Non-exponential Reward Discounting in Reinforcement Learning

Raja Farrukh Ali

Department of Computer Science
Kansas State University
rfali@ksu.edu

Abstract

Reinforcement learning methods typically discount future rewards using an exponential scheme to achieve theoretical convergence guarantees. Studies from neuroscience, psychology, and economics suggest that human and animal behavior is better captured by the hyperbolic discounting model. Hyperbolic discounting has recently been studied in deep reinforcement learning and has shown promising results. However, this area of research is seemingly understudied, with most extant and continuing research using the standard exponential discounting formulation. My dissertation examines the effects of non-exponential discounting functions (such as hyperbolic) on an agent's learning and aims to investigate their impact on multi-agent systems and generalization tasks. A key objective of this study is to link the discounting rate to an agent's approximation of the underlying hazard rate of its environment through survival analysis.

Introduction

The reinforcement learning (RL) paradigm has shown promise as a path towards important aspects of rational utility in autonomous agents. Reward, specifically reward maximization, has been hypothesized as being *enough* to learn intelligent behavior (Silver et al. 2021). For a learning agent to acquire multiple abilities simultaneously (e.g. planning, motor control, language, etc.), the singular goal of reward maximization may be enough to generate complex behavior, rather than learning and reasoning over specialized problem formulations for each ability. How we treat this reward signal is thus central to our quest for intelligent agents.

In the RL problem setting, the objective is to maximize cumulative rewards over time, known as return. There are different approaches to calculating this return; aggregating undiscounted rewards over a finite number of steps, calculating a discounted sum (infinite rewards sum to a finite number), or calculating the average reward per time-step. The discounted reward formulation remains the most common in contemporary research, in which a discount factor $0 \leq \gamma < 1$ exponentially reduces or *discounts* the present value of future rewards, r_t at step t, as $\gamma^t r_t$. Reward discounting prioritizes sooner rewards over later rewards and enables a convergence proof for the infinite horizon case.

The functional form of the discounting function directly influences the solutions learned. Evidence from psychology and economics shows that human and animal preferences for future rewards can be modeled more accurately using hyperbolic discounting ($\Gamma_k(t) = \frac{1}{1+kt}$ for $k > 0$). Preference reversals can also occur with time, which can be modeled by hyperbolic discounting but not exponential discounting. Fedus et al. (2019) show that a deep RL agent which acts via hyperbolic discounting is indeed feasible, while approximating hyperbolic (and other non-exponential) discounting function using familiar temporal difference (TD) learning methods like Q-learning. This approximation is made possible by learning many Q-values simultaneously, each for a different discount factor, and in doing so, the agent also learns over multiple horizons which is shown to be an effective auxiliary task. Exponential discounting is consistent with a prior belief that there exists a known constant risk to the agent (Sozou 1998). However, hyperbolic and non-exponential discounting is more appropriate when an agent holds uncertainty over the environment's hazard rate (defined as the per-time-step risk the agent incurs as it acts in the environment). Hyperbolic discounting is theorized to be most beneficial when the hazard rate characterizing the environment is unknown.

My dissertation research focuses on the use of non-exponential discounting functions and explores them under environment conditions that are inherently hazardous (dynamic hazard), where the severity of the hazard is unknown, and where the agent is unsure about its survival. My current research focus involves investigating this approach in multi-agent systems and on generalization tasks, thereafter investigating the connection between discounting and hazard rate via survival analysis. The dissertation's central research question is thus:

> *Can we develop a learning representation that accounts for the empirical hazard rate of an environment over time and utilizes it for reward discounting?*

Related Work

Sozou (1998) propose a per-time-step death via the hazard rate. Alexander and Brown (2010) suggest a TD-based hyperbolic discounting solution. Kurth-Nelson and Redish (2009) propose the modeling of hyperbolic discounting via

distributed exponential discounting. Fedus et al. (2019) extend this formulation to deep reinforcement learning by approximating hyperbolic discounting from exponential discounting and evaluate their approach using a value-based method on episodic, finite-horizon tasks. Pitis (2019) consider a state-action dependent discount factor. Another perspective is the continuing (infinite horizon) vs. episodic (finite horizon) formulation of the problem. White (2017) suggest a transition-based discounting method to unify episodic and continuing task specifications. Naik et al. (2019) argue that discounting is fundamentally incompatible with function approximation for control in continuing tasks and suggest the use of average reward in continuing tasks.

Research Plan and Contributions

Multi-Agent Systems

This dimension of my research hypothesizes that agents in a cooperative multi-agent setting can benefit from using non-exponential (e.g., hyperbolic) discounting, where each agent in the team discounts future rewards using a different discount factor. This can be thought of as the team learning over multiple horizons simultaneously such that the learned team policy is robust to the unknown hazard rate characterizing the environment. Instead of each team member discounting exponentially using a fixed γ, the team discounts over the entire horizon i.e. $\gamma \in [0, 1)$, effectively setting some agents to be myopic and other agents to be far-sighted or strategic. We replace the single discount factor γ with a hazard distribution $\mathcal{H}$ such that at the beginning of each episode, a hazard $\lambda \in [0, \infty]$ is sampled from the hazard distribution $\mathcal{H}$ for each agent. This approach can be integrated with the different multi-agent reinforcement learning (MARL) paradigms such as independent learning, value function factorization, and centralized training decentralized execution (CTDE), however each would require a different problem formulation.

Generalization

We study the effects of hyperbolic discounting on the generalization ability of an agent by evaluating it on procedurally generated environments via policy gradient methods (Nafi, Ali, and Hsu 2022). Because an agent in such environments is tested on levels that it has not seen during training (which contributes to its uncertainty regarding the environment's hazard rate), hyperbolic discounting may be preferable to exponential discounting. We implement hyperbolic discounting-based advantage estimation in which the agent learns the advantage function over multiple horizons simultaneously, and results show improved performance over baseline policy-gradient methods. Tangentially to this work, we also explore the auxiliary task of learning over multiple horizons for off-policy, value-based methods in procedurally generated environments by incorporating recent architectural improvements and implementation techniques (Ali et al. 2023). Results indicate that learning over multiple horizons alone is not sufficient and the choice of behavior policy enforced through the discounting function (hyperbolic or exponential) has an effect on the agent's performance.

Survival Analysis

An agent's policy should be cognizant of the environment's underlying hazard rate and adapt accordingly. Human behavior is greatly influenced by how hazardous the situation is (e.g., war vs. peacetime). Our plans can change drastically based on how long we expect to remain alive (e.g., a terminal diagnosis). Known, constant hazard implies exponential discounting and unknown hazard implies non-exponential discounting. As hyperbolic discounting is more robust in scenarios where the hazard rate is unknown, an agent may employ it to learn over multiple horizons. However, if the agent can approximate the underlying hazard rate of the environment, then it can equivalently select a single, hazard-appropriate exponential discount factor γ. The aim of this research direction is to use survival analysis to estimate hazard rate using both statistical methods (Kaplan-Meier, Cox regression) and machine learning methods. Future work may include incorporating other non-exponential discounting functions, especially studying priors for the gamma distribution of the hazard rate (such as Erlang distributions), as well as elucidating practical applications where certain discounting functions may be undesirable.

References

Alexander, W. H.; and Brown, J. W. 2010. Hyperbolically discounted temporal difference learning. *Neural computation*, 22(6): 1511–1527.

Ali, R. F.; Duong, K.; Nafi, N. M.; and Hsu, W. 2023. Multi-Horizon Learning in Procedurally-Generated Environments for Off-Policy Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37.

Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H. 2019. Hyperbolic discounting and learning over multiple horizons. *Conference on Reinforcement Learning and Decision Making (RLDM)*.

Kurth-Nelson, Z.; and Redish, A. D. 2009. Temporal-difference reinforcement learning with distributed representations. *PLoS One*, 4(10): e7362.

Nafi, N. M.; Ali, R. F.; and Hsu, W. 2022. Hyperbolically Discounted Advantage Estimation for Generalization in Reinforcement Learning. In *Decision Awareness in Reinforcement Learning Workshop, ICML*.

Naik, A.; Shariff, R.; Yasui, N.; Yao, H.; and Sutton, R. S. 2019. Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv:1910.02140*.

Pitis, S. 2019. Rethinking the discount factor in reinforcement learning: A decision theoretic approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7949–7956.

Silver, D.; Singh, S.; Precup, D.; and Sutton, R. S. 2021. Reward is enough. *Artificial Intelligence*, 299: 103535.

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

White, M. 2017. Unifying task specification in reinforcement learning. In *International Conference on Machine Learning*, 3742–3750. PMLR.