# Reinforcement Learning Dynamics in the Infinite Memory Limit

Extended Abstract

Wolfram Barfuss
School of Mathematics, University of Leeds, UK & Max Planck Institute for Mathematics in the Sciences, Leipzig, GER
barfuss@mis.mpg.de

## ABSTRACT

Reinforcement learning algorithms have been shown to converge to the classic replicator dynamics of evolutionary game theory, which describe the evolutionary process in the limit of an infinite population. However, it is not clear how to interpret these dynamics from the perspective of a learning agent. In this paper we propose a data-inefficient batch-learning algorithm for temporal difference Q learning and show that it converges to a recently proposed deterministic limit of temporal difference reinforcement learning. In a second step, we state a data-efficient learning algorithm, that uses a form of experience replay, and show that it retains core features of the batch learning algorithm. Thus, we propose an agent-interpretation for the learning dynamics: What is the infinite population limit of evolutionary dynamics is the infinite memory limit of learning dynamics.

## KEYWORDS

Multi-agent reinforcement learning; Evolutionary game theory; Batch learning; Experience replay; Stochastic games

**ACM Reference Format:**
Wolfram Barfuss. 2020. Reinforcement Learning Dynamics in the Infinite Memory Limit. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

## 1 INTRODUCTION

The link between evolutionary game theory and multi-agent reinforcement learning has proven itself useful to gain improved, qualitative insights into the resulting collective learning dynamics of a multi-agent system [5]. The relationship between the two fields is as follows: one population with a frequency over phenotypes in the evolutionary setting corresponds to one agent with a frequency over actions in the learning setting [22]. In their seminal work, Börgers and Sarin showed how one of the most basic reinforcement learning update schemes, Cross learning [7], converges to the deterministic replicator dynamics of evolutionary games theory [6]. Likewise, the convergence to the replicator dynamics has been shown for single-state Q learning [21, 23].

This deterministic - sometimes also called *evolutionary* - limit can be taken in multiple ways. In continuous time, the learning rate is sent to zero [21, 23]. In discrete time, the batch size of a batch learning algorithm is sent to infinity [4, 9–11]. In essence, both ways assume that policy updates occur on much slower time scales than actual interactions with other agents and the environment.

*Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.* 

So far the majority of learning dynamics studies focused on single-state repeated games. Previous work on learning dynamics which consider more realistic multi-state environments combine replicator dynamics with switching dynamics between cell partitions of the state space of the dynamical system [12, 13, 24]. They consider an average reward setting, whereas in Q learning a discounted reward is commonly used. Only recently, an analytical method to derive the deterministic, discrete-time limit of temporal difference reinforcement learning with discounted rewards was proposed [2]. However, it is unclear how the temporal difference batch learning algorithm must be constructed, such that its learning trajectories through policy space can converge to the ones of the deterministic learning equations under large batch size. This lack of connection between algorithmic implementation and analytical equations make an agent-interpretation of this deterministic limit difficult.

Taken together, it is therefore still an open question how to interpret the deterministic, *evolutionary*, limit in the context of reinforcement learning agents. The classic replicator equations have a clear interpretation. They model the dynamics of an infinite population evolving under the pressures of selection [15].

## 2 METHODS

In this work, we propose to interpret the deterministic, *evolutionary*, limit of reinforcement learning as learning in the infinite memory limit. We do so by comparing the deterministic temporal difference reinforcement learning dynamics (DetRL) [2] against three algorithms (see below). As a testbed, we use two environments: a single-agent two-actions two-states environment, modeling an intertemporal risk-reward dilemma [3]; and a two-agents two-actions two-states Matching Pennies game, [12], which presents a challenge of coordination. We let each algorithm update its policy for a 100 times.

## 3 ALGORITHMS

Research activity on batch reinforcement learning has grown substantially in recent years [17]. In this work, we exclusively use the tabular case (without function approximation) and thus, focus on the issue of data efficiency [25].

### 3.1 Sample-Batch

First, we propose a novel, data-inefficient, sample-batch reinforcement learning algorithm for Q learning (SBATCH). Key to its performance is the separation of the state-action values into two data structures, one for the value estimation inside the batch, the other for acting outside the batch. Since the agent interacts physically with the environment during the interaction phase, SBATCH requires many interaction steps and is therefore highly data-inefficient.

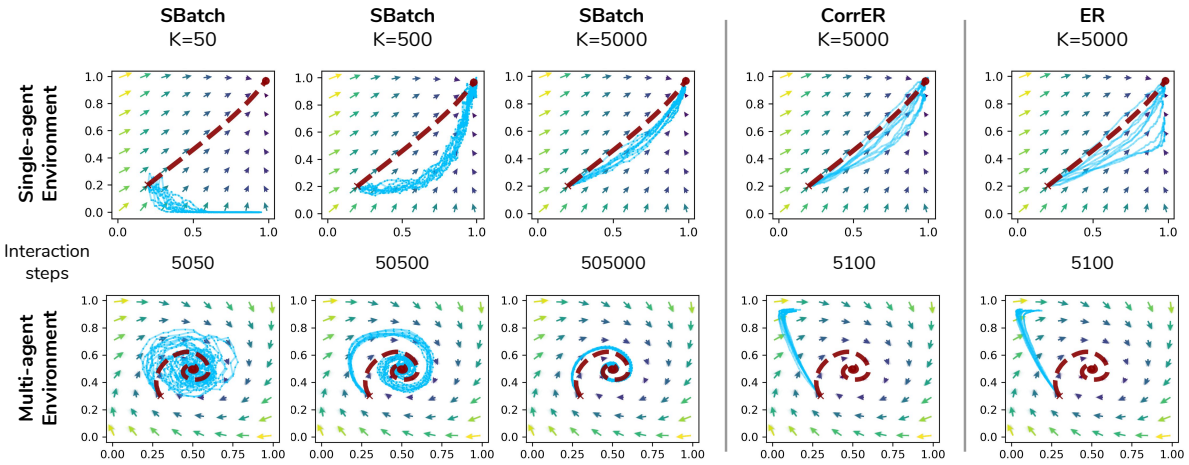


**Figure 1: Comparison between the DetRL dynamics (dark red dashed line) and the SBATCH, CorrER and ER algorithms (light blue straight lines) for various batch sizes $K$; shown are policy spaces of one state, i.e., the probability of playing action 1 on the x-axes for agent 1 and on the y-axes for agent 2; on top for a single-agent intertemporal risk-reward dilemma; on the bottom for the multi-agent two-state Matching pennies game. We let each algorithm adapt its behavioral policy a 100 times. Resulting interaction steps required with the environment are shown in the middle. SBatch matches the deterministic limit increasingly well under increasing batch size $K$, however at the cost of increasing interaction steps, which make SBatch highly data-inefficient. CorrER and ER require only 5100 interaction steps for a batch size of $K = 5000$ and are therefore highly data-efficient. For the single-agent environment the trajectories of CorrER match well with the ones of the deterministic limit, in contrast to the ones of ER. However, for the multi-agent environment CorrER fails to gain similarity to the deterministic limit because it does not take into account the actions of the other agents.**

Fig. 1 shows that SBATCH converges indeed to the learning trajectories of DetRL, both, for the single- and the multi-agent environment. However, it requires many interactions with the environment, proportional to the batch size, and is therefore highly data-inefficient.

### 3.2 Correlated Experience Replay

Second, we transform the data-inefficient batch learning algorithm into a data-efficient learning algorithm, which uses a form of experience replay. Thus, we shift the batch of actual interactions with the environment into the memory of the agent. In contrast to popular uses of experience replay with neural network function approximation [17], we use correlated experiences to be replayed to the Q update of the agent. Therefore we term it CorrER.

Fig. 1 shows that it retains core features of the batch learning algorithm, in contrast to an experience replay variant without correlated experiences (ER), when used within a single-agent environment. However, when used in the multi-agent environment, CorrER trajectories were not closer to the one of DetRL than the ones of ER because CorrER does not take into account the actions of the other agents. This suggests that DetRL - interpreted in the infinite memory limit - represent a form of joint-action learning.

## 4 CONCLUSION

Taken together, we provide an individual agent interpretation for the dynamics of learning. The deterministic limit of reinforcement learning, which results from a time-scale separation of interaction and adaptation, is like learning under infinite joint-action memory. Although derived from independent, model-free learners, this suggests that these dynamics represent a form of model-based joint-action reinforcement learning.

Especially when the evolutionary process is understood as a form of social, cultural learning [14] we can state this equivalence as follows. Evolutionary imitation learning from others' experience in an infinite population of equals resembles individual learning from own experience under infinite memory of joint-action observations. What is the infinite population limit of evolutionary dynamics is an infinite memory limit of learning dynamics.

This result is of potential use for broadening the scope of previous research in ecology and economics, where an infinite population approximation is often used to study the convergence to equilibria [8, 16, 18–20]. Our work suggests that these results apply also to finite populations with large memory.

Furthermore, we hope that our results contribute to a better understanding of the dynamics of collective reinforcement learning. Such an understanding is crucial in order to put the study of collective learning dynamics into practical use for overcoming critical challenges of multi-agent reinforcement learning, such as nonstationarities, the curse of dimensionality of the joint-state-action space, the increased number of hyper parameters, coordination needs, and the possibility of social dilemmas [1].

## REFERENCES

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

*Reinforcement Learning* (2020).
[2] 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.
[3] 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), 2354. https://doi.org/10.1038/s41467-018-04738-z
[4] Alex J. Bladon and Tobias Galla. 2011. Learning dynamics in public goods games. *Physical Review E* 84, 4 (oct 2011). https://doi.org/10.1103/physreve.84.041132
[5] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary Dynamics of Multi-Agent Learning: A Survey. *Journal of Artificial Intelligence Research* 53 (aug 2015), 659–697. https://doi.org/10.1613/jair.4818
[6] Tilman Börgers and Rajiv Sarin. 1997. Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory* 77, 1 (nov 1997), 1–14. https://doi.org/10.1006/jeth.1997.2319
[7] John G. Cross. 1973. A Stochastic Learning Model of Economic Behavior. *The Quarterly Journal of Economics* 87, 2 (may 1973), 239. https://doi.org/10.2307/1882186
[8] Michael Doebeli and Christoph Hauert. 2005. Models of cooperation based on the Prisoner's Dilemma and the Snowdrift game. *Ecology letters* 8, 7 (2005), 748–766.
[9] Tobias Galla. 2009. Intrinsic Noise in Game Dynamical Learning. *Physical Review Letters* 103, 19 (nov 2009). https://doi.org/10.1103/physrevlett.103.198702
[10] Tobias Galla. 2011. Cycles of cooperation and defection in imperfect learning. *Journal of Statistical Mechanics: Theory and Experiment* 2011, 08 (aug 2011), P08007. https://doi.org/10.1088/1742-5468/2011/08/p08007
[11] Tobias Galla and J. Doyne Farmer. 2013. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences* 110, 4 (jan 2013), 1232–1236. https://doi.org/10.1073/pnas.1109672110
[12] Daniel Hennes, Michael Kaisers, and Karl Tuyls. 2010. RESQ-learning in stochastic games. In *Adaptive and Learning Agents Workshop at AAMAS*. 8.
[13] Daniel Hennes, Karl Tuyls, and Matthias Rauterberg. 2009. State-coupled replicator dynamics. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*. 789–796.
[14] Joseph Henrich and Robert Boyd. 2002. On modeling cognition and culture. *Journal of Cognition and Culture* 2, 2 (2002), 87–112.
[15] Josef Hofbauer and Karl Sigmund. 1998. *Evolutionary games and population dynamics*. Cambridge university press.
[16] Josef Hofbauer and Karl Sigmund. 2003. Evolutionary game dynamics. *Bulletin of the American mathematical society* 40, 4 (2003), 479–519.
[17] Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. In *Reinforcement learning*. Springer, 45–73.
[18] Martin A Nowak and Karl Sigmund. 2004. Evolutionary dynamics of biological games. *science* 303, 5659 (2004), 793–799.
[19] Jorge M Pacheco, Francisco C Santos, Max O Souza, and Brian Skyrms. 2009. Evolutionary dynamics of collective action in N-person stag hunt dilemmas. *Proceedings of the Royal Society B: Biological Sciences* 276, 1655 (2009), 315–321.
[20] Fernando P Santos, Francisco C Santos, Ana Paiva, and Jorge M Pacheco. 2015. Evolutionary dynamics of group fairness. *Journal of theoretical biology* 378 (2015), 96–102.
[21] 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). https://doi.org/10.1103/physreve.67.015206
[22] Karl Tuyls and Ann Nowé. 2005. Evolutionary game theory and multi-agent reinforcement learning. *The Knowledge Engineering Review* 20, 1 (2005), 63–90.
[23] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. 2003. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. ACM, 693–700.
[24] Peter Vrancx, Karl Tuyls, and Ronald Westra. 2008. Switching dynamics of multi-agent learning. In *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent systems (AAMAS 2008)*. 307–313.
[25] Shangtong Zhang and Richard Sutton. 2018. A Deeper Look at Experience Replay. *arXiv preprint arXiv:1712.01275* (2018).