Enabling Option Learning in Sparse Rewards with Hindsight Experience Replay

Extended Abstract

Gabriel Romio
Universidade do Vale do Rio dos Sinos
São Leopoldo, RS, Brazil
gromio@edu.unisinos.br

Mateus Begnini Melchiades
Universidade do Vale do Rio dos Sinos
São Leopoldo, RS, Brazil
mateusbme@edu.unisinos.br

Bruno Castro da Silva
University of Massachusetts, Amherst
Amherst, MA, USA
bsilva@cs.umass.edu

Gabriel de Oliveira Ramos
Universidade do Vale do Rio dos Sinos
São Leopoldo, RS, Brazil
gdoramos@unisinos.br

ABSTRACT

Hierarchical Reinforcement Learning (HRL) algorithms such as Option-Critic (OC) and Multi-updates Option Critic (MOC) advance the learning of reusable options but struggle in multi-goal environments with sparse rewards. We propose MOC-HER, which integrates the Hindsight Experience Replay (HER) mechanism into MOC. By relabeling goals from achieved outcomes, MOC-HER addresses sparse reward environments that are intractable for the original MOC. However, for object manipulation tasks, where rewards are determined by object placement rather than agent-centric states, standard relabeling is often insufficient. We introduce Dual Objectives Hindsight Experience Replay (2HER), which augments goal relabeling with virtual goals from the agent's effector positions. This encourages both effective object interaction and task completion. In robotic manipulation tasks, MOC-2HER achieves success rates up to 90%, compared to under 11% for MOC and MOC-HER.

KEYWORDS

Hierarchical Reinforcement Learning; Options; Sparse Rewards; Multi-Goal Environments; Hindsight Experience Replay

ACM Reference Format:
Gabriel Romio, Mateus Begnini Melchiades, Bruno Castro da Silva, and Gabriel de Oliveira Ramos. 2026. Enabling Option Learning in Sparse Rewards with Hindsight Experience Replay: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/HIWW4582

1 INTRODUCTION

Hierarchical Reinforcement Learning (HRL) seeks to structure complex tasks by decomposing them into subtasks through temporal abstraction, enabling decision-making at multiple time scales [3, 5, 7, 12]. The Options Framework is a key mechanism in this context, representing reusable skills as sub-policies that execute extended action sequences [11]. Methods such as Option-Critic [2] and its extension, Multi-updates Option Critic (MOC) [6], enable the automatic learning of hierarchical policies and improve performance in complex environments. However, their effectiveness is significantly reduced in sparse-reward settings [5].

The Hindsight Experience Replay (HER) technique [1] improves learning in sparse-reward multi-goal reinforcement learning [9] by recomputing rewards using achieved goals, enabling agents to extract informative feedback from unsuccessful trajectories. Based on this principle, we integrate HER into the HRL scenario, using MOC as the underlying framework. However, object manipulation tasks present additional challenges, since the reward signal depends exclusively on the final state of the object. While HER is effective in standard reinforcement learning, it is insufficient in hierarchical settings, where an agent must learn and coordinate multiple policies [8]. To address this limitation, we propose Dual Objectives Hindsight Experience Replay (2HER), which introduces an additional objective by capturing agent–object interactions. These effector-based goals provide auxiliary success signals, enabling more structured option learning.

We evaluate MOC-HER and MOC-2HER in Fetch Robotics environments [9], a standard benchmark for multi-goal robotic manipulation. The results demonstrate that both HER and 2HER can solve sparse-reward environments that are challenging for conventional HRL approaches. Specifically, in object manipulation tasks, our 2HER-augmented method achieved a success rate of up to 90%, a significant improvement over the 11% achieved by the original algorithm. To the best of our knowledge, this is the first approach to tackle sparse rewards using the MOC framework.[1]

2 METHOD

We incorporated a HER mechanism into the MOC framework to improve performance in multi-goal environments with sparse rewards while preserving MOC's original architecture. As the agent interacts with the environment, each state transition is stored in a replay buffer containing the executed action, the active option, the observed state, the resulting state, and the received reward. The observation space includes a *desired goal* (the target) and an *achieved goal* (the position reached by the agent).

[1] A full version of this work can be found in [10].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/HIWW4582

After each episode, we process the stored trajectory with HER. To define the new goals, we use the strategy $\mathbb{S} = \texttt{future}$, chosen for its strong performance in the original HER experiments [1]. According to this strategy, for each transition $(s_t, o_t, a_t, r_t, s_{t+1})$ in the original trajectory, we randomly sample a set of k additional goals $\{g'_1, ..., g'_k\}$ from future states of the episode. The reward is then reevaluated for each transition based on these new goals, utilizing the reward function inherent to the selected environment. For the sparse reward setting considered, the value is 0 if the achieved goal matches the desired goal in a given transition, and -1 otherwise.

We introduce the 2HER extension to improve learning in object manipulation scenarios with sparse rewards. In such tasks, the agent typically receives a reward signal only when an object reaches its target position, making it difficult to learn the prerequisite interaction and manipulation behaviors. The 2HER extension addresses this limitation by creating a second type of virtual goal in addition to the conventional HER goal based on the object's future positions. Specifically, while standard HER relabels goals using future object states, 2HER additionally samples virtual goals from the agent's future end-effector positions, which replace the object position in the observation space. In summary, the agent is retrospectively rewarded for having moved the manipulator to a location where the object could have been.

By combining rewards from two distinct hindsight goals, one for agent-object interaction and another for task completion, we create a composite learning signal. This signal incentivizes the agent to engage with the object, facilitating the discovery of interaction dynamics while simultaneously learning to move the object to its target. The task completion reward, $r_{\text{goal}}(s_{t+1})$, is 0 if $d(p_{\text{obj}}(s_{t+1}), g) \leq \epsilon$, and -1 otherwise. Similarly, the interaction reward, $r_{\text{object}}(s_{t+1})$, is 0 if $d(p_{\text{agent}}(s_{t+1}), p_{\text{obj}}(s_{t+1})) \leq \epsilon$, and -1 otherwise. In this notation, p_{obj} and p_{agent} denote object and agent positions, g the goal position, d the Euclidean distance, and ϵ the distance threshold defining success. The overall reward combines these two components as $r(s_t, a_t, s_{t+1}) = (1 - C_r)\, r_{\text{goal}}(s_{t+1}) + C_r\, r_{\text{object}}(s_{t+1})$, with hyperparameter $C_r \in [0, 1]$ balancing their relative contributions.

Subsequently, the revisited HER buffer is merged with the original MOC experience buffer to update all options, after which both buffers are cleared to be repopulated as the training continues. To improve learning efficiency and stability, we introduce two additional modifications to HER: (i) hindsight relabeling is restricted to trajectories exhibiting meaningful object interaction, identified by a minimum object displacement between the initial and final states, and (ii) the number of relabeled goals per transition is scheduled by initializing k with a high value and gradually decaying it over training to balance learning speed and stability.[2]

3 EXPERIMENTS

We evaluate MOC-HER and MOC-2HER in the Fetch tasks from Gymnasium Robotics [9]. In these environments, the goal changes in each episode, and the reward follows a sparse formulation. An episode is considered successful only if the object is on the target at the end of the episode [1]. Furthermore, we integrated HER and 2HER into the IOC [4] algorithm, establishing a new baseline for comparison against our approaches.

[2]Our full code is available at https://github.com/ramos-ai/MOC_2HER.

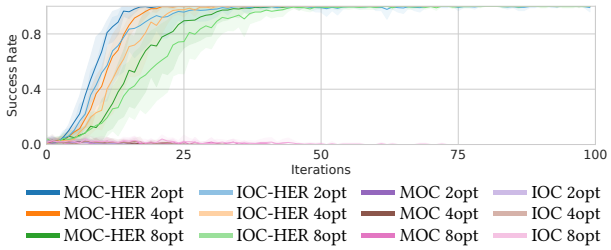

Figure 1: Comparison of different approaches in FetchReach.

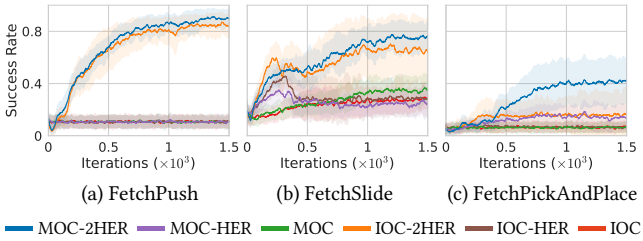

Figure 2: Comparison of different approaches with 4 options in FetchPush (a), FetchSlide (b) and FetchPickAndPlace (c).

Our first experiment compares MOC-HER and IOC-HER with the standard algorithms in the FetchReach positioning task [9], considering different numbers of options. The results are shown in Figure 1, where the shaded areas denote the standard deviation across 10 runs. The HER-based extensions consistently solve the task, with MOC-HER converging faster than IOC-HER. In contrast, the MOC and IOC algorithms fail to learn a viable policy within the analyzed training horizon. Moreover, the required number of training steps increases with the number of options, reflecting the additional complexity of coordinating multiple hierarchical policies.

Our subsequent experiments focus on object manipulation tasks, with the 2HER extension enabled and a secondary objective incorporated into the hindsight process. We compare MOC-2HER and IOC-2HER with the MOC and IOC baselines and their single-objective HER variants. Results for FetchPush, FetchSlide, and FetchPickAndPlace tasks with four options are presented in Figure 2. While MOC, IOC, and their HER variants exhibit limited and inconsistent performance, the 2HER-augmented approaches consistently outperform their respective baselines across all benchmarks.

4 CONCLUSION

In this work, we addressed the challenge of applying HRL in multi-goal environments with sparse rewards. We introduced MOC-HER, integrating HER into the MOC framework, and 2HER, a novel mechanism generating virtual goals for both the object and the actuator. Experiments demonstrate that algorithms augmented with HER and 2HER significantly outperform their original counterparts.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable feedback. This work was supported by Kunumi Institute. The authors thank the institution for its financial support and commitment to advancing scientific research. This research was also partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico

- CNPq (grants 313845/2023-9, 443184/2023-2, 445238/2024-0, and 404800/2025-4).

REFERENCES

[1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, and Wojciech Zaremba. 2017. Hindsight Experience Replay. In *Advances in Neural Information Processing Systems*, Vol. 30.

[2] Pierre-Luc Bacon, Jean Harb, and Doina Precup. 2017. The Option-Critic Architecture. *Proc. of the AAAI Conference on Artificial Intelligence* 31, 1 (2017).

[3] Peter Dayan and Geoffrey E Hinton. 1993. Feudal reinforcement learning. *Advances in neural information processing systems* 5 (1993).

[4] Khimya Khetarpal, Martin Klissarov, Maxime Chevalier-Boisvert, Pierre-Luc Bacon, and Doina Precup. 2020. Options of Interest: Temporal Abstraction with Interest Functions. *Proc. of the AAAI Conference on Artificial Intelligence* 34, 04 (2020), 4444–4451.

[5] Martin Klissarov, Akhil Bagaria, Ziyan Luo, George Konidaris, Doina Precup, and Marlos C Machado. 2025. Discovering Temporal Structure: An Overview of Hierarchical Reinforcement Learning. *arXiv preprint arXiv:2506.14045* (2025).

[6] Martin Klissarov and Doina Precup. 2021. Flexible Option Learning. In *Advances in Neural Information Processing Systems*, Vol. 34. 4632–4646.

[7] Ofir Nachum, Haoran Tang, Xingyu Lu, Shixiang Gu, Honglak Lee, and Sergey Levine. 2019. Why does hierarchy (sometimes) work so well in reinforcement learning? *arXiv:1909.10618* (2019).

[8] Shubham Pateria, Budhitama Subagdja, Ah-Hwee Tan, and Chai Quek. 2021. Hierarchical Reinforcement Learning: A Comprehensive Survey. *Comput. Surveys* 54 (2021).

[9] Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Peter Welinder, et al. 2018. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv:1802.09464* (2018).

[10] Gabriel Romio, Mateus Begnini Melchiades, Bruno Castro da Silva, and Gabriel de Oliveira Ramos. 2026. Enabling Option Learning in Sparse Rewards with Hindsight Experience Replay. arXiv:2602.13865 [cs.AI] https://arxiv.org/abs/2602.13865

[11] Richard S. Sutton and Andrew G. Barto. 2020. Reinforcement learning: An introduction (2nd ed.). The MIT Press.

[12] Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112, 1 (1999), 181–211.