Constrained Policy Optimization

Avinash G. Kori — ED15B006

Summary

In this manuscript authors proposed new update rule for trust region based policy optimization for Constrained Markov Decision Processes (CMDP). They also provided the motivation for the use of CMDP, along with convergence guarantees for their new constrained policy update algorithm. Authors also proposed the approximation for the theoretical constraints and objective with surrogate functions, which are easy to estimate from samples, and based on their theoretical analysis, they provided bound for updates in worst-case performance and worst-case constraint violation with values that depend on a hyperparameter of the algorithm.

Major strengths of the paper:

- Manuscript is theoretically sound, Authors provided tight convergence bounds for proposed policy update algorithm
- New theoretical tighter bounds on the difference in returns between two arbitrary policies is presented, exploiting the Jensons and total variational - KL divergence inequality. These are useful as (i) It helps tightening the connection between theory and practice in reinforcement learning, (ii) It also helps to consider CPO as a trust region optimization method
- For all the practical applications, objective function is well approximated by linearizing around a given policy, and KL divergence constraints are approximated by second order expansion
- Approximated function formulation turns out to be constrained convex optimization problem, which was further simplified by solving Lagrangian dual of the primal problem
- In all the previously existing algorithms like (PDO) intermediate policies were not guaranteed to satisfy constraints, only the policy at convergence was guaranteed. But the proposed method CPO computes new dual variables from scratch at each update to exactly enforce constraints.
- Comparative study between various policy gradient methods (TRPO, PDO & CPO) is well established
- The proposed algorithm was tested with and without cost shaping (i.e with and without proposed modification in cost function) along with two different constraints, (i) The agent is rewarded for running in a wide circle, but is constrained to stay within a safe region smaller than the radius of the target circle (ii) The agent is rewarded for collecting green apples, and constrained to avoid red bombs. In both the cases algorithm was tested on various Mujoco environments (like Point-Circle, Ant-Circle, Humanoid-Circle, Point-Gather, and Ant-Gather) and the algorithm provides good results in maintaining the limiting constrained value with low returns
- Manuscript has clear description of the methods, well written, and posses a novel approach for CMDP optimization

Major weaknesses of the paper:

- Proposed method involves second order oracle information: costlier operation and sometime may not be available, this case is not addressed (Hessian can be simulated using gradient information using somesort of BFGS with Rank 1 matrix update)
- Clear way of generating surrogate functions is not discussed in the manuscript
- All the experiments were conducted using very similar environments, authors should consider experimenting on different kinds of environments like: Car Racing, Mario, and many other cases
- All the constraints used were in terms of distance, authors should also consider experimenting with variance, correlation or some other statistics based constraints