Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning

Prakash K. Naikade
prna00001@stud.uni-saarland.de

1 Summary

The paper investigates an underlying reinforcement learning (RL) security threat in which an attacker poisons the environment, causing the agent to implement a target policy. This paper demonstrates that poisoning transition dynamics optimally is far more difficult than poisoning rewards, and that the attack may not always be possible. The paper lays the theoretical groundwork for environment poisoning against RL in a variety of new attack dimensions like, adversarial manipulation of transition dynamics; attack against RL agents maximizing average reward in an undiscounted infinite horizon; and analyzing different attack costs for offline planning and online learning settings. The paper also provides sufficient technical criteria for the assault to be successful, as well as lower and upper bounds on the attack cost. According to experimental data attacker can easily teach any target policy to the RL agents under mild conditions.

1.1 Overview of attacks

Paper presents attacks in two settings: an offline setting where the agent is doing planning in the poisoned environment and an online setting where the agent is learning a policy using a regret-minimization framework with poisoned feedback.
Goal of the attack against an offline planning agent is defined as,

$$\rho(, \widehat{M}) \geq \rho(\pi, \widehat{M}) + \epsilon, \quad \forall \pi \neq \pi^{\dagger} \tag{1}$$

Where, ϵ is margin parameter and target policy is ϵ-robust optimal in the poisoned Markov Decision Process (MDP) $\widehat{M}$.
Cost of the attack against an offline planning agent is defined in terms of ℓ_p-norm of differences in reward functions (i.e., $\overline{R}$ and $\widehat{R}$) and in transition dynamics (i.e., $\overline{P}$ and $\widehat{P}$),

$$\left\|\widehat{R} - \overline{R}\right\|_p = \left(\sum_{s,a}\left(\left|\widehat{R}(s,a) - \overline{R}(s,a)\right|\right)^p\right)^{1/p}, for\ reward\ functions \tag{2}$$

$$\left\|\widehat{P} - \overline{P}\right\|_p = \left(\sum_{s,a}\left(\sum_{s'}\left|\widehat{P}(s,a,s') - \overline{P}(s,a,s')\right|\right)^p\right)^{1/p}, for\ transition\ dynamics \tag{3}$$

Goal of the attack against an online learning agent is defined as notion of *average mismatch* of learner's actions in time horizon T,

$$(T) = \frac{1}{T} \cdot \left(\sum_{t=0}^{T-1} a_t \neq (s_t)\right) \tag{4}$$

The aim of the attacker is to ensure that (T) is $o(1)$, or, alternatively, the total number of time steps where there is a mismatch is $o(T)$.

Preprint. Under review.

Attacks in Offline Setting via Poisoning Rewards is formulated as an optimization problem,

$$\begin{aligned} \min_{R} \quad & \left\| R - \overline{R} \right\|_p \\ \text{s.t.} \quad & \widehat{R} = R \text{ and } \widehat{\mu} = \overline{\mu} \end{aligned} \tag{P1}$$

The optimization problem (P1) is always feasible, and the cost of the optimum solution is bounded by,

$$\frac{\overline{\alpha}}{2} \cdot \left\| \overline{\chi}_\epsilon \right\|_\infty \leq \left\| \widehat{R} - \overline{R} \right\|_p \leq \left\| \overline{\chi}_\epsilon \right\|_p \tag{5}$$

An optimization problem for **Attacks in Offline Setting via Poisoning Dynamics** is formulated as,

$$\min_{P, \mu, \mu^{sa}} \quad \left\| P - \overline{P} \right\|_p \tag{P2}$$

If there exists a solution $\widehat{P}$ to the optimization problem (P2), its cost satisfies

$$\left\| \widehat{P} - \overline{P} \right\|_p \cdot \left\| \overline{V} \right\|_\infty \geq \frac{\cdot \overline{\alpha}}{2} \cdot \left\| \overline{\chi}_0 \right\|_\infty \tag{6}$$

If for every state s and action a, it holds that either $\overline{\beta}(s,a) \geq \epsilon \cdot (1 + \overline{D})$ OR $\overline{\chi}_\epsilon(s,a) = 0$, then the optimization problem (P2) has a solution $\widehat{P}$ whose cost is upper bounded by,

$$\left\| \widehat{P} - \overline{P} \right\|_p \leq 2 \cdot \left\| \overline{\Lambda} \right\|_p \tag{7}$$

An attacker changes the environment feedback (reward r_t or state s_{t+1}) to attack an online learning agent.

2 Strengths

This paper presents environmental poisoning by manipulating both rewards or transition dynamics as compared to previous existing work which only manipulates rewards to perform poison attack on RL agents. This is important step in new direction as attacking specific applications could be efficiently done by manipulating transition dynamics instead of only rewards.

3 Weaknesses

The experimental results show that the examined attack models are effective, but they do not tell which types of learning algorithms are most sensitive to the attack tactics studied in the paper. Relaxing the assumptions about the attacker's understanding of the underlying MDP may result in more resilient attack tactics.

4 Possible Improvements

While the paper treats reward and transition attack models separately, attacking rewards and transitions at the same time may result in more robust and cost-effective solutions. This might have been solved by merging the corresponding constraints and specifying the cost function as a weighted sum of the cost functions used in (P1) and (P2).

5 Possible Extensions

Important possible extension could be expanding the attack models, such as attacking rewards and transitions simultaneously, and widening the range of attack goals, such as under partial specification of target policy.