Truly Black-box Attack on Reinforcement Learning via Environment Poisoning

Hang Xu, Zinovi Rabinovich

Nanyang Technological University, Singapore

ALA 2021 - Workshop at AAMAS 2021

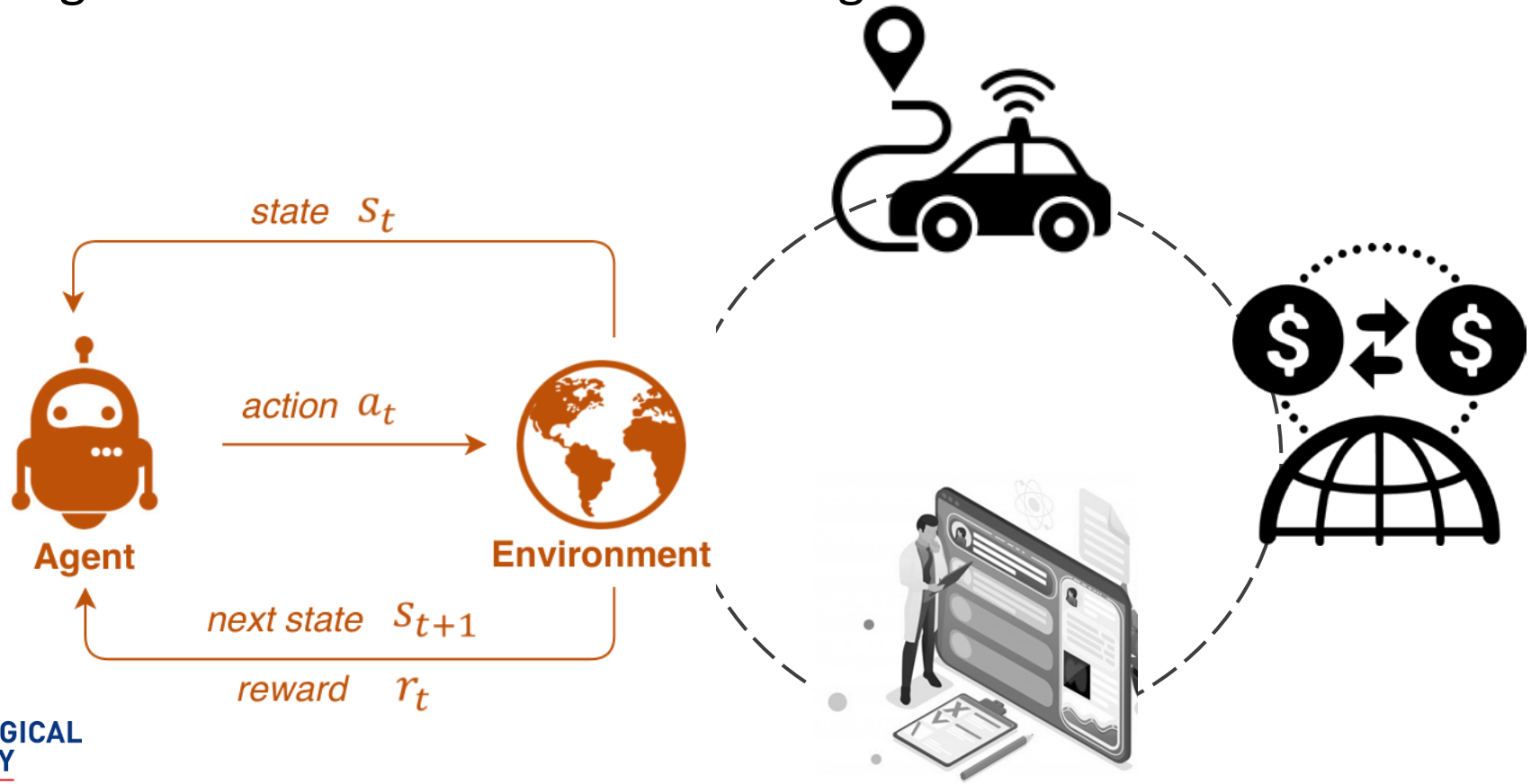
Overview

- **Topic:** Training-time attack on Reinforcement Learning (RL) in *double black-box* settings
- **Goal:** To induce a black-box RL agent to learn a target policy in a black-box environment
- **Solution:** Manipulate the environment dynamics during agent's learning progress
- **Problem of interest:**
  - Unknown agent's learning algorithm
  - Unknown environment dynamics

  How to induce the unknown agent to learn a target policy in unknown environment?

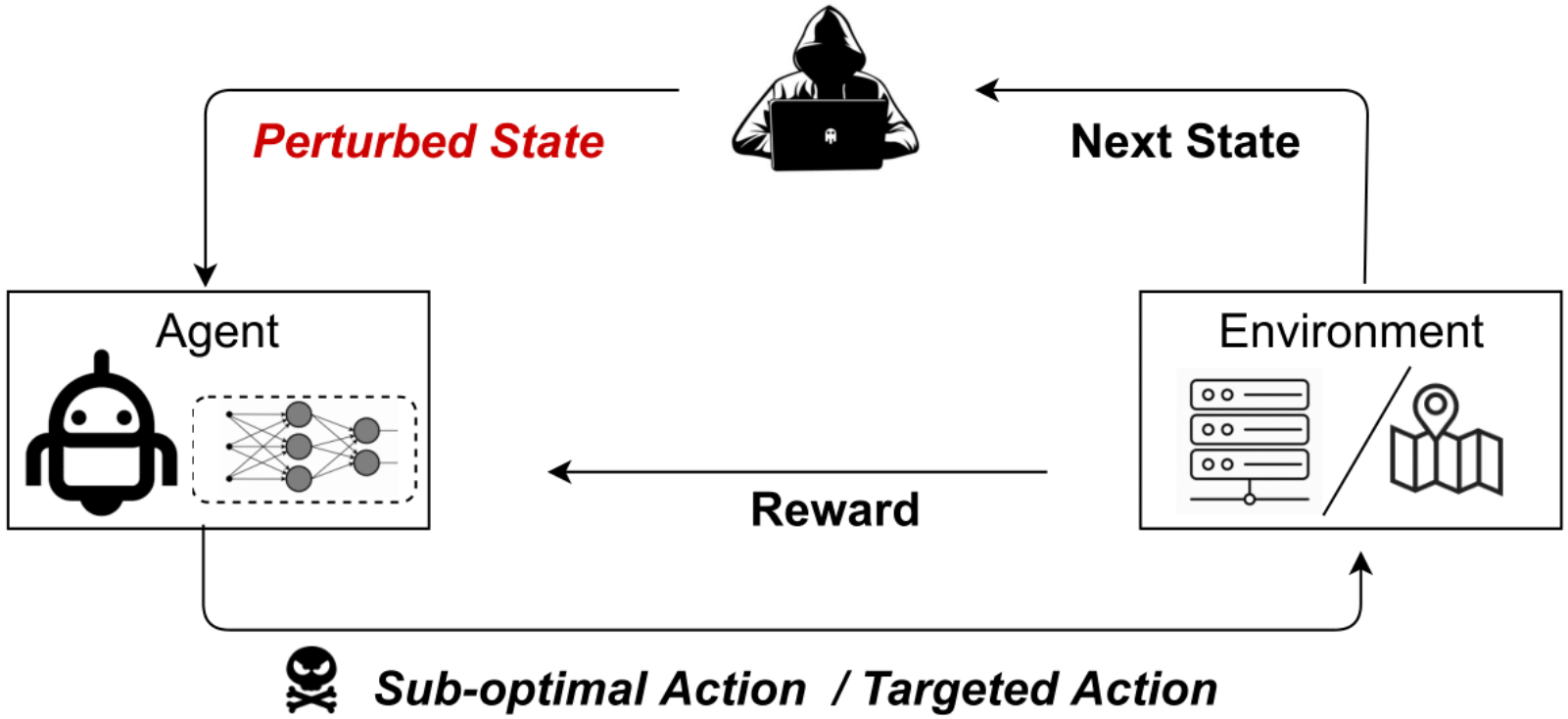
Background: Why study attacks on RL

- To understand the nature of adversarial vulnerabilities on RL
- To find potential mitigation procedures
- Categories: Test-Time Attack & Training-Time Attack



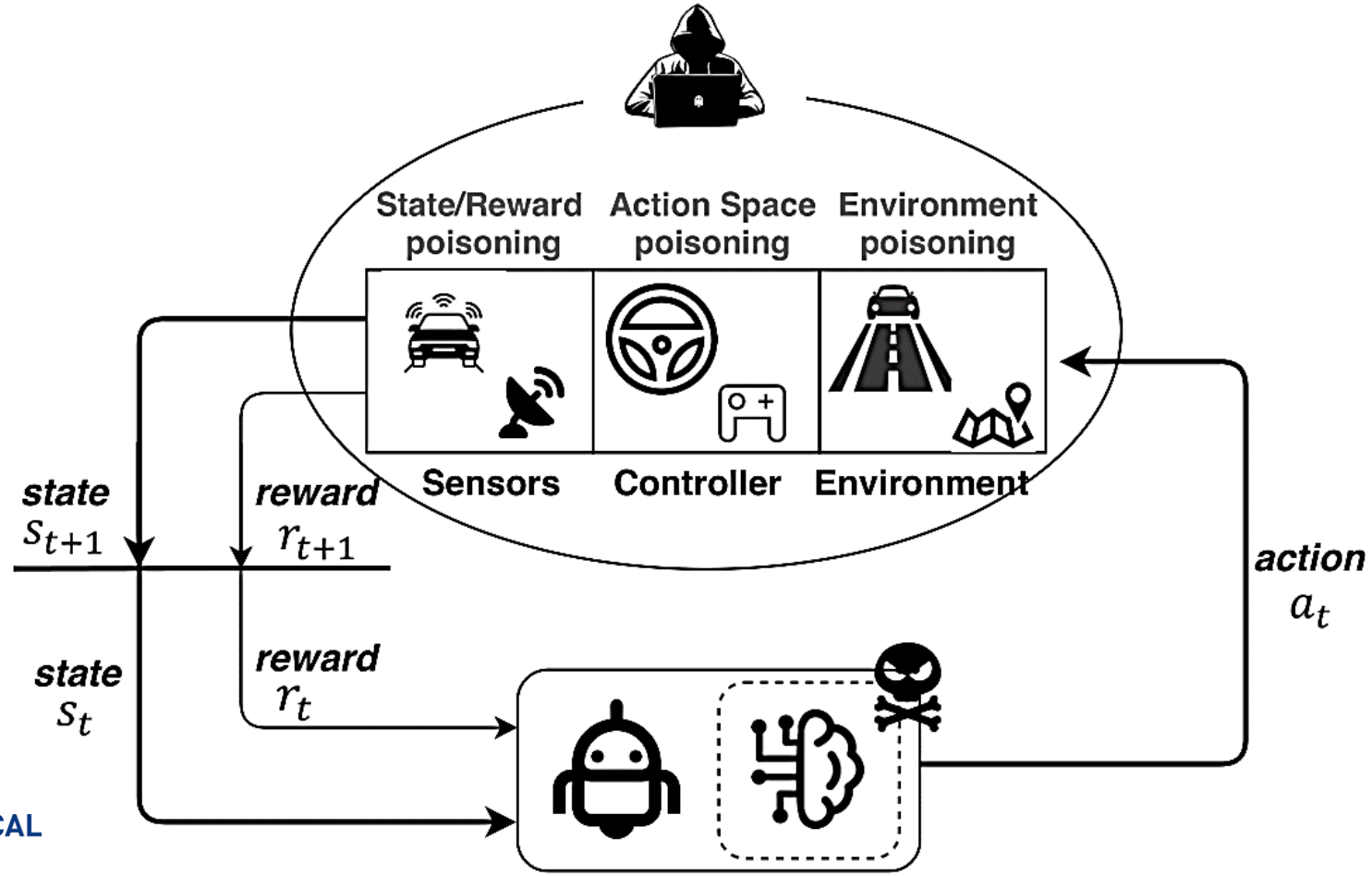
Background: Test-Time Attack

- Objective: To degrade the deployment performance of the policy
- Mislead agent's action decision by perturbing agent's input state



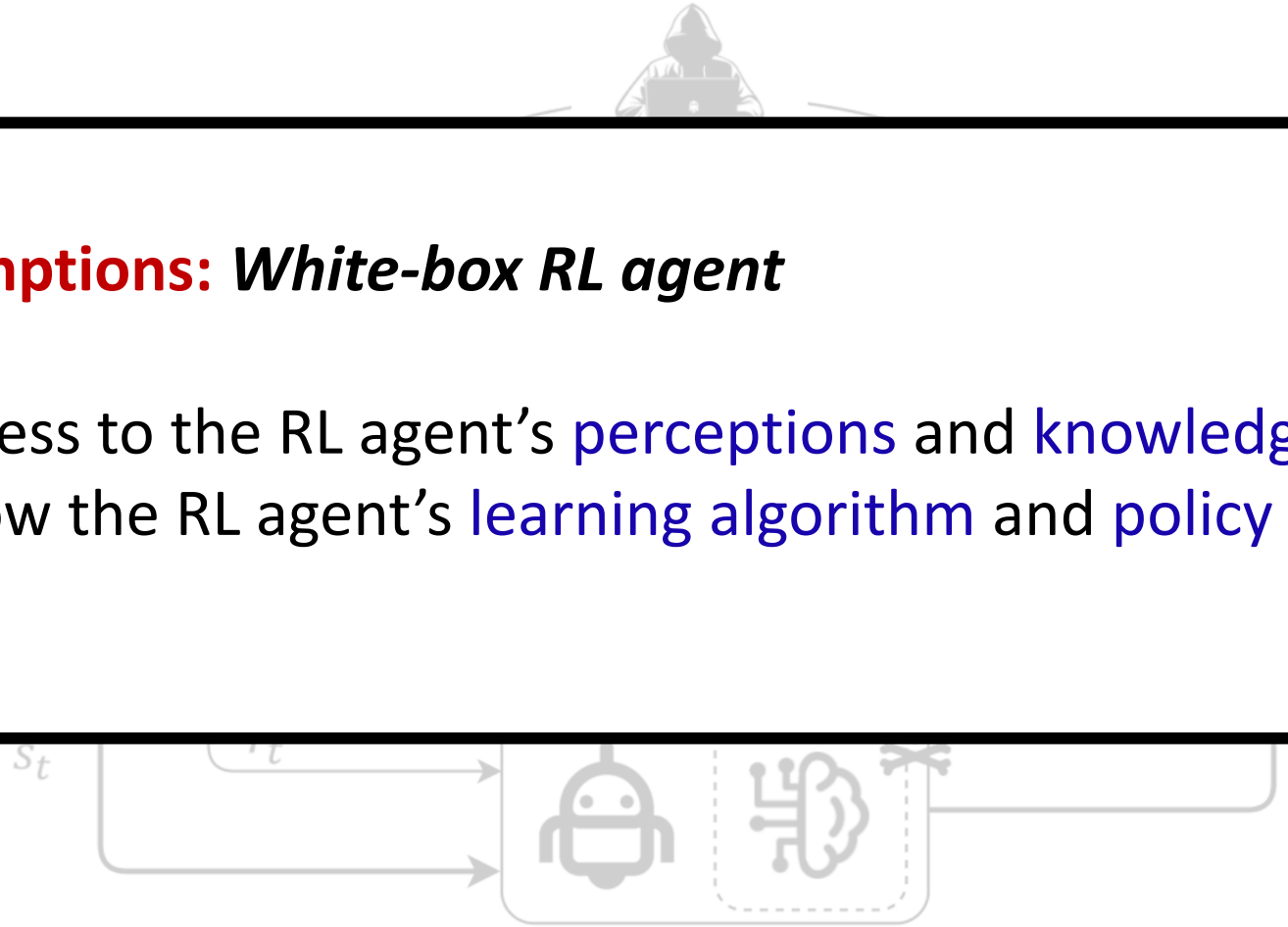
Background: Training-Time Attack

- Objective: To change the RL agent's policy itself
- Manipulate the agent's policy by poisoning its reward or environment at training time



Background: Training-Time Attack

- To change the RL agent's policy itself
- Manipulate the agent's policy by poisoning its reward or environment at training time



Assumptions: ***White-box RL agent***

1. Access to the RL agent's perceptions and knowledge
2. Know the RL agent's learning algorithm and policy model

TEPA: *Transferable Environment-Poisoning Attack*

Enable the training-time attack effective for a **black-box** RL agent:

- Environment-dynamics poisoning attack
  - Tweak the physical properties of the agent's external environment
  - Manipulate the environment's response to the agent's action (i.e. dynamics)

- Transferability property of poisoned training environment
  - Design an attack strategy on a white-box proxy agent
  - Transfer the strategy to attack a black-box victim

Hang Xu, Rundong Wang, Lev Raizman, and Zinovi Rabinovich. 2021. Transferable Environment Poisoning: Training-time Attack on Reinforcement Learning. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems.

TEPA: *Transferable Environment-Poisoning Attack*

Assumption of TEPA: white-box training environment

- knowledge of environment dynamics (i.e., state transition probabilities)
- Knowledge of stationary state distribution

constraints scalability TEPA to sophisticated or real-world scenarios

Prior knowledge	White-Box	TEPA	Our Work
Victim's learning algorithm and policy model	Yes	No	No
Interaction information between victim and environment	Yes	No	No
Dynamics model of victim's training environment	Yes	Yes	No

Hang Xu, Rundong Wang, Lev Raizman, and Zinovi Rabinovich. 2021. Transferable Environment Poisoning: Training-time Attack on Reinforcement Learning. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems.

Motivation

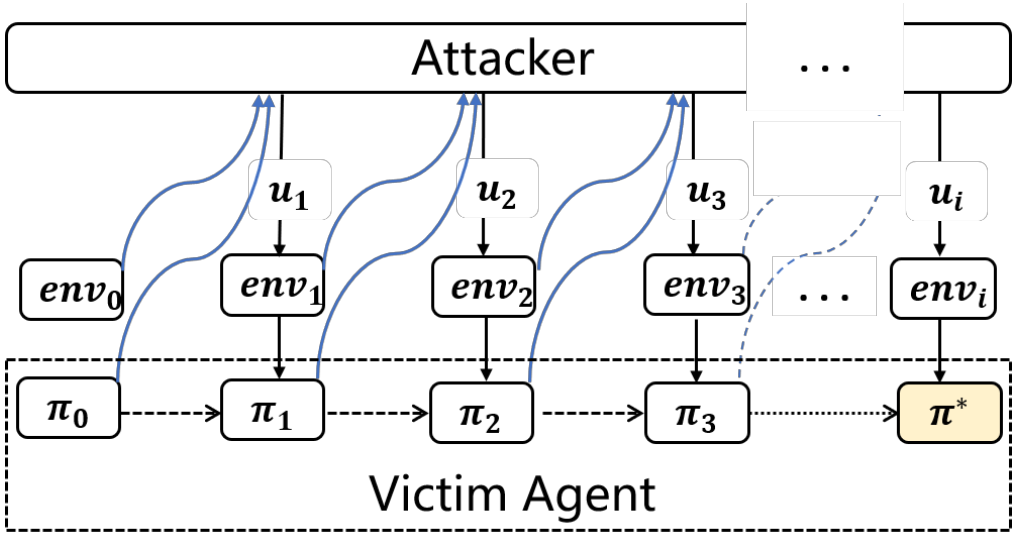
We study the training-time attack on RL in double black-box settings

- RL agent is black-box: unknown learning algorithm & policy models
- Environment is black-box: unknown dynamics model

To force a black-box RL agent to learn a target policy in a black-box environments.

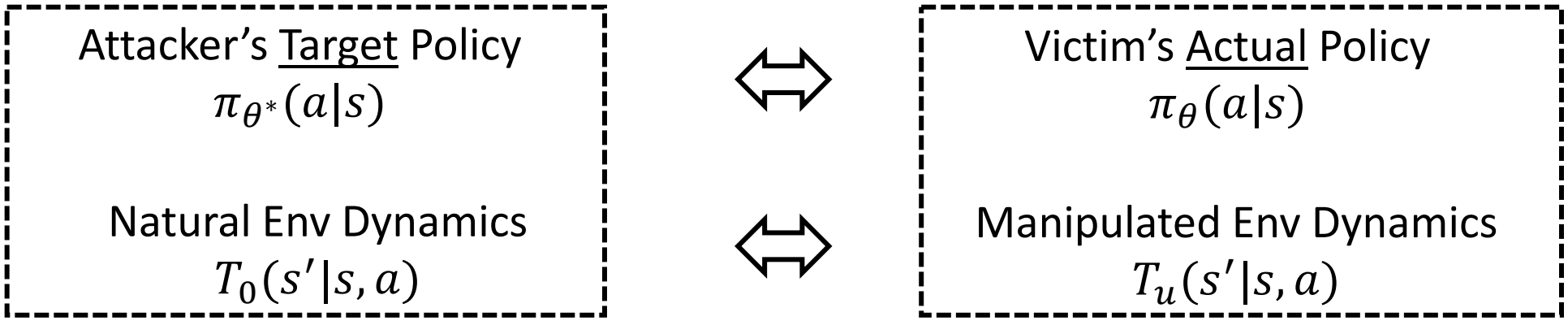
Introduction

- Inherit the attack framework from TEPA
- Pursue the same attack objective as TEPA



- Victim-level objectives:
  - pursue its own benefits, maximize the cumulative rewards
- Attack-level objectives:
  - Force the victim to follow a target policy designed by the attacker
  - Control cumulative changes of environments

Problem Statement



$$P^*(s',a'|s,a) = T_0(s'|s,a)\pi_{\theta^*}(a'|s') \qquad P_{u,\theta}(s',a'|s,a) = T_u(s'|s,a)\pi_\theta(a'|s')$$

Optimization objectives: minimize cumulative attack cost

$$min \sum_{i=1}^{\infty} \gamma^i c_i$$

$$s.t.$$

$$c_i(\theta, u) = \Delta(P_{u_i,\theta_i}(s',a'|s,a)||P^*(s',a'|s,a))$$

Problem Statement

To learn an attack strategy $\sigma(u_i|u_{1:i-1}, \theta_{i-1})$ which

$$min \sum_{i=1}^{\infty} \gamma^i c_i$$

$$s.t.$$

$$c_i(\theta, u) = \Delta(P_{u_i,\theta_i}(s', a'|s, a)||P^*(s', a'|s, a))$$

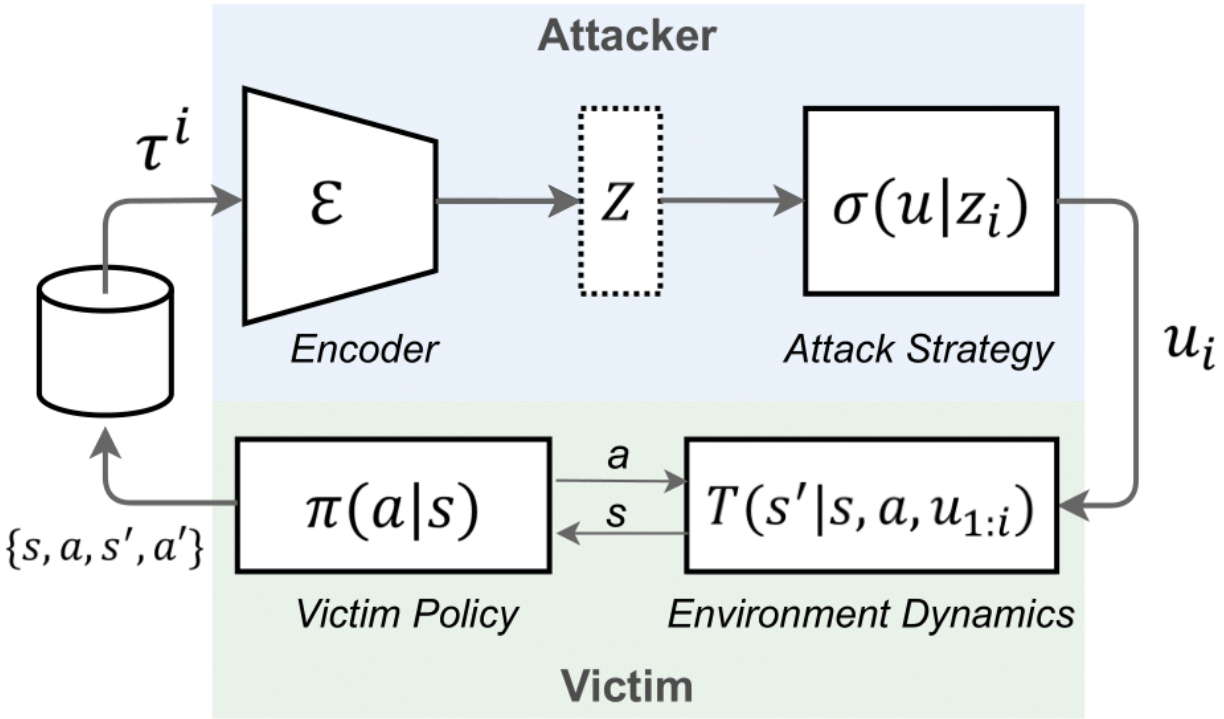
$$P_{u,\theta}(s', a'|s, a) = T_u(s'|s, a)\pi_\theta(a'|s')$$

$$P^*(s', a'|s, a) = T_0(s'|s, a)\pi_{\theta^*}(a'|s')$$

Problem to solve in double black-box settings:

- Can not know the condition information of the attack strategy
  - ➢ *capture latent representation of victim's policy and dynamics features*
- $c_i(\theta, u)$ cannot be computed directly
  - ➢ *approximate attack cost in latent space*

Approach



The **attack procedure** consists of two stages:

1. represent the victim's stochastic process based on its trajectories
2. decide attack actions responding to the representation information

Training of attack strategy includes two parts:

1. Latent representation learning
2. Attack strategy learning

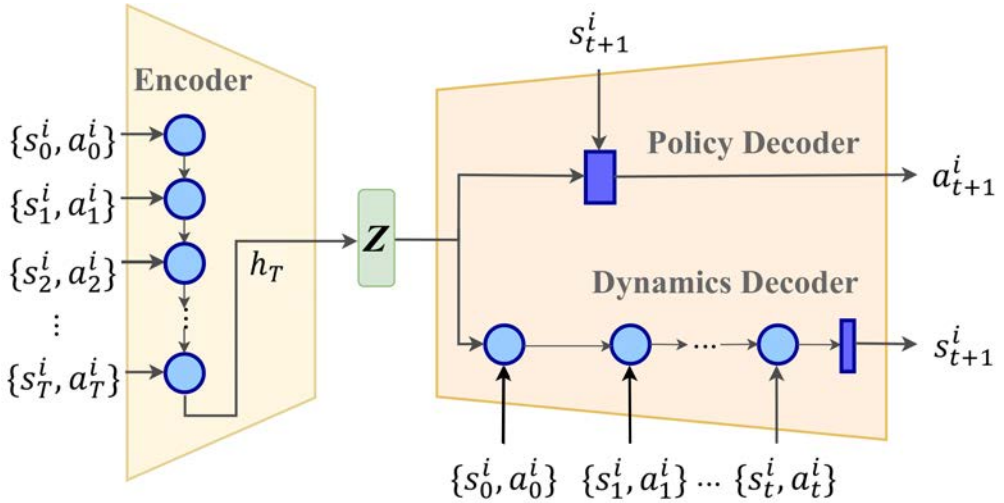
Approach

1. Latent representation learning

learns a latent representation Z that captures information of

- Victim's environment dynamics from $< s, a, s' >$
- Victim's policy feature from $< s, a >$

Using **Encoder Dual-Decoder Network**



2. Attack strategy learning

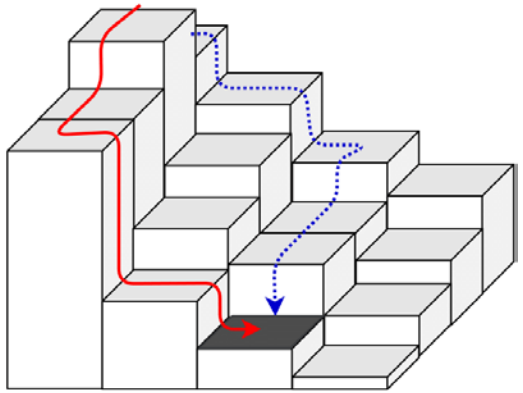
- To measure the similarity between $P(s', a'|s, a)$ and $P^*(s', a'|s, a)$, we approximate their deviation by the embedding z and z^* in the latent space.
- We use **Cosine Similarity** to measure distance between z and z^*, the attack cost c_i as:

$$c_i(z^i, z^*) = 1 - \frac{z^i \cdot z^*}{\|z^i\| \, \|z^*\|}$$

Experiments

- **Objectives:** Evaluate attack performance in double black-box settings with the comparison of baseline TEPA

- **Environment**
  - 3D grid world [4] simulates mountain or rugged terrain
  - Elevation changing is a mechanism to change the environment transition dynamics

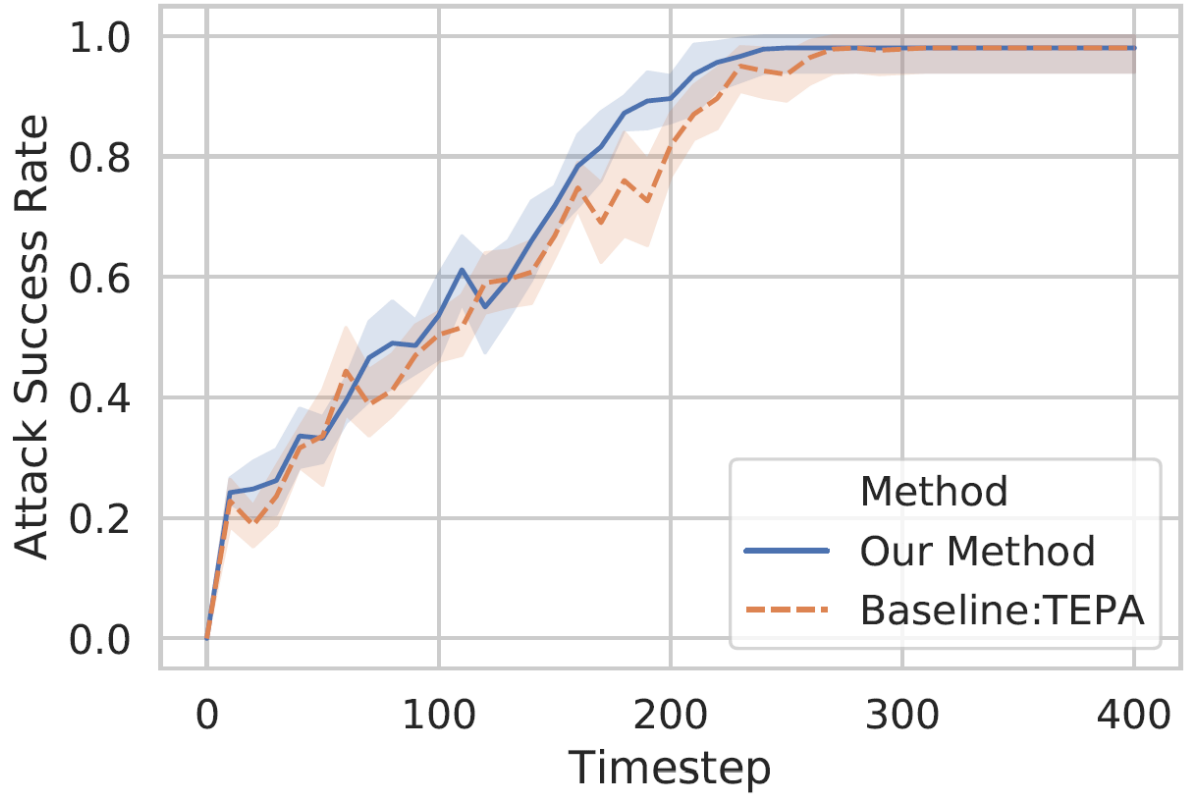


- Red line: attacker's target path
- Blue line: agent's optimal path
- Shallow cell: destination

- Hang Xu et al. "Transferable Environment Poisoning: Training-time Attack on Reinforcement Learning", AAMAS, 2021
- Zinovi Rabinovich et al. "Cultivating desired behaviour: Policy teaching via environment-dynamics tweaks", AAMAS, 2010

Results

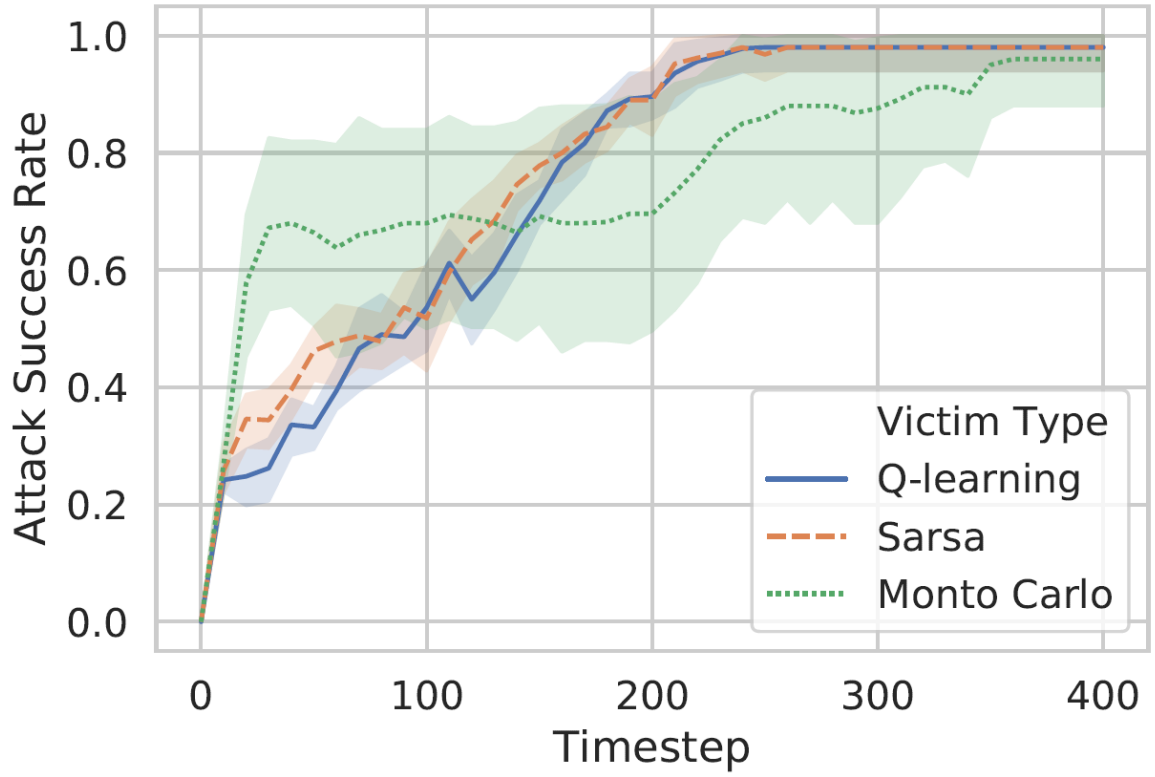
- Attack success rate is quite similar to the baseline TEPA
- Successfully forces the black-box victim to learn the target policy in a black-box environment



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Results

- Attack strategy, learned on one black-box victim in the unknown training environment, is also effective for different black-box victims in the same training environment.
- Attack strategy is insensitive to victims' learning algorithms and thereby is effective for different algorithm-based victims.



Discussion

- Baseline TEPA:
  - requires dynamics model of training environments
  - learns an attack strategy on a white-box proxy agent and then transfers the strategy to a black-box victim.

- Our work:
  - removes the constraints of using a transfer-learning capable white-box proxy agent and environment
  - trains the attack strategy directly on the black-box victim.
  - attack strategy is effective for various black-box RL agents

Summary

In this in-process work:

- We study the environment-poisoning attacks on training-time RL in the **double black-box** setting, where **victim's policy and environment dynamics** are unknown.
- We propose to learn the **latent representation** of victim's stochastic process from its trajectories, and thereby train an adaptive attack strategy **directly** on a black-box victim.
- We present **preliminary experiment** results to show that the attack strategy is generally effective for **different** black-box victims in one unknown environment.

Thank you.

Email: HANG017@e.ntu.edu.sg