Multi-Horizon Learning in Procedurally-Generated Environments for Off-Policy Reinforcement Learning (Student Abstract)

Raja Farrukh Ali, Kevin Duong, Nasik Muhammad Nafi, William Hsu

Department of Computer Science, Kansas State University
{rfali, kevduong, nnafi, bhsu}@ksu.edu

Abstract

Value estimates at multiple timescales can help create advanced discounting functions and allow agents to form more effective predictive models of their environment. In this work, we investigate learning over multiple horizons concurrently for off-policy reinforcement learning by using an advantage-based action selection method and introducing architectural improvements. Our proposed agent learns over multiple horizons simultaneously, while using either exponential or hyperbolic discounting functions. We implement our approach on Rainbow, a value-based off-policy algorithm, and test on Procgen, a collection of procedurally-generated environments, to demonstrate the effectiveness of this approach, specifically to evaluate the agent's performance in previously unseen scenarios.

Introduction

A reinforcement learning agent learns to maximize the rewards it receives over its expected lifetime. But what if its estimate of the expected lifetime is biased? While RL algorithms usually plan for a single, fixed, *distant* horizon by setting the discount factor γ to a value closer to 1, the agent may not live long enough as it had previously expected. Thus, it might have fared better by basing its policy on estimates over multiple time horizons; short, intermediate, and long. For example, humans make plans for the immediate future (work day), short-term (professional goals), and long-term (retirement savings), but a person's policy (based on the valuation of distant rewards) may change drastically because of a terminal diagnosis (curtailment of expected lifetime). Since the goal of an RL agent is to optimize its return (cumulative reward) through a combination of prediction and control, the prediction method used to estimate the value function (estimated return) plays a central role in the agent's performance.

The discount factor γ determines the timescale of the return. When γ is closer to 0, the agent becomes short-sighted and only maximizes near-term reward, whereas when γ reaches closer to 1, the agent values rewards far into the future. In this work, we explore multi-horizon learning in procedurally generated environments, which are designed to test an agent's ability to learn a robust, generalizable policy.

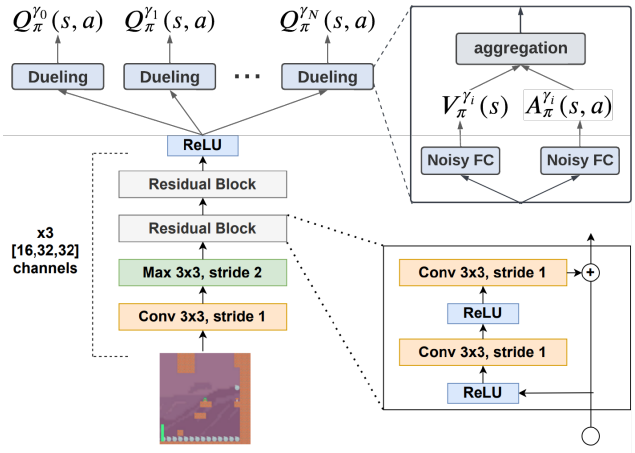

Figure 1: Multi-Horizon network architecture. Output layers predict Q-values for different discount factors using an individual output block for each gamma. Advantages from the dueling head for each gamma value are used for action selection.

The ideas behind our approach are based on work by Fedus et al. (2019) and we propose novel contributions by leveraging recent advances such as hyperbolic advantage (Nafi, Ali, and Hsu 2022) and an efficient Rainbow variant (Schmidt and Schmied 2021). We evaluate both exponential and hyperbolic discounting functions with advantage-based action selection in the multi-horizon setting and compare and contrast these with learning over a fixed, single-horizon setting.

Methodology

We use Rainbow (Hessel et al. 2018), a value-based RL method to evaluate multi-horizon learning in off-policy reinforcement learning. We evaluate the agent on the Procgen benchmark (Cobbe et al. 2020), which is designed to study sample efficiency and generalization in reinforcement learning. The agent is generally trained on a smaller number of levels and expected to perform well in diverse unseen levels. The single-horizon agent uses vanilla Rainbow and learns over a fixed horizon with a single $\gamma = 0.99$ and discounts rewards exponentially. For multi-horizon, the agent

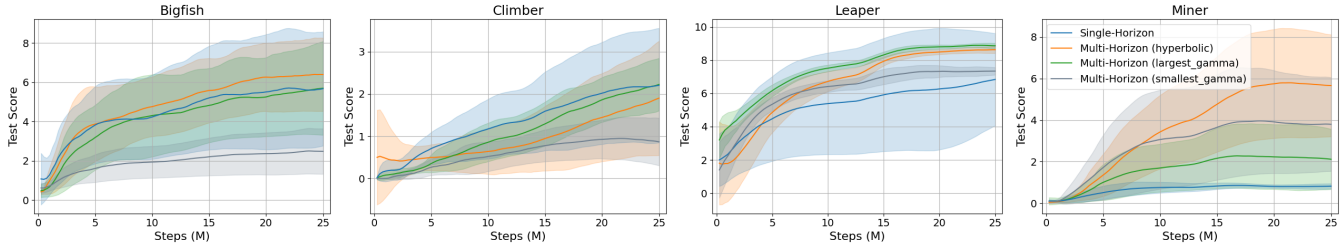

Figure 2: Test performance of Single-Horizon and Multi-Horizon variants (actions based on advantage values corresponding to smallest gamma, largest gamma, or hyperbolically discounted advantage) of Rainbow over a subset of 4 Procgen environments.

simultaneously learns Q-values for n_γ, but has the choice between acting either using a hyperbolically discounted advantage, $A^{\Gamma}_{\pi}(s,a)$, or exponentially discounted advantage, $A^{\gamma_n}_{\pi}(s,a)$. This exponentially discounted advantage can correspond either to the smallest gamma value (γ_0) making the agent's actions myopic, or the largest gamma value (γ_N), resulting in a far-sighted agent. Unlike Fedus et al. (2019), we use the dueling architecture as part of Rainbow and use the advantage from the dueling head for action selection (Figure 1). For calculating hyperbolically discounted advantage, as introduced for on-policy RL methods by Nafi, Ali, and Hsu (2022), we use the hyperbolic function evaluation $\Gamma_k(t) = \frac{1}{1+kt} = \int_0^1 \gamma^{kt} d\gamma$ and approximate hyperbolically discounted advantage using a Riemann sum over the discrete interval $G = [\gamma_0, \gamma_1, ..., \gamma_N]$ as:

$$A^{\Gamma}_{\pi}(s,a) \approx \sum_{\gamma^i \in G} w(\gamma_i) A^{\gamma^i}_{\pi}(s,a)$$

where weights $w(\gamma_i) = (\gamma_{i+1} - \gamma_i)$. In our experiments, we set n_γ = 5 and k=0.1 [γ = 0.906, 0.951, 0.972, 0.985, 0.99], and evaluated the agent using the 'Easy' distribution mode of Procgen with 5 seeds per environment.

Results

Results indicate that agents which model value estimates over multiple timescales generally perform better than their single-horizon counterparts (Figure 2). Moreover, advanced discounting functions like hyperbolic that are leveraged through the use of multiple timescales perform better or at par with exponentially-discounted, multi-horizon (largest gamma) variants. Agents which learn over multiple horizons but act myopically (smallest gamma) can also sometimes fare better than single horizon baselines, by virtue of their ability to learn from farther horizons and act shortsightedly. The reason why no particular method performs well across all environments can be attributed to the fact that an agent's prior belief of the risk in the environment influences the specific discounting function used (Fedus et al. 2019). However, for trials across the full suite of 16 environments over 25M timesteps (Figure 3), the performance profile of multi-horizon (hyperbolic) is higher (and better) than the single-horizon and multi-horizon (largest gamma) variants.

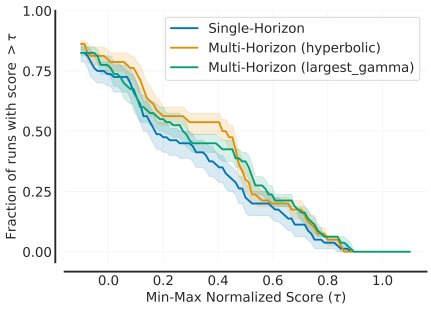

Figure 3: Min-Max normalized performance profiles (Agarwal et al. 2021) across all 16 Procgen environments.

Conclusions

We report initial results on the use of multi-horizon learning using advantage-based action selection in procedurally generated environments. Our contribution can be seen as validating the impact of multi-horizon learning on an agent's policy through learning accurate value estimates, achieving higher sample efficiency and better generalization.

References

Agarwal, R.; Schwarzer, M.; Castro, P. S.; Courville, A.; and Bellemare, M. G. 2021. Deep Reinforcement Learning at the Edge of the Statistical Precipice. *NeurIPS*.

Cobbe, K.; Hesse, C.; Hilton, J.; and Schulman, J. 2020. Leveraging procedural generation to benchmark reinforcement learning. In *ICML*, 2048–2056. PMLR.

Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H. 2019. Hyperbolic discounting and learning over multiple horizons. In *RLDM*.

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

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

Schmidt, D.; and Schmied, T. 2021. Fast and Data-Efficient Training of Rainbow: an Experimental Study on Atari. In *Deep RL Workshop NeurIPS*.