Hyperbolic Discounting in Hierarchical Reinforcement Learning

Anonymous authors
Paper under double-blind review

Abstract

Decisions often require balancing immediate gratification against long-term benefits. In Reinforcement Learning (RL), this balancing act is influenced by temporal discounting, which quantifies the devaluation of future rewards. Prior research indicates that human decision-making aligns more closely with hyperbolic discounting than the conventional exponential discounting used in RL. As artificial agents become more advanced and pervasive, particularly in multi-agent settings alongside humans, the need for appropriate discounting models becomes critical. Although hyperbolic discounting has been proposed for single-agent learning along with multi-agent reinforcement learning (MARL), it is still underexplored in more advanced settings such as the hierarchical reinforcement learning (HRL). We introduce and formulate hyperbolic discounting in HRL, establishing theoretical and practical foundations across various frameworks, including option critic and Feudal Networks methods. We evaluate hyperbolic discounting on diverse tasks, comparing it to the exponential discounting baseline. Our results show that hyperbolic discounting achieves higher returns in 50 of scenarios and performs on par with exponential discounting in 95 of tasks, with significant improvements in sparse reward and coordination-intensive environments. This work opens new avenues for robust decision-making processes in the development of advanced RL systems.

1 Introduction

Hierarchical reinforcement learning (HRL) extends the capabilities of RL, by proposing a divide-and-conquer approach. In this approach, the complex, difficult to solve problem, is abstracted into multiple smaller problems. These abstracted problems are generally easier to solve and their solutions can be reused to solve different problems. This approach has previously been successfully utilized (Georgievski, I.; Aiello, M, 2015) to speed up many offline planning algorithms where the dynamics of the environment are known in advance. This compositionality has been identified (Sacerdoti, E.D., 1973) as one of the key building blocks of artificial intelligence. Humans intuitively harness compositionality in order to tackle complex problems. Efficiently using such abstractions has proven to make significant contributions towards solving various important open RL problems such as reward-function specification, exploration, sample efficiency, transfer learning, lifelong learning and interpretability.

In reinforcement learning (RL), the goal of maximizing rewards is central to learning intelligent behavior (Silver et al., 2021). This involves prioritizing reward maximization to generate complex behaviors without specialized problem formulations. The treatment of the reward signal is crucial in developing intelligent agents. Human and animal behavior often shows a preference for immediate rewards over delayed ones (O'Donoghue Rabin, 2000), rooted in temporal discounting, where the value of rewards diminishes over time. In RL, discounting influences the time-preference for rewards, enforces shortest path strategies, and represents the probability of termination (Puterman, 2014). Discounting plays a pivotal role, particularly in infinite horizon objectives, to ensure well-defined long-term reward goals (Sutton Barto, 2018). Rewards are typically discounted exponentially, meaning that a reward obtained t time steps in the future is discounted by a factor of γ^t(Bellman, 1957b; Sutton Barto, 1998). This approach establishes a fixed learning horizon for the agent: a smaller γ value prioritizes short-term rewards, while a larger γ value emphasizes long-term rewards. However, human and animal behavior often follows hyperbolic discounting patterns (Mazur, 1987), characterized by the hyperbolic function $\frac{1}{1+kt}$, where k > 0 represents the hyperbolic discounting rate. Unlike exponential models, hyperbolic discounting accounts for preference reversal over time (Green et al., 1994) and offers better alignment with decision-making scenarios involving multiple reward variables, such as delay length, reward magnitude, and probability (Green Myerson, 2004).

In this work, we posit that incorporating hyperbolic discounting into HRL can enhance agents' adaptability to diverse partners by aligning their decision-making with human temporal preferences. This can improve the robustness and flexibility of HRL systems and foster more effective human-AI collaboration by making agents' behavior more predictable and intuitive. We explore hyperbolic discounting for Hierarchical learning, focusing on its impact on agent interactions. We compare hyperbolic discounting against exponential discounting as a baseline, noting our agent-centric perspective. Our findings reveal that hyperbolic discounting consistently outperforms exponential discounting, yielding higher returns, especially in environments with sparse rewards and the need for intricate coordination. These improvements are evident across various learning modalities and are particularly pronounced in the grid-world environments. The main contributions of this work are:

- We establish theoretical and empirical foundations for incorporating hyperbolic discounting across two HRL algorithms, covering option critic and Feudal networks methods.
- We propose and conduct a comprehensive comparative analysis of two hyperbolic discounting schemes against the traditional exponential model: one computes a hyperbolic value estimate, and the other averages multiple value estimates using normally distributed exponential discount factors. We perform empirical evaluations across two HRL tasks, demonstrating the advantages of hyperbolic discounting in various settings.

2 Preliminaries

2.1 Survival and Hazard Rate

We start by motivating against the use of a single, fixed discount factor. In survival analysis (Cox, 1972), the primary focus is on analyzing and modeling the time until specific events occur, such as death. Sozou (1998) extend this by formalizing time preferences, showing that future rewards should be discounted according to the probability that an agent will not survive to collect them due to encountered risks or hazards. This survival probability is defined as s(t) = P (agent is alive|at time t). The present value of a future reward rt is discounted by s(t), i.e., v(rt) = s(t)rt. If s(t) = 1, the reward is not discounted. The hazard rate, h(t), is defined as the negative rate of change of the log-survival probability, $h(t) = \frac{-dt}{d} lns(t)$. For a constant hazard rate λ, the survival rate is $s(t) = e$, leading to an exponential discount function $s(t) = \gamma^t$ with $\gamma = e^{\lambda}$. Increasing hazard leads to myopic behavior (as $\lambda \to \infty$, $\gamma \to 0$), and decreasing hazard leads to strategic behavior (as $\lambda \to 0$, $\gamma \to 1$). When the hazard rate is uncertain, the survival rate is computed by integrating over a prior distribution $p(\lambda)$, $s(t) =$. For an exponential prior $p(\lambda) = \frac{1}{k} exp(\lambda/k)$, the expected survival rate becomes hyperbolic, $s(t) = \frac{1}{1+kt} \equiv \Gamma_k(t)$, where $\Gamma_k(t)$ is the hyperbolic discount function. Different priors over the hazard rate yield different discount functions (Sozou, 1998).

2.2 Hazardous Markov Games

For the hierarchical setting, we formalize our problem based on the Markov Game (Littman, 1994), generalized to include partial observability. Moreover, to consider distributions over the hazard rate, and use non-exponential discounting functions, we remove the discount factor γ and introduce two additions; a hazard distribution, and a general discount function. Concretely, the Hazardous

Markov Game (HMG) for N agents is defined by the tuple $G = <N, S, O_{ii \in N}, A_{ii \in N}, \Omega, P, r>$, with agents $i \in N = 1, ..., N$, state space S, joint observation space $O = O1 \times ... \times O_N$, and joint action space $A = A_1 \times ... \times A_N$. Each agent i only perceives local observations $o_i \in O_i$, which depend on the state and joint action via the observation function $\Omega : S \times A \rightarrow \Delta(O)$. The transition function $P : S \times A \rightarrow \Delta(S)$ returns a distribution over states given a state and a joint action $A = (a_1, a_2, ..., a_N)$. $r : S \times A \rightarrow R$ is the shared reward function, with $r(s, a_1, a_2, ..., a_N)$ representing the reward received by all agents after taking actions $a_1, a_2, ..., a_N$ in state s. H is the hazard distribution from which a hazard rate $\lambda \in [0, \infty)$ is sampled at the beginning of each episode. Finally, instead of γ, we consider d(t), which is a general discount function, of which exponential and hyperbolic will be special cases. The objective is to jointly optimize the discounted cumulative reward $G = E_{E_{st,A_t}[\sum_t^\infty d(t) r_t]}$ where A_t is the joint action at timestep t and $\lambda \sim H$.

3 Hyperbolic Discounting in HRL

We now discuss the theoretical foundations that can allow us to derive temporal-difference learning solutions while using hyperbolic discounting.

3.1 Value-Based Methods

We first show how exponentially discounted Q-values can be used to derive hyperbolic discounted Q-values, building on prior work (Fedus et al., 2019). The Bellman equation (Bellman, 1957a) is written as:

$$Q_\pi^{\gamma^t}(s, a) = E_{\pi,p}[R(s, a) + \gamma Q_\pi(s', a')] \tag{1}$$

where expectation $E_{\pi;P}$ denotes sampling $a \sim \pi(|s), s' \sim P(|s; a)$, and $a_0 \sim \pi(|s_0)$.

We start by estimating the value function where rewards are discounted hyperbolically instead of the common exponential scheme. We refer to the hyperbolic Q-values as Q_π^Γ:

$$Q_\pi^{\Gamma_k}(s, a) = E_\pi[\Sigma \Gamma_k(t) R(s_t, a_t) | s, a] \tag{2}$$

We establish a connection between hyperbolic $Q_\pi^{\Gamma_k}$-values and values obtained through standard Q-learning. The hyperbolic discount Γ_k can be represented as the integral of a specific function f(γ, t) for γ = [0, 1):

$$\int_{\gamma=0}^{1} \gamma^{kt} d\gamma = \frac{1}{1 + kt} = \Gamma_k(t) \tag{3}$$

The integration of the function $f(\gamma, t) = \gamma^{kt}$ across the domain $\gamma \in [0, 1)$ results in the hyperbolic discount factor $\Gamma_k(t)$. This integration, incorporating an infinite set of exponential discount factors γ, reveals that $\Gamma_k(t)$ functions as the standard exponential discount factor, linking the concept to traditional Q-learning. This approach suggests that by aggregating an infinite collection of γ values, hyperbolic discounts can be derived for each respective time step t. For a hyperbolic discount function $\Gamma_k(t)$, the hyperbolic Q-values can be written as:

$$Q^{\Gamma \pi(s,a)} = E_\pi[\Sigma \Gamma_k(t) R(s_t, a_t) | s, a] \tag{4}$$

$$= E_\pi[\Sigma_t (\int_{\gamma=0}^{1} \gamma^{k^t} d\gamma) R(s_t, a_t) | s, a] \tag{5}$$

$$= \int_{\gamma=0}^{1} E_\pi[\Sigma_t R(s_t, a_t)(\Gamma^k)^t | s, a] d\gamma \tag{6}$$

$$= \int_{\gamma=0}^{1} Q_\pi^{(\gamma^{kt})}(s, a) d\gamma \tag{7}$$

4 Experiments

4.1 methods

We introduce two novel discounting methods: hyperbolic discounting and averaged horizon discounting. The latter is a special case of the former, where the agent learns over multiple discount factors γ.

Hyperbolic Discounting Following Fedus et al. (2019), we implement hyperbolic discounting in HRL using a multi-headed value output structure, where each head corresponds to a distinct discount factor. We approximate the hyperbolic value function by integrating multiple value estimates via a Riemann sum:

$$Q_{\pi}^{\Gamma}(s,a) = \sum_{\gamma_i \in G} (\gamma_{i+1} - \gamma_i) w(\gamma_i) Q_{\pi}^{\gamma^i}(s,a) \tag{8}$$

Here, $G = [\gamma_0, \gamma_1, ..., \gamma_n]$ is the set of discount factors, with Q^{γ_i} denoting the Q-values for each γ_i.

4.2 Setup

We evaluate the effectiveness of the proposed hyperbolic method across two HRL algorithms: option-critic and Feudal networks. These methods are tested in two distinct HRL environments: one low-dimensional state-space environment (grid-world) and one high-dimensional state-space environment (Atari games). Each environment presents unique challenges to assess the scalability and the generalization capabilities of the algorithms.

4.3 Results

We present results for the two proposed discounting methods and the baseline for each of the two HRL algorithms across four benchmarks, comparing their performance. We show results for four-room grid world and Atari games. Figure 1 and Figure 2 show the comparison of hyperbolic against exponential discounting for the two methods across four-room grid world, while Figure 3 shows the the hyperbolic discounting using Feudal networks in Atari Pong games. Generally, performance differences are noticeable in four-room and Atari, with one of the proposed variants performing better, while performance differences in option critic network are minimal.

5 Discussion

Please see Appendix A for related works. We introduce hyperbolic discounting for HRL settings. Our experiments revealed improvements in performance, stability, and sample efficiency with non exponential discounting methods, which outperformed traditional exponential discounting on more than 50 of the tasks. Hyperbolic discounting emerged as the most reliable method, showing smaller standard deviations and enhanced performance across various algorithms. The structural differences in algorithms influenced the impact of non-exponential discounting, with some benefiting more than others. Future research could explore ensemble methods to further improve non exponential discounting functions. These findings highlight the potential of non-exponential discounting in reinforcement learning, promoting more efficient and effective decision-making in real-world applications.

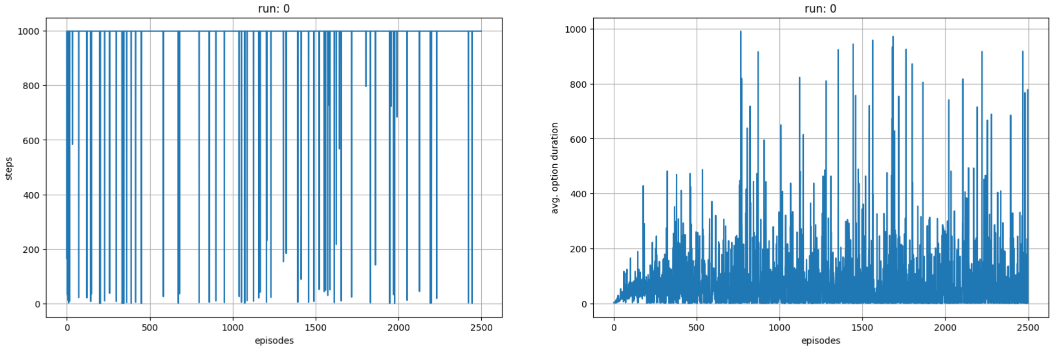

Figure 1: Four-room grid world Results: Proposed hyperbolic discounting policies

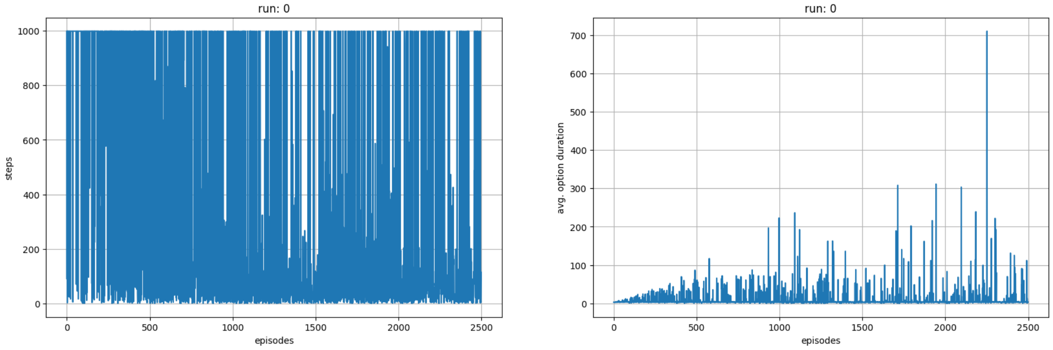

Figure 2: Four-room grid world Results: Proposed exponential discounting policies

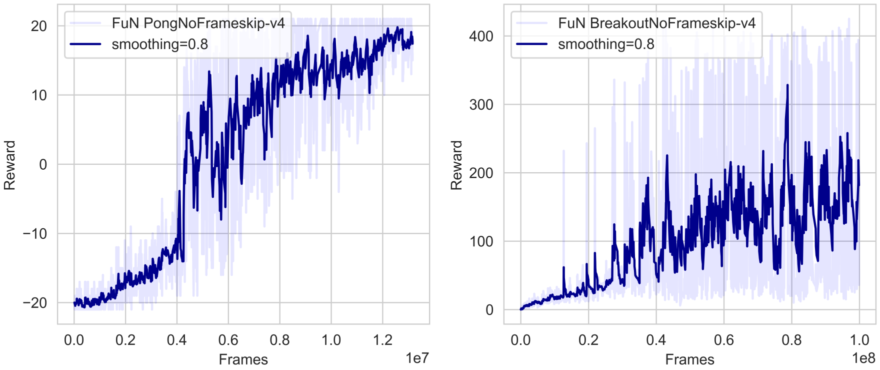

Figure 3: Hyperbolic discounting using Feudal networks tested in Pong Atari games

There are avenues of future work. The results of Bowling et al. (2023) address scenarios with a constant exponential discount factor, not considering hyperbolic discounting. Since hyperbolic discounting, as approximated by Fedus et al. (2019) and extended here, uses multiple constant exponential discount factors, further theoretical analysis would be beneficial, such as Pitis (2023). Moreover, it would be interesting to study effects of reward discounting in human-AI teams where long-term decision trade-offs are involved.

References

Sacerdoti, E.D. Planning in a Hierarchy of Abstraction Spaces. Artif. Intell. 1973, 5, 115–135.

Georgievski, I.; Aiello, M. HTN planning: Overview, comparison, and beyond. Artif. Intell. 2015, 222, 124–156.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. Artificial Intelligence, 299:103535, 2021.

Ted O'Donoghue and Matthew Rabin. The economics of immediate gratification. Journal of behavioral decision making, 13(2):233–250, 2000.

Martin L Puterman. Markov decision processes: discrete stochastic dynamic programming. John Wiley Sons, 2014.

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction, Second Edition. MIT press Cambridge, 2018.

James E. Mazur. An adjusting procedure for studying delayed reinforcement. In Quantitative analyses of behavior, volume 5, pp. 55–73. Lawrence Erlbaum Associates, Inc.,

Leonard Green, Nathanael Fristoe, and Joel Myerson. Temporal discounting and preference reversals in choice between delayed outcomes. Psychonomic Bulletin Review, 1(3):383–389, 1994.

David R Cox. Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2):187–202, 1972.

Peter D Sozou. On hyperbolic discounting and uncertain hazard rates. Proceedings of the Royal Society of London. Series B: Biological Sciences, 265(1409):2015–2020, 1998.

Peter D Sozou. On hyperbolic discounting and uncertain hazard rates. Proceedings of the Royal Society of London. Series B: Biological Sciences, 265(1409):2015–2020, 1998.

William Fedus, Carles Gelada, Yoshua Bengio, Marc G Bellemare, and Hugo Larochelle. Hyperbolic discounting and learning over multiple horizons. arXiv preprint arXiv:1902.06865, 2019.

Richard Bellman. Dynamic Programming. Princeton University Press, Princeton, NJ, USA, 1957a.

Michael Bowling, John D Martin, David Abel, and Will Dabney. Settling the reward hypothesis. In International Conference on Machine Learning, pp. 3003–3020. PMLR, 2023.

Silviu Pitis. Consistent aggregation of objectives with diverse time preferences requires non-markovian rewards. Advances in Neural Information Processing Systems, 36, 2023.

Michael T Bixter and Christian C Luhmann. The social contagion of temporal discounting in small social networks. Cognitive Research: Principles and Implications, 6(1):13, 2021.

Appendix

A Related Work

We focus on the aspect of discounting preferences in social settings which involves more than one individual (we refer the reader to Fedus et al. (2019) for an in depth review of discounting in individual human preferences). In controlled studies, discounting future rewards has been mostly studied as a personal preference parameter, where each individual is given a questionnaire to evaluate their valuation of future rewards. These studies show how decisions involving immediate versus long-term benefits are influenced by temporal discounting—where individuals place less value on delayed rewards. Recent studies have expanded this concept to decisions made in group settings, like dyads or small groups, revealing that direct interactions can lead to aligned preferences among participants, making them more similar in patience level over time. Bixter Luhmann (2021) study whether such social influences could also be indirect, such as through mutual acquaintances within a group. Focusing on hypothetical monetary rewards, the research involved groups of three where one member's decision preferences before collaboration were linked to another's preferences after collaborating with an intermediary. Findings highlighted that decision-making tendencies regarding time can spread through a social network's connections, showing the presence of indirect social influence in a controlled setting.