Paper Review: Q-PROP: SAMPLE-EFFICIENT POLICY GRADIENT WITH AN OFF-POLICY CRITIC

Summary:

The authors of this paper develop methods for a policy gradient algorithm, called Q-Prop, that leverages on- and off-policy methods to improve sample efficiency and create stability. To guide the optimal policy updates, Q-Prop uses a critic network to estimate the state-action value function. This method allows Q-Prop to use off-policy data to update the policy, which improves sample efficiency and leads to faster convergence. The paper uses OpenAI Gym physics simulator MuJoCo to demonstrate that Q-Prop outperforms TRPO and DDPG on several benchmark environments. Overall, the paper demonstrates the effectiveness of Q-Prop in improving the sample efficiency and convergence speed of policy gradient algorithms.

Contributions:

The main issue current Deep RL algorithms have is that they require an extremely high sample to train that it becomes infeasible to implement in the real world. The contribution of this paper is demonstrating that Q-Prop has a faster convergence rate compared to traditional on-policy policy gradient methods, this drastic changes makes real world applications attainable. Additionally, the paper also proposes a new method for computing the advantage function, something unique to this paper, which further improves the stability and convergence of the policy updates in Q-Prop.

Strengths and Weaknesses:

I believe the papers' major strength is in its ability to clearly setup and explain the problem they are attempting to solve. The introduction and background section were very well written to educate the reader on the Monte Carlo and Policy Gradient methodologies. While the paper provides strong evidence for faster convergence, I believe this paper falls short by failing to test Q-Prop in any real-world environment.

Experimental Validity:

The experimental evaluation in this paper is thorough and uses seven benchmark environments to compare Q-Prop to several policy gradient algorithms. The paper fails to explain the goal of each environment that it is put in and assumes that the reader already knows this. Overall, the experimental results demonstrate that theoretically Q-Prop outperforms existing algorithms in terms of sample efficiency and convergence speed.

How can this work be extended:

As discussed previously, one necessary extension of this work would be to apply Q-Prop to real world environments, such as robotics or real time natural language processing. Another

possible extension could be to study the use of Q-Prop in a system with other RL techniques to take advantage of its faster convergence time. Additionally, it would be interesting to explore the use of Q-Prop in multi-agent RL settings, such as a game of chess or poker. The work done in this paper is bringing efficiency and practicality to RL applications and that's where it would be exciting to see this research headed to next.