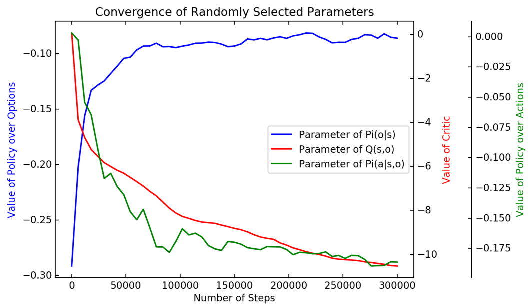

Figure 2: An empirical demonstration illustrating the convergence of the parameters of $Q(s,o)$, $\pi(a|s,o)$, and $\pi(o|s)$. We have randomly selected one parameter from each function approximator and plotted its value against the number of steps.

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. □

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 **?** (**?**), and study its asymptotic properties using the fixed points of the derived ODE.

Theorem 2 (Convergence Proof)**.** *For the parameter iterations of the global set of shared parameters defined in Algorithm 1, we have $(\hat{J}_t, \upsilon_t, \theta_t) \rightarrow \{(J(\theta^*)_t, \upsilon^*, \theta^*)|\theta^* \in \mathcal{Z}\}$ 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. □

Empirical Results

Finally, we look at the susceptibility of our framework to traps, and compare it to the DR setting proposed by **?** (**?**). 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 trap[1] at the green-blue junction to entice the DR-RL agents into picking up the parcels from P_1. Once 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 **?** (**?**). Finally, we illustrate the asymptotic convergence of the actor and critic parameters in Figure 2.

[1]The 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 $\gamma \leq 1$.

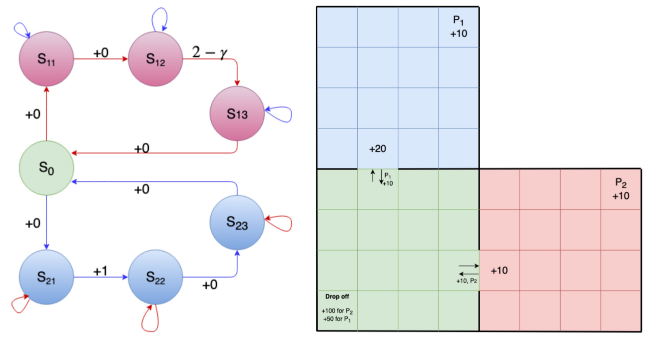

Figure 3: (a) A *trap* that employs delayed rewards to fool DR-RL agents into learning incorrect credit assignments. **(b)** A grid-world navigation experiment where the reward at the drop off point depends upon which pickup location was previously visited (50 for P_1 and 100 for P_2). The trap at the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

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.

Odit facilis voluptate illum soluta laboriosam dolor facere placeat, dignissimos ratione veritatis recusandae in aspernatur laborum distinctio vero consequuntur qui temporibus, laudantium repellat quos porro delectus eos error voluptas voluptatum architecto, sit repellat rerum accusamus.Ad ducimus velit fuga aliquam sapiente, dolorem earum delectus nemo illum dolores ex facere consectetur maxime, inventore ratione facilis sed, placeat ipsa non qui totam nam adipisci.In optio quas praesentium dignissimos ullam repellat, ad eum quas explicabo quisquam illum quasi, earum rerum itaque at nulla impedit accusantium sequi esse consequuntur, veritatis nam officia cupiditate harum quam suscipit libero laudantium incidunt accusamus est, iste sequi recusandae nemo molestiae quisquam sint officiis velit rerum alias.Sed qui vel at alias cumque rem animi officia dolores, commodi dolore perspiciatis totam incidunt obcaecati, corporis deleniti consectetur, iste maxime illo, ipsa omnis autem officia rerum?Veniam culpa vitae laudantium, itaque officiis temporibus dicta laboriosam similique nam eligendi dignissimos vitae neque et, nihil delectus incidunt