Multi-level Certified Defense Against Poisoning Attacks in Offline Reinforcement Learning

Shijie Liu[1], Andrew C. Cullen[1], Paul Montague[2], Sarah M. Erfani[1], Benjamin I. P. Rubinstein[1]
[1]The University of Melbourne, Melbourne, Australia
[2]Defence Science and Technology Group, Adelaide, Australia

Background and Motivation

- **Offline Reinforcement Learning** (RL): trains policies using pre-collected data without direct environmental interaction
- **Poisoning Attack**: manipulates a subset (attack budget) of the training dataset, causing RL policies to perform poorly or dangerously
- **Robustness** of RL agent is important, especially in high stakes fields, e.g., healthcare, robotics, autonomous driving

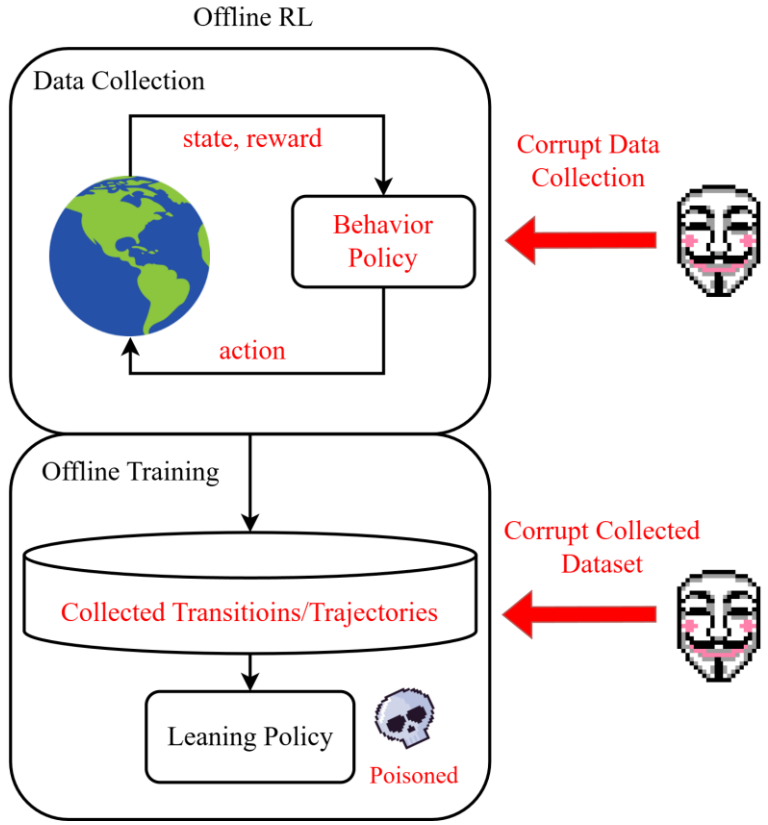

Figure 1: In offline RL learning process, the historical data is collected by the behavior policy and used by the learning policy. The adversary could corrupt the data collection via manipulating the state, reward, action, and the behavior policy. Or they could corrupt the collected dataset via altering the data records.

Existing Work & Its Limitations

- Empirical vs **Certified** Robustness
- Existing Certified Defenses
  - Theoretical approaches often **simplify** scenarios unrealistically (linear MDPs, discrete action spaces)
  - Practical approaches provide computable lower bound or certified radius, however, they don't apply to general RL settings - including **stochastic** environments and **continuous** action space
  - Certification limited to specific states/trajectories, rather than the **robustness of the policy**
- Research Gap: Certified defenses against **multi-level poisoning attacks** in offline RL that provides both **action and policy level robustness guarantees** and applies to **general RL settings**.

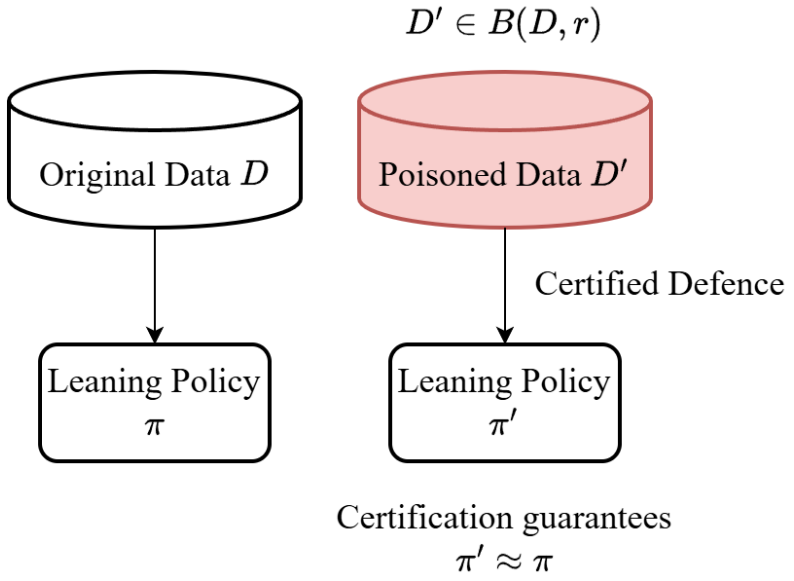

Figure 2: Certified defenses ensure that the poisoned policy π' achieves a similar performance as the clean policy π for any poisoned dataset D' within the attack budget $B(D, r)$.

Approach: Multi-level Certified Defense

- Against Multi-level Poisoning Attacks
  - **Trajectory-level Poisoning:** Corrupts data **during** collection; changes propagate through trajectories
  - **Transition-level Poisoning:** Corrupts dataset **after** collection; changes affect only specific transitions
- Provide Multi-level Robustness Certification
  - **Action-level Robustness:** Actions remain stable in critical states: $\pi(s_t) = \pi'(s_t)$
  - **Policy-level Robustness:** Expected cumulative reward (ECR) remain stable: $J(\pi') = E[\sum_t \gamma^t r_t \mid \pi'] \geq K(J(\pi))$

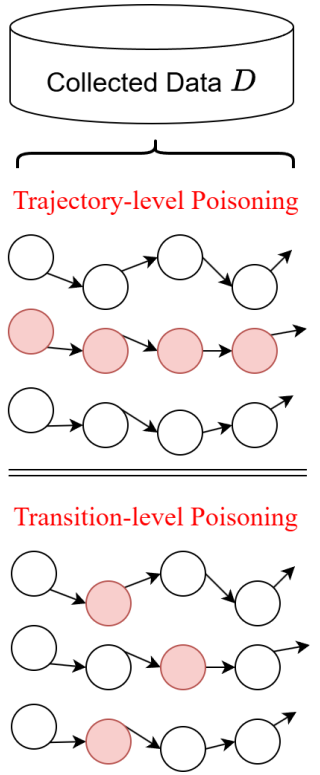

Figure 3: Trajectory-level poisoning vs transition-level poisoning.

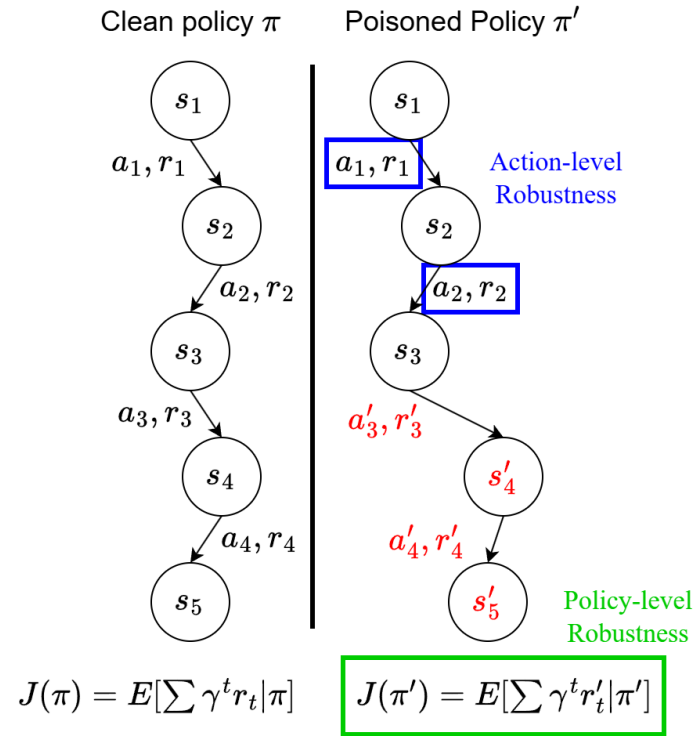

Figure 4: Action-level robustness ensures the actions a_1, a_2 in critical states s_1, s_2 remains the same; Policy-level robustness ensures the expected cumulative reward J(π') lower bounded by a function K of J(π).

Approach: Multi-level Certified Defense

1. Obtain **Differential Privacy (DP)** guarantee via Randomized Training Process
   - Adds controlled Gaussian noise to the gradients in the training process
   - Employs Sampled Gaussian Mechanism and DP-Federated Averaging
2. Derive **Robustness Certification** from DP Guarantees
   - Utilize proposed **DP outcomes guarantees** that converts DP guarantee to **output value bounds,** where the processed result from the poisoned dataset $\mathrm{f}(\mathrm{D}')$ remains bounded by $\mathrm{K}(\mathrm{f}(\mathrm{D}'))$
   - **Action-level Robustness**: bound the inferred score $I_{a_t}(s_t, \pi')$ of the actions, such that the desired action a_t ensures to achieves the highest score

$$K^{-1}\left(I_{a_l}(s_t, \pi)\right) \leq I_{a_l}(s_t, \pi') \leq K\left(I_{a_l}(s_t, \pi)\right)$$

   - **Policy-level Robustness**: lower bound on the ECR as

$$J(\pi') \geq \underline{J}_r(\pi') = K^{-1}\big(J(\pi)\big)$$

Experimental Results

Defence	Environment		Action Space		Certification	
	Deterministic	Stochastic	Discrete	Continuous	Action-level	Policy-level
COPA [Wu et al.,]	✓	✗	✓	✗	✓	✗
Our Method	✓	✓	✓	✓	✓	✓

- Experimental Setup
  - Atari: discrete action space; MuJoCo: continuous action space
  - Stability Ratio: proportion of states with certified radii >= poisoning threshold
  - ECR Lower Bound: lower bound value of ECR for a poisoning size
- Highlights
  - Certified radii ~**5 times larger** compared to prior works
  - Performance drops ≤ 50% with up to **7% poisoned data**, compared to just **0.008%** in prior work

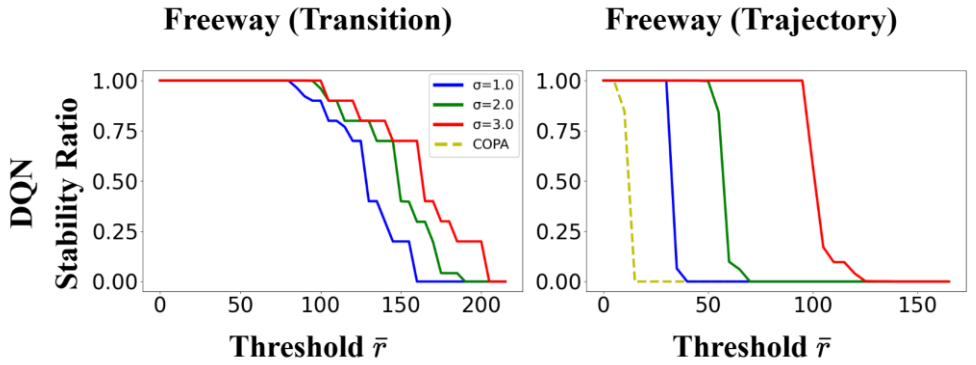

Figure 5: Action-level robustness. Red, green, and blue lines denotes our method trained in different noise levels, while yellow line denotes compared prior work COPA.

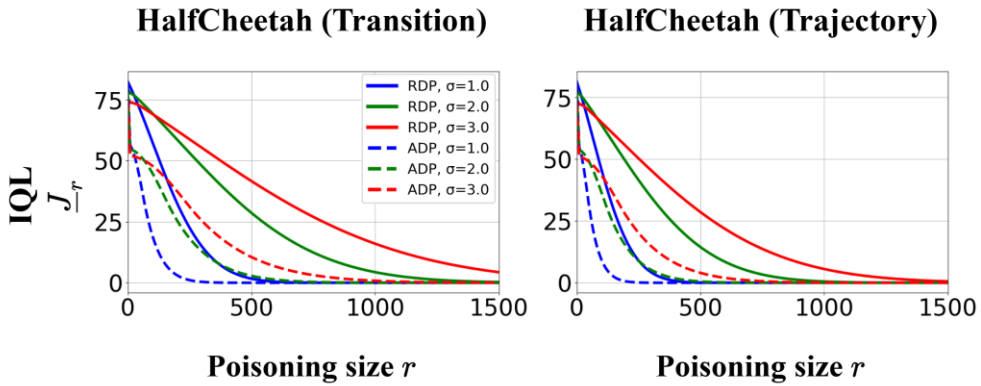

Figure 6: Policy-level robustness. Red, green, and blue lines denotes our method trained in different noise levels, solid and dashed lines represents training under different DP mechanisms.

[Wu et al.,] Fan Wu, Linyi Li, Chejian Xu, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, DingZhao, and Bo Li. COPA: Certifying Robust Policies for Offline Reinforcement Learningagainst Poisoning Attacks, March 2022.

Thank You!

Multi-level Certified Defense Against Poisoning Attacks in Offline Reinforcement Learning

Shijie Liu[1], Andrew C. Cullen[1], Paul Montague[2], Sarah M. Erfani[1], Benjamin I. P. Rubinstein[1]

[1]The University of Melbourne, Melbourne, Australia

[2]Defence Science and Technology Group, Adelaide, Australia