Experience-replay Innovative Dynamics

Extended Abstract

Tuo Zhang
University of Birmingham
Birmingham, United Kingdom
txz257@student.bham.ac.uk

Leonardo Stella
University of Birmingham
Birmingham, United Kingdom
l.stella@bham.ac.uk

Julian Barreiro-Gomez
Khalifa University
Abu Dhabi, United Arab Emirates
julian.barreirogomez@ku.ac.ae

ABSTRACT

Multi-agent reinforcement learning (MARL) has achieved groundbreaking success in recent years. Yet, several open problems remain, including nonstationarity and instability. Evolutionary game theory (EGT) provides a theoretical framework to tackle instability by leveraging the properties of its most well-known model, namely, the replicator dynamics, for theoretical guarantees of convergence to Nash equilibria. However, these guarantees do not hold true in certain settings, e.g., zero-sum games. In contrast, innovative dynamics, such as the Brown-von Neumann-Nash (BNN) or Smith, retain the convergence guarantees in these settings. We develop a novel MARL algorithm based on innovative dynamics with a sampling process that resembles experience replay. We show that our approach is theoretically grounded as other state-of-the-art MARL algorithms, but most importantly it outperforms other approaches in the case of nonstationary environments.

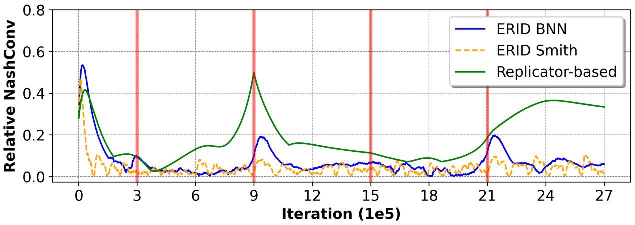
KEYWORDS

Evolutionary game theory, multi-agent systems, reinforcement learning.

ACM Reference Format:
Tuo Zhang, Leonardo Stella, and Julian Barreiro-Gomez. 2025. Experience-replay Innovative Dynamics: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

In recent years, multi-agent reinforcement learning (MARL) has demonstrated groundbreaking success across various domains such as real-time strategy games and Go, as well as robot control, cyber-physical systems, finance, and sensor networks, where numerous agents interact within complex environments [1, 4, 10, 11, 14, 21]. Despite its success, MARL still faces several open problems, including nonstationarity and instability. The former is induced by the change in policy as the agents act and learn concurrently. Indeed, the rewards that each agent receives are determined not only through its policy, but also through the policies of other agents [25, 26]. The latter affects the ability of MARL algorithms to achieve optimality. To this end, providing theoretical guarantees of convergence under general conditions is paramount.

Evolutionary game theory (EGT) studies the evolution of strategic interactions in a population of decision-makers, where the fitness of a strategy increases based on the success of that strategy in a given environment [15, 16, 23]. EGT has played a critical role in the analysis and evaluation of MARL algorithms in complex multi-agent environments. Formal connections between EGT and MARL dates back to late '90s [2], when it was demonstrated that, with a sufficiently small learning rate, the learning trajectories of *cross learning* [5] – a stateless MARL algorithm – converge to the trajectories of the *replicator dynamics*, the most well-studied model in EGT. This formal link has attracted increasing interest as it allows researchers to analyse MARL and its stochastic learning processes through the deterministic framework of replicator dynamics.



;

Figure 1: Policy NashConv of BNN-based and Smith-based ERID, and Replicator-based learning in nonstationary RPS.

Nevertheless, the majority of these studies have concentrated solely on replicator dynamics and its variants. However, it is well known that replicator dynamics do not to converge in certain game settings. We refer to the two main families of games as identified in [7]: *strictly stable games* and *null-stable games*. In strictly stable games, replicator dynamics can asymptotically converge to the Nash equilibrium, whereas in null-stable games they form closed orbits around the Nash equilibrium at a proximity that depends on the initial conditions. A particularly important class of null-stable games is zero-sum games. As a result, approaches based on replicator dynamics often rely on time-averaging to ensure convergence to the Nash equilibrium [8, 13, 22]. However, this method has a significant limitation because of the cumulative nature of the time average, which affects the ability of the dynamics to adapt to nonstationarities. Specifically, when the environment changes, the policy of the agents may require exponential time to adapt to the new conditions as is the case, e.g., of feedback-evolving games [18–20, 24, 27].

To overcome the limitations of replicator dynamics, we turn our attention to learning algorithms based on innovative dynamics [6], a family of dynamics that includes Brown-von Neumann-Nash (BNN) [3] and Smith [17] dynamics. Innovative dynamics, in contrast to replicator dynamics, converge to the Nash equilibrium in null-stable games [7]. However, their application to learning tasks is not as straightforward as with replicator dynamics. In dynamic environments with discrete changes and stochastic processes, replicator dynamics allow multi-step sampling to remain unbiased with

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

respect to the underlying fitness-based equation. Hence, similar update mechanisms cannot directly be used in innovative dynamics, which makes it necessary to find an alternative approach to ensure unbiased sampling.

Contribution. In this paper, we introduce a novel algorithm, *Experience-replay Innovative Dynamics (ERID)*, based on innovative dynamics via experience-replay. We show that the learning trajectories of ERID converge to the corresponding base dynamics through the choice of the protocol factor. This enables ERID to benefit from the convergence guarantees of these dynamics in strictly stable and null-stable games. Experience-replay is a reinforcement learning mechanism used to enhance learning efficiency and stability by storing and reusing past experiences during the training process [9]. Its basic principle involves maintaining a memory buffer that records state transitions, actions, and rewards, which are then sampled randomly during training to decorrelate consecutive experiences and smooth the learning process [12]. In our case, experience-replay is used as a batch that moves in time along the trajectory. Its aim is to reduce sample variance by mixing rewards from each step with historical rewards. This mitigates the influence of the non-linear revision protocol and ensures that the algorithm aligns with the desired underlying dynamics.

As a motivating example, consider the nonstationary rock paper scissors (RPS) game. We use Relative NashConv, a normalised metric that measures the distance from the NE, to compare a Replicator-based approach with ERID. As depicted in Fig. 1, ERID outperforms these approaches in the presence of nonstationarities.

In order to use the stored experiences for the policy updates, it is essential to calculate the average rewards associated with each action and the overall average reward across all actions. These two quantities are defined as:

$$\bar{r}_i = \begin{cases} \dfrac{1}{|I_i|} \displaystyle\sum_{j \in I_i} b_j, & \text{if } I_i \neq \emptyset, \\ 0, & \text{otherwise,} \end{cases} \qquad \bar{r} = \frac{1}{K} \sum_{j=1}^{K} b_j. \tag{1}$$

In EGT, each revision protocol ρ_{ij} corresponds to a specific set of evolutionary dynamics, and determines how the probabilities of choosing different strategies change over time. In our MARL framework, we introduce the protocol factor η_{ij}, which is computed based on the average rewards as defined in (1). For each specific set of dynamics, the corresponding ρ_{ij} can be mapped to a specific η_{ij}. The specific mapping involves replacing the fitness values in ρ_{ij} with the corresponding reward values. We can now present the update formula for the policy $\pi_i(t)$:

$$\pi_i(t+1) \leftarrow \pi_i(t) + \alpha \left(\sum_{j=1}^{M} \pi_j(t) \eta_{ji} - \pi_i(t) \sum_{j=1}^{M} \eta_{ij} \right), \tag{2}$$

where α is the learning rate. The pseudocode of the algorithm corresponding to the above policy update is given in Algorithm 1.

Algorithm 1 Experience-replay Innovative Dynamics

```
Require: Initial policy $\pi_0$, buffer size $K$, learning rate $\theta$
Initialize: Policy $\pi_0 \leftarrow$ initial strategy, buffer $B \leftarrow$ empty
for $t = 1, 2, \ldots$ do
    if $t > K$ then
        $B \leftarrow \text{shift}(B, (a, r))$
    else
        $B \leftarrow B \cup (a, r)$
    $\bar{r}_i \leftarrow \text{getAverageReward}(B, i)$
    $\bar{r} \leftarrow \text{getOverallAverageReward}(B)$
    $\pi_t \leftarrow \text{updatePolicy}(\pi_t, \bar{r}_i, \bar{r}, \theta)$
```

To validate our approach, let us consider the biased RPS game. The only difference between the biased RPS game and the standard RPS game is that the payoff of the rock-paper matchups is scaled by a factor of 2. Figure 2 shows a comparison between the evolutionary dynamics and ERID in the biased RPS game. The top-left plot depicts the trajectories of the BNN dynamics, while the bottom-left plot shows the evolution of the Smith dynamics. On the right-hand side of the figure, the top-right plot depicts the results from simulations using the BNN-based ERID algorithm with a step size of $1e - 5$ and a buffer size of 1000, whereas the bottom-left plot shows the policy of the Smith-based ERID algorithm. We observe that the ERID-generated trajectories on the right closely follow the trajectories of the corresponding innovative dynamics. Since both BNN and Smith dynamics are known to converge to the Nash equilibrium in zero-sum games, we can see that the corresponding ERID trajectories also converge to the NE, given small perturbations due to the stochastic nature of the learning algorithm.

CONCLUSION

In this paper, we have proposed a novel algorithm based on innovative dynamics, ERID, which is able to adapt to dynamic changes in the environment more effectively than traditional approaches based on replicator dynamics and their time-averaged counterpart. We have demonstrated that our algorithm converges to the corresponding dynamics, making it theoretically grounded, and showed that it outperforms replicator-based algorithms in nonstationary environments.

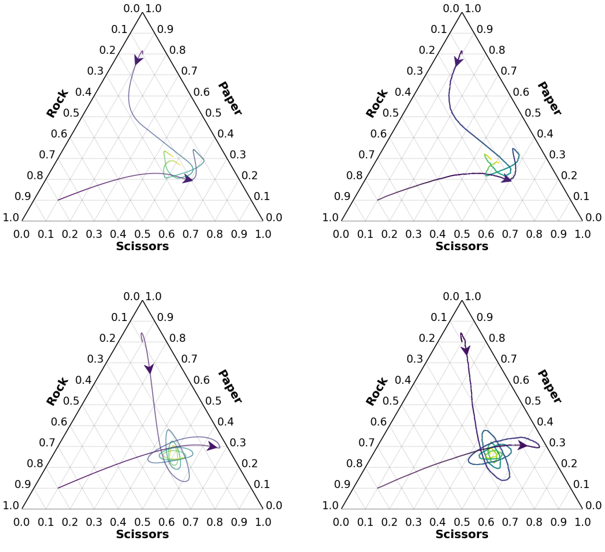


Figure 2: Innovative dynamics (left) vs policy trajectories (right) of the corresponding ERID algorithm.

REFERENCES

[1] Jeffrey L Adler and Victor J Blue. 2002. A cooperative multi-agent transportation management and route guidance system. *Transportation Research Part C: Emerging Technologies* 10, 5-6 (2002), 433–454.

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

[3] George W Brown and John Von Neumann. 1950. *Solutions of games by differential equations*. Rand Corporation.

[4] Jorge Cortes, Sonia Martinez, Timur Karatas, and Francesco Bullo. 2004. Coverage control for mobile sensing networks. *IEEE Transactions on robotics and Automation* 20, 2 (2004), 243–255.

[5] John G Cross. 1973. A stochastic learning model of economic behavior. *The quarterly journal of economics* 87, 2 (1973), 239–266.

[6] Josef Hofbauer. 2011. Deterministic evolutionary game dynamics. (2011).

[7] Josef Hofbauer and William H Sandholm. 2009. Stable games and their dynamics. *Journal of Economic theory* 144, 4 (2009), 1665–1693.

[8] Josef Hofbauer, Sylvain Sorin, and Yannick Viossat. 2009. Time average replicator and best-reply dynamics. *Mathematics of Operations Research* 34, 2 (2009), 263–269.

[9] Sascha Lange, Thomas Gabel, and Martin Riedmiller. 2012. Batch reinforcement learning. In *Reinforcement learning: State-of-the-art*. Springer, 45–73.

[10] Jae Won Lee, Jonghun Park, O Jangmin, Jongwoo Lee, and Euyseok Hong. 2007. A multiagent approach to q-learning for daily stock trading. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 37, 6 (2007), 864–877.

[11] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971* (2015).

[12] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[13] William H Sandholm, Emin Dokumacı, and Ratul Lahkar. 2008. The projection dynamic and the replicator dynamic. *Games and Economic Behavior* 64, 2 (2008), 666–683.

[14] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[15] J Maynard Smith. 1974. The theory of games and the evolution of animal conflicts. *Journal of theoretical biology* 47, 1 (1974), 209–221.

[16] J Maynard Smith and George R Price. 1973. The logic of animal conflict. *Nature* 246, 5427 (1973), 15–18.

[17] Michael J Smith. 1984. The stability of a dynamic model of traffic assignment—an application of a method of Lyapunov. *Transportation science* 18, 3 (1984), 245–252.

[18] Leonardo Stella, Wouter Baar, and Dario Bauso. 2022. Lower network degrees promote cooperation in the prisoner's dilemma with environmental feedback. *IEEE Control Systems Letters* 6 (2022), 2725–2730.

[19] Leonardo Stella and Dario Bauso. 2023. The impact of irrational behaviors in the optional prisoner's dilemma with game-environment feedback. *International Journal of Robust and Nonlinear Control* 33, 9 (2023), 5145–5158.

[20] Andrew R Tilman, Joshua B Plotkin, and Erol Akçay. 2020. Evolutionary games with environmental feedbacks. *Nature communications* 11, 1 (2020), 915.

[21] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[22] Yannick Viossat and Andriy Zapechelnyuk. 2013. No-regret dynamics and fictitious play. *Journal of Economic Theory* 148, 2 (2013), 825–842.

[23] Jörgen W Weibull. 1997. *Evolutionary game theory*. MIT press.

[24] 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.

[25] Yaodong Yang and Jun Wang. 2020. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583* (2020).

[26] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control* (2021), 321–384.

[27] Tuo Zhang, Harsh Gupta, Kumar Suprabhat, and Leonardo Stella. 2023. A multi-agent reinforcement learning approach to promote cooperation in evolutionary games on networks with environmental feedback. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 2196–2201.