Paper Review - "*Bridging the Gap Between Value and Policy Based Reinforcement Learning, Nachum et al, 2017*"

Summary: The authors exploit a relationship between policy optimization under entropy regularization and softmax value consistency to obtain a new form of stable off-policy learning. They contribute new observations to the study based on which they develop two novel off-policy RL algorithms (PCL and unified PCL) that minimize a notion of soft consistency error along multi-step action sequences extracted from both on- and off-policy traces. They observe that under the new objective actor-critic, it can be unified in a single model that coherently fulfills both roles.

Main Contributions: The authors develop two key observations as described above based on which they develop the algorithm Path Consistency Leaning (PCL) and Unified PCL. PCL attempts to minimize what they call squared 'soft consistency' error over a set of sub-trajectories. PCL exploits off-policy trajectories by maintaining a replay buffer. Since soft consistency errors can also be expressed in terms of Q-values they introduce a new approach called Unified PCL. It optimizes the same objective as PCL but differs by combining the policy and value functions into a single model. This is one of the main differences as it presents a new actor-critic paradigm where the policy(actor) is not distinct from the values (critic).

Strengths: The authors drew connections to other actor-critic methods like A2C and A3C and also Q-Learning. They mention that their approach is better but also robust enough that it can also be generalized as Q-Learning. They mention that we can interpret the notion of temporal consistency proposed in the paper as a generalization of the one-step temporal consistency given by hard-max Q-values.

Weakness: As mentioned in the paper, one of the major limitations of the algorithm is seen in stochastic settings. The squared inconsistency objective approximated by Monte Carlo samples is a biased estimate of the true squared inconsistency.

Experiments: They evaluated the proposed algorithms across several tasks and compared them to A3C and Q-Learning. What they observed here is that PCL consistently either matches or beats the performance of these baselines. The gap is hard to observe in simpler tasks like Copy, Reverse, etc but a more noticeable gap is observed in harder tasks like ReversedAddition. Next, they compared PCL and Unified PCL to find that the use of a single unified model for both values and policy is competitive with PCL on more difficult tasks. Lastly, for more difficult tasks they also experiment with seeding the replay buffer with 10 randomly sampled trajectories.

Extensions: Similar to Q-Learning, the PCL approach also faces that the squared inconsistency objective approximated by Monte Carlo samples is a biased estimate of the true squared inconsistency. To be able to correct that we can try certain remedies mentioned in the paper "Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path" where they prove that a linear parameterization is equivalent to Least-Squares Policy Iteration. We can see how this helps improve the current model.