dynamical systems perspective on individual learning agents employing the widely-occurring principle of temporal-difference reinforcement learning [49] in which the agents treat their observations as if they were the actual states of the environment. Temporal-difference learning is not only a computational technique [50], it also occurs in biological agents through the dopamine reward prediction error signal [16, 43]. We focus on agents who employ either so-called memoryless policies, at which they choose their actions based solely on their current observation [45], or they use a short and fixed history of current and past observations and actions to base the current action upon. This has the advantage of being simple to act upon [57], and they are easy to realize at no or little additional computational cost.

To highlight our dynamics' broad applicability, we study the emerging learning behavior across five partially observable environment classes. Detailed results can be found in the full paper [11]. We find various effects caused by partial observability, which generally depend on the environment and its representation. For instance, partial observability can lead to better learning outcomes faster in a single-agent renewable resource harvesting task, stabilize a chaotic learning process in a multi-state zero-sum game and even overcome social dilemmas. Compared to fully observant agents, partially observant learning often requires more exploration and less weight on future rewards to obtain favorable learning outcomes. Furthermore, our method allows applying dynamical systems theory to partially observable multi-agent learning. We find that partial observability can cause the emergence of catastrophic limit cycles, a critical slowing down of the learning processes between reward regimes, and the separation of the learning dynamics into fast and slow eigendirections.

3 EXAMPLE: EMERGENCE OF COOPERATION

The emergence of cooperative and sustainable behavior in social dilemmas is a crucial research challenge for evolutionary biology, the social and sustainability sciences [8–10, 19, 32, 38, 47]. 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. 1 A, cf., 33, 34]. In the pure Prisoner's Dilemma, defection is the Nash equilibrium, which leads to a suboptimal reward for both agents. 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. Here, we consider the situation when the agents are uncertain by a certain noise level about the type of game they face at each decision point.

Fig. 1 shows how homogeneous uncertainty (where both agents are uncertain) can overcome the social dilemma through the emergence of a stable, mutually high rewarding fixed point above a critical level of observational noise. However, heterogeneous uncertainty (where only agent 2 is uncertain) leads to reward inequality. In both cases, the bifurcation is accompanied by a critical slowing down of the convergence speed. Interestingly, the type of phase transitions appears to be different between the two scenarios. Under homogeneous uncertainty, it seems to be discontinuous, whereas, under heterogeneous uncertainty, it seems to be continuous.

4 CONCLUSION

We hope that the presented dynamics become a formal yet practical, lightweight, and robust tool for researchers in biology, social science, and machine learning to systematically investigate the effect of uncertainty of interacting agents. Python code to reproduce all results is available at https://github.com/wbarfuss/POLD.

REFERENCES

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

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

[3] Adib Bagh and Yoko Kusunose. 2020. On the Economic Value of Signals. *The BE Journal of Theoretical Economics* 20, 1 (2020).

[4] Wolfram Barfuss. 2020. Reinforcement Learning Dynamics in the Infinite Memory Limit.. In *AAMAS*. 1768–1770.

[5] Wolfram Barfuss. 2020. Towards a unified treatment of the dynamics of collective learning. In *AAAI Spring Symposium: Challenges and Opportunities for Multi-Agent Reinforcement Learning*.

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

[7] Wolfram Barfuss, Jonathan F. Donges, and Jürgen Kurths. 2019. Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E* 99, 4 (2019), 043305.

[8] Wolfram Barfuss, Jonathan F Donges, Steven J Lade, and Jürgen Kurths. 2018. When optimization for governing human-environment tipping elements is neither sustainable nor safe. *Nature communications* 9, 1 (2018), 1–10.

[9] Wolfram Barfuss, Jonathan F Donges, Vítor V Vasconcelos, Jürgen Kurths, and Simon A Levin. 2020. 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 (2020), 12915–12922.

[10] Wolfram Barfuss, Jonathan F Donges, Marc Wiedermann, and Wolfgang Lucht. 2017. Sustainable use of renewable resources in a stylized social–ecological network model under heterogeneous resource distribution. *Earth System Dynamics* 8, 2 (2017), 255–264.

[11] Wolfram Barfuss and Richard P Mann. 2022. Modeling the effects of environmental and perceptual uncertainty using deterministic reinforcement learning dynamics with partial observability. *Physical Review E* 105, 3 (2022), 034409.

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

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

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

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

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

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

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

[19] Fabian Geier, Wolfram Barfuss, Marc Wiedermann, Jürgen Kurths, and Jonathan F Donges. 2019. The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources. *The European Physical Journal Special Topics* 228, 11 (2019), 2357–2369.

[20] Gerd Gigerenzer and Wolfgang Gaissmaier. 2011. Heuristic decision making. *Annual Review of Psychology* 62 (2011), 451–482.

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

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

[23] Thomas L Griffiths, Falk Lieder, and Noah D Goodman. 2015. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science* 7, 2 (2015), 217–229.

[24] Joseph Y Halpern. 2017. *Reasoning about uncertainty*. MIT Press.

[25] Eric A. Hansen, Daniel S. Bernstein, and Shlomo Zilberstein. 2004. Dynamic Programming for Partially Observable Stochastic Games. In *National Conference on Artifical Intelligence (AAAI)*. 709–715.

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

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

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

[29] Josef Hofbauer and Karl Sigmund. 1998. *Evolutionary games and population dynamics*. Cambridge University Press.

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

[31] Mykel J Kochenderfer. 2015. *Decision making under uncertainty: theory and application*. MIT Press.

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

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

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

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

[36] John Loch and Satinder P. Singh. 1998. Using Eligibility Traces to Find the Best Memoryless Policy in Partially Observable Markov Decision Processes. In *International Conference on Machine Learning (ICML)*. 323–331.

[37] Vincent AWJ Marchau, Warren E Walker, Pieter JTM Bloemen, and Steven W Popper. 2019. *Decision making under deep uncertainty: from theory to practice*. Springer Nature.

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

[39] Frans A Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs*. Vol. 1. Springer.

[40] Fabio Panozzo, Nicola Gatti, and Marcello Restelli. 2014. Evolutionary Dynamics of Q-Learning over the Sequence Form. In *Conference on Artificial Intelligence (AAAI)*. 2034–2040.

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

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

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

[44] Herbert Alexander Simon. 1997. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press.

[45] Satinder P. Singh, Tommi Jaakkola, and Michael I Jordan. 1994. Learning without state-estimation in partially observable Markovian decision processes. In *Machine Learning Proceedings 1994*. Elsevier, 284–292.

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

[47] Felix M Strnad, Wolfram Barfuss, Jonathan F Donges, and Jobst Heitzig. 2019. Deep reinforcement learning in World-Earth system models to discover sustainable management strategies. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 29, 12 (2019), 123122.

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

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

[50] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning*. MIT Press.

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

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

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

[54] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. 2003. A Selection-Mutation Model for Q-learning in Multi-agent Systems. In *International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 693–700.

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

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

[57] John K. Williams and Satinder P. Singh. 1998. Experimental Results on Learning Stochastic Memoryless Policies for Partially Observable Markov Decision Processes. In *International Conference on Neural Information Processing Systems (NeurIPS)*. 1073–1079.