Weekly Report (RL Group in UoA)

September 16, 2024

Recap on Plan

1. We focus on the robust constrained RL setting
2. We aim to develop a stronger adversary than the existing ones and train a more robust RL agent against it.

Recap on Motivation

1. Robust RL under safety setting adopts State-Adversarial Constrained MDP (SA-CMDP) $\mathcal{M} :=< \mathcal{S}, \mathcal{A}, R, C, P, \gamma, \nu >$ where $\nu : \mathcal{S} \to \mathcal{S}$ modifies agent's observation of states.
2. This adversary only modifies the agent's observation but not the true state, so does not impact the system dynamics.
3. The existence of an adversary amplifies the conventional model with the ability to represent malicious attacks and system malfunctions.

Recap on Motivation

1. The existing models usually suppose at each step t, the adversary attacks with a probability δ[3] (sometimes set to 1[1, 2]) and a 'radius' ϵ which bounds the distance between the perturbed state and the original state.
2. However, we think this existing paradigm cannot lead to a real trustworthy agent in SA-CMDP.
3. The reasons will be given in the following slides.

Our observation

1. Take the existing work[1] on SA-CMDP as an example.
2. The algorithm in this paper performs well (satisfies the constraint) with 'scenarios with stable dangerous levels across an execution'.
3. However, we observe it performs not well in the 'scenarios with unstable dangerous levels'.

Our observation

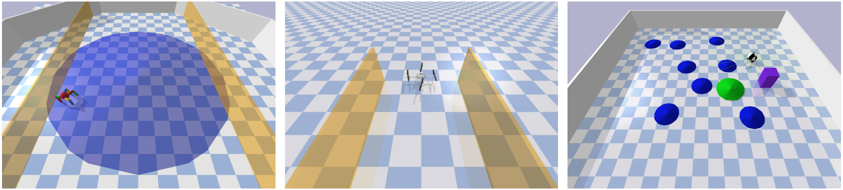

Figure: Circle Run and Reach

Our assumptions

- **How are the tasks different?** Different states in execution have different importance w.r.t a robust safe RL agent.
- **What does the difference imply?** The agent is more fragile toward attacks in some states.
- **What does the difference refer to?** Malicious attackers often flood the system when it is more fragile.
- **Where is the room for improvement?** Developing an adversary focuses on the 'fragile states'.

Our claims

1. We call the adversary attacks consistently across the entire execution **cons.adv.** ;
2. We propose a new adversary floods the agent only when the agent is fragile (**flood.adv.**);
   - The trained existing adversary cannot beat flood.adv. Against flooding adversary, the trained RL-agent is more robust.
   - Flood.adv. leads to similar performance but converges faster.
   - (At least) Flood.adv. leads to similar performance with fewer attacks.

Our TODO

1. Implement the flood.adv. and verify the claims by experiments.
2. Carefully check if this idea is novel (I believe the answer is affirmative at least in SA-CMDPs).
3. Implement 1st versions: a self-adaptive adversary who decides whether to attack based on q-value for cost[1]
4. Implement 2nd version: a model-based one that directly modifies the environment.

[1] Challenge: How do we assess the q-value for a state without a given action?

Small questions

1. Is it reasonable to set up a higher ϵ for flood.adv.
2. How does the attacking radius influence the result?
3. Can rl-agent handle flood-adv actually?
4. We need to take attack budget (numbers of attacks performed) as a metric as well?
5. Experiment on a new scenario (more realistic) is preferred, such as network defence.

Zuxin Liu, Zijian Guo, Zhepeng Cen, Huan Zhang, Jie Tan, Bo Li, and Ding Zhao.
On the robustness of safe reinforcement learning under observational perturbations.
arXiv preprint arXiv:2205.14691, 2022.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary.
Robust deep reinforcement learning with adversarial attacks.
arXiv preprint arXiv:1712.03632, 2017.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh.
Robust deep reinforcement learning against adversarial perturbations on state observations.
Advances in Neural Information Processing Systems, 33:21024–21037, 2020.