Understanding the Limits of Poisoning Attacks in Episodic Reinforcement Learning

Anshuka Rangi[1] , **Haifeng Xu**[2] , **Long Tran-Thanh**[3] and **Massimo Franceschetti**[1]
[1]University of California San Diego, USA
[2]University of Virginia, USA
[3]University of Warwick, UK
{arangi, mfranceschetti}@ucsd.edu, hx4ad@virginia.edu, long.tran-thanh@warwick.ac.uk

Abstract

To understand the security threats to reinforcement learning (RL) algorithms, this paper studies poisoning attacks to manipulate *any* order-optimal learning algorithm towards a targeted policy in episodic RL and examines the potential damage of two natural types of poisoning attacks, i.e., the manipulation of *reward* and *action*. We discover that the effect of attacks crucially depend on whether the rewards are bounded or unbounded. In bounded reward settings, we show that only reward manipulation or only action manipulation cannot guarantee a successful attack. However, by combining reward and action manipulation, the adversary can manipulate any order-optimal learning algorithm to follow any targeted policy with $\tilde{\Theta}(\sqrt{T})$ total attack cost, which is order-optimal, without any knowledge of the underlying MDP. [1] In contrast, in unbounded reward settings, we show that reward manipulation attacks are sufficient for an adversary to successfully manipulate any order-optimal learning algorithm to follow any targeted policy using $\tilde{O}(\sqrt{T})$ amount of contamination. Our results reveal useful insights about what can or cannot be achieved by poisoning attacks, and are set to spur more works on the design of robust RL algorithms.

1 Introduction

Learning algorithms have been widely used in web services [Zhao *et al.*, 2018], conversational AI [Dhingra *et al.*, 2016], UAV coordination [Venugopal *et al.*, 2021], medical trials [Badanidiyuru *et al.*, 2018], and crowdsourcing systems [Rangi and Franceschetti, 2018]. The distributed nature of these applications makes these algorithms prone to third party attacks. For example, in web services decision making critically depends on reward collection, and this is prone to attacks that can impact observations and monitoring, delay or temper rewards, produce link failures, and generally modify or delete information through hijacking of communication links [Agarwal *et al.*, 2016; Cardenas *et al.*, 2008]. Making these systems secure requires an understanding of the regime where the systems may be vulnerable, and designing ways to mitigate these attacks. This paper focuses on the former aspect, namely understanding of the regime where the systems can be attacked, in episodic Reinforcement Learning (RL).

We consider a *man in the middle* (MITM) attack. In this attack, there are three entities: the environment, the learner (RL algorithm), and the adversary. The learner interacts with the environment for T episodes, and each episode has H steps. In episode $t \leq T$ at step $h \leq H$, the learner observes the state $s_t(h) \in \mathcal{S}$ of the environment, selects an action $a_t(h) \in \mathcal{A}$, the environment then generates a reward $r_t(s_t(h), a_t(h))$ and changes its state based on an underlying Markov Decision Process (MDP), and attempts to communicate the new state to the learner. However, an adversary acts as a "man in the middle" between the learner and the environment. It can observe and may *manipulate the action* $a_t(h)$ to $a_t^o(h) \in \mathcal{A}$ which will generate reward $r_t(s_t(h), a_t^o(h))$ corresponding to the manipulated action. Additionally, the adversary may also *intercept the reward* $r_t(s_t(h), a_t^o(h))$ by adding contamination noise $\epsilon_{t,h}(s_t(h), a_t(h))$. With both attacks, the learner ends up observing the contaminated reward $r_t^o(s_t(h), a_t(h)) = r_t(s_t(h), a_t^o(h)) + \epsilon_{t,h}(s_t(h), a_t(h))$. The cost of attack is measured as the *amount of contamination* $\sum_{t,h} |\epsilon_{t,h}(s_t(h), a_t(h))|$ and *number of action manipulations* $\sum_{t,h} \mathbf{1}(a_t(h) \neq a_t^o(h))$, respectively. Notably, with the wide application of RL today, MITM attack is a realistic concern to the vulnerability of RL algorithms and is thus important to understand. For instance, RL-based UAV coordination to reduce poaching activities in conservation areas is naturally subject to poachers' poisoning attacks, which can falsify the reward feedback (i.e., reward manipulation) and executed actions (i.e., action manipulation) [Venugopal *et al.*, 2021]; similarly, RL algorithms for recommender systems are subject to attacks from hackers or competitors [Zhao *et al.*, 2018].

Reward poisoning attack is a special case of the MITM attack where $a_t^o(h) = a_t(h)$, and has been widely studied in both RL and Multi-Armed Bandits (MAB) settings [Jun *et al.*, 2018; Rakhsha *et al.*, 2020; Rangi *et al.*, 2021b]. Likewise, action manipulation attack is another special case of the MITM attack where $\epsilon_{t,h}(s_t(h), a_t(h)) = 0$, and has been previously studied for MAB setting [Liu and Lai, 2020]. Another variant of action manipulation attack, recently studied in RL [Rakhsha *et al.*, 2020], is manipulation of the transi-

[1]Throughout the paper, $\tilde{\Theta}(\cdot)$ and $\tilde{O}(\cdot)$ notation omits the logarithmic terms.

Settings	Reward	Attack	Bound on Attack Cost
White-box in infinite horizon [2] RL	Unbounded	Reward Manipulation	$\tilde{O}(\sqrt{T})$ [Rakhsha *et al.*, 2020] [3]
White-box in infinite horizon RL	Unbounded	Dynamics Manipulation (under sufficient conditions only)	$\tilde{O}(\sqrt{T})$ [Rakhsha *et al.*, 2020] [3]
Black-box in infinite horizon RL	Unbounded	Reward Manipulation (under a setting with L learner)	$\tilde{O}(T \log L + L\sqrt{T})$ [Rakhsha *et al.*, 2021] [3]
Black-box in episodic RL	Unbounded	Reward Manipulation	$\tilde{O}(\sqrt{T})$ (This work)
Black-box and White Box in RL	Bounded	Reward Manipulation	Infeasible (This work)
Black-box and White Box in RL	Bounded	Action Manipulation	Infeasible (This work)
Black-box and White Box in RL	Bounded	Reward and Action Manipulation	$\tilde{\Theta}(\sqrt{T})$ (This work)

Table 1: Comparison of the attack cost in the episodic RL and infinite RL setting when the adversary does not know the RL algorithm.
[2]Performance of algorithms in infinite horizon RL is studied by analyzing the performance over first T rounds [Wu *et al.*, 2021].
[3]This result is not explicitly stated, and is derived for order-optimal algorithms from the reference.

tion dynamics. This can be considered as manipulating the action $a_t(h)$ to another action, not necessarily in $\mathcal{A}$. Therefore, the attacker has strictly stronger power there than the action manipulation as in our setting. MITM attacks has also been previously considered in cyber-physical systems [Rangi *et al.*, 2021a; Khojasteh *et al.*, 2020].

We consider MITM attacks in two different settings: *unbounded* rewards and *bounded* rewards, which turns out to differ fundamentally. In unbounded reward setting, the contamination $\epsilon_{t,h}(s_t(h), a_t(h))$ is unconstrained whereas in bounded reward setting, the contaminated reward $r_t^o(s_t(h), a_t^o(h))$ is constrained to be in the interval $[0, 1]$, just like the original rewards $r_t(s_t(h), a_t^o(h))$. This constrained situation limits the attacker's contamination at every round, and turns out to be provably more difficult to attack. In each setting, we study attack strategy in more realistic black-box setting in which the attacker does not know, and needs to learn, the underlying MDP as well. A similar study between the two settings has been performed for MAB in [Rangi *et al.*, 2021b], where it is shown that bounded rewards setting is more difficult to attack in comparison to unbounded reward setting. This paper extends the study from MAB to RL, and provides new insights on the feasibility of attacks.

1.1 Summary of Contributions

We consider poisoning attacks with the objective of forcing the learner to execute a target policy π^+. More specifically, for all $h \leq H$ and $s \in \mathcal{S}$, if $\pi_h(s) \neq \pi_h^+(s)$, then the attack aims to induce values satisfying

$$\tilde{V}_h^{\pi}(s) < \tilde{V}_h^{\pi^+}(s), \tag{1}$$

where a) policy π of an agent is a collection of H functions $\{\pi_h : \mathcal{S} \to \mathcal{A}\}$; and b) value function $\tilde{V}_h^{\pi}(s)$ is the expected reward of state s under policy π using contaminated reward observation, between step h until step H.

We propose the first set of efficient *black-box* online attacks to any order-optimal episodic RL algorithms in both bounded and unbounded reward settings. Specifically, for the bounded reward setting, we show that mere reward manipulation does not guarantee successful attacks; namely there exist an MDP and a target policy π^+ which cannot be attacked — i.e., (1) cannot be achieved — by manipulating only the rewards. Similarly, we show that action manipulation attack does not suffice by showing the existence of an MDP and a target policy π^+ which cannot be attacked by manipulating actions. Hence, to guarantee a successful attack, the attacker needs a combined the power of reward and action manipulation. Indeed, we propose an MITM attack in bounded reward setting, which requires $\tilde{O}(\sqrt{T})$ amount of reward contamination and $\tilde{O}(\sqrt{T})$ number of action manipulations to attack any order-optimal learning algorithm. We also show that this attack cost, namely sum of amount of reward contamination and number of action manipulations, is order-optimal.

Next we move to unbounded reward settings. Reward manipulation attack for unbounded rewards has been studied in [Rakhsha *et al.*, 2020]. However, the work investigated *infinite-horizon* RL and proposed a *white-box* attack. Extending this research agenda, we propose *black-box* attack for *episodic* RL, and show that our proposed attack can attack any order-optimal learning algorithm in $\tilde{O}(\sqrt{T})$ amount of contamination. An interesting conceptual message from our results is that bounded reward setting is more difficult to attack than the unbounded reward setting. Our results are compared with the relevant literature in Table 1.

2 Problem Formulation

We consider the episodic Markov Decision Process (MDP), denoted by $(\mathcal{S}, \mathcal{A}, H, \mathcal{P}, \mu)$, where $\mathcal{S}$ is the set of states with $|\mathcal{S}| = S$, $\mathcal{A}$ is the set of actions with $|\mathcal{A}| = A$, H is the number of steps in each episode, $\mathcal{P}$ is the transition metric such that $\mathbb{P}(.|s, a)$ gives the transition distribution over the next state if action a is taken in the current state s, and $\mu : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the expected reward of state action pair (s, a). To avoid cumbersome notations, we work with the stationary MDPs here with the same reward and transition functions at each $h \leq H$. However all our analysis extend trivially to non-stationary MDPs, with just more involved notations.[4]

An RL agent (or learner) interacts with the MDP for T episodes, and each episode consists of H steps. In each episode of the MDP, an initial state $s_t(1)$ can be fixed or selected from any distribution. In episode t and step h, the

[4]Moreover, if an online learning is already vulnerable for this "nice" special case, then it must be vulnerable in general as well since the stationary MDP is a special case of general MDPs.

learner observes the current state $s_t(h) \in \mathcal{S}$, selects an action $a_t(h) \in \mathcal{A}$, and incurs a noisy reward $r_{t,h}(s_t(h), a_t(h))$. Also, we have $\mathbb{E}[r_{t,h}(s_t(h), a_t(h))] = \mu(s_t(h), a_t(h))$. Our results can also be extended to the setting where reward is a function of step $h \leq H$. Finally, both $\mathcal{P}$ and μ are unknown to the learner and the attacker.

We consider episodic RL under MITM attacks. The attacker can manipulate the action $a_t(h)$ selected by the learner to another action $a_t^o(h) \in \mathcal{A}$. The MDP thus undergoes transition to next state based on the action $a_t^o(h)$, namely the next state is drawn from the distribution $\mathbb{P}(.|s_t(h), a_t^o(h))$. The reward observation $r_t(s_t(h), a_t^o(h))$ is generated. If $a_t^o(h) \neq a_t(h)$, then the episode t and step h is said to be *under action manipulation attack*. Hence, the *number of action manipulations* is $\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}(a_t^o(h) \neq a_t(h))$.

The adversary can also intercept the realized reward $r_t(s_t(h), a_t^o(h))$ and contaminate it by adding noise $\epsilon_{t,h}(s_t(h), a_t(h))$. Learner thus observes reward

$$r_{t,h}^o(s_t(h), a_t(h)) = r_{t,h}(s_t(h), a_t^o(h)) + \epsilon_{t,h}(s_t(h), a_t(h)), \tag{2}$$

where the contamination $\epsilon_{t,h}(s_t(h), a_t(h))$ added by the attacker is a function of the entire history, including all the states visited previously and all the actions selected previously by the learner and the attacker. If $\epsilon_{t,h}(s_t(h), a_t(h)) \neq 0$, then the episode t and step h is said to be *under reward manipulation attack*. Hence, the *number of reward manipulations* is $\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}(\epsilon_{t,h}(s_t(h), a_t(h)) \neq 0)$, and the *amount of contamination* is $\sum_{t=1}^{T}\sum_{h=1}^{H} |\epsilon_{t,h}(s_t(h), a_t(h))|$. Notably, *reward manipulation attack* is a special case of MITM attack, where the adversary cannot manipulate the action (i.e., $a_t(h) = a_t^o(h)$). *Action manipulation attack* is also special case of MITM, in which the adversary cannot contaminate the reward observation (i.e., $\epsilon_{t,h}(s_t(h), a_t(h)) = 0$).

A (deterministic) policy π of an agent is a collection of H functions $\{\pi_h : \mathcal{S} \to \mathcal{A}\}$. The value function $V_h^\pi(s)$ is the expected reward under policy π, starting from state s at step h, until the end of the episode, namely

$$V_h^\pi(s) = \mathbb{E}\big[\sum_{h'=h}^{H} \mu(s_{h'}, \pi_{h'}(s_{h'}))|s_h = s\big], \tag{3}$$

where $s_{h'}$ denotes the state at step h'. Likewise, the Q-value function $Q_h^\pi(s, a)$ is the expected reward under policy π, starting from state s and action a, until the end of the episode, namely

$$Q_h^\pi(s, a) = \mathbb{E}[\sum_{h'=h+1}^{H} \mu(s_{h'}, \pi_{h'}(s_{h'}))|s_h = s, a_h = a] + \mu(s, a), \tag{4}$$

where $a_{h'}$ denotes the action at step h'. Since S, A and H are finite, there exists an optimal policy π^* such that $V_h^{\pi^*}(s) = \sup_\pi V_h^\pi(s)$. The regret $R^{\mathcal{A}}(T, H)$ of any algorithm $\mathcal{A}$ is the difference between the total expected true reward from the best fixed policy π^* in the hindsight, and the expected true reward over T episodes, namely

$$R^{\mathcal{A}}(T, H) = \sum_{t=1}^{T} \big(V_1^{\pi^*}(s_t(1)) - V_1^{\pi_t}(s_t(1))\big). \tag{5}$$

The objective of the learner is to minimize the regret $R^{\mathcal{A}}(T, H)$. In contrast, the objective of the attacker is to poison the environment to teach/force the learner to execute a target policy π^+ by achieving the following objective: for all $h \leq H$, $s \in \mathcal{S}$ and any policy π,

$$\text{if } \pi_h(s) \neq \pi_h^+(s), \text{ then } \tilde{V}_h^\pi(s) < \tilde{V}_h^{\pi^+}(s), \tag{6}$$

where $\tilde{V}_h^\pi(s)$ is the expected reward in states based on the reward observation in (2) under policy π. Consequently, the policy π^+ will be executed for $\Omega(T)$ times under the attack.

3 Attacks in Bounded Reward Setting

In this section, we investigate bounded reward setting, i.e., $r_{t,h}(s, a) \in [0, 1]$ with mean $\mu(s, a) \in (0, 1]$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.[5] We first show in subsection 3.1 that there exist MDPs and target policies π^+ such that the objective of the attacker, namely (6), cannot be achieved by only reward manipulation attack or only action manipulation attack. In subsection 3.2, we show that combined reward and action manipulation suffice for a successful attack.

3.1 Insufficiency of (Only) Reward or Action Manipulation

Similar to [Rakhsha *et al.*, 2020], attacker is subject to a constraint that the reward manipulation (i.e., $\epsilon_{t,h}(s_t(h), a_t(h)) \neq 0$) and action manipulation (i.e., $a_t(h) \neq a_t^o(h)$) can occur only if the selected action is different from the desired action, namely $a_t(h) \neq \pi_h^+(s_t(h))$.Intuitively, it can also be interpreted as the attacker can manipulate rewards (or actions) whenever the learner's action is not in accordance with the attacker's desired policy π^+. Therefore, in the reward manipulation attack, the attacker is subject to following constraints

$$\begin{aligned} &r_t^o(s_t(h), a_t(h)) = r_t(s_t(h), a_t(h)) \text{ if } a_t(h) = \pi_h^+(s_t(h)), \\ &\text{and } r_t^o(s_t(h), a_t(h)) \in [0, 1], \\ &\text{or equivalently,} \end{aligned} \tag{7}$$

$$\begin{aligned} &\epsilon_t(s_t(h), a_t(h)) = 0 \text{ if } a_t(h) = \pi_h^+(s_t(h)), \text{ and} \\ &\epsilon_t(s_t(h), a_t(h)) \in [-r_t(s_t(h), a_t(h)), 1 - r_t(s_t(h), a_t(h))]. \end{aligned} \tag{8}$$

This constraint is crucial for obtaining sub-linear attack cost due to the following reason. Suppose the objective in (6) is achieved by contaminating the reward of action $\pi_h^+(s_t(h))$. Now since the learner would execute the policy π^+ with high probability, namely $\Omega(T)$ times, the total contamination will thus grow linearly with T. This constraint (or strategy of not contamination $\pi_h^+(s_t(h))$) is also applied in the previous literature in RL [Rakhsha *et al.*, 2020] and MAB [Jun *et al.*, 2018; Ma *et al.*, 2019b; Rangi *et al.*, 2021b], where the reward manipulation are performed only on non-desirable actions. Similarly, in the action manipulation attack, we assume that

$$a_t^o(h) = a_t(h) \text{ if } a_t(h) = \pi_h^+(s_t(h)). \tag{9}$$

That is, the action can be manipulated only if the selected action is not the same as the target action of the policy.

Our first result establishes that only the reward manipulation or only the action manipulation cannot always guarantee successful attacks in bounded reward setting.

[5] The $\mu(s, a) \neq 0$ is to avoid minor technicality issues due to tie breaking.

Theorem 1. *In bounded reward setting,*

1. *there exists an MDP and a target policy π^+ such that any reward manipulation attack satisfying* (7) *cannot be successful, namely achieve the objective in* (6)*.*
2. *there exists an MDP and a target policy π^+ such that any action manipulation attack satisfying* (9) *cannot be successful, namely achieve the objective in* (6)*.*

The proof of Theorem 1 proceeds by constructing a carefully crafted MDP such that there exists a target policy π^+ and another policy $\tilde{\pi}$ such that $V_h^{\tilde{\pi}}(s) > V_h^{\pi^+}(s)$ irrespective of action or reward manipulation. This implies that the constructed target policy π^+ can not be induced in the constructed MDP.[6]

3.2 Efficient Attack by Combining Reward & Action Manipulation

We now show that the attacker can achieve its objective by combining the strength of both reward manipulation and action manipulation attacks in bounded reward setting. This is the first attack strategy of its kind, using both reward and action manipulation, proposed in the literature.

Given the target policy π^+, for all $s_t(h) \in \mathcal{S}$, $a_t(h) \in \mathcal{A}$ and $h \leq H$, we consider the following attack strategy

$$a_t^o(h) = \begin{cases} a_t(h) & \text{if } a_t(h) = \pi_h^+(s), \\ \pi_h^+(s) & \text{if } a_t(h) \neq \pi_h^+(s), \end{cases} \tag{10}$$

$$r_t^o(s_t(h), a_t(h)) = \begin{cases} r_t(s_t(h), a_t(h)) & \text{if } a_t(h) = \pi_h^+(s), \\ 0 & \text{if } a_t(h) \neq \pi_h^+(s). \end{cases} \tag{11}$$

In the above attack, the adversary manipulates both the action and the reward observation when $a_t(h) \neq \pi_h^+(s)$. Specifically, the adversary manipulates the action to $\pi_h^+(s)$ to control the transition dynamics, and at the same time manipulates the reward observation to zero so that the action $a_t(h)$ appears to be sub-optimal in comparison to the action $\pi_h^+(s)$.In this attack, we carefully "coordinate" action and reward manipulation to achieve two goals simultaneously: (1) the target policy is optimal; (2) the target policy is the greedy solution of the resultant MDP, namely the action yielding maximum reward at current step h is also the action yielding maximum cumulative reward between h and H.

The following theorem shows that novel attack in (10) and (11) can achieve the objective in (6), and does not require learning the parameters of the MDP. Additionally, the attack cost is $\tilde{O}(\sqrt{T})$ for any order-optimal learning algorithm.

Theorem 2. *Consider any learning algorithm $\mathcal{A}$ such that its regret in the absence of attack is*

$$R^{\mathcal{A}}(T, H) = \tilde{O}(\sqrt{T}H^{\alpha}) \qquad \forall T \geq t_0, \tag{12}$$

with probability at least $1 - \delta$, where $\alpha \geq 1$ is a numerical constant. For any sub-optimal target policy π^+, if an adversary follows the strategy in (10) *and* (11) *to attack the algorithm $\mathcal{A}$, then with probability at least $1-\delta$ the following hold simultaneously:*

1. *The attacker achieves its objective in* (6) *and moreover $\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}(a_t(h) = \pi_h^+(s_t(h))) = \Omega(T)$;*
2. *The number of reward manipulation attacks, namely $\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}(\epsilon_{t,h}(s_t(h), a_t(h)) \neq 0)$ is $\tilde{O}\big(\sqrt{T}H^{\alpha}/\min_{h,s}\mu(s, \pi_h^+(s))\big)$*
3. *The amount of reward contamination, namely $\sum_{t=1}^{T}\sum_{h=1}^{H} |\epsilon_{t,h}(s_t(h), a_t(h))|$, is $\tilde{O}\big(\sqrt{T}H^{\alpha}/\min_{h,s}\mu(s, \pi_h^+(s))\big)$*
4. *The number of action manipulation attacks, namely $\sum_{t=1}^{T}\sum_{h=1}^{H} \mathbf{1}(a_t^o(h) \neq a_t(h))$, is $\tilde{O}\big(\sqrt{T}H^{\alpha}/\min_{h,s}\mu(s, \pi_h^+(s))\big)$.*

In [Chen *et al.*, 2021], if the attacker can observe learner's action (as in our setting), BARBAR-RL has a regret $\tilde{O}(\sqrt{T} + C^2)$ where C is the total contamination in rewards and transition dynamics. Since *action and reward manipulation* is a special case of *transition dynamics and reward manipulation*, the bound of BARBAR-RL will also hold in our setting. If attacker is absent ($C = 0$), then the regret of the algorithm is $\tilde{O}(\sqrt{T})$, and (12) holds. If $C = o(\sqrt{T})$, then the regret of BARBAR-RL is $o(T)$, and claim 1 in Theorem 2 does not hold. Thus, there exists an order-optimal algorithm such that attacker's objective is not achieved in $o(\sqrt{T})$ attack cost. Hence, the attack cost of proposed strategy is order-optimal.

The order-optimal algorithm, in (12), has three parameters α, t_0 and δ. The parameter α is used to capture the dependence of regret on H; parameter t_0 captures the fact that the regret bound holds for sufficiently large T; parameter δ captures the fact that the regret bound holds with high probability. Some examples of order-optimal algorithm in episodic RL can be found in [Jin, 2018] and references therein. We present results on attack cost for order-optimal algorithms, however this is not a limitation of our attack strategy. The analysis can be extended for any no-regret learning algorithm and a corresponding bound on attack cost can be obtained.

Our result directly focus on the more realistic and also harder black-box attacks; for completeness, we also discuss how to design a white-box attack by manipulating rewards and actions in Appendix B. This can also be used for designing a black-box attack in unbounded reward setting.

4 Reward Poisoning in Unbounded Settings

In this section, we investigate the unbounded reward setting. Formally, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we assume $r_{t,h}(s, a)$ follows a sub-Gaussian distribution with mean $\mu(s, a)$ and standard deviation σ. While the rewards, drawn from sub-Gaussians, may be unbounded, we assume their mean $\mu(s, a) \in [-M, M]$ is bounded within some known interval for any (s, a) pair. Our main result in this section is the design of a black-box attack to *any* order-optimal episodic RL algorithm with $\tilde{O}(\sqrt{T})$ amount of reward contamination. To our knowledge, this is the first efficient attack to episodic RL algorithms in unbounded reward settings.

[6] We note that such *worst-case* analysis is typical for establishing lower bounds on regret, which usually identifies a difficult instance and shows the impossibility of a good regret (see, e.g., [Bogunovic *et al.*, 2021]).

Formally, we consider the black-box attack setting in which the attacker doesn't know about the expected reward and the transition dynamics of the underlying MDP. In this setting, we propose an attack which learns about the MDP, and has almost the same attack cost as the white-box attack, with an additional $O(\sqrt{\log T})$ factor.

Given the input parameters π^+ and $\epsilon > 0$, our proposed attack strategy is presented in Algorithm 1. In Algorithm 1, the attacker utilizes its estimate of $\hat{\mu}(s,a)$, $\hat{\mu}^{UCB}(s,a)$ and $\hat{\mu}^{LCB}(s,a)$, defined in (13), (14) and (15) respectively, to contaminate the reward observations. These estimates are initialized using the fact that $\mu(s,a) \in [-M, M]$, and are updated in each episode $t \leq T$ and at each step $h \leq H$. The parameters $\hat{\mu}^{UCB}(s,a)$ and $\hat{\mu}^{LCB}(s,a)$ are Upper Confidence Bound (UCB) and Lower Confidence Bound (LCB) of $\mu(s,a)$. Therefore, we show that, with high probability,

$$\hat{\mu}^{LCB}(s,a) \leq \mu(s,a) \leq \hat{\mu}^{UCB}(s,a). \tag{17}$$

In this attack, the reward observations are contaminated only if the action selected by the learner is not the same as the action desired by the target policy, namely $a_t(h) \neq \pi_h^+(s_t(h))$. In this scenario, the reward observation $r_t^o(s_t(h), a_t(h))$ is defined in (16). In (16), LCB estimates of the expected rewards are used to get the LCB estimate of value function of *target policy* and UCB estimates of the expected rewards are used to get UCB value function over *all policy*. This ensures the target policy is optimal, and reward observation consist of a large negative bias sufficient to achieve the objective. This reward manipulation strategy in (16) ensures that the target policy π^+ is the optimal policy based on the observed reward observations, namely for all $h \leq H$, and $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $a \neq \pi_h^+(s)$, we have

$$\tilde{Q}_h^{\pi}(s,a) \leq \tilde{Q}_h^{\pi^+}(s, \pi_h^+(s)) - \epsilon, \tag{18}$$

where $\tilde{Q}_h^{\pi}(s,a)$ is the expected reward in state s for action a for the reward observation under policy π. These values will not be same as the ones defined in (3) and (4) since the reward observations are manipulated. We remark that the $r_t^o(s_t(h), a_t(h))$ can be computed through a backward induction procedure starting from horizon H. At any step h in the episode, the definition of $r_t^o(s_t(h), a_t(h))$ depends linearly on the Q-values at h, which then depends linearly on $r_t^o(s_t(h), a_t(h))$. Therefore, $r_t^o(s_t(h), a_t(h))$ at any horizon h can be computed by solving a linear system.

We briefly discuss the key steps in this process of obtaining (18). We show that with high probability

$$\begin{aligned}&\hat{\mu}^{LCB}(s_t(h), \pi_h^+(s_t(h))) + (H-h) \min_{s,a \in \mathcal{S} \times \mathcal{A}} \hat{\mu}^{LCB}(s,a) \\ &\leq \tilde{Q}_h^{\pi^+}(s_t(h), \pi_h^+(s_t(h))).\end{aligned} \tag{19}$$

Additionally, we have that with high probability

$$\begin{aligned}&(H-h) \max_{s,a \in \mathcal{S} \times \mathcal{A}} \hat{\mu}^{UCB}(s,a) \\ &\geq \mathbb{E}_{s' \sim \mathbb{P}(s' | s_t(h), a_t(h))}[\tilde{V}_{h+1}^{\pi^+}(s')].\end{aligned} \tag{20}$$

Combining (19) and (20), we have that with high probability, the rewards contamination in Algorithm 1 ensures (17).

The following theorem shows that our proposed black box attack has $\tilde{O}(\sqrt{T})$ amount of contamination.

Theorem 3. *Consider any learning algorithm $\mathcal{A}$ such that its regret in the absence of attack is $R^{\mathcal{A}}(T,H) = \tilde{O}(\sqrt{T}H^{\alpha}), \quad \forall T \geq t_0$*

$$R^{\mathcal{A}}(T,H) = \tilde{O}(\sqrt{T}H^{\alpha}), \quad \forall T \geq t_0 \tag{21}$$

with probability at least $1-\delta$ where $\alpha \geq 1$ is a numerical constant. For any sub-optimal target policy π^+, $\epsilon > 0$ and $T \geq t_0^2$, if an attacker follows strategy in Algorithm 1, then with probability at least $1 - \delta - 2/(HSAT)$ the following hold simultaneously:

1. *The attacker achieves its objective in* (6) *and moreover $\sum_{t=1}^{T} \sum_{h=1}^{H} \mathbf{1}(a_t(h) = \pi_h^+(s_t(h))) = \Omega(T)$;*
2. *The number of reward manipulations is $\tilde{O}\big(\sqrt{T}H^{\alpha}/\epsilon\big)$.*
3. *The total amount of reward contamination is $\tilde{O}\big(\sqrt{T}H^{\alpha+1}(\epsilon + \sqrt{\log(HTSA)})/\epsilon\big)$.*

Algorithm 1 Black box attack strategy

Initialization: Input parameters are $\epsilon > 0$ and policy π^+. For all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$, attacker initializes $\hat{\mu}(s,a) = 0$, $\hat{\mu}^{UCB}(s,a) = M$, $\hat{\mu}^{LCB}(s,a) = -M$ and $N(s,a) = 0$.
for episode $t \leq T$ **do**
  Observe the initial state $s_t(1)$
  for $h \leq H$ **do**
    Observe the selected action $a_t(h)$, reward $r(s_t(h), a_t(h))$ and next state $s_t(h+1)$.
    Update

$$\begin{aligned}&\hat{\mu}(s_t(h), a_t(h)) \\ &= \frac{\hat{\mu}(s_t(h), a_t(h)) N(s_t(h), a_t(h)) + r(s_t(h), a_t(h))}{N(s_t(h), a_t(h)) + 1},\end{aligned} \tag{13}$$

$$\hat{\mu}^{UCB}(s_t(h), a_t(h)) = \hat{\mu}(s_t(h), a_t(h)) + \sigma \sqrt{\frac{4 \log(2THSA)}{N(s_t(h), a_t(h)) + 1}}, \tag{14}$$

$$\hat{\mu}^{LCB}(s_t(h), a_t(h)) = \hat{\mu}(s_t(h), a_t(h)) - \sigma \sqrt{\frac{4 \log(2THSA)}{N(s_t(h), a_t(h)) + 1}}, \tag{15}$$

    and $N(s_t(h), a_t(h)) = N(s_t(h), a_t(h)) + 1$.
    if $a_t(h) = \pi_h^+(s_t(h))$ **then**
      Do not contaminate.
    else
      Contaminate the reward observation such that

$$\begin{aligned}&r_t^o(s_t(h), a_t(h)) \\ &= \hat{\mu}^{LCB}(s_t(h), \pi_h^+(s_t(h))) - \epsilon \\ &+ (H-h) \min_{s,a \in \mathcal{S} \times \mathcal{A}} \hat{\mu}^{LCB}(s,a) \\ &- (H-h) \max_{s,a \in \mathcal{S} \times \mathcal{A}} \hat{\mu}^{UCB}(s,a).\end{aligned} \tag{16}$$

    end if
  end for
end for

Additionally, the proposed black-box attack has an additional cost $O(\sqrt{\log T})$ in comparison to the white-box attack in unbounded reward setting.

Remark 1. *In unbounded setting, reward poisoning attack has been studied in black-box setting recently in RL for infinite horizon by [Rakhsha* et al.*, 2021]. They consider attacking L online learners whereas we only has one learner. The attack objective of [Rakhsha* et al.*, 2021] is to force all these learning algorithms to execute the target policy π^+. We now highlight the key differences between our attack strategy and the strategy of [Rakhsha* et al.*, 2021]. The attack cost in [Rakhsha* et al.*, 2021] is $\tilde{O}(T \log L)$, which is linear in T. On contrary, the attack cost of our proposed attack is $\tilde{O}(\sqrt{T})$. This difference occurs because the attack in [Rakhsha* et al.*, 2021] estimates both the expected reward and the transition dynamics of the MDP. This is done by an explore-then-exploit form of strategy which leads to $\tilde{O}(T \log L)$ attack cost. On contrary, our strategy focuses on estimating expected rewards only (and not transition dynamics). It compensates for this lack of knowledge of transition dynamics by adding a negative bias $O(\sqrt{\log T})$ to the reward observation. However, this additional cost is minimal in comparison to the cost of learning transition dynamics in [Rakhsha* et al.*, 2021]. This key technical insight allows us to reduce the attack cost of $O(T)$ in [Rakhsha* et al.*, 2021] to $\tilde{O}(\sqrt{T})$ in current work.*

5 Additional Related Work

Reward manipulation attack has been studied extensively in MAB [Jun *et al.*, 2018; Liu and Shroff, 2019; Rangi *et al.*, 2021b], where the attacker's objective is to mislead the learner to choose a suboptimal action. Action manipulation attack has also been studied in MAB [Liu and Lai, 2020], and the number of action manipulations required by the attacker is $O(\log T)$. All these attacks are studied in a Black-box setting. In this work, we show that unlike MAB setting, reward manipulation (only) and action manipulation (only) are not sufficient to successfully attack Episodic RL setting with bounded rewards. [Bogunovic *et al.*, 2021] considered the "possibility" of reward poisoning attack in Linear Bandits, as a step towards showing lower bound for designing robust algorithms. While their results can be extended to prove a similar regret lower bound for RL, this only means that there *exists* an RL instance such that the attacker can successfully attack any no-regret algorithm for this particular instance. However, this existence result is significantly different from our perspective of designing attacks, in which we design strategies that can successfully attack an *arbitrary* episodic RL instance.

In online RL setting, studies related to poisoning attacks have only started recently, and have primarily focused on white-box settings, where the attacker has complete knowledge of the underlying MDP models, with unbounded rewards [Rakhsha *et al.*, 2020]. In such white-box attacks, [Rakhsha *et al.*, 2020] show that reward poisoning attack requires $\tilde{\Theta}(\sqrt{T})$ amount of contamination to attack any order-optimal learning algorithm; they also show that dynamic manipulation attack can achieve the same success with similar amount of cost in unbounded reward setting under some sufficient conditions. [Zhang *et al.*, 2020] study the feasibility of the reward poisoning attack in white box setting for Q-learning, and the attacker is constrained by the amount of contamination. In a slightly different thread, [Huang and Zhu, 2020] analyse the degradation of the performance of Temporal difference learning and Q-learning under falsified rewards.

To our knowledge, [Rakhsha *et al.*, 2021] is the only work that studies poisoning attack in black-box setting for policy teaching in *infinite-horizon* RL. However, they focused on the settings with L online learners, and the objective of their attacker is to force all these learners to execute a target policy π^+. They proposed an attack with $\tilde{O}(T \log L + L\sqrt{T})$ amount of contamination when L is large enough. However, our work focuses on attacking a *single* learner and thus our setting is not comparable to [Rakhsha *et al.*, 2021]. However, we can indeed apply our attack repeatedly to different learners to obtain an effective attack strategy for the setup of [Rakhsha *et al.*, 2021], which leads to an attack cost of $\tilde{O}(L\sqrt{T})$ (note however, the attack of [Rakhsha *et al.*, 2021] cannot work for small L, e.g., $L = 1$ as in our setup) in *episodic* RL. This improves their attack cost by an additive amount $O(T \log L)$. Our more efficient attack is due to a more efficient design for the adversary to explore and learn the MDP, which is discussed at length in Remark 1 in Section 4 .

Test-time adversarial attacks against RL has also been studied. Here, however, the policy π of the RL agent is pre-trained and fixed, and the objective of the attacker is to manipulate the perceived state of the RL agent in order to induce undesired action [Huang *et al.*, 2017; Lin *et al.*, 2017; Kos and Song, 2017; Behzadan and Munir, 2017]. Such test-time attacks do not modify the the policy π, whereas training-time attacks we study in this paper aims at poisoning the learned policy directly and thus may have a longer-term bad effects. There have also been studies on reward poisoning against *Batch RL* [Ma *et al.*, 2019a; Zhang *et al.*, 2009] where the attacker can modify the pre-collected batch data set at once. The focus of the present work is on *online* attack where the poisoning is done on the fly.

6 Conclusion and Future Directions

This paper tries to understand poisoning attacks in RL. Towards that end, we propose a reward manipulation attack for unbounded reward setting which successfully fool any order-optimal RL algorithm to pull a target policy with $\tilde{O}(\sqrt{T})$ attack cost. Extending the study to bounded reward setting, we show that the adversary cannot achieve its objective using either reward manipulation or action manipulation attack even in white-box setting, where the information about the MDP is assumed to be known. Hence, to contaminate a order-optimal RL algorithm, the adversary needs to combine the power of reward manipulation and action manipulation. Indeed, we show that an attack that uses both reward manipulation and action manipulation can achieve adversary's objective with $\tilde{\Theta}(\sqrt{T})$ attack cost, which is order-optimal. We also studied the in-feasibility of the attack under the constraint that the adversary can attack only if $a_t(h) \neq \pi_h^+(s)$.

References

[Agarwal *et al.*, 2016] A. Agarwal, S. Bird, M. Cozowicz, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

[Badanidiyuru *et al.*, 2018] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3), 2018.

[Behzadan and Munir, 2017] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *MLDM*. Springer, 2017.

[Bogunovic *et al.*, 2021] Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *AISTATS*, pages 991–999. PMLR, 2021.

[Cardenas *et al.*, 2008] Alvaro A Cardenas, Saurabh Amin, and Shankar Sastry. Secure control: Towards survivable cyber-physical systems. In *International Conference on Distributed Computing Systems Workshops*. IEEE, 2008.

[Chen *et al.*, 2021] Yifang Chen, Simon S Du, and Kevin Jamieson. Improved corruption robust algorithms for episodic reinforcement learning. *arXiv preprint arXiv:2102.06875*, 2021.

[Dhingra *et al.*, 2016] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. Towards end-to-end reinforcement learning of dialogue agents for information access. *arXiv preprint arXiv:1609.00777*, 2016.

[Huang and Zhu, 2020] Yunhan Huang and Quanyan Zhu. Manipulating reinforcement learning: Poisoning attacks on cost signals. *arXiv preprint arXiv:2002.03827*, 2020.

[Huang *et al.*, 2017] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

[Jin, 2018] Chi et al. Jin. Is q-learning provably efficient? In *Proceedings of the 32nd Nuerips*, 2018.

[Jun *et al.*, 2018] Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu. Adversarial attacks on stochastic bandits. In *NeurIPS*, pages 3640–3649, 2018.

[Khojasteh *et al.*, 2020] Mohammad Javad Khojasteh, Anatoly Khina, Massimo Franceschetti, and Tara Javidi. Learning-based attacks in cyber-physical systems. *IEEE Transactions on Control of Network Systems*, 2020.

[Kos and Song, 2017] Jernej Kos and Dawn Song. Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*, 2017.

[Lin *et al.*, 2017] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the 26th IJCAI*, 2017.

[Liu and Lai, 2020] Guanlin Liu and Lifeng Lai. Action-manipulation attacks against stochastic bandits: Attacks and defense. *IEEE Transactions on Signal Processing*, 68:5152–5165, 2020.

[Liu and Shroff, 2019] Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pages 4042–4050, 2019.

[Ma *et al.*, 2019a] Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy poisoning in batch reinforcement learning and control. In *Advances in Neural Information Processing Systems*, pages 14570–14580, 2019.

[Ma *et al.*, 2019b] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: attacks and defenses. In *International Joint Conference on Artificial Intelligence*. AAAI Press, 2019.

[Rakhsha *et al.*, 2020] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning*. PMLR, 2020.

[Rakhsha *et al.*, 2021] Amin Rakhsha, Xuezhou Zhang, Xiaojin Zhu, and Adish Singla. Reward poisoning in reinforcement learning: Attacks against unknown learners in unknown environments. *preprint arXiv:2102.08492*, 2021.

[Rangi and Franceschetti, 2018] Anshuka Rangi and Massimo Franceschetti. Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *AAMAS*, pages 1345–1352, 2018.

[Rangi *et al.*, 2021a] Anshuka Rangi, Mohammad Javad Khojasteh, and Massimo Franceschetti. Learning based attacks in cyber physical systems: Exploration, detection, and control cost trade-offs. In *Learning for Dynamics and Control*, pages 879–892. PMLR, 2021.

[Rangi *et al.*, 2021b] Anshuka Rangi, Long Tran-Thanh, Haifeng Xu, and Massimo Franceschetti. Secure-ucb: Saving stochastic bandits from poisoning attacks via limited data verification. *arXiv preprint arXiv:2102.07711*, 2021.

[Venugopal *et al.*, 2021] Aravind Venugopal, Elizabeth Bondi, Harshavardhan Kamarthi, Keval Dholakia, Balaraman Ravindran, and Milind Tambe. Reinforcement learning for unified allocation and patrolling in signaling games with uncertainty. In *AAMAS*, 2021.

[Wu *et al.*, 2021] Yue Wu, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. *arXiv preprint arXiv:2102.07301*, 2021.

[Zhang *et al.*, 2009] Haoqi Zhang, David C Parkes, and Yiling Chen. Policy teaching through reward function learning. In *ACM conference on Electronic commerce*, 2009.

[Zhang *et al.*, 2020] X. Zhang, Y. Ma, A. Singla, and X. Zhu. Adaptive reward-poisoning attacks against reinforcement learning. *arXiv preprint arXiv:2003.12613*, 2020.

[Zhao *et al.*, 2018] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. Deep reinforcement learning for page-wise recommendations. In *12th ACM Conference on Recommender Systems*, 2018.