Defense Against Reward Poisoning Attacks in Reinforcement Learning

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

1 Summary

This paper investigates defense strategies against reward poisoning attacks in reinforcement learning (RL). The paper proposed an optimization framework for developing optimal defense policies against reward poisoning attacks, specifically poisoning attacks that alter an agent's reward structure in order to force the agent to follow a target policy and when the attack parameter is known and unknown. The paper also examines the value of employing such defense tactics, offering characterisation results that specify provable performance guarantees. The paper also provides valuable insights into some bounds about true and unpoisoned reward, lower bounds on the expected return of the defense policies, and upper bounds on how sub-optimal these defense policies are in comparison to the attacker's target policy. Then the paper demonstrates the effectiveness of proposed defense approach by using simulation-based experiments.

1.1 Overview

The proposed optimization problem of maximizing the worst case performance of the agent is,

$$\max_{\pi} \min_{R} \rho^{\pi} \qquad \text{s.t.} \quad \widehat{R} = \mathcal{A}(R, \pi_{\dagger}, \epsilon_{\dagger}). \qquad \text{(agent does know } \epsilon_{\dagger})$$

and,

$$\max_{\pi} \min_{R,\epsilon} \rho^{\pi} \qquad \text{s.t.} \quad \widehat{R} = \mathcal{A}(R, \pi_{\dagger}, \epsilon), \qquad 0 < \epsilon \leq \epsilon_{\mathcal{D}}, \qquad \text{(agent does not know } \epsilon_{\dagger})$$

where $\overline{M}$ is *original* MDP with unpoisoned reward function $\overline{R}$, i.e., $\overline{M} = (S, A, \overline{R}, P, \gamma, \sigma)$,
$\widehat{M}$ is *poisoned* MDP with poisoned reward function $\widehat{R}$, i.e., $\widehat{M} = (S, A, \widehat{R}, P, \gamma, \sigma)$,
the score of policy π under $\overline{R}$ is $\overline{\rho}^{\pi}$, whereas its score under $\widehat{R}$ is $\widehat{\rho}^{\pi}$,
agent has access to the poisoned reward vector $\widehat{R} = \mathcal{A}(\overline{R}, \pi_{\dagger}, \epsilon_{\dagger})$,
but $\overline{R}$, $\pi_{\dagger}$, and $\epsilon_{\dagger}$ are not given to the agent,
$\pi_{\dagger}$ is obtainable by solving the optimization problem $\arg\max_{\pi} \widehat{\rho}^{\pi}$ as $\pi_{\dagger}$ is uniquely optimal in $\widehat{M}$,
$\overline{R}$ is unknown to the agent,
the agent uses defence parameter ($\epsilon_{\mathcal{D}}$) as an upper bound on attack parameter ($\epsilon_{\dagger}$).

2 Strengths

This paper presents basis for designing an effective defense strategy against reward poisoning attacks by exploiting the underlying structure of these attacks, and provides some key insights into this problem.

3 Weaknesses

For the purpose of determining optimal defense policies, the paper did not take into account any prior knowledge that an agent might have about the underlying reward function or the attacker.

Preprint. Under review.

4 Possible Improvements

While this paper analyzed defense strategies in the tabular setting only but largescale RL problems typically rely on function approximation, therefore these defense strategies should be analyzed in this setting also. While selecting $\epsilon_{\mathcal{D}}$ optimization framework should consider not only the cost that the attacker has for different choices of $\epsilon_{\dagger}$ but also full game-theoretic characterization of the parameter selection problem.

5 Possible Extensions

Some of the bounds associated with the notion of attack influence are dependent on the poisoned reward function in the paper but to have more thorough examination, it would be very interesting to develop bounds that are not dependent on the poisoned reward function, or to prove that this is not possible. Also, important extension could be to analyze more defence objectives other than just maximization of the agent's worst-case utility.