Hyperbolically Discounted Advantage Estimation for Generalization in Reinforcement Learning

Anonymous Authors[1]

Abstract

In reinforcement learning (RL), agents typically discount future rewards using an exponential scheme. However, studies have shown that humans and animals instead exhibit hyperbolic time-preferences and thus discount future rewards hyperbolically. In the quest for RL agents that generalize well to previously unseen scenarios, we study the effects of hyperbolic discounting on generalization tasks and present Hyperbolic Discounting for Generalization in Reinforcement Learning (HDGenRL). We propose a hyperbolic discounting-based advantage estimation method that makes the agent aware of and robust to the underlying uncertainty of survival and episode duration. On the challenging RL generalization benchmark Procgen, our proposed approach achieves up to 200% performance improvement over the PPO baseline that uses classical exponential discounting. We also incorporate hyperbolic discounting into another generalization-specific approach (APDAC), and results indicate further improvement in APDAC's generalization ability. This denotes the effectiveness of our approach as a plug-in to any existing methods in aiding generalization.

1. Introduction

Generalization refers to the capability of an agent to perform in similar but unseen environments and is currently an active research challenge. Training deep RL algorithms is a data-intensive task and given a sufficiently large set of samples, they can learn a specific skill (Mnih et al., 2015; 2016; Haarnoja et al., 2018). However, they tend to overfit even with large training samples (Cobbe et al., 2020; 2021; Grigsby & Qi, 2020; Justesen et al., 2018). To facilitate research on this issue, newer benchmarks like Procgen (Cobbe et al., 2020) have been introduced that present a set of procedurally generated environments to evaluate generalization through held-out levels used only in testing. In this work, we consider the problem of generalization in the context of such procedurally generated environments.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

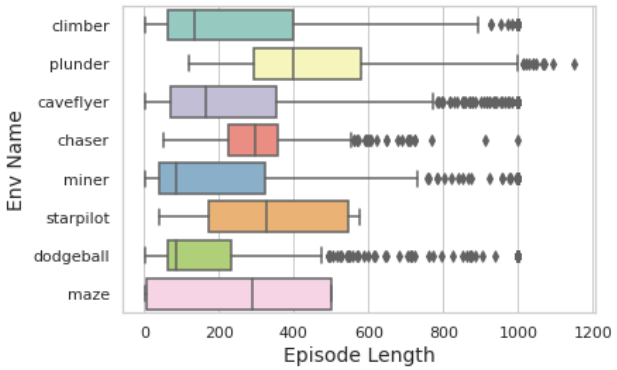

Figure 1. Distribution of episode lengths for eight Procgen environments, estimated based on 1000 episodes randomly sampled from the test levels for each game. This depicts how much an episode varies in terms of duration for each of these games.

Standard RL algorithms specify a discount factor, $0 \leq \gamma \leq 1$, that discounts the future reward r_t at step t as $\gamma^t r_t$ (Sutton et al., 1998). Such discounting guarantees the theoretical convergence of the value function and stabilize the optimization. The value of γ sets a fixed effective horizon for an agent such that all rewards beyond that point in time will be ignored. At the same time, the notion of effective horizon induces a prior belief that there exists a known, fixed risk or hazard rate in the environment (Fedus et al., 2019). We argue that this assumption does not comply with the set of diverse, ever-changing levels in a procedurally generated environment, wherein an environment's dynamics and the attributes change across levels, hence introducing significant uncertainty and stochasticity compared to the generic RL environments. Figure 1 shows the distribution of episode length measured over 1000 episodes sampled randomly from the test levels for 8 Procgen environments. The episodes were generated using a learned policy trained on 25M time steps. Even with such a mature policy, we see that the episode length varies significantly for each enviornment.

This indicates the uncertainty over the hazard rate (survival rate) and why a fixed effective horizon may fail to better assess expected future rewards. Existing literature indicates that when an agent holds uncertainty over the environment's hazard rate, a non-exponential (such as hyperbolic) discounting factor is more suitable (Sozou, 1998; Fedus et al., 2019). Hyperbolic discounting of rewards also aligns with the time-preference of animals and humans (Mazur, 1997).

To address this uncertainty regarding an environment's hazard rate and to mitigate the drawbacks of the fixed effective horizon, we propose to discount future rewards hyperbolically for generalization tasks and present our method HDGenRL (Hyperbolic Discounting for Generalization in Reinforcement Learning). We present a simple extension to calculate hyperbolically-discounted advantage estimates that can be used with policy gradient methods. We evaluate our approach on all sixteen environments from the Procgen benchmark and our approach significantly outperforms the Proximal Policy Optimzation (PPO) baseline that uses exponential discounting (Schulman et al., 2017). We further integrate our discounting scheme with Attention-based Partially Decoupled Actor-Critic (APDAC) (Nafi et al., 2021), a recent generalization specific method, and results indicate the potential of our approach to be used with any existing method that employs classical exponential discounting.

2. Related Work

Generalization in Deep RL. There has been a lot of emphasis lately on building intelligent agents that avoid overfitting and can generalize well to unseen data (Rajeswaran et al., 2017; Justesen et al., 2018; Grigsby & Qi, 2020; Cobbe et al., 2019). Methods that have been used to some success include regularization techniques like dropout, batch normalization, data augmentation (Igl et al., 2019; Hu et al., 2021). Raileanu & Fergus (2021) leverages fully separated policy and value function networks to achieve generalization, while Cobbe et al. (2021) introduces a phase-wise training.

Hyperbolic Discounting. Hyperbolic discounting has been studied in the fields of behavior psychology, economics, neuroscience and lately, to a limited extent, in reinforcement learning. Sozou (1998) proposed a per-time-step death via the hazard rate, whereas Dasgupta & Maskin (2005) proposed that uncertainty over the timing of rewards lead to preference reversals as exhibited in hyperbolic discounting. Alexander & Brown (2010) proposed a temporal difference (TD) based hyperbolically discounting solution. Although TD learning relies on exponential discounting in its calculation, naive modifications to it to discount hyperbolically have been shown to be inconsistent. Kurth-Nelson & 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, Rainbow (Hessel et al., 2018), on the ALE benchmark.

3. Background

3.1. Discount Factor and Effective Horizon

In standard RL algorithms, the agent's objective is to maximize the sum of the discounted rewards over the future. Formally, the expected discounted return G_t is:

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ..., \tag{1}$$

where $0 \leq \gamma \leq 1$ is called the discount factor (Sutton et al., 1998). The discount factor γ determines the current value of the future rewards. The horizon refers to how many steps into the future the agent takes into account and values the reward it receives. If $\gamma = 0$, then the agent becomes myopic and only maximizes immediate rewards. Conversely when $\gamma \to 1$, the agent values future rewards as well, making the agent far-sighted. Thus a particular value of γ sets a fixed effective horizon till which point the rewards are considered and beyond that the rewards are discarded.

3.2. Uncertain Hazard and Hyperbolic Discounting

An alternative perspective connects the discount factor to the notion of an agent not surviving to collect the reward due to encountering a risk or *hazard* (Sozou, 1998; Fedus et al., 2019). If $s(t)$ is the probability that the agent survives until time t, then the present value of a future reward r_t shall be discounted by the probability that the agent will survive till time t to collect it. Hence $v(r_t) = s(t)r_t$. The negative rate of change of the log-survival at time t is called the *hazard rate* $h(t)$:

$$h(t) = -\frac{d}{dt} \ln s(t). \tag{2}$$

By solving eq. 2 as $s(t) = e^{-\lambda t}$, and setting $s(t) = \gamma^t$ (exponential discounting) as in $v(r_t) = s(t)r_t$, the relation between hazard rate $\lambda \in [0, \infty]$ and $\gamma \in [0, 1]$ can be expressed as $\gamma = e^{-\lambda}$. A single value of γ thus represents a known constant hazard rate $\lambda = -\ln(\gamma)$. However, to reflect the uncertainty in the hazard rate, we can relax this assumption of knowing the exact true hazard rate and replace it with a hazard prior $p(\lambda)$ such that: $s(t) = \int_{\lambda=0}^{\infty} p(\lambda) e^{-\lambda t} d\lambda$. It is further shown by (Sozou, 1998) that for an exponential hazard prior $p(\lambda) = \frac{1}{k} e^{(-\lambda/k)}$, the survival rate of the agent becomes hyperbolic.

$$s(t) = \frac{1}{1 + kt} \equiv \Gamma_k(t) \tag{3}$$

where k is the hyperbolic exponent with value $k > 0$.

4. POMDP Formulation

The problem of generalization recognizes a distribution of POMDP, $p(m)$ where $m \in M$, and each instance m is defined by $(\mathcal{S}_m, \mathcal{O}_m, \mathcal{A}, \mathcal{T}_m, \mathcal{R}_m, \Omega_m, \gamma_m)$, where $\mathcal{S}_m$ is the set of states, $\mathcal{O}_m$ is the set of observations, $\mathcal{A}$ is the set of actions, $\mathcal{T}_m(s'|s,a)$ are the transition probabilities, $\Omega_m(o|s',a)$ are the conditional observation probabilities, $\mathcal{R}_m(s,a)$ is the reward function, and γ_m is the discount factor sampled from the hazard distribution instead of a constant one. Limited number of POMDPs are exposed during training. Let's say we have $\mathcal{M}_{train} = \{m_1, m_2, ..., m_n\}$, where $m_i \sim p$ and $i \in \{1, 2, ..., n\}$, the goal is to optimize the policy π_θ using the objective function $J(\pi_\theta) = \mathbb{E}_{p,\pi,\mathcal{T}_m}\left[\sum_{t=0}^{T} \gamma^t \mathcal{R}_m(s_t, a_t)\right]$ over the full distribution of POMDPs. Each environment corresponds to a distribution of POMDPs, $p(m)$ and each level of the game is analogous to a sampled POMDP instance. The model is trained on 200 levels and tested on the full distribution of levels which is significantly larger than the training set. This enables the evaluation of the model's generalization capability beyond those 200 training levels.

5. Methodology

We aim to learn a value function and consequently a good advantage estimate to guide the policy optimization, in order to achieve a generalizable policy. We identify that hyperbolic discounting is more suitable than exponential discounting on the generalization task and propose to optimize the policy using hyperbolically discounted advantage estimate.

5.1. Avoiding the Curse of Fixed Effective Horizon

As discussed in Section 3.1, a fixed discount factor in the case of exponential discounting can impose a single effective horizon for the agent. Thus the agent's value function estimate relies on a prior belief about the length of the episode. Raileanu & Fergus (2021) demonstrate that even with the same starting state, the length of the episode can differ significantly. The episode length can significantly change the value estimate of the earlier states in a trajectory based on the later reward. For example, the final reward of an episode will be highly discounted and perceived as small if the episode length is too big. However, if the episode length is small, then the same final reward will contribute much more to the value estimate. Thus, due to the fixed effective horizon, an agent may fail to correctly anticipate the worth of future rewards in case of highly-varied episode length (see Figure 1). As we can not restrict the length of an episode, we propose to relax this fixed effective horizon by considering multiple horizons simultaneously. Thus, we need a value estimate that considers multiple discount factors γs while calculating the value estimate.

5.2. Modelling the Unknown Hazard

Unseen levels in a highly diverse environment imply unknown hazard. Since the agent is unsure about the environment's hazard rate across unseen levels, we model this uncertainty by injecting a hazard distribution in the POMDP through the variable discount factor, γ_m (see Section 4). Thus an episode from a new level comes with a hazard rate λ_m, where $\gamma_m = e^{-\lambda_m}$. Since an agent cannot accurately estimate the value of γ_m for each new level in a model-free setup, and hyperbolic discounting is better able to capture this uncertain hazard, we use hyperbolically discounted advantage estimate to learn a robust policy for unseen levels.

5.3. Hyperbolically-discounted Advantage Estimation

We extend the basic idea of hyperbolic discounting-based value function to estimate hyperbolically-discounted advantage, so that it can be integrated with policy gradient methods. Fedus et al. (2019) presents an approach that leverage multiple discount factors from exponential discounting scheme to approximate the hyperbolic discounting function. We adopt this approximation approach but instead use it for advantage estimation with respect to multiple effective horizons.

Policy gradient objective for PPO (Schulman et al., 2017):

$$J_\pi(\theta) = \hat{\mathbb{E}}_t\Big[min\big(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big]$$

where $r_t(\theta) = \frac{\pi_{(\theta)}(a_t|s_t)}{\pi_{(\theta)_{old}}(a_t|s_t)}$, and $\hat{A}_t$ is the estimation of the advantage function at timestep t. Advantage is defined as $A(s_t, a_t) = Q(s_t, a_t) - V(s_t)$. Leveraging the hyperbolic function evaluation $\Gamma_k(t) = \frac{1}{1+kt} = \int_0^1 \gamma^{kt} d\gamma$, we propose to estimate hyperbolically-discounted advantage as follows:

$$\begin{aligned} A_\pi^\Gamma(s_t, a_t) &= \int_0^1 A_\pi^{(\gamma^k)^t}(s_t, a_t) d\gamma \\ &= \int_0^1 \Big[Q_\pi^{(\gamma^k)^t}(s_t, a_t) - V_\pi^{(\gamma^k)^t}(s_t)\Big] d\gamma \\ &= \int_0^1 \Big[Q_\pi^{(\gamma^k)^t}(s_t, a_t)\Big] d\gamma - \int_0^1 \Big[V_\pi^{(\gamma^k)^t}(s_t)\Big] d\gamma \end{aligned}$$

$Q_\pi^{(\gamma^k)^t}(s_t, a_t)$ can be decomposed into $r_t + V_\pi^{(\gamma^k)^t}(s_{t+1})$. Thus based on the value function calculated over all the γ^k where $0 \leq \gamma \leq 1$, we can estimate the hyperbolically-discounted advantage. Note that the effective discount factor is γ^k, not the original γ. From a practical perspective, we consider a finite set of γ (consequently γ^k) to approximate the advantage. Using a multihead architecture, where each head corresponds to the value function for each γ^k, we minimize the average of the losses calculated for these multiple γ^k. Each loss function corresponding to a γ^k is defined as:

$$L_v^{\gamma^k}(\theta) = \hat{\mathbb{E}}_t\Big[\Big(V_\theta^{\gamma^k}(s_t) - \hat{V}_{targ}^{\gamma^k}\Big)^2\Big]$$

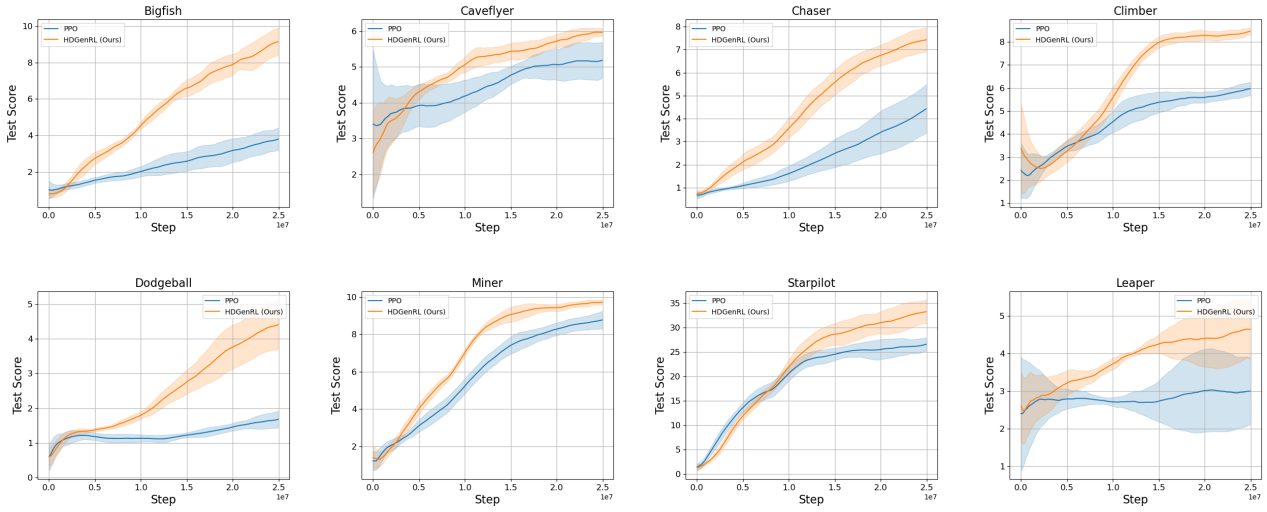

Figure 2. Test performance of our proposed HDGenRL and PPO (Schulman et al., 2017) over eight Procgen environments.

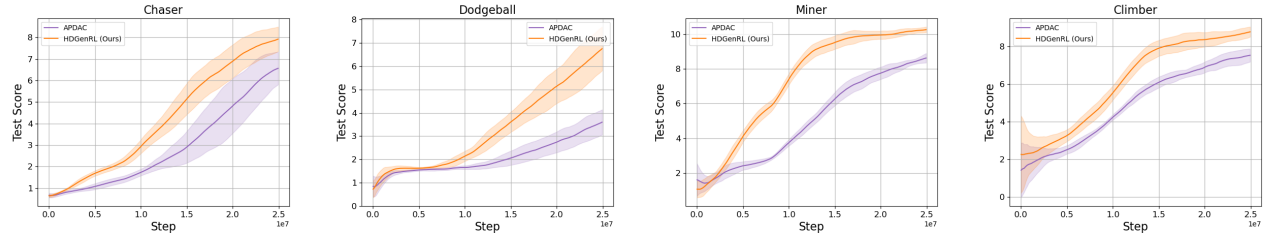

Figure 3. Test performance of APDAC (Nafi et al., 2021) and corresponding HDGenRL version over four Procgen environments.

6. Experiments and Results

Following previous works on Procgen, we use the IMPALA-CNN architecture as the actor-critic model for the PPO baseline in our experiments (Cobbe et al., 2020). This CNN architecture has three similar blocks each with 5 convolutional layers. To implement our proposed approach HDGenRL, we augment the same architecture with five value heads corresponding to five different γ. Then we calculate the advantage value for each of the value predictions. Finally, we integrate these advantage values to obtain the approximation for the hyperbolic advantage.

We train the model for 25M time steps. Figure 2 shows the experimental results on the test distribution of levels for 8 out of the 16 environments from Procgen, and presents rolling mean test scores and standard deviations calculated over five trials. The results indicate that the proposed HDGenRL significantly outperforms the PPO baseline on the test levels. Since PPO was not specifically designed for generalization, we further compare our approach against APDAC, a recent generalization-specific approach, to get a robust performance comparison on the generalization task.

Figure 3 shows that our proposed hyperbolic discounting based counterpart performs much better than APDAC on the test distribution of four selected games.

7. Discussion

This work presents a hyperbolic discounting based method of estimating advantages and applies it on the generalization task. We argue that since the underlying hazard rate in a procedurally generated environment is more uncertain, having an agent that discount future rewards hyperbolically would perform better on unseen levels. Through out the training, the agent learns the advantage estimate simultaneously over multiple horizons through the exponential discount factors $\gamma_0, \gamma_1, ..., \gamma_n$, which has been shown to be an approximation of hyperbolic discounting. We evaluate our proposed method of hyperbolically discounted advantage estimation on PPO and APDAC, and results show that the modified agent performs well on more than half of the tasks. We plan on extending this work by testing the proposed method against more recent state-of-the-art methods, varying number of discount factors, and other ablation studies.

References

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

Cobbe, K., Klimov, O., Hesse, C., Kim, T., and Schulman, J. Quantifying generalization in reinforcement learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1282–1289. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cobbe19a.html.

Cobbe, K., Hesse, C., Hilton, J., and Schulman, J. Leveraging procedural generation to benchmark reinforcement learning. In *International conference on machine learning*, pp. 2048–2056. PMLR, 2020.

Cobbe, K. W., Hilton, J., Klimov, O., and Schulman, J. Phasic policy gradient. In *International Conference on Machine Learning*, pp. 2020–2027. PMLR, 2021.

Dasgupta, P. and Maskin, E. Uncertainty and hyperbolic discounting. *American Economic Review*, 95(4):1290–1299, 2005.

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

Grigsby, J. and Qi, Y. Measuring visual generalization in continuous control from pixels. *CoRR*, abs/2010.06740, 2020. URL https://arxiv.org/abs/2010.06740.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., and Silver, D. Rainbow: Combining improvements in deep reinforcement learning. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Hu, T., Wang, W., Lin, C., and Cheng, G. Regularization matters: A nonparametric perspective on overparametrized neural network. In *International Conference on Artificial Intelligence and Statistics*, pp. 829–837. PMLR, 2021.

Igl, M., Ciosek, K., Li, Y., Tschiatschek, S., Zhang, C., Devlin, S., and Hofmann, K. Generalization in reinforcement learning with selective noise injection and information bottleneck. *Advances in neural information processing systems*, 32, 2019.

Justesen, N., Torrado, R. R., Bontrager, P., Khalifa, A., Togelius, J., and Risi, S. Illuminating generalization in deep reinforcement learning through procedural level generation. *arXiv preprint arXiv:1806.10729*, 2018.

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

Mazur, J. E. Choice, delay, probability, and conditioned reinforcement. *Animal Learning & Behavior*, 25(2):131–147, 1997.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Nafi, N. M., Glasscock, C., and Hsu, W. Attention-based partial decoupling of policy and value for generalization in reinforcement learning. In *Deep RL Workshop NeurIPS 2021*, 2021.

Raileanu, R. and Fergus, R. Decoupling value and policy for generalization in reinforcement learning. In *International Conference on Machine Learning*, pp. 8787–8798. PMLR, 2021.

Rajeswaran, A., Lowrey, K., Todorov, E., and Kakade, S. Towards generalization and simplicity in continuous control. *arXiv preprint arXiv:1703.02660*, 2017.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Sutton, R. S., Barto, A. G., et al. Introduction to reinforcement learning. 1998.