

Hierarchical Average Reward Policy Gradient Algorithms (Student Abstract)

Akshay Dharmavaram
 Birla Institute of Technology and
 Science, Pilani
 f20150039@goa.bits-pilani.ac.in
 +91-9604079793

Matthew Riemer
 IBM Research
 mdriemer@us.ibm.com

Shalabh Bhatnagar
 Indian Institute of Science, Bangalore
 shalabh@iisc.ac.in

Abstract

Option-critic learning is a general-purpose reinforcement learning (RL) framework that aims to address the issue of long term credit assignment by leveraging temporal abstractions. However, when dealing with extended timescales, discounting future rewards can lead to incorrect credit assignments. In this work, we address this issue by extending the hierarchical option-critic policy gradient theorem for the average reward criterion. Our proposed framework aims to maximize the *long-term* reward obtained in the steady-state of the Markov chain defined by the agent’s policy. Furthermore, we use an ordinary differential equation based approach for our convergence analysis and prove that the parameters of the intra-option policies, termination functions, and value functions, converge to their corresponding optimal values, with probability *one*. Finally, we illustrate the competitive advantage of learning options, in the average reward setting, on a grid-world environment with sparse rewards.

Introduction

Humans routinely employ high-level temporal abstractions for everyday decision making. Bacon, Harb, and Precup (2017) investigate the use of learning temporally extended abstractions to augment the exploration and credit assignment capabilities of the actor-critic framework. However, employing a discount factor to bound the cumulative rewards can inadvertently lead to incorrect credit assignment. We address this by extending the framework proposed by Riemer et al. (2020) for the average reward (AR) criterion¹.

Figure 3(a) is a motivating example that illustrates how simple *traps* can beguile the discounted rewards (DR) framework into learning a sub-optimal credit assignment. It illustrates two different Markov chains, resulting from two disparate policies (π_R and π_B). π_R always chooses red and π_B always chooses blue. When a DR-RL agent is at S_0 , it prefers the sub-optimal policy π_B , because $\forall \gamma < 1$:

$$v_{\pi_R}(S_{11}) = \frac{\gamma(2 - \gamma)}{(1 - \gamma^4)} < \frac{1}{(1 - \gamma^4)} = v_{\pi_B}(S_{21})$$

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹See an extended version of this paper including the appendix at <https://arxiv.org/pdf/1911.08826.pdf>

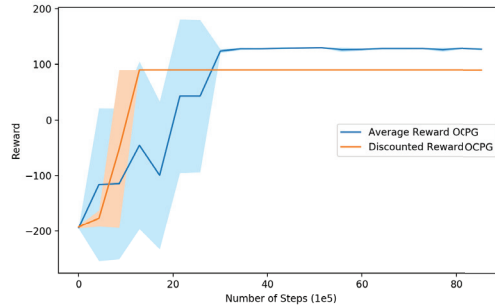


Figure 1: Mean and standard deviations of the learning curves for the average reward and discounted reward OCPG agents, in the grid-world delivery experiment.

Policy-Gradient with Function Approximation

First, we illustrate how to extend the OCPG framework by Riemer et al. (2020) for the AR criterion. Apart from addressing the AR criterion, our framework also presents a simplified and intuitive approach to dealing with hierarchical option-critic algorithms (Riemer, Liu, and Tesauro 2018) by introducing the concept of o^0 and o^N .

Theorem 1 (Hierarchical Average Reward Option-Critic Policy Gradient Theorem). *Given an N level hierarchical set of Markov options with stochastic option policies at each level π^ℓ and termination functions at each level β^ℓ differentiable in their parameters θ , the gradient of the expected reward per step with respect to θ is:*

$$\sum_{s, o^{0:N-1}, s'} \mu_\Omega(s, o^{0:N-1}, s') \left(\sum_a \frac{\partial \pi(a|s, o^{0:N-1})}{\partial \theta} \right. \\
Q_U(s, o^{0:N-1}, a) + \sum_{o^{0:N-1}} \sum_{\ell=1}^{N-1} \left[\prod_{k=N-1}^{\ell} \beta^k(s', o^{0:k}) \right. \\
\left. \frac{\partial \pi^\ell(o^\ell|s', o^{0:\ell-1})}{\partial \theta} Q_\Omega(s', o^{0:\ell}) P_{\pi, \beta}(o^{0:\ell-1}|s', o^{0:\ell-1}) \right. \\
\left. - \frac{\partial \beta^\ell(s', o^{0:\ell})}{\partial \theta} A_\Omega(s', o^{0:\ell}) \prod_{k=N-1}^{\ell+1} \beta^k(s', o^{0:k}) \right] \Bigg),$$

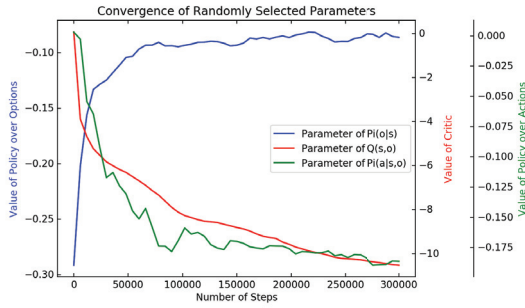


Figure 2: An empirical demonstration illustrating the convergence of the parameters of $Q(s, o)$, $\pi(a|s, o)$, and $\pi(o|s)$. We have randomly selected one parameter from each function approximator and plotted its value against the steps.

where μ_Ω is the stationary distribution of the Markov chain defined by the hierarchical policy, and $P_{\pi, \beta}$ is the probability while at the next state, and terminating the options for the last state, that the agent arrives at a particular new set of option selections.

Proof. The proof for this theorem is in the Appendix. \square

Two-Timescale Convergence

Next, we prove that the aforementioned parameters, θ , asymptotically converge to their optimal values, when employing a linear approximation $\forall Q_\Omega$. We analyze our framework using the ordinary differential equation (ODE) approach, delineated by Bhatnagar et al. (2009), and study its asymptotic properties using the fixed points of the ODE.

Theorem 2 (Convergence Proof). *For the parameter iterations of the global set of shared parameters defined in Algorithm 1, we have $(\hat{J}_t, v_t, \theta_t) \rightarrow \{(J(\theta^*)_t, v^*, \theta^*) | \theta^* \in \mathcal{Z}\}$ as $t \rightarrow \infty$ with probability one, where \mathcal{Z} corresponds to the set of local maxima of a performance function whose gradient is $E[\delta_t^\pi \psi(s_t, a_t) | \theta]$*

Proof. The proof for this theorem is in the Appendix. \square

Empirical Results

Finally, we look at the susceptibility of our framework to traps, and compare it to the DR setting proposed by Riemer et al. (2020). Figure 3(b) depicts a grid world environment characterized by sparse rewards. An agent must navigate to either one of the pickup locations, P_1 or P_2 , in order to retrieve a parcel; and must subsequently deliver the parcel to the drop off location. The agent gets a reward of +100 for every parcel from P_2 , and +50 for every parcel from P_1 . The optimal policy for an agent would naturally involve picking up the parcels from P_2 .

We introduce a trap¹ at the green-blue junction to entice the DR-RL agents into picking up the parcels from P_1 . Once

¹The reward of +20 was primarily chosen for illustrating the potential pitfalls when employ a $\gamma \leq 0.9$. Similar traps can be created for any $\gamma \leq 1$.

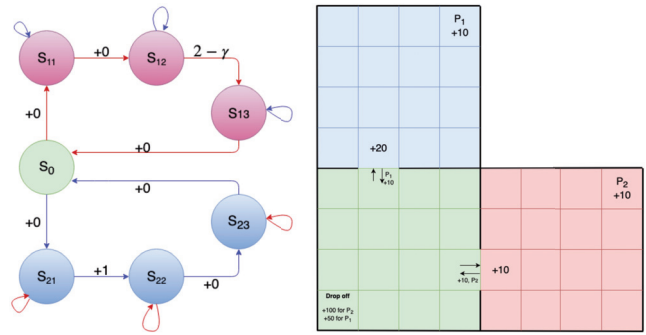


Figure 3: (a) A trap that employs delayed rewards to fool DR-RL agents into learning incorrect credit assignments. (b) A grid-world navigation experiment where the reward at the drop off point depends upon which pickup location was previously visited (50 for P_1 and 100 for P_2). The trap at the blue-green junction misguides agents towards the sub-optimal pickup location, P_1 .

the agent reaches the blue zone, it obtains a reward of +20 as opposed to a reward of +10 at the red-green junction. In Figure 1, we plot the rewards obtained per cycle for both the AR-RL agent and a DR-RL agent, and show that the hierarchical AR policy gradient performs better than its DR counterpart proposed by Riemer et al. (2020). Finally, we illustrate the asymptotic convergence of the actor and critic parameters in Figure 2.

Conclusion and Future Work

In this work, we propose a novel method for maximizing the long term steady-state reward, by learning intra-option policies, termination functions, and value functions end-to-end. These algorithms can be used in infinite-horizon control problems that exhibit an inherent cyclic structure, like inventory-management, queuing and traffic light control. A detailed empirical analysis for a cyclical infinite-horizon application would be necessary to demonstrate the viability of our approach in complex environments. Additionally, while the proofs provided here leverage a linear approximation for each of the $Q_\Omega(s, o^{0:\ell})$, it would also be interesting to investigate the convergence properties of a non-linear critic.

References

Bacon, P.-L.; Harb, J.; and Precup, D. 2017. The option-critic architecture. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Bhatnagar, S.; Sutton, R. S.; Ghavamzadeh, M.; and Lee, M. 2009. Natural actor-critic algorithms. *Automatica* 45(11):2471–2482.

Riemer, M.; Cases, I.; Rosenbaum, C.; Liu, M.; and Tesauro, G. 2020. On the role of weight sharing during deep option learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Riemer, M.; Liu, M.; and Tesauro, G. 2018. Learning abstract options. In *Advances in Neural Information Processing Systems*, 10424–10434.