Paper Review-"*Q-Prop: Sample-Efficient Policy Gradient with An Off-Policy Critic, Gu et al,2016*"

Summary: The authors build upon to develop methods that combine the stability of policy gradient with the efficiency of off-policy RL algorithms. They presented a Taylor expansion of the off-policy critic as a control variate to prove that the new algorithm is much more stable and sample efficient than any other on-policy or off-policy algorithm as it effectively combines the benefits of both. The algorithm, called Q-Prop, uses an off-policy critic network to estimate the Q-function and a trust region approach to limit the distance between the updated policy and the old policy. It works by first training a critic network to estimate the Q-function for a given policy. This critic network is trained off-policy using previously collected data, which allows Q-Prop to make use of past experience and improve sample efficiency.

Main Contributions: They introduce a new algorithm called Q-prop. The core idea behind this algorithm is to use the first-order Taylor expansion of critic as a control variate. This in turn results in an analytical gradient term through the critic and a Monte Carlo policy gradienr term. They mention tah this insight comes from the fact that the critic Q_w can be trained using off-poilcy data. They introduce 2 variants of Q-prop - adaptive Q-Prop and conservative Q-Prop. The main difference between the two is how they update the critic network during training. In Adaptive Q-Prop, the critic network is updated using a variant of the Q-learning update rule that is adapted to the off-policy setting. In Conservative Q-Prop, the critic network is updated using a conservative update rule that ensures that the Q-function estimate remains lower-bounded by the minimum Q-value observed during training.

Strengths: Q-Prop tries to get the best of both on-policy and off-policy methods. It has improved stability, monotic learning over algorithms like DDPG and has better sample efficiency over algorithms like TRPO. It also relatively simple to implement with less hyperparameters to tune which contributes to the stability.

Weakness: As mentioned in the paper, the compute time per episode is bound by the critic training at each iteration which is slow and if the data collection is very fast, it creates a bottleneck for the algorithm. Another limitation is the model's robustness to bad critics i.e estimating when an off-policy critic is reliable or not is still a fundamental problem of the algorithm.

Experiments: Q-prop was evaluated on various continuous control environments. They first compare the standard Q-prop algorithm with it's variants. They show that results are inline with the theory that conservative Q-Prop achieves a much more stable performance than others and all the Q-prop variants outperform TRPO in terms of sample efficiency which is SOTA. Next, they evaluate the two variants against other model-free algorithms on the common problem of HalfCheetah-v1. They show that conservative Q-Prop outperforms the best TRPO and VPG methods. It is notes that DDPG has inconsistent results and vary a lot on hyperparameter tuning. With a certain set of hyperparameter, DDPG did beat Q-Prop but it is very tough to arrive to the right set of hyperparameters hence DDPG is very unstable whereas Q-Prop is very stable. They also the evaluated Q-Prop in different domains and conclude that it is very stable.

Extensions: We can add some regualrization terms to the policy optimization objective in order to encourage exploration and prevent premature convergence to suboptimal policies. Another possible extension is to try adding model-based approach instead of model free to improve sample efficiency.