Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Conservative Safety Critics for Exploration]
Date: [20-09-2024]

Make sure your critique Address the following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

Problem Statement:
The paper addresses the issue of *safe exploration in reinforcement learning (RL)*. Specifically, the goal is to ensure that the agent explores the environment without encountering catastrophic failures during training. In standard RL settings, agents can perform unsafe actions that lead to significant failures, particularly during early learning stages when the policy is not fully trained. The challenge is to balance the need for effective exploration with maintaining safety constraints, such that the probability of catastrophic failures is minimized throughout the training process.

2 Key contributions of the paper

1. The introduction of the *Conservative Safety Critics (CSC)* framework for safe exploration in reinforcement learning. The CSC framework estimates the safety of states and actions by conservatively overestimating the probability of catastrophic failures.
2. The paper provides *theoretical guarantees* that bound the probability of failures at each iteration during training, ensuring that the likelihood of catastrophic failures is minimized.
3. The use of a *KL-divergence constraint* in policy updates, which ensures that the state distribution induced by the new policy is not drastically different from the old policy, thereby limiting unsafe behavior during exploration.
4. The development of a safe RL algorithm with a *provable trade-off* between task performance and safety, offering bounds on failure probability while ensuring convergence rates that are not worse than standard RL methods.

3 Proposed algorithm/framework

Algorithm 1 CSC: safe exploration with conservative safety critics

Initialize V^0 (task value fn.), Q^0 (safety critic), policy π^0, D_{env}, thresholds δ, ξ.
Set $\mathcal{V}_c^0(\mu)$ (denotes avg. failures in the previous epoch).
for each until convergence **do**
    Execute actions in the environment. Collect on-policy samples.
    for episode e in $1, \ldots, M$ **do**
        Set $\epsilon = (1 - \xi) \times \mathcal{V}_c^{old}(\mu)$
        Sample $a \sim \pi^{old}(s)$. If $c(s, a) \leq \epsilon$, execute a. Else, resample a.
        Obtain next state s', r, c as $R(s, a)$, $C(s, a)$.
        $D_{\text{env}} \leftarrow D_{\text{env}} \cup \{(s, a, r, s', c)\}$
    end for
    Store the average episodic failures $\mathcal{V}_c^{old}(\mu) \leftarrow \sum \mathcal{V}_c^e$
    Update D_{env} with off-policy/online data if available.
    for step i in $1, \ldots, N$ **do**
        Gradient updates on Q and Optionally add Entropy regularization
        Gradient descent on Lagrange multiplier λ
    end for
    $\pi^{old} \leftarrow \pi$
end for

4 How the proposed algorithm addressed the problem

- **Conservative safety critic:** The algorithm leverages a *conservative safety critic* $Q_c(s, a)$ to assess the safety of actions before executing them. By using a threshold ϵ that depends on the estimated average episodic failures $\mathcal{V}_c^{old}(\mu)$, the agent ensures that it only executes actions that are predicted to be sufficiently safe, thereby preventing the occurrence of catastrophic failures during exploration.

- **Resampling unsafe actions:** If an action is deemed unsafe (i.e., if $Q_c(s, a) > \epsilon$), the algorithm resamples a different action, thus encouraging the exploration of safer actions while still enabling learning.

- **KL-divergence constraint:** The KL-divergence constraint limits the update of the policy by ensuring that the new policy does not diverge too much from the previous one. This reduces the risk of sudden unsafe behaviors during policy updates, further enhancing safe exploration.

- **Safety-aware policy optimization:** The algorithm balances task performance and safety by performing gradient updates on both the value function and the safety critic, while also adjusting the Lagrange multiplier λ to penalize unsafe actions more strongly, ensuring that the probability of failures remains low.

- **Offline and on-policy learning:** The use of both offline and on-policy data allows for effective learning while maintaining safety constraints, as offline data can provide additional safety information even when on-policy exploration is restricted.