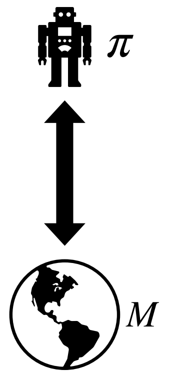
Optimal Attack and Defense on Reinforcement Learning

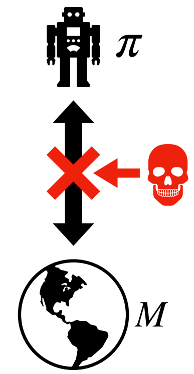
Jeremy McMahan, Young Wu, Xiaojin Zhu, and Qiaomin Xie

University of Wisconsin-Madison

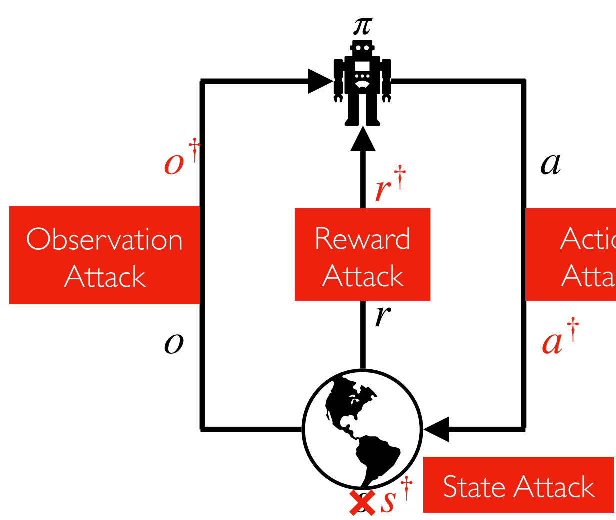
Introduction



An external *attacker* can hack the interaction of π and M.



Attack Surfaces



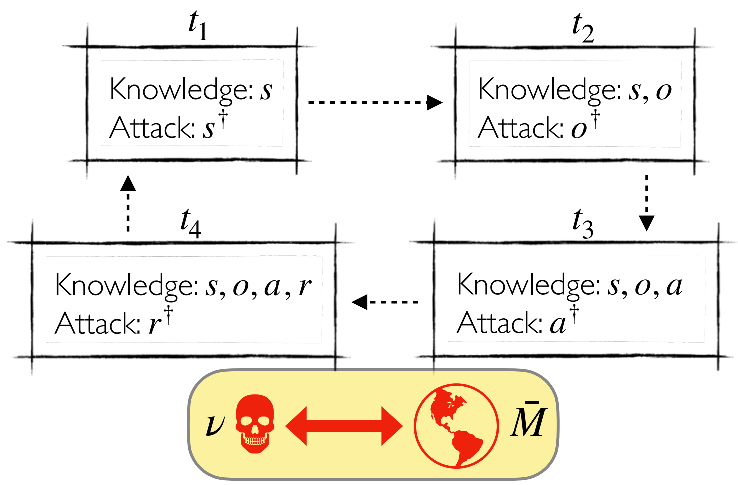
Attack Problem

Attacker has its own reward $g(s_t, a_t, r_t)$ that depends on the victim's.

Definition (Attack): Given π, the attacker wishes to compute,

$$\nu^* \in \arg\max_{\nu} \mathbb{E}_M^{\pi,\nu}\left[\sum_{h=1}^{H} g(s_h, a_h, r_h)\right]$$

Reduction to RL



Proposition: Solving $\overline{M}$ yields an optimal attack.

Defense Problem

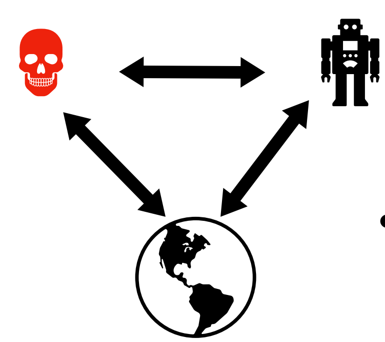
Let $(V_1^{\pi,\nu}, V_2^{\pi,\nu})$ denote the victim's and attacker's value, respectively.

Definition (Defense): The agent wishes to compute,

$$\pi^* \in \arg\max_{\pi} \min_{\nu \in BR(\pi)} \mathbb{E}_M^{\pi,\nu}\left[\sum_{h=1}^{H} r(s_h, a_h)\right]$$

$$BR(\pi) := \arg\max_{\nu \in N} V_2^{\pi,\nu}$$

Reduction to MARL

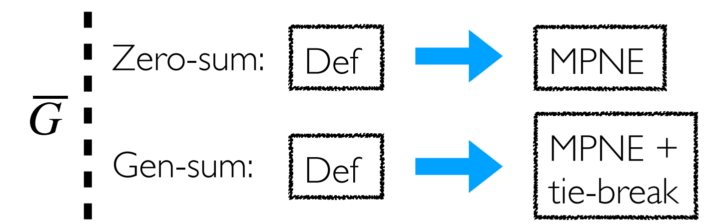


- Interaction forms a two-player turn-based Markov Game $\overline{G}$
- Defenses correspond to a Weak Stackelberg Equilibrium (WSE)

Tractable Solutions

Proposition: The defense problem is as hard as solving POMDPs. Thus, the defense problem is NP-hard to even approximate.

Key: Disallow Observation Attacks.



Both Efficiently Solvable!

Efficient Algorithms

Generalized Rollback:

1. Victim determines Attacker's best response to any action a:

$$BR_h(s,a) = \arg\max_{a^\dagger \in \overline{\mathcal{A}}(s,a)} \left[g_h(s,a,r_h(s,a)) + \mathbb{E}_{s' \sim P_h(s,a^\dagger)}\left[V_{h+1,2}^*(s', \pi_{h+1}^*(s'))\right]\right]$$

2. Victim picks a based on the worst-case best-response:

$$V_{h,1}^*(s) = \max_{a \in \mathcal{A}} \min_{a^\dagger \in BR_h(s,a)} \left[r_h(s,a^\dagger) + \mathbb{E}_{s' \sim P_h(s,a^\dagger)}\left[V_{h+1,1}^*(s')\right]\right]$$

Conclusions

- Optimal attacks can be efficiently computed for all attack surfaces.
- The defense problem is NP-hard to even approximate.
- Absent observation attacks, optimal defenses can be efficiently computed.