Non-Exponential Reward Discounting in Reinforcement Learning

Raja Farrukh Ali

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

Abstract

In reinforcement learning (RL), reward discounting is a fundamental component of a Markov Decision Process, in which agents typically discount future rewards using an exponential scheme that helps achieve theoretical convergence guarantees. Studies from neuroscience, psychology, and economics suggest that humans and animals instead discount future rewards hyperbolically. Hyperbolic discounting has been studied in deep reinforcement learning recently and has shown promising results. However, this area of research is seemingly understudied, with most ongoing research using the standard exponential discounting formulation. My dissertation examines the effects of non-exponential discounting functions (such as hyperbolic) and aims to investigate its 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, such as the ability to adapt to uncertainty regarding risk and reward, over multiple time horizons and in multi-agent systems. Reward, specifically reward maximization, has recently been hypothesized as being *enough* to learn intelligent behavior (Silver et al. 2021) in which it has been suggested that in pursuit of intelligence that can master multiple abilities simultaneously (e.g. planning, motor control, language etc.), instead of learning and reasoning over specialized problem formulations for each ability, the singular goal of reward maximization may be enough to generate complex behavior exhibiting these multiple abilities. A followup commentary argues that scalar rewards may not be enough to achieve such intelligence and instead suggest multi-objective models of reward maximization i.e. vector-valued rewards (Vamplew et al. 2022). However, there is a consensus that the reward signal itself is fundamental in the quest for artificial intelligence.

In the RL problem setting, while the objective is to maximize cumulative rewards over time, called return, there are different ways to calculate this return; summing rewards over a finite number of steps, calculating a discounted sum (infinite rewards sum to a finite number), or the average reward per time-step. The discounted reward formulation remains the most followed one 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. However, is exponential discounting the only discounting function that should be considered?

The functional form of the discounting function establishes strong priors over the solutions learned. Studies from psychology, neuroscience and economics suggest that humans and animals instead discount future returns hyperbolically ($\Gamma_k(t) = \frac{1}{1+kt}$ for $k > 0$). Fedus et al. (2019) show that a deep RL agent that acts via hyperbolic discounting is indeed feasible, while approximating hyperbolic (and other non-exponential) discounting function using familiar temporal difference learning methods like Q-learning. Exponential discounting has been shown to be 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 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, with further plans of linking the discount function to the underlying hazard via survival analysis. This dissertation's central research question is thus:

> *Can we develop a learning representation that uses reward discounting to account for the empirical hazard rate of an environment over time?*

Related Work

Sozou (1998) proposed a per-time-step death via the haz-

ard rate. Alexander and Brown (2010) proposed a temporal difference (TD) based hyperbolically discounting solution. Kurth-Nelson and Redish (2009) proposed the modeling of hyperbolic discounting via distributed exponential discounting. Fedus et al. (2019) were the first to extend this formulation to deep reinforcement learning by approximating hyperbolic discounting from exponential discounting and evaluated their approach using a value-based method on episodic, finite-horizon tasks. Pitis (2019) consider a state, action dependent discount factor. Another debate in RL 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 specification. 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 discounting, such as hyperbolic, 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 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 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 $\mathcal{H}$. 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

In this work, we study the effects of hyperbolic discounting on generalization tasks and present HDGenRL, Hyperbolic Discounting for Generalization in Reinforcement Learning (Nafi, Ali, and Hsu 2022). Our key contribution is a hyperbolic discounting-based advantage estimation that makes the agent aware of and robust to the underlying uncertainty of survival and episode duration. We consider the problem of generalization in the context of procedurally generated environments and argue that exponential reward discounting limits the agent's performance in unseen scenarios. We demonstrate that the hazard-rate characterizing these procedurally generated environments is unknown, which motivates for the use of hyperbolic discounting, and results show significant improvement over baseline policy-gradient methods. This collaboration work has already been presented as a workshop paper.

Survival Analysis

An agent's policy should be cognizant of the environment's underlying hazard rate and adapt accordingly. Human behavior is shaped greatly 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 hazard implies exponential discount and unknown hazard implies non-exponential discount (Fedus et al. 2019). As hyperbolic discounting is more robust in scenarios where the hazard-rate is unknown, an agent may employ this and 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 research work may include incorporating other non-exponential discounting functions, especially studying the gamma prior distribution of the hazard rate (such as Erlang distribution), 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.

Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H. 2019. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865*.

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.

Vamplew, P.; Smith, B. J.; Källström, J.; Ramos, G.; Rădulescu, R.; Roijers, D. M.; Hayes, C. F.; Heintz, F.; Mannion, P.; Libin, P. J.; et al. 2022. Scalar reward is not enough: A response to Silver, Singh, Precup and Sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2): 1–19.

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