Adaptive Reward-Poisoning Attacks against Reinforcement Learning

Xuezhou Zhang 1 Yuzhe Ma 1 Adish Singla 2 Xiaojin Zhu 1

Abstract

In reward-poisoning attacks against reinforcement learning (RL), an attacker can perturb the environment reward $r_t$ into $r_t + \delta_t$ at each step, with the goal of forcing the RL agent to learn a nefarious policy. We categorize such attacks by the infinity-norm constraint on $\delta$: We provide a lower threshold below which reward-poisoning attack is infeasible and RL is certified to be safe; we provide a corresponding upper threshold above which the attack is feasible. Feasible attacks can be further categorized as non-adaptive where $\delta$ depends only on $(s_t, a_t, s_{t+1})$, or adaptive where $\delta$ depends further on the RL agent’s learning process at time $t$. Non-adaptive attacks have been the focus of prior works. However, we show that under mild conditions, adaptive attacks can achieve the nefarious policy in steps polynomial in state-space size $|S|$, whereas non-adaptive attacks require exponential steps. We provide a constructive proof that a Fast Adaptive Attack strategy achieves the polynomial rate. Finally, we show that empirically an attacker can find effective reward-poisoning attacks using state-of-the-art deep RL techniques.

1. Introduction

In many reinforcement learning (RL) applications the agent extracts reward signals from user feedback. For example, in recommendation systems the rewards are often represented by user clicks, purchases or dwell time (Zhao et al., 2018; Chen et al., 2019); in conversational AI, the rewards can be user sentiment or conversation length (Dhingra et al., 2016; Li et al., 2016). In such scenarios, an adversary can manipulate user feedback to influence the RL agent in nefarious ways. Figure 1 describes a hypothetical scenario of how conversational AI can be attacked. One real-world example is that of the chatbot Tay, which was quickly corrupted by a group of Twitter users who deliberately taught it misogynistic and racist remarks shortly after its release (Neff & Nagy, 2016). Such attacks reveal significant security threats in the application of reinforcement learning.

\begin{align*}
\text{a}_t: & \quad \text{`Hello! You look pretty!'} & \quad \text{Thank you!} & \quad r_t = +1 \\
\text{a}_t: & \quad \text{`Hello! You look pretty!'} & \quad \text{Hey, don’t say that!} & \quad r_t = -1
\end{align*}

Figure 1. Example: an RL-based conversational AI is learning from real-time conversations with human users. The chatbot says “Hello! You look pretty!” and expects to learn from user feedback (sentiment). A benign user will respond with gratitude, which is decoded as a positive reward signal. An adversarial user, however, may express anger in his reply, which is decoded as a negative reward signal.

In this paper, we formally study the problem of \textit{training-time attack on RL via reward poisoning}. As in standard RL, the RL agent updates its policy $\pi_t$ by performing action $a_t$ at state $s_t$ in each round $t$. The environment Markov Decision Process (MDP) generates reward $r_t$ and transits the agent to $s_{t+1}$. However, the attacker can change the reward $r_t$ to $r_t + \delta_t$, with the goal of driving the RL agent toward a target policy $\pi_t \rightarrow \pi^\dagger$.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{A chain MDP with attacker’s target policy $\pi^\dagger$}
\end{figure}

Figure 2 shows a running example that we use throughout the paper. The episodic MDP is a linear chain with five states, with left or right actions and no movement if it hits the boundary. Each move has a -0.1 negative reward, and $G$ is the absorbing goal state with reward 1. Without attack, the optimal policy $\pi^*$ would be to always move right. The attacker’s goal, however, is to force the agent to learn the nefarious target policy $\pi^\dagger$ represented by the arrows in Figure 2. Specifically, the attacker wants the agent to move left and hit its head against the wall whenever the agent is at the left-most state.
Our main contributions are:

1. We characterize conditions under which such attacks are guaranteed to fail (thus RL is safe), and vice versa;
2. In the case where an attack is feasible, we provide upper bounds on the attack cost in the process of achieving $\pi^*$;
3. We show that effective attacks can be found empirically using deep RL techniques.

2. Related Work

Test-time attacks against RL Prior work on adversarial attacks against reinforcement learning focused primarily on test-time, where the RL policy $\pi$ is pre-trained and fixed, and the attacker manipulates the perceived state $s_t$ to $s'_t$ in order to induce undesired action (Huang et al., 2017; Lin et al., 2017; Kos & Song, 2017; Behzadan & Munir, 2017). For example, in video games the attacker can make small pixel perturbation to a frame (Goodfellow et al., 2014)) to induce an action $\pi(s'_t) \neq \pi(s_t)$. Although test-time attacks can severely impact the performance of a deployed and fixed policy $\pi$, they do not modify $\pi$ itself. For ever-learning agents, however, the attack surface includes $\pi$. This motivates us to study training-time attack on RL policy.

Reward Poisoning: Reward poisoning has been studied in bandits (Jun et al., 2018; Peltila et al., 2019; Altschuler et al., 2019; Liu & Shroff, 2019; Ma et al., 2018), where the authors show that adversarially perturbed reward can mislead standard bandit algorithms to pull a suboptimal arm or suffer large regret.

Reward poisoning has also been studied in batch RL (Zhang & Parkes, 2008; Zhang et al., 2009; Ma et al., 2019) where rewards are added to a pre-collected batch data set by some behavior policy, and the attacker modifies the batch data. Because all data are available to the attacker at once, the batch attack problem is relatively easier. This paper instead focuses on the online RL attack setting where reward poisoning must be done on the fly.

(Huang & Zhu, 2019) studies a restricted version of reward poisoning, in which the perturbation only depend on the current state and action: $\delta_t = \phi(s_t, a_t)$. While such restriction guarantees the convergence of Q-learning under the perturbed reward and makes the analysis easier, we show both theoretically and empirically that such restriction severely harms attack efficiency. Our paper subsumes their results by considering more powerful attacks that can depend on the RL victim’s Q-table $Q_t$. Theoretically, our analysis does not require the RL agent’s underlying $Q_t$ to converge while still providing robustness certificates; see section 4.

Reward Shaping: While this paper is phrased from the adversarial angle, the framework and techniques are also applicable to the teaching setting, where a teacher aims to guide the agent to learn the optimal policy as soon as possible, by designing the reward signal. Traditionally, reward shaping and more specifically potential-based reward shaping (Ng et al., 1999) has been shown able to speed up learning while preserving the optimal policy. (Devlin & Kudenko, 2012) extend potential-based reward shaping to be time-varying while remains policy-preserving. More recently, intrinsic motivations (Schmidhuber, 1991; Oudeyer & Kaplan, 2009; Barto, 2013; Bellemare et al., 2016) was introduced as a new form of reward shaping with the goal of encouraging exploration and thus speed up learning. Our work contributes by mathematically defining the teaching via reward shaping task as an optimal control problem, and provide computational tools that solve for problem-dependent high-performing reward shaping strategies.

3. The Threat Model

In the reward-poisoning attack problem, we consider three entities: the environment MDP, the RL agent, and the attacker. Their interaction is formally described by Alg 1.

The environment MDP is $\mathcal{M} = (S, A, R, P, \mu_0)$ where $S$ is the state space, $A$ is the action space, $R : S \times A \times S \rightarrow \mathbb{R}$ is the reward function, $P : S \times A \times S \rightarrow \mathbb{R}$ is the transition probability, and $\mu_0 : S \rightarrow \mathbb{R}$ is the initial state distribution. We assume $S$, $A$ are finite, and that a uniformly random policy can visit each $(s, a)$ pair infinitely often.

We focus on an RL agent that performs standard $Q$-learning defined by a tuple $\mathcal{A} = \{Q_0, \varepsilon, \gamma, \{\alpha_t\}\}$, where $Q_0$ is the initial $Q$ table, $\varepsilon$ is the random exploration probability, $\gamma$ is the discounting factor, $\{\alpha_t\}$ is the learning rate scheduling as a function of $t$. This assumption can be generalized: in the additional experiments provided in appendix ??, we show how the same framework can be applied to attack general RL agents, such as DQN. Denote $Q^*$ as the optimal $Q$ table that satisfies the Bellman’s equation:

$$Q^*(s, a) = \mathbb{E}_{P(s'|s, a)} \left[ R(s, a, s') + \gamma \max_{a' \in A} Q^*(s', a') \right]$$

and denote the corresponding optimal policy as $\pi^*(s) = \arg \max_{a} Q^*(s, a)$. For notational simplicity, we assume $\pi^*$ is unique, though it is easy to generalize to multiple optimal policies, since most of our analyses happen in the space of value functions.

The Threat Model The attacker sits between the environment and the RL agent. In this paper we focus on white-box attacks: the attacker has knowledge of the environment MDP and the RL agent’s $Q$-learning algorithm, except for their future randomness. Specifically, at time $t$ the attacker observes the learner $Q$-table $Q_t$, state $s_t$, action $a_t$, the environment transition $s_{t+1}$ and reward $r_t$. The attacker
Algorithm 1 Reward Poisoning against Q-learning

**PARAMETERS:** Agent parameters $\mathcal{A} = (Q_0, \varepsilon, \gamma, \{\alpha_t\})$, MDP parameters $\mathcal{M} = (S, A, R, P, \mu_0)$.

1: for $t = 0, 1, \ldots$ do
2: agent at state $s_t$, has Q-table $Q_t$.
3: agent acts according to $\varepsilon$-greedy behavior policy
   \[ a_t \left\{ \begin{array}{ll}
   \arg \max_{a} Q_t(s_t, a), & \text{w.p. } 1 - \varepsilon \\
   \text{uniform from } A, & \text{w.p. } \varepsilon.
   \end{array} \right. \tag{2} \]
4: environment transits $s_{t+1} \sim P(\cdot | s_t, a_t)$, produces reward $r_t = R(s_t, a_t, s_{t+1})$.
5: attacker poisons the reward to $r_t + \delta_t$.
6: agent receives $(s_{t+1}, r_t + \delta_t)$, performs Q-learning update:
   \[ Q_{t+1}(s_{t+1}, a_t) \leftarrow (1 - \alpha_t)Q_t(s_t, a_t) + \alpha_t \left( r_t + \delta_t + \gamma \max_{a' \in A} Q_t(s_{t+1}, a') \right) \tag{3} \]
7: environment resets if episode ends: $s_{t+1} \sim \mu_0$.
8: end for

If the target policy $\pi^*$ always specifies a singleton action or does not care on all states, then $Q^*$ is a convex set. But in general when $1 < |\pi^*(s)| < |A|$ on any $s$, $Q^*$ will be a union of convex sets but itself can be in general non-convex.

### 4. Theoretical Guarantees

Now, we are ready to formally define the optimal attack problem. At time $t$, the attacker observes an attack state (N.B. distinct from MDP state $s_t$):
\[ \xi_t := (s_t, a_t, s_{t+1}, r_t, Q_t) \in \Xi \tag{4} \]
which jointly characterizes the MDP and the RL agent. The attacker’s goal is to find an attack policy $\phi : \Xi \rightarrow [-\Delta, \Delta]$, where for $\xi_t \in \Xi$ the attack action is $\delta_t := \phi(\xi_t)$, that minimizes the number of rounds on which the agent’s $Q_t$ disagrees with the attack target $Q^*$:
\[ \min_{\phi} \quad \mathbb{E}_{\phi} \sum_{t=0}^{\infty} 1[Q_t \notin \mathbb{Q}^*], \tag{5} \]
where the expectation accounts for randomness in Alg 1. We denote $J(\phi) = \mathbb{E}_{\phi} \sum_{t=0}^{\infty} 1[Q_t \notin \mathbb{Q}^*]$ the total attack cost, and $J_T(\phi) = \mathbb{E}_{\phi} \sum_{t=0}^{T} 1[Q_t \notin \mathbb{Q}^*]$ the finite-horizon cost. We say the attack is feasible if (5) is finite.

Next, we characterize attack feasibility in terms of poison magnitude constraint $\Delta$, as summarized in Figure 3. Proofs to all the theorems can be found in the appendix.

#### 4.1. Attack Infeasibility

Intuitively, smaller $\Delta$ makes it harder for the attacker to achieve the attack goal. We show that there is a threshold $\Delta_1$ such that for any $\Delta < \Delta_1$ the RL agent is eventually safe, in that $\pi_t \rightarrow \pi^*$ the correct MDP policy. This implies that (5) is infinite and the attack is infeasible. There is a potentially larger $\Delta_2$ such that for any $\Delta < \Delta_2$ the attack is also infeasible, though $\pi_t$ may not converge to $\pi^*$.

While the above statements are on $\pi_t$, our analysis is via the RL agent’s underlying $Q_t$. Note that under attack the rewards $r_t + \delta_t$ are no longer stochastic, and we cannot utilize the usual Q-learning convergence guarantee. Nonetheless, we show that $Q_t$ is bounded in a polytope in the Q-space.

**Theorem 1** (Boundedness of Q-learning). Assume that $\delta_t < \Delta$ for all $t$, and the stepsize $\alpha_t$ satisfies $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$. Let $Q^*$ be defined as (1). Then, for any attack sequence $\{\delta_t\}$, there exists $N \in \mathbb{N}$ such that, with probability 1, for all $t \geq N$, we have
\[ Q^*(s, a) - \frac{\Delta}{1 - \gamma} \leq Q_t(s, a) \leq Q^*(s, a) + \frac{\Delta}{1 - \gamma}. \tag{6} \]

Remark 1: The bounds in Theorem 1 are in fact tight. The lower and upper bound can be achieved by setting $\delta_t = -\Delta$ or $+\Delta$ respectively.
We immediately have the following two infeasibility certificates.

**Corollary 2 (Strong Infeasibility Certificate).** Define
\[
\Delta_1 = (1 - \gamma) \min_{s} \left[ Q^*(s, \pi^*(s)) - \max_{a \neq \pi^*(s)} Q^*(s, a) \right] / 2.
\]
If \( \Delta < \Delta_1 \), there exist \( N \in \mathbb{N} \) such that, with probability 1, for all \( t > N \), \( \pi_t = \pi^* \). In other words, eventually the RL agent learns the optimal MDP policy \( \pi^* \) despite the attacks.

**Corollary 3 (Weak Infeasibility Certificate).** Given attack target policy \( \pi^+ \), define
\[
\Delta_2 = (1 - \gamma) \max_{s} \left[ Q^*(s, \pi^+(s)) - \max_{a \in \pi^+(s)} Q^*(s, a) \right] / 2.
\]
If \( \Delta < \Delta_2 \), there exist \( N \in \mathbb{N} \) such that, with probability 1, for all \( t > N \), \( \pi_t(s) \neq \pi^+(s) \) for some \( s \in S^1 \). In other words, eventually the attacker is unable to enforce \( \pi^+ \) (though \( \pi_t \) may not settle on \( \pi^* \) either).

Intuitively, an MDP is difficult to attack if its margin \( \min_s [Q^*(s, \pi^*(s)) - \max_{a \neq \pi^*(s)} Q^*(s, a)] \) is large. This suggests a defense: for RL to be robust against poisoning, the environmental reward signal should be designed such that the optimal actions and suboptimal actions have large performance gaps.

### 4.2. Attack Feasibility

We now show there is a threshold \( \Delta_3 \) such that for all \( \Delta > \Delta_3 \), the attacker can enforce \( \pi^+ \) for all but finite number of rounds.

**Theorem 4.** Given a target policy \( \pi^+ \), define
\[
\Delta_3 = \frac{1 + \gamma}{2} \max_{s \in S^1} \max_{a \neq \pi^+(s)} Q^*(s, a) - \min_{a \in \pi^+(s)} Q^*(s, a)
\]
where \( \max(x, 0) \) is the same conditions on \( \alpha_t \) as in Theorem 1. If \( \Delta > \Delta_3 \), there is a feasible attack policy \( \phi_{\Delta_3}^{\text{as}} \). Furthermore, \( J_\infty(\phi_{\Delta_3}^{\text{as}}) \leq O(L^5) \), where \( L \) is the covering number.

| Guarantee \( \pi_t \to \pi^+ \) | \( \exists \phi \), s.t. \( \pi_t \to \pi^+ \) Non-adaptive attack achieves \( J_\infty = O(\exp(|S|)) \). |
|---|---|
| \( \Delta_1 \) | Adaptive attack achieves \( J_\infty = O(|S|^2) \). |

**Algorithm 2** The Non-Adaptive Attack \( \phi_{\Delta_3}^{\text{as}} \)

**PARAMETERS:** target policy \( \pi^+ \), agent parameters \( M = (S, A, R, P, \mu_0) \), maximum magnitude of poisoning \( \Delta \).

**def Init(\( \pi^+, A, M \)):**

1: Construct a Q-table \( Q' \), where \( Q'(s, a) \) is defined as
\[
\begin{cases}
Q^*(s, a) + \Delta/(1 + \gamma), & \text{if } s \in S^1, a \in \pi^+(s) \\
Q^*(s, a) - \Delta/(1 + \gamma), & \text{if } s \in S^1, a \notin \pi^+(s) \\
Q^*(s, a), & \text{if } s \notin S^1
\end{cases}
\]
2: Calculate a new reward function
\[
R'(s, a) = Q'(s, a) - \gamma E_{P(s'|s,a)} \left[ \max_{a'} Q'(s', a') \right].
\]
3: Define the attack policy \( \phi_{\Delta_3}^{\text{as}} \) as:
\[
\phi_{\Delta_3}^{\text{as}}(s, a) = R'(s, a) - \gamma E_{P(s'|s,a)} [R(s, a, s)], \forall s, a.
\]

**def Attack(\( \xi_\ell \))**:

1: Return \( \phi_{\Delta_3}^{\text{as}}(s_t, a_t) \)

**Theorem 4** is proved by constructing an attack policy \( \phi_{\Delta_3}^{\text{as}}(s_t, a_t) \), detailed in Alg. 2. Note that this attack policy does not depend on \( Q_\ell \). We call this type of attack **non-adaptive attack**. Under such construction, one can show that Q-learning converges to the target policy \( \pi^+ \). Recall the covering number \( L \) is the upper bound on the minimum sequence length starting from any \( (s, a) \) pair and follow the MDP until all (state, action) pairs appear in the sequence (Even-Dar & Mansour, 2003). It is well-known that \( \varepsilon \)-greedy exploration has a covering time \( L \leq O(e^{|S|}) \) (Kearns & Singh, 2002). Prior work has constructed examples on which this bound is tight (Jin et al., 2018). We show in Appendix ?? that on our toy example \( \varepsilon \)-greedy indeed has a covering time \( O(e^{|S|}) \). Therefore,
We now show that there is a fast adaptive attack \( \phi \in \pi \). Denote since we are in an MDP, there exists a common (partial) hitting time tightly upper bounds the optimal hitting time, bounded by \(|S|\). The FAA attack policy \( \phi_{FAA}^{\varepsilon} \) is defined in Alg. 3.

Conceptually, the FAA algorithm ranks the target states in descending order by their distance to the starting states, and focusing on attacking one target state at a time. Of central importance is the temporary target policy \( \nu_i \), which is designed to navigate the agent to the currently focused target state \( s_{(i)}^{\dagger} \), while not altering the already achieved target actions on target states of earlier rank. This allows FAA to achieve a form of program invariance: after FAA achieves the target policy in a target state \( s_{(i)}^{\dagger} \), the target policy on target state \( i \) will be preserved indefinitely. We provide a more detailed walk-through of Alg. 3 with examples in appendix ??.

**Definition 3.** Define the shortest \( \varepsilon \)-distance from \( s \) to \( s' \) as
\[
d_{\varepsilon}(s, s') = \min_{\pi \in \Pi} \mathbb{E}_{\pi|s}[T]
\]
s.t. \( s_0 = s, s_T = s' \), \( s_i \neq s', \forall t < T \)

where \( \pi_{\varepsilon} \) denotes the epsilon-greedy policy based on \( \pi \). Since we are in an MDP, there exists a common (partial) policy \( \pi_{s'} \) that achieves \( d_{\varepsilon}(s, s') \) for all source state \( s \in S \). Denote \( \pi_{s'} \) as the navigation policy to \( s' \).

**Definition 4.** The \( \varepsilon \)-diameter of an MDP is defined as the longest shortest \( \varepsilon \)-distance between pairs of states in \( S \):
\[
D_{\varepsilon} = \max_{s,s' \in S} d_{\varepsilon}(s, s')
\]

**Theorem 5.** Assume that the learner is running \( \varepsilon \)-greedy Q-learning algorithm on an episodic MDP with \( \varepsilon \)-diameter \( D_{\varepsilon} \) and maximum episode length \( H \), and the attacker aims at \( k \) distinct target states, i.e. \(|S| = k \). If \( \Delta \) is large enough that the Clip\( \Delta () \) function in Alg. 3 never takes effect, then \( \phi_{FAA}^{\varepsilon} \) is feasible, and we have
\[
J_\infty(\phi_{FAA}^{\varepsilon}) \leq k\frac{|S||A|H}{1-\varepsilon} + \frac{|A|}{1-\varepsilon} \left| \frac{1}{\varepsilon} \right|^k D_{\varepsilon},
\]

How large is \( D_{\varepsilon} \)? For MDPs with underlying structure as undirected graphs, such as the grid worlds, it is shown that the expected hitting time of a uniform random walk is bounded by \( O(|S|^2) \) (Lawler, 1986). Note that the random hitting time tightly upper bounds the optimal hitting time.

**Algorithm 3 The Fast Adaptive Attack (FAA)**

**PARAMETERS:** target policy \( \pi \), margin \( \eta \), agent parameters \( A = (Q_0, \varepsilon, \gamma, \{a_t\}) \), MDP parameters \( M = (S, A, R, P, \mu_0) \).

**def Init(\( \pi \), \( A \), \( M \), \( \eta \)):**

1: Given \((s_t, a_t, Q_t)\), define the hypothetical Q-update function without attack as \( Q_{t+1}(s_t, a_t) = (1-\alpha_t)Q_t(s_t, a_t) + \alpha_t \{r_t + \gamma(1-EOE) \max_{a \in A} Q_t(s_{t+1}, a')\} \).
2: Given \((s_t, a_t, Q_t)\), denote the greedy attack function at \( s_t \) w.r.t. a target action set \( A_{s_t} \), as \( g(A_{s_t}) \), defined as
\[
\left\{
\begin{array}{ll}
\frac{1}{\alpha_t} \max_{a \notin A_{s_t}} Q_t(s_t, a) - Q_{t+1}(s_t, a) + \eta & \text{if } a_t \in A_{s_t} \\
\frac{1}{\alpha_t} \max_{a \in A_{s_t}} Q_t(s_t, a) - Q_{t+1}(s_t, a) + \eta & \text{if } a_t \notin A_{s_t}.
\end{array}
\right.
\]
3: Define \( \text{Clip}_\Delta(\delta) = \min(\max(\delta - \Delta), \Delta) \).
4: Rank the target states in descending order as \( \{s_{(1)}^{\dagger}, ..., s_{(k)}^{\dagger}\} \), according to their shortest \( \varepsilon \)-distance to the initial state \( E_{s_0} \sim \mu_0 \) \( d^\varepsilon(s, s_{(i)}) \).
5: for \( i = 1, ..., k \) do
6: Define the temporary target policy \( \nu_i \) as
\[
\nu_i(s) = \left\{
\begin{array}{ll}
\pi_{s_{(i)}^{\dagger}}(s) & \text{if } s \notin \{s_{(j)}^{\dagger} : j \leq i\} \\
\pi^\dagger(s) & \text{if } s \in \{s_{(j)}^{\dagger} : j \leq i\}.
\end{array}
\right.
\]

7: end for

**def Attack(\( \varepsilon \)):**

1: for \( i = 1, ..., k \) do
2: if \( \arg \max_a Q_t(s_{(i)}^{\dagger}, a) \neq \pi^\dagger(s_{(i)}^{\dagger})\) then
3: Return \( \delta_t \leftarrow \text{Clip}_\Delta(g(\nu_i(s_{(i)}))) \).
4: end if
5: end for
6: Return \( \delta_t \leftarrow \text{Clip}_\Delta(g(\{\pi^\dagger(s_{(i)})\})) \).

a.k.a. the \( \varepsilon \)-diameter \( D_{\varepsilon} \), and they match when \( \varepsilon = 1 \). This immediately gives us the following result:

**Corollary 6.** If in addition to the assumptions of Theorem 5, the maximal episode length \( H = O(|S|) \), then \( J_\infty(\phi_{FAA}^{\varepsilon}) \leq O(e^k|S|^2|A|) \) in Grid World environments. When the number of target states is small, i.e. \( k \leq O(\log |S|) \), \( J_\infty(\phi_{FAA}^{\varepsilon}) \leq O(\text{poly}(|S|)) \).

**Remark 2:** Theorem 5 and Corollary 6 can be thought of as defining an implicit \( \Delta_4 \), such that for any \( \Delta > \Delta_4 \), the clip function in Alg. 3 never take effect, and \( \phi_{FAA}^{\varepsilon} \) achieves polynomial cost.
In this section, we present our key computational insight: the attacker can find an effective attack policy by relaxing the attack problem (5) so that the relaxed problem can be effectively solved with RL. Concretely, consider the higher-level attack MDP $N' = (\Xi, \Delta, \rho, \tau)$ and the associated optimal control problem:

- The attacker observes the attack state $\xi_t \in \Xi$.
- The attack action space is $\{\delta_t \in \mathbb{R} : |\delta_t| \leq \Delta\}$.
- The original attack loss function $1[I(Q_t \not\in Q^*)]$ is a 0-1 loss that is hard to optimize. We replace it with a continuous surrogate loss function $\rho$ that measures how close the current agent Q-table $Q_t$ is to the target Q-table set:

$$\rho(\xi_t) = \sum_{s \in S^1} \max_{a \in \pi^t(s)} Q_t(s, a) - \max_{a \in \pi^t(s)} Q^*(s, a) + \eta$$

(12)

where $\eta > 0$ is a margin parameter to encourage that $\pi^t(s)$ is strictly preferred over $A|\pi^t(s)$ with no ties.

- The attack state transition probability is defined by $\tau(\xi_{t+1} | \xi_t, \delta_t)$. Specifically, the new attack state $\xi_{t+1} = (s_{t+1}, a_{t+1}, s_{t+2}, r_{t+1}, Q_{t+1})$ is generated as follows:
  - $s_{t+1}$ is copied from $\xi_t$ if not the end of episode, else $s_{t+1} \sim \mu_0$.
  - $a_{t+1}$ is the RL agent’s exploration action drawn according to (2), note it involves $Q_{t+1}$.
  - $s_{t+2}$ is the RL agent’s new state drawn according to the MDP transition probability $P(\cdot | s_{t+1}, a_{t+1})$.
  - $r_{t+1}$ is the new (not yet poison) reward according to MDP $R(s_{t+1}, a_{t+1}, s_{t+2})$.
  - The attack $\delta_t$ happens. The RL agent updates $Q_{t+1}$ according to (3).

With the higher-level attack MDP $N'$, we relax the optimal attack problem (5) into

$$\phi^* = \arg \min_{\phi} \mathbb{E}_\phi \sum_{t=0}^{\infty} \rho(\xi_t)$$

(13)

One can now solve (13) using Deep RL algorithms. In this paper, we choose Twin Delayed DDPG (TD3) (Fujimoto et al., 2018), a state-of-the-art algorithm for continuous action space. We use the same set of hyperparameters for TD3 across all experiments, described in appendix ??.
We perform the experiments on the chain MDP in Figure 2. Recall that on this example, $\Delta_3 = 0.132$ and $\Delta_4 = 1$ (implicit). We evaluate across 4 different $\Delta$ values, $[0.1, 0.2, 0.5, 1]$, covering the range from $\Delta_3$ to $\Delta_4$. The result is shown in Figure 4.

We are able to make several interesting observations:

1. All attacks are feasible ($y$-axis $\ll T$), even when $\Delta$ falls under the thresholds $\Delta_3$ and $\Delta_4$ for corresponding methods. This suggests that the feasibility thresholds are not tight.

2. For non-adaptive attacks, as $\Delta$ increases the best-found attack policies $\phi_{\text{SAS}}$ achieve small improvement, but generally incur a large attack cost.

3. Adaptive attacks are very efficient when $\Delta$ is large. At $\Delta = 1$, the best adaptive attack $\phi_{\text{FAA} + \text{TD3}}$ achieves a cost of merely 13 (takes 13 steps to always force $\pi^\dagger$ on the RL agent). However, as $\Delta$ decreases the performance quickly degrades. At $\Delta = 0.1$ adaptive attacks are only as good as non-adaptive attacks. This shows an interesting transition region in $\Delta$ that our theoretical analysis does not cover.

6.2. Adaptive Attacks are Faster

In this experiment, we empirically verify that, while both are feasible, adaptive attacks indeed have an attack cost $O(\text{Poly}[S])$ while non-adaptive attacks have $O(\epsilon[S])$. The $0$-1 costs $1[\pi_t \neq \pi^\dagger]$ are in general incurred at the beginning of each $t = 0 \ldots T$ run. In other words, adaptive attacks achieve $\pi^\dagger$ faster than non-adaptive attacks. We use several chain MDPs similar to Figure 2 but with increasing number of states $|S| = 3, 4, 5, 6, 12$. We provide a large enough $\Delta = 2 \gg \Delta_4$ to ensure the feasibility of all attack policies. The result is shown in Figure 5. The best-found non-adaptive attack $\phi_{\text{SAS} + \text{TD3}}$ is approximately straight on the log-scale plot, suggesting attack cost $J$ growing exponentially with MDP size $|S|$. In contrast, the two adaptive attack policies $\phi_{\text{FAA}}^\dagger$ and $\phi_{\text{FAA} + \text{TD3}}^\dagger$ actually achieve attack cost linear in $|S|$. This is not easy to see from this log-scaled plot; we reproduce Figure 5 without the log scale in the

Figure 5. Attack performances on the chain MDPs of different lengths. Each curve shows mean $\pm 1$ standard error over 1000 independent test runs.
We use three MDPs: a 6-state chain MDP, a 12-state chain MDP, and a 10 × 10 grid world MDP. The 10 × 10 MDP has two separate target policies π† 1 and π† 2, see Figure 6.

For evaluation, we compute the number of target actions achieved |{s ∈ S† : πt(s) ∈ π†(s)}| as a function of t. This allows us to look more closely into the progress made by an attack. The results are shown in Figure 7.

First, observe that across all 4 experiments, attack policy φTD3 found by out-of-the-box TD3 never succeeded in achieving all target actions. This indicates that TD3 alone cannot produce an effective attack. We hypothesize that this is due to a lack of effective exploration scheme: when the target states are sparse (|S†| ≪ |S|) it can be hard for TD3 equipped with Gaussian exploration noise to locate all target states. As a result, the attack policy found by vanilla TD3 is only able to achieve the target actions on a subset of frequently visited target states.

Hand-crafted φFAA is effective in achieving the target policies, as is guaranteed by our theory. Nevertheless, we found that φFAA+TD3 always improves upon φTD3. Recall that we use FAA as the initialization and then run TD3. This indicates that TD3 can be highly effective with a good initialization, which effectively serves as the initial exploration policy that allows TD3 to locate all the target states.

Of special interest are the two experiments on the 10 × 10 Grid World with different target policies. Conceptually, the advantage of the adaptive attack is that the attacker can perform explicit navigation to lure the agent into the target states. An efficient navigation policy that leads the agent to all target states will make the attack very efficient. Observe that in Figure 6, both target policies form a chain, so that if the agent starts at the beginning of the chain, the target actions naturally lead the agent to the subsequent target states, achieving efficient navigation.

Recall that the FAA algorithm prioritizes the target states farthest to the starting state. In the 10 × 10 Grid World, the farthest state is the top-left grid. For target states S† 1, the top-left grid turns out to be the beginning of the target chain. As a result, φFAA is already very efficient, and φFAA+TD3 couldn’t achieve much improvement, as shown in 7c. On the other hand, for target states S† 2, the top-left grid is in the middle of the target chain, which makes φFAA not as efficient. In this case, φFAA+TD3 makes a significant improvement, successfully forcing the target policy in about 500 steps, whereas it takes φFAA as many as 1000 steps, about twice as long as φFAA+TD3.

7. Conclusion

In this paper, we studied the problem of reward-poisoning attacks against reinforcement-learning agents. Theoretically, we provide robustness certificates that guarantee the truthfulness of the learned policy when the attacker’s constraint is stringent. When the constraint is loose, we show that by being adaptive to the agent’s internal state, the attacker can force the target policy in polynomial time, whereas a naive non-adaptive attack takes exponential time. Empirically, we formulate that the reward-poisoning problem as an optimal control problem on a higher-level attack MDP, and developed computational tools based on DRL that is able to find efficient attack policies across a variety of environments.

Acknowledgments

This work is supported in part by NSF 1545481, 1623605, 1704117, 1836978 and the MADLab AF Center of Excellence FA9550-18-1-0166.
References


