Flexible Deliberation Costs via Inversely Proportional Decay in the Option-Critic Architecture

Augusto Antônio Fontanive Leal[1], **Mateus Begnini Melchiades**[1], **Gabriel de Oliveira Ramos**[1]

[1]Universidade do Vale do Rio dos Sinos (UNISINOS)

{aaleal@edu.,mateusbme@edu.,gdoramos@}unisinos.br

Abstract. *In this paper, we propose a flexible deliberation cost for the option-critic architecture, defined as an inverse function of option duration. Unlike fixed penalties, this adaptive cost reduces hyperparameter sensitivity, enhances option specialization and stability, and prevents degeneration into primitive actions, leading to more coherent behaviors.*

Resumo. *Propõe-se um custo de deliberação flexível para a arquitetura option-critic, definido como uma função inversa da duração das options. Esse custo adaptativo reduz a sensibilidade a hiperparâmetros, melhora a especialização e estabilidade das options, evita a degeneração em ações primitivas e promove comportamentos mais coerentes.*

1. Introduction

The Option-Critic architecture [Bacon et al. 2017] is built upon reinforcement learning [Sutton and Barto 2018, p. 1-2] (where problems are typically formalized as Markov Decision Processes (MDPs), $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ [Szepesvári 2010, p. 2]), and considers temporal abstractions as a framework for autonomously learning temporally extended actions (options) [Sutton et al. 1999][Sutton et al. 1999, p. 183]. Option-Critic jointly learns intra-option policies, termination functions, and the policy over options via policy gradients, offering greater scalability and data efficiency than subgoal-based methods. To prevent options from degenerating into primitive actions, a deliberation cost η was introduced [Harb et al. 2018, p. 3165], penalizing frequent switches and encouraging longer options, though highly sensitive to the chosen hyperparameter [Harutyunyan et al. 2019].

In this paper, we introduce a flexible deliberation cost for the Option-Critic, formulated as an inversely proportional function of option duration. Our approach dynamically adapts to option terminations, reducing sensitivity to hyperparameters while discouraging excessive switching and rewarding sustained commitments. This formulation advances the literature on temporal abstraction by eliminating the need for costly hyperparameter tuning and promoting more stable and specialized options. Empirically, the method yields consistent improvements across environments: returns improved by about 35% in Cartpole and 25% in MuJoCo Ant (Table 1). Results demonstrate option specialization and higher rewards compared to the existing deliberation cost strategies.

2. Proposed Method

Our approach dynamically adjusts the deliberation cost based on the time spent in an option, making the penalty inversely proportional to its duration. To formalize this, we

introduce κ, the number of consecutive steps the agent remains committed to the same option: $\kappa \leftarrow 0$ when a new option is selected, and $\kappa \leftarrow \kappa + 1$ otherwise. This mechanism ensures that κ accurately reflects the current commitment duration and can be implemented without additional memory or structural modifications.

The adaptive deliberation cost is defined as $\eta(\kappa) = \frac{1}{1+\kappa}$, so frequent switching keeps η high, penalizing rapid option changes, while persisting in the same option reduces the penalty, encouraging temporally extended behaviors. In the Option-Critic gradient, this is incorporated as $\nabla_\vartheta J(\vartheta) \propto \frac{\partial \beta_{\omega,\vartheta}(s)}{\partial \vartheta}\big(Q_\Omega(s,\omega) - V_\Omega(s) + \eta(\kappa)\big)$, embedding the adaptive cost directly into learning and promoting option specialization while reducing sensitivity to hyperparameters. The term in red highlights our modification to the original gradient, which previously ended with a fixed η. Intuitively, frequent switching increases η, discouraging erratic option changes, whereas committing to a successful option reduces the cost over time, promoting more coherent and effective behaviors.

3. Methodology

Experiments were conducted in tabular, discrete-action, and continuous-action domains. In the tabular Four-Rooms environment [Sutton et al. 1999], the agent independently learned options to minimize steps to the goal, starting from the east doorway; ten runs per setting were performed with a learning rate of 3×10^{-3}, and the goal was relocated after 1000 steps to test adaptability, as illustrated in Figure 1 (left). In discrete-action domains, we used Cartpole [Barto et al. 1983] and Lunar Lander, each trained for ten runs per seed. Cartpole used a learning rate of 8×10^{-5} for 3000 episodes (max 500 steps), and Lunar Lander used 7×10^{-4} for 2000 episodes (max 1000 steps). For continuous actions, the algorithm was adapted to Gaussian policies in MuJoCo Ant [Schulman et al. 2018] and Half Cheetah [Wawrzyński 2009], each trained for ten runs with a learning rate of 1×10^{-4}. Ant ran for 7000 episodes (max 1000 steps) and Half Cheetah for 3000 episodes (max 1000 steps).

We evaluated three baselines: the original Option-Critic [Bacon et al. 2017], a variant with multiple fixed deliberation costs [Harb et al. 2018], and the proposed inversely proportional decay. For the fixed-cost variant, results were averaged over several runs to reduce hyperparameter sensitivity. The main metric is the average return over the last 100 episodes.

4. Numerical Results

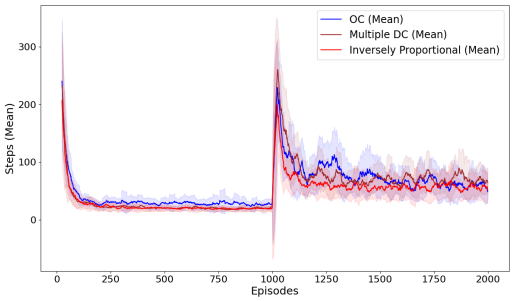
As shown in Table 1, the adaptive approach yielded the highest rewards, improving learning efficiency, option specialization, and decision stability across tasks.

Table 1. Comparison of rewards in the last 100 episodes

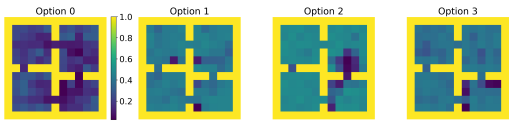
Environment	Original	Multiple DC	Inversely Prop.
Cartpole	356.59	374.75	**482.11**
Lunar Lander	191.29	84.26	**225.70**
MuJoCo Ant	1361.42	1305.14	**1702.99**
MuJoCo Half Cheetah	1422.07	1475.04	**1671.32**

In the Four-Room domain, the inversely proportional decay approach achieved higher rewards in the first 1000 steps and faster recovery after the goal change, Figure 1

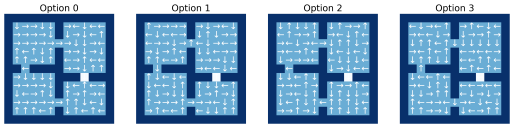
(a). Termination maps, Figure 1 (b), show that options tend to persist when approaching meaningful subgoals, such as doorways, with darker colors indicating lower termination probabilities in these regions. This behavior allows options to span across subgoal areas rather than ending prematurely. Option 0 defines a main route, while options 1, 2, and 3 adjust depending on the agent's position, complementing each other to handle deviations efficiently. The greedy policy after training, Figure 1 (c), demonstrates how these complementary options combine to guide the agent coherently from start to goal. Together, the results indicate that adaptive deliberation costs foster stable, interpretable, and temporally extended behaviors that exploit emergent subgoals.



(a) Performance in first 1000 steps and recovery after goal change.



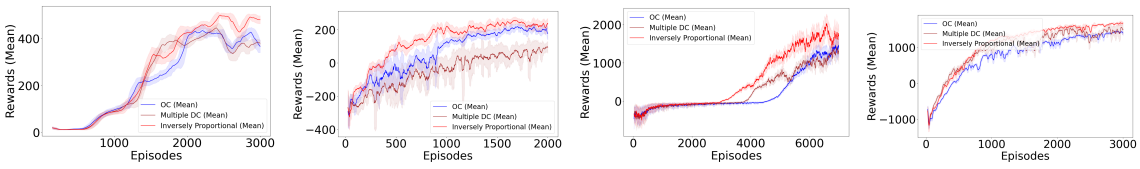
(b) Termination map with goal at the rightmost door.



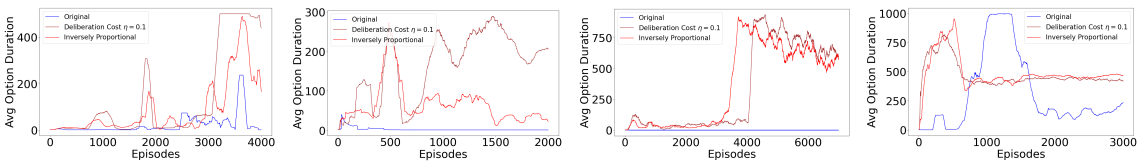
(c) Greedy policy after training.

Figure 1. Overall results for the Four Rooms environment. Shaded areas represent the standard deviation.

In both discrete and continuous action space environments, Figure 2 (top), the adoption of a flexible deliberation cost with inversely proportional decay consistently outperformed both the original Option-Critic algorithm and the version with multiple fixed deliberation costs. In Lunar Lander and Ant, this approach also led to faster convergence, indicating its ability to adapt to the temporal dynamics of each task. In environments such as Ant, where complex movement coordination is required, encouraging longer option durations promoted more stable and natural locomotion.



(a) Cartpole **(b) Lunar Lander** **(c) MuJoCo Ant** **(d) Half Cheetah**



(e) Cartpole **(f) Lunar Lander** **(g) MuJoCo Ant** **(h) Half Cheetah**

Figure 2. Top: learning curves. Bottom: average option duration per episode in Cartpole, Lunar Lander, MuJoCo Ant, and MuJoCo Half Cheetah. Shaded areas represent the standard deviation.

The reduction in option switching, as shown in Figure 2 (bottom), reflects improved temporal abstraction with the use of inversely proportional deliberation cost and

indicates that observed option durations correspond to meaningful behavioral changes rather than degenerate into primitive actions. This results in more stable policies with longer option durations and greater computational efficiency, as fewer resources are spent on frequent option changes. Sustaining temporally extended actions allows the agent to exploit specialized behaviors, contributing to higher performance and more coherent trajectories over time.

5. Conclusion

This work introduced a flexible deliberation cost formulated as an inversely proportional function of option duration, aimed at overcoming the limitations of fixed deliberation cost configurations in the Option-Critic architecture. Our approach consistently improved performance and mitigated the degeneration of options into primitive actions. Future work will explore smoothing the decay of the deliberation cost to further stabilize learning while preserving adaptability, preventing abrupt cost changes, and expanding comparisons with state-of-the-art Deep RL and Policy Gradient methods.

Acknowledgments

This research was partially supported by CNPq (grants 443184/2023-2, 313845/2023-9, 445238/2024-0)

References

Bacon, P.-L., Harb, J., and Precup, D. (2017). The option-critic architecture. *Proceedings of the AAAI 2017*, 31(1).

Barto, A. G., Sutton, R. S., and Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846.

Harb, J., Bacon, P.-L., Klissarov, M., and Precup, D. (2018). When waiting is not an option: Learning options with a deliberation cost. *Proc. of AAAI 2018*, 32(1).

Harutyunyan, A., Dabney, W., Borsa, D., Heess, N., Munos, R., and Precup, D. (2019). The termination critic. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 2231–2240.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. (2018). High-dimensional continuous control using generalized advantage estimation.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, 2nd edition.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1):181–211.

Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer Cham, 1 edition.

Wawrzyński, P. (2009). A cat-like robot real-time learning to run. In *Adaptive and Natural Computing Algorithms*, pages 380–390, Berlin. Springer Berlin Heidelberg.