Review-16

Conservative Safety Critics For Exploration

In various critical scenarios a Reinforcement Learning (RL) agent may desire safe exlporation scenarios in order to evade catastrophic states. A natural alternative is conservative safety behavior which allows the agent to learn safety estimate of states. Conservative Safety Critics (CSC) learn a conervative safety estimate of states through a critic which overstimates the probability of failure using Conservative Q-Learning (CQL). The work provides theoretical characterizations on the trade-off between safety and policy improvement and derives provable convergence guarantees. CSC demonstrates lower failure rates while achieving competitive task rewards.

CSC makes use of CQL to overestimate the critic Q-function Q_C which indicates the safety estimate corresponding to the probability of failure. Actions from the agent's policy are sampled using rejection sampling wherein the sampled action is only executed if Q_C is less than a threshold ϵ. Policy optimization is modeled as an optimization problem which is carried out using primal dual descent. Policy parameters are updated using gradient ascent on a KL-constrained objective based on Fisher Information Matrix while the Lagrange multiplier λ is updated using gradient descent. Parameters of the conservative critic are updated using gradient descent on estimates from on-policy and off-policy experiences. Following CSC, the work presents theoretical insights into the trade-off between policy improvement and safety. Expected probability of failure $V_C^{\pi_\phi}(\mu)$ under policy π_ϕ is upper bounded during every policy update iteration. Additionally, convergence rate for policy gradient updates has an upper bounded which is as good as standard RL. Theoretical claims of CSC on reward-safety trade-off and convergence rate are validated using empirical evaluations on a suite of robot control tasks.

CSC is evaluated on navigation, manipulation and locootion tasks in comparison to challenging baselines. EXperiments depict lower failure rates for CSC while maintaining competitive task rewards. Additionally, ablation studies carried out on safety thresholds validate the theoretical claims. While the proposed approach is effective and suitable for safe exploration, its analysis poses two limitations. Firstly, CSC overstimates failure probabilities which motivate conservative exploration. Such a scheme hinders the agent from obtaining higher task rewards and may be problematic from its theoretical viewpoint. An increase in the overestimation gap may lead to a decrease in the upper bound of $V_C^{\pi_\phi}(\mu)$. This would lead to inaccurate estimates sub-optimal safety as the bound will not hold true. Lastly, at zero safety threshold the CSC agent should demonstrate no failures which is not the case in experiments. This could be further improved by utilizing a more robust safety framework which does not depend on the threshold heuristic of $Q_C \geq \epsilon$.

The CSC framework makes an apt use of CQL for learning safe estimates of states. The work suggests two new directions for future work. Firstly, the use of a heuristic for selecting actions from policy could be further improved by making use of a more robust scheme. Lastly, the provision of offline learning could be extended by pretraining the agent on a larger number of failure cases in absence of task rewards. This would motivate generalized behavior during policy execution in unforeseen circumstances.