Timing Optimization in Dynamic Discrete Action Space Lifelong Reinforcement Learning

Extended Abstract

Pangjing Wu*
College of Computer Science and Software Engineering,
Hohai University
Nanjing, China
Department of Computing,
The Hong Kong Polytechnic University
Hong Kong, China
pang-jing.wu@connect.polyu.hk

Xiaodong Li†
College of Computer Science and Software Engineering,
Hohai University
Nanjing, China
xiaodong.li@hhu.edu.cn

ABSTRACT

Dynamic discrete action spaces pose a core challenge for lifelong reinforcement learning (LRL), where agents must continually adapt as the set of available actions expands over time. Existing approaches typically follow fixed update cycles, which either introduce new actions prematurely or delay policy improvement. To overcome this challenge, we propose *Timing Optimization Lifelong Reinforcement Learning* (TO-LRL), a framework that treats the timing of action space expansion as a decision variable. Instead of passively following predefined cycles, TO-LRL leverages regret-driven signals of exploration sufficiency to determine the optimal timing for expanding the action space, ensuring that new options are introduced only when they can accelerate policy improvement rather than destabilize it. We establish regret bounds for both bandit and MDP settings, showing that lifelong regret depends jointly on interaction budgets and the effective action space. Empirically, we evaluate TO-LRL on lifelong multi-armed bandits, treasure hunting, and algorithmic trading. Across these tasks with dynamic discrete action spaces, TO-LRL consistently outperforms fixed-cycle and adaptive baselines, achieving lower cumulative regret, faster convergence, and more stable long-term returns. These findings highlight timing optimization as a principled and effective strategy for dynamic discrete action LRL.

KEYWORDS

Dynamic Discrete Action Space; Lifelong Learning; Reinforcement Learning

ACM Reference Format:
Pangjing Wu and Xiaodong Li. 2026. Timing Optimization in Dynamic Discrete Action Space Lifelong Reinforcement Learning: 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/NLQA5070

*This work was done when the author studied at Hohai University.
†Corresponding Author.

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/NLQA5070

1 INTRODUCTION

Lifelong reinforcement learning (LRL) has become an increasingly important paradigm for building adaptive agents that can operate in non-stationary environments. Unlike classical reinforcement learning (RL), which assumes a stationary setting with a fixed set of states and actions, LRL must handle environments where tasks, dynamics, or available actions evolve [6, 9, 11]. Such adaptability is critical for domains ranging from robotics to finance, where the decision space itself is dynamic.

Gap. Existing work addresses dynamic discrete action space (DDAS) through representation transfer or continual adaptation, including zero-shot and embedding-based action generalization [4, 5, 12], incremental action space growth [1, 2, 8], and decoupled or replay-based policy architectures [7, 10, 14]. However, **nearly all existing methods leverage fixed cycles** to incorporate new actions. It introduces timing dilemmas: (i) expanding too frequently prevents sufficient exploration of actions, thus inflating exploration costs; (ii) expanding too infrequently delays the use of beneficial actions and slows policy improvement. In short, existing methods primarily optimize *how* to learn under a given action space, while the question of *when* to expand remains unexplored.

Key idea. We treat *expansion timing* as a first-class decision variable in DDAS-LRL. Instead of relying on predefined schedules or heuristic triggers, TO-LRL explicitly evaluates whether the *current action set has been sufficiently explored*. To this end, TO-LRL employs optimism-based exploration and tracks an aggregated uncertainty signal across actions. When exploration is inadequate, this signal exhibits high variance and non-stationarity; as uncertainty diminishes, it stabilizes. TO-LRL uses this stabilization as a principled indicator of *exploration sufficiency*. By aligning expansion decisions with the evolving confidence in the learning process, TO-LRL introduces new actions when they are most likely to accelerate policy improvement rather than increase exploration costs or disrupt learning stability.

Contributions. Our contributions are three-fold: (i) We propose *TO-LRL*, which schedules action space expansion via exploration sufficiency rather than a fixed cycle. (ii) We provide regret analysis of the TO-LRL in DDAS scenarios, highlighting how regret depends jointly on interaction budgets and the *effective* action space size accumulated over time. (iii) We demonstrate consistent empirical gains on three representative DDAS-LRL scenarios.

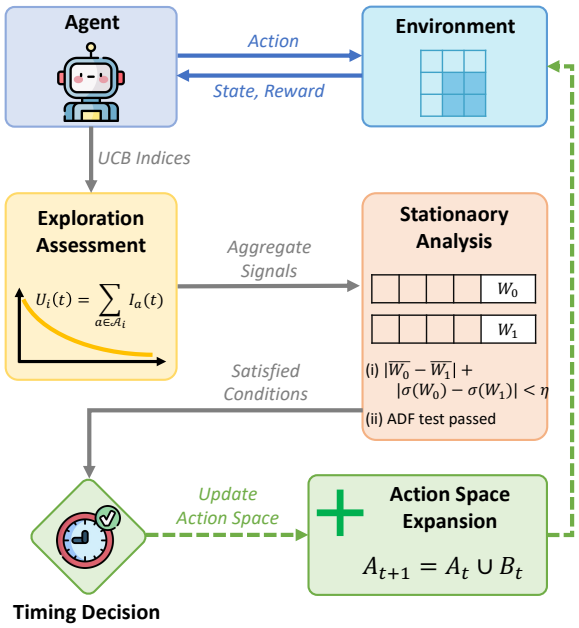

Figure 1: The framework of TO-LRL.

2 TO-LRL FRAMEWORK

To bridge the aforementioned gap, we propose **TO-LRL**, a framework that treats the timing of action space expansion as a decision variable rather than a pre-defined schedule. As illustrated in Figure 1, it consists of two interacting components:

- **Exploration Assessment Module:** evaluates whether the current action set has been sufficiently explored.
- **Timing Decision Module:** triggers expansion only when exploration reaches statistical stability.

This design ensures that new actions are introduced precisely when their inclusion accelerates learning rather than destabilizes it.

2.1 Exploration Assessment Module

For each action $a \in \mathcal{A}_i$, the agent maintains the UCB index,

$$I_a(t) = \widehat{\mu}_a(t) + U_a(t), \tag{1}$$

where $\widehat{\mu}_a(t)$ denotes the empirical mean reward and $U_a(t)$ represents the confidence bonus. To capture the exploration status of the entire stage, we aggregate the indices by,

$$\mathcal{U}_i(t) = \sum_{a \in \mathcal{A}_i} I_a(t). \tag{2}$$

At the start of a stage, $\mathcal{U}_i(t)$ fluctuates due to high uncertainty; as exploration proceeds, $U_a(t)$ diminishes and $\mathcal{U}_i(t)$ stabilizes. The transition from non-stationarity to stationarity thus encodes whether the agent has sufficiently explored the current action set.

2.2 Timing Decision Module

The expansion decision relies on detecting the stabilization of the signal sequence $\{\mathcal{U}_i(t)\}_{t \geq 1}$. Let $W_0 = \{\mathcal{U}_i(t-2l+1), \ldots, \mathcal{U}_i(t-l)\}$ and $W_1 = \{\mathcal{U}_i(t-l+1), \ldots, \mathcal{U}_i(t)\}$ as two adjacent sliding windows of length l. Exploration sufficiency is declared when,

$$\left|\overline{W_1} - \overline{W_0}\right| + \left|\sigma(W_1) - \sigma(W_0)\right| < \eta, \tag{3}$$

$$\text{ADF}(W_0 \cup W_1) \text{ rejects the unit-root hypothesis.} \tag{4}$$

Here $\eta > 0$ is a stability threshold, $\overline{W_j}$ and $\sigma(W_j)$ denote the mean and variance of window W_j, and ADF$(\cdot)$ refers to the Augmented Dickey–Fuller test for stationarity. The expansion time for stage i is thus defined as,

$$\tau_i = \min\{t : \mathcal{U}_i(t) \text{ satisfies (3) and (4)}\}. \tag{5}$$

Once τ_i is reached, the timing module expands the action space as $\mathcal{A}_{i+1} = \mathcal{A}_i \cup \mathcal{B}_i$. This rule embodies the principle *"expand only after optimism stabilizes,"* aligning timing decisions with the confidence structure of the exploration process. By aligning expansion timing with the statistical stability of exploration, TO-LRL converts a fixed heuristic schedule into an adaptive process.

3 EMPIRICAL HIGHLIGHTS

We evaluate TO-LRL on three DDAS-LRL scenarios: **(i) lifelong multi-armed bandits** with new arms appearing over time, **(ii) treasure hunting** where expansions correspond to more candidate locations, and **(iii) algorithmic trading**, which introduces additional order sizes or price levels dynamically. Baselines include two fixed-cycle methods and three adaptive heuristics that update based on reward trends or exploitation-phase detection [3, 8, 13].

Bandits. Across lifelong bandit settings, TO-LRL consistently achieves the lowest cumulative regret, reducing regret by roughly 10–20% compared to the strongest adaptive baseline. The gains are more pronounced when the exploration budget is tight, reflecting TO-LRL's ability to delay expansion until existing arms are sufficiently explored.

Treasure hunting. TO-LRL converges faster and more smoothly than fixed-cycle and heuristic adaptive methods. By avoiding premature expansions in short-horizon, sparse-reward episodes, it reduces search costs and improves cumulative reward.

Trading. In algorithmic trading, TO-LRL yields more stable long-horizon returns. Fixed schedules either disrupt profitable policies by expanding too early or too late, while TO-LRL expands only after the existing policy stabilizes.

4 CONCLUSION

This work introduced TO-LRL, a framework that elevates timing to a first-class decision variable in DDAS-LRL. By coupling UCB-based exploration with stationarity tests, TO-LRL adaptively schedules action space expansions to minimize lifelong regret. Our analysis establishes formal regret bounds, and experiments from ideal to practical scenarios demonstrate consistent gains in efficiency, stability, and cumulative reward. These findings show that *when* to expand the action space is as critical as *how* to act, positioning timing optimization as a new principle for advancing LRL.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China, No. 2024YFC3210800. We thank Ms. Jiaqi Ye for helpful discussions.

REFERENCES

[1] Craig Boutilier, Alon Cohen, Avinatan Hassidim, Yishay Mansour, Ofer Meshi, Martin Mladenov, and Dale Schuurmans. 2018. Planning and learning with stochastic action sets. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 4674–4682.

[2] Yash Chandak, Georgios Theocharous, Chris Nota, and Philip Thomas. 2020. Lifelong learning with a changing action set. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 3373–3380.

[3] Will Dabney, Georg Ostrovski, and Andre Barreto. 2020. Temporally-Extended ε-Greedy Exploration. In *International Conference on Learning Representations*.

[4] Ayush Jain, Norio Kosaka, Kyung-Min Kim, and Joseph J Lim. 2021. Know your action set: Learning action relations for reinforcement learning. In *International Conference on Learning Representations*.

[5] Ayush Jain, Andrew Szot, and Joseph Lim. 2020. Generalization to New Actions in Reinforcement Learning. In *International Conference on Machine Learning*. PMLR, 4661–4672.

[6] Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. 2022. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research* 75 (2022), 1401–1476.

[7] Anthony Kobanda, Rémy Portelas, Odalric-Ambrym Maillard, and Ludovic Denoyer. 2025. Hierarchical Subspaces of Policies for Continual Offline Reinforcement Learning. In *ICLR-MCDC 2025-Workshop on Modularity for Collaborative, Decentralized, and Continual Deep Learning*.

[8] Travis Mandel, Yun-En Liu, Emma Brunskill, and Zoran Popović. 2017. Where to add actions in human-in-the-loop reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[9] Sindhu Padakandla. 2021. A survey of reinforcement learning algorithms for dynamically varying environments. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–25.

[10] Chaofan Pan, Jiafen Liu, Yanhua Li, Linbo Xiong, Fan Min, Wei Wei, Tianrui Li, and Xin Yang. 2024. ARC-RL: Self-Evolution Continual Reinforcement Learning via Action Representation Space. https://openreview.net/forum?id=M9p2SIq0Oj

[11] Chaofan Pan, Xin Yang, Yanhua Li, Wei Wei, Tianrui Li, Bo An, and Jiye Liang. 2025. A Survey of Continual Reinforcement Learning. *arXiv preprint arXiv:2506.21872* (2025).

[12] Brandon Trabucco, Mariano Phielipp, and Glen Berseth. 2022. Learning Transferable Policies By Inferring Agent Morphology. In *ICLR 2022 Workshop on Generalizable Policy Learning in Physical World*.

[13] Lieping Zhang, Liu Tang, Shenglan Zhang, Zhengzhong Wang, Xianhao Shen, and Zuqiong Zhang. 2021. A self-adaptive reinforcement-exploration Q-learning algorithm. *Symmetry* 13, 6 (2021), 1057.

[14] Tiantian Zhang, Kevin Zehua Shen, Zichuan Lin, Bo Yuan, Xueqian Wang, Xiu Li, and Deheng Ye. 2023. Replay-enhanced Continual Reinforcement Learning. *Transactions on Machine Learning Research* (2023).