Online Defense Strategies for Reinforcement Learning Against Adaptive Reward Poisoning

Andi Nika
MPI-SWS

Adish Singla
MPI-SWS

Goran Radanovic
MPI-SWS

Abstract

We consider the problem of defense against reward-poisoning attacks in reinforcement learning and formulate it as a game in T rounds between a defender and an adaptive attacker in an adversarial environment. To address this problem, we design two novel defense algorithms. First, we propose *Exp3-DARP*, a defense algorithm that uses Exp3 as a hyperparameter learning subroutine, and show that it achieves order-optimal $\tilde{\Theta}(T^{1/2})$ bounds on our notion of regret with respect to a defense that always picks the optimal parameter in hindsight. We show that the order of T in the bounds cannot be improved when the reward arrival process is adversarial, even if the feedback model of the defense is stronger. However, assuming that the environment is stochastic, we propose *OMDUCB-DARP* that uses estimates of costs as proxies to update the randomized strategy of the learner and are able to substantially improve the bounds proportional to how smoothly the attacker's strategy changes. Furthermore, we show that weaker types of defense, that do not take into account the attack structure and the poisoned rewards, suffer linear regret with respect to a defender that always selects the optimal parameter in hindsight when faced with an adaptive attacker that uses a no-regret algorithm to learn the behavior of the defense. Finally, we support our theoretical results with experimental evaluations on three different environments, showcasing the efficiency of our methods.

1 INTRODUCTION

One of the key aspects of designing novel machine learning methods is their robustness to adversarial attacks. The virtual environments of today are increasingly relying on complex learning algorithms for decision-making, thus the study of security threats to such algorithms is paramount. We consider training-time reward poisoning attacks to reinforcement learning. Training-time attacks have been previously studied in supervised learning (Rosenfeld et al., 2020), in reinforcement learning (Rakhsha et al., 2021a; Zhang et al., 2020b), and in the simpler bandit setting (Liu and Shroff, 2019; Rangi et al., 2022a). These types of attacks are characterized by their intervention in the training data set. They can either modify, delete or insert new data points into it. A naive algorithm that is oblivious to such attacks will inevitably adopt a suboptimal behavior and thus, employing defense strategies against them becomes necessary.

There have been various approaches to the defense problem against data poisoning attacks. Randomization over the training/test data, both in reinforcement learning (Kumar et al., 2021; Wu et al., 2022) and in supervised learning (Lecuyer et al., 2019; Cohen et al., 2019; Rosenfeld et al., 2020) is a recently studied technique, where the prediction of the defense corresponds to the one with the highest probability when random noise is applied to the data point. Each prediction is associated with a certificate of its certainty. On the other hand, robust statistics techniques can be used to detect outliers in the dataset by evaluating their sample variances against an appropriately chosen threshold (Zhang et al., 2021b). Furthermore, the utilization of the specific attack structure to solve an inverse worst-case optimization problem (e.g. in offline reinforcement learning (Banihashem et al., 2021)) is another type of defense that uses the optimization problem to compute robust policies.

All the techniques mentioned above are first and foremost parametric in nature, in that the defense takes as input the dataset and a *defense hyperparameter* and outputs a prediction. In randomized smoothing, the defense parameter corresponds to the noise variance, in outlier detection it corresponds to the threshold, and in the third described method, to the attack parameter that the defense uses to solve the inverse optimization problem. While in certain scenarios, knowledge of the optimal parameter might be available, it is usually a strong assumption, even more so when the attack adaptively changes its strategy.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

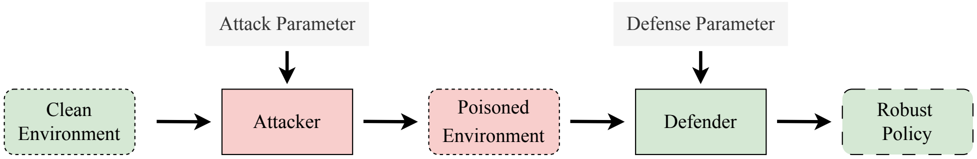


Figure 1: Illustration of our attack and defense models. The attacker observes the clean environment, chooses an attack parameter, and poisons the environment using its attack model. Subsequently, the defender observes the poisoned environment, chooses a defense parameter, and uses its defense model to compute a robust policy.

Motivated by such examples, we consider parametric defenses against parametric adaptive reward poisoning attacks in reinforcement learning and propose online learning strategies for defense parameter selection. The attacker poisons the reward function, according to the attack parameter, in order to optimize its utility, while the learner observes the poisoned reward and commits to a robust policy, to optimize its return with respect to the original reward. Figure 1 illustrates both the attack and defense models we use. We formulate this sequential interaction as a multi-agent learning problem, in both adversarial and stochastic environments, and propose specific online learning strategies for both cases, that yield sublinear regret bounds with respect to a defender that always picks the optimal defense parameter in hindsight. Furthermore, we show that other types of defense that only take into account partial information about the attack are hopeless against such adaptive attackers.

1.1 Our Contributions.

Our main contributions can be summarized as follows.

- We propose a game-theoretic framework for the problem of defense against adaptive reward poisoning attacks in RL and introduce a new notion of regret with respect to a defense that always selects the optimal parameter in hindsight. We argue that this notion is stronger than the standard notion of regret used in the online learning literature.
- Our attack model is stronger than previously considered, in that: i) it observes the environment; ii) at the end of each round, it observes the learner's policy and its strategy of selecting the defense parameter; iii) it has an unlimited budget; iv) it employs a no-regret learning strategy to select the attack parameter.
- For adversarial environments we propose Exp3-DARP (Exp3 for Defense Against Reward Poisoning), a defense algorithm that uses Exp3 as a parameter learning subroutine, and show that it incurs $\tilde{\Theta}(\sqrt{T})$ regret in T rounds. Furthermore, we show that, even when allowed to observe full information feedback, the learner cannot improve these bounds in adversarial environments.
- For stochastic environments we propose OMDUCB-DARP (Optimistic Mirror Descent with Upper Confidence Bounds-DARP), a novel online learning algorithm that achieves $\tilde{O}(T^{1/3})$, $\tilde{O}(T^{1/4})$ and $O(\log T)$ regret bounds on our notion of regret, depending on the rate of change in the attacker's strategy.
- Moreover, we prove that knowledge of the attack model and the poisoned reward is critical in achieving sublinear regret bounds under our notion of regret, by showing linear lower bounds for weaker types of defenses.
- Finally, we evaluate our methods in three synthetic environments and compare them to an agent that is oblivious to attacks, an agent that uses fixed defense, and a naive learner that only uses delayed feedback, and show that our methods are substantially superior to them.

1.2 Other Related Work

Our work is related to the following two strands of machine learning research.

Adversarial attacks and robustness. Arguably, the closest to this paper is the line of work on adversarial attacks and robustness against these attacks. In recent years, adversarial attacks have been extensively studied in machine learning (Szegedy et al., 2013; Biggio et al., 2013; Nguyen et al., 2015; Papernot et al., 2017; Biggio et al., 2012; Xiao et al., 2012; Li et al., 2016), including RL (Huang et al., 2017; Lin et al., 2017; Sun et al., 2020a,b; Ma et al., 2019; Rakhsha et al., 2021a; Everitt et al., 2017; Wang et al., 2020; Huang and Zhu, 2019; Rangi et al.). One of the most common types of adversarial attacks are test-time attacks that manipulate an already trained ML model or RL policy (Szegedy et al., 2013; Biggio et al., 2013; Nguyen et al., 2015; Papernot et al., 2017; Huang et al., 2017; Lin et al., 2017; Sun et al., 2020a). We focus on poisoning attacks, which instead manipulate the learning agent during training by poisoning its input.

Data-poisoning attacks in machine learning have drawn much attention over the last years (Biggio et al., 2012; Mei and Zhu, 2015; Xiao et al., 2015; Alfeld et al., 2016; Wang and Chaudhuri, 2018; Li et al., 2016). Reward poisoning attacks to RL have been previously considered by Zhang et al. (2020b); Banihashem et al. (2021); Rakhsha et al. (2021b), while general data-poisoning attacks that include transition poisoning have been considered by Rakhsha et al. (2021a); Zhang et al. (2021a); Rangi et al. (2022b). While these works primarily focus on studying various types of attacks in various reinforcement learning settings, we take the point of view of the defense, while at the same time preserving all the characteristics of a strong adaptive attack, that has full knowledge of the environment.

Furthermore, since our main focus in this work is on defenses, our work also relates to the rich literature on robustness against poisoning attacks in ML (Paudice et al., 2018; Zhang et al., 2018; Charikar et al., 2017; Diakonikolas et al., 2019), and more specifically in RL (Lykouris et al., 2021; Zhang et al., 2021a; Kumar et al., 2021; Wu et al., 2022; Rangi et al., 2022a). We differ from the latter works in that we take a meta-optimization perspective in which the learner utilizes a defense method tailored to a given attack model, and aims at learning the optimal parameters of the defense. Additionally, we also take a game-theoretic perspective on the problem, where the attacker is assigned a utility function, and can also learn its attack parameters. As is the case for test-time adversarial attacks, our work broadly relates to the literature that studies robustness to test-time attacks in RL (Pattanaik et al., 2017; Zhang et al., 2020a, 2021a).

Adversarial online learning and learning in games. The adversarial online learning literature is very rich, with several no-regret algorithms such as Multiplicative Weights (Littlestone and Warmuth, 1994; Freund and Schapire, 1997; Arora et al., 2012), Mirror Descent (Nemirovskij and Yudin, 1983), Follow The Regularized Leader (Kalai and Vempala, 2005) and Exp3 (Auer et al., 2002) for the bandit feedback case. All of these methods and their theoretical properties are well-understood (see Cesa-Bianchi and Lugosi (2006) for an excellent overview). However, the vanilla versions of such algorithms do not take into account the potential structure of the data coming from the adversary (Rakhlin et al., 2011). We model our problem as a multi-agent online learning problem. When there is more than one player playing a static game using the same family of algorithms, then it is possible to say something about the sequence of feedback vectors that each player observes, an idea explored by Rakhlin and Sridharan (2013a,b), where the family of Mirror Descent type algorithms is studied. In this case, it is shown that we can get faster convergence rates when all players are playing no-regret algorithms that satisfy some desired properties, from the $\tilde{\Theta}(\sqrt{T})$ in the fully adversarial case, to $\tilde{O}(T^{1/4})$ in (Syrgkanis et al., 2015), to $O(T^{1/6})$ in (Chen and Peng, 2020) for two-player games, to the near-optimal bounds of $O(\text{poly}(\log T))$ regret for the Optimistic Hedge algorithm in general-sum multi-player games (Daskalakis et al., 2021). However, the full utilization of such features is impossible in adversarial environments, which implies worst-case $O(\sqrt{T})$ bounds.

Time-varying two-player zero-sum games have been considered by Zhang et al. (2022). Unknown games with correlated payoff have been previously considered by Sessa et al. (2019), where they use learning methods that leverage the structure of Gaussian processes, to obtain worst-case $O(\sqrt{T})$ bounds, without any assumptions on the other players. We use the Exp3 procedure for adversarial environments, thus incurring worst-case bounds on the regret. However, utilizing the fact that the second player's strategy is changing smoothly, we manage to provide tighter bounds for the learner when the environment is stochastic, by introducing a learning algorithm that uses a rationale similar to Optimistic Mirror Descent (Chiang et al., 2012), tailored for time-varying stochastic games.

2 PROBLEM SETUP

In this section, we introduce the background of the problem and lay out the necessary definitions. Then we formalize the general sequential interaction between the attacker and defender that will be used throughout the paper.

2.1 Preliminaries

Let $M = (S, A, P, R, \gamma, \nu)$ be an MDP with state space S and action space A, reward function $R : S \times A \to [0, 1]$, transition model $P : S \times A \times S \to S$, where $P(s, a, s')$ denotes the probability of transitioning to state s' given that action a is taken when in state s, discount factor $\gamma \in [0, 1)$ and initial state distribution ν. We assume the state and action spaces to be finite and thus denote by $|S|$ and $|A|$ their cardinalities, respectively. We also denote by $\mathcal{R} = [0, 1]^{|S|\cdot|A|}$ the set of all possible reward vectors.

A stochastic policy π is a mapping $\pi : S \to \mathcal{P}(A)$, where $\mathcal{P}(A)$ denotes the set of all probability distributions over A. As usual, $\pi(a|s)$ denotes the probability of taking action a in state s. We denote by Π the set of all stochastic policies on M and by Π^{det} its subset of all deterministic policies. Given policy π, we define its score as $\rho(\pi, R) = (1 - \gamma)\mathbb{E}\left[\sum_{\tau=1}^{\infty} \gamma^{\tau-1} R(s_\tau, a_\tau)|\pi, \nu\right]$, where the expectation is with respect to the randomness of π and ν. Note that for any given $\pi \in \Pi$, we also have that $\rho(\pi, R) \in [0, 1]$.

Given an MDP $\overline{M} = (S, A, \overline{R}, \overline{P}, \gamma, \nu)$ with the aforementioned characteristics, a data-poisoning attack to offline RL is to be understood as a mapping of $\overline{M}$ to another MDP $\widehat{M} = (\widehat{S}, \widehat{A}, \widehat{R}, \widehat{P}, \gamma, \nu)$, with $S = \widehat{S}$ and $A = \widehat{A}$. In other words, the attack modifies either the reward function, the transition matrix, or both. Usually, the attack modifies $\overline{R}$ and/or $\overline{P}$ in order to enforce a target policy to the learning agent (Rakhsha et al., 2021a). The type of attack that only modifies the reward function, in which case we have $\overline{P} = \widehat{P}$, is called a *reward-poisoning attack*.

2.2 Interaction Between Attacker And Defender

We consider reward poisoning attacks in an offline learning setting. The attacker's goal is to deceive the learner into adopting a deterministic target policy $\pi_\dagger \in \Pi^{\text{det}}$ by poisoning the reward vector $\overline{R}$. We assume the attack is a function $\mathcal{A} : \mathcal{R} \times \mathcal{E} \to \mathbb{R}^{|S|\cdot|A|}$, parametrized by the finite set $\mathcal{E} = \{\epsilon_1, \ldots, \epsilon_E\} \subset \mathbb{R}^{n_1}$, for some $n_1, E \in \mathbb{N}$, where, given reward vector $\overline{R} \in \mathcal{R}$ and a parameter $\epsilon \in \mathcal{E}$, it outputs the poisoned reward vector $\widehat{R} = \mathcal{A}(\overline{R}, \epsilon)$.[1] We denote by $\mathfrak{A}$ the class of such parametric reward-poisoning attacks.

Similarly, a defense $\mathcal{D} \in \mathfrak{D}$ is a mapping $\mathcal{D} : \mathbb{R}^{|S|\cdot|A|} \times \Theta \times \mathfrak{A} \rightarrow \Pi$, parametrized by the finite set $\Theta = \{\theta_1, \ldots, \theta_D\} \subset \mathbb{R}^{n_2}$, for some $n_2, D \in \mathbb{N}$, such that, given attack model $\mathcal{A} \in \mathfrak{A}$ and parameter $\theta \in \Theta$, it takes as input the poisoned reward $\widehat{R} \in \mathbb{R}^{|S|\cdot|A|}$ and outputs a policy $\pi = \mathcal{D}(\widehat{R}, \theta, \mathcal{A})$, hoping that it performs well under the true reward $\overline{R}$. When $\mathcal{A}$ and $\mathcal{D}$ are fixed, we will abuse notation and write $\pi = \mathcal{D}(\widehat{R}, \theta)$, to avoid overloading.

Let $\mathcal{A} \in \mathfrak{A}$ and $\mathcal{D} \in \mathfrak{D}$ be given. We formulate the interaction between the learner and an adaptive attacker as a sequential game. Each round $t \geq 1$ is characterized by a new MDP $\overline{M}_t = (S, A, \overline{R}_t, P, \gamma, \nu)$, where everything is fixed except the reward vector chosen by the environment. At the beginning of round t, the attacker and defender select $\epsilon_t \in \mathcal{E}$ and $\theta_t \in \Theta$ according to their randomized strategies[2] $\phi_t^{\mathcal{A}}$ and $\phi_t^{\mathcal{D}}$, respectively. The attacker observes the true environment $\overline{M}_t$ and modifies its reward $\overline{R}_t$ into $\widehat{R}_t = \mathcal{A}(\overline{R}_t, \epsilon_t)$, thus aiming to enforce $\pi_\dagger$ to the learner.

Next, the learner observes the poisoned reward $\widehat{R}_t$ and uses its defense model $\mathcal{D}$ to compute policy $\mathcal{D}(\widehat{R}_t, \theta_t) = \pi_t$. The objective of the learner is to maximize the score $\rho(\pi_t, \overline{R}_t)$ of its policy with respect to the true reward $\overline{R}_t$. At the end of the round, the attacker observes θ_t, $\phi_t^{\mathcal{D}}$ and π_t and computes its own cost[3], which we denote by $\sigma(\pi_t, \widehat{R}_t)$. The learner, on the other hand, observes only[4] the score of policy π_t with respect to the true reward vector $\overline{R}_t$. Both players use the observed feedback to update their randomized strategies $\phi_t^{\mathcal{A}}$ and $\phi_t^{\mathcal{D}}$. Algorithm 1 illustrates the whole interaction.

We model the aforementioned interaction as a multi-agent learning problem, where the learner's goal is to maximize its cumulative utility over a fixed horizon T. We do not make any further assumptions on the attacker, except assuming that it learns the defense's behavior based on previous observations. We measure the learner's performance against a defender that always selects the parameter θ that maximizes its cumulative utility. To that end, we define regret as

$$Reg_{\mathcal{D}}(T) := \max_{\theta} \sum_{t=1}^{T} \rho(\mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon_t), \theta), \overline{R}_t) - \mathbb{E}\left[\sum_{t=1}^{T} \rho(\mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon_t), \theta_t), \overline{R}_t)\right], \quad (1)$$

where the expectation is with respect to any potential randomness coming from the sequence $\phi_1^{\mathcal{D}}, \ldots, \phi_T^{\mathcal{D}}$.

Algorithm 1 Interaction between attacker and defender.

```
Input: Attack model $\mathcal{A}$; defense model $\mathcal{D}$; attacker's strategy; defender's strategy.
for $t = 1, 2, 3, \ldots, T$ do:
    Attacker and defender simultaneously select their actions, $\epsilon_t$ and $\theta_t$, respectively, based on their strategies.
    $\overline{R}_t$ is chosen by the environment.
    $\widehat{R}_t = \mathcal{A}(\overline{R}_t, \epsilon_t)$ is revealed to the defender.
    $\pi_t = \mathcal{D}(\widehat{R}_t, \theta_t)$ is computed.
    Attacker observes $\pi_t$ and $\theta_t$ (also has access to the defender's strategy) and incurs utility $\sigma(\pi_t, \widehat{R}_t)$.
    Defender observes utility value $\rho(\pi_t, \overline{R}_t)$.
    Both players update their strategies.
end for
```

Remark 1. *Note that our problem can be formulated as a two-player general-sum time-varying game. Zero-sum time-varying games have been previously considered by Cardoso et al. (2019) and Zhang et al. (2022), where a unified notion of NE-regret (or Dynamic NE-regret) is used. However, this notion of regret is tailored for zero-sum games, and our setting is a general-sum one, thus we cannot apply it. Furthermore, in our setting, there is no interest in maximizing social welfare, since the players have a conflict of interest. Therefore, the metric we consider in our setting is the individual regret of the defender.*

3 LEARNING THE DEFENSE PARAMETER

In this section, we first introduce Exp3-DARP (*Exp3 for Defense Against Reward-Poisoning*), a defense algorithm that employs Exp3 as a parameter learning subroutine for adversarial environments. We show order-optimal bounds on its regret and prove that, even under full information feedback, the order of T cannot be improved due to the adversarial nature of rewards. Next, motivated by the rationale of OMD, we introduce OMDUCB-DARP (*Optimistic Mirror Descent with Upper Confidence Bounds-DARP*), a defense learning algorithm for stochastic environments, and prove tighter bounds on its regret. Pseudocodes can be found in Section A, while proofs of stated results are in Sections C and D of the appendix.

3.1 Exp3-DARP For Adversarial Environments

Assume that the sequence of reward functions $\overline{R}_1, \overline{R}_2, \ldots$ is adversarially chosen by the environment. Furthermore, assume that the attacker employs a no-regret[5] learning algorithm to update $\phi_t^{\mathcal{A}}$ at each round t, and uses some learning factor $\eta^{\mathcal{A}} \in (0, 1]$. Note that the attacker can pick from a

[1] Note that we do not use $\pi_\dagger$ as a parameter of $\mathcal{A}$, for brevity, since it is fixed throughout the paper.

[2] We will specify these strategies, which depend on the feedback the players observe, in the next section.

[3] See Section 5 for an instantiated attack cost.

[4] In Section 3.2 we will consider a stronger feedback model for the learner in order to achieve tighter bounds.

[5] We instantiate the regret of the attacker in Section 5.

wide range of online learning methods since its feedback model is very strong. In order to minimize its regret, the learner needs to learn the optimal defense parameter over time, that would be efficient against an attacker that adapts to its behavior.

We consider the bandit feedback case, where the learner only observes the score of the policy it commits to with respect to the true reward function. One might hope to utilize the potentially benign sequence of scores that depends on the attacker's strategy, and use refined methods of online learning in regularized games (e.g. Optimistic Mirror Descent) in order to incur better than worst-case regret bounds. However, as Theorem 2 shows, this is not possible, as long as the environment is adversarial. Thus, we will use the Exp3 algorithm of Auer et al. (2002), designed for adversarial environments with bandit feedback, as the parameter learning subroutine, to update the strategy of the learner.

We initialize $\phi_1^{\mathcal{D}}$ as the uniform distribution, fix step-size $\eta^{\mathcal{D}} \in (0, 1]$, and feed them both to Algorithm 1. Then, for every $t \geq 1$, the strategy $\phi_{t+1}^{\mathcal{D}}$ is updated using Exp3. We provide the update rules and the pseudocode of Exp3-DARP in the appendix for completion. We state the following result that gives order-optimal bounds for this setting.

Theorem 1. *(Bubeck and Cesa-Bianchi, 2012) Let $\mathcal{A} \in \mathfrak{A}$ and $\mathcal{D} \in \mathfrak{D}$. Moreover, given $T, D \in \mathbb{N}$, set $\eta^{\mathcal{D}} = \min\{1, \sqrt{(D \ln D)/((e-1)T)}\}$. Then, we have $Reg_{\mathcal{D}}(T) \leq O(\sqrt{TD \ln D})$. On the other hand, there exists an attack $\mathcal{A} \in \mathfrak{A}$ and a distribution β of rewards $\overline{R}_1, \dots, \overline{R}_T$, such that, for any defense $\mathcal{D} \in \mathfrak{D}$, the expected regret $\mathbb{E}_\beta Reg_{\mathcal{D}}(T)$ is at least $\Omega(\sqrt{TD})$.*

Remark 2. *In the case of compact action spaces, instead of Exp3, one can use the Adversarial Zooming algorithm of Podimata and Slivkins (2021) as a learning subroutine. We believe that, under the additional assumption of Lipschitz continuity of the defense parameters with respect to the learner's utilities, one would still be able to achieve sublinear regret bounds.*

Next, we answer the question: *Can we get better bounds (in terms of T) if the defense is made stronger and uses more information about the attack?* Unfortunately, the answer is no. One can always find a reward sequence that makes the regret at least $\Omega(T)$, irrespective of the learning method and the attack structure. That is what our next result shows.

Theorem 2. *Assume that, at the end of round t, the learner can observe the attacker's strategy and the true reward function $\overline{R}_t$. Then, for any defense $\mathcal{D} \in \mathcal{D}$ and any sequence $\phi_1^{\mathcal{D}}, \dots, \phi_T^{\mathcal{D}}$, there exists an attack $\mathcal{A} \in \mathfrak{A}$, such that we have $Reg_{\mathcal{D}}(T) = \Omega(\sqrt{T \log D})$.*

3.2 OMDUCB-DARP For Stochastic Environments

The adversarial nature of the environment implies limitations on the regret analysis. However, the environment can often be stochastic and thus, estimates of it are possible. Therefore, we now turn our focus to such environments.

We assume that the environment is stochastic, that is, we assume that the reward functions $\overline{R}_1, \overline{R}_2, \dots, \overline{R}_T$ are i.i.d. random vectors. This, in turn, implies that, for every $\theta \in \Theta$ and $\epsilon \in \mathcal{E}$, the scores $\rho(\mathcal{D}(\mathcal{A}(\overline{R}_1, \epsilon), \theta)\overline{R}_1)$, $\dots$ $\rho(\mathcal{D}(\mathcal{A}(\overline{R}_T, \epsilon), \theta), \overline{R}_T)$ are i.i.d. random variables with mean $\mathbb{E}\left[\rho(\mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon), \theta), \overline{R}_t)\right]$, for any $t \leq T$. Furthermore, the feedback model of the defense that we consider now is stronger than the one in the previous section. That is, at the end of each round $t \geq 1$, the learner observes the true reward function $\overline{R}_t$ and the attacker's strategy $\phi_t^{\mathcal{A}}$ at round t. The algorithm that we propose and the results of this section depend on this assumption. Now let us introduce the cost of the learner with respect to a particular pair (θ, ϵ) as $\omega(\epsilon, \theta, \overline{R}_t) := 1 - \rho(\mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon), \theta), \overline{R}_t)$ and let $\omega(\epsilon, \theta) := \mathbb{E}[\omega(\epsilon, \theta, \overline{R}_t)]$ denote its expected value. We will now consider the problem of cost minimization for the learner, which is equivalent to the problem of utility maximization when the costs are defined as above. We do this to avoid unnecessary complications in the analysis.

Let us define $\overline{G}_t = [\omega(\epsilon_i, \theta_j, \overline{R}_t)]_{i=1,j=1}^{E,D}$ to be the game matrix at time t and $G = [\omega(\epsilon_i, \theta_j)]_{i=1,j=1}^{E,D}$ to be the expected game matrix. Let $\mathcal{G}$ denote the distribution of the random matrix G_t with mean G. Our goal is to design a learning method that incurs sublinear expected regret, defined as

$$Reg_{\mathcal{D}}^*(T) = \mathbb{E}_{\overline{G}_t \sim \mathcal{G}, \epsilon_t \sim \phi_t^{\mathcal{A}}, \theta_t \sim \phi_t^{\mathcal{D}}} \sum_{t=1}^{T} \overline{G}_t[\epsilon_t, \theta_t] - \min_{\phi} \mathbb{E}_{\overline{G}_t \sim \mathcal{G}, \epsilon_t \sim \phi_t^{\mathcal{A}}, \theta \sim \phi} \sum_{t=1}^{T} \overline{G}_t[\epsilon_t, \theta] \,. \quad (2)$$

Note that the usual notion of regret in regularized games is the quantity above inside the expectation with respect to $\mathcal{G}$. However, this notion of regret is defined when the players are playing a fixed game. The game matrix in our setting changes over time. Thus, it is reasonable to take expectations with respect to the randomness of the environment.

We now introduce OMDUCB-DARP, a novel learning algorithm that is designed to incur tighter than worst-case bounds on the expected value of our regret. To that end, let $f : [0,1]^D \to \mathbb{R}$ be a 1-strongly convex function with respect to $\|\cdot\|_1$ and let $\mathcal{B}_f(\cdot, \cdot)$ denote the Bregman divergence with respect to f. Motivated by the rationale of Optimistic Mirror Descent, tailored to yield better bounds in regularized games, and by the Upper Confidence Bound (UCB) paradigm, based on the concentration of the costs around their mean, we introduce the following update for the learner, for every $t \geq 1$:

$$\widetilde{\phi}_{t+1}^{\mathcal{D}} = \arg\min_{\phi} \eta \langle \phi, \widehat{G}_t^T \phi_t^{\mathcal{A}} \rangle + \mathcal{B}_f(\phi, \widetilde{\phi}_t^{\mathcal{D}}) \,, \quad (3)$$

$$\phi_{t+1}^{\mathcal{D}} = \arg\min_{\phi} \eta \langle \phi, \widetilde{G}_{t+1}^T \phi_t^{\mathcal{A}} \rangle + \mathcal{B}_f(\phi, \widetilde{\phi}_{t+1}^{\mathcal{D}}) \,, \quad (4)$$

where $\widetilde{G}_t$ is an $E \times D$-dimensional matrix with entries

$$\widetilde{G}_t[\epsilon, \theta] = \frac{1}{t-1} \sum_{k=1}^{t-1} \overline{G}_k[\epsilon, \theta] ,$$

and

$$\widehat{G}_t[\epsilon, \theta] := \widetilde{G}_t[\epsilon, \theta] + \sqrt{\frac{\log(\pi^2 ED/(3\delta))}{2(t-1)}} \tag{5}$$

denotes the optimistic estimate[6] on the mean $G[\epsilon, \theta]$, for all $\epsilon \in \mathcal{E}$ and $\theta \in \Theta$, and some $\delta \in (0,1)$. As usual, $\langle \cdot, \cdot \rangle$ denotes the inner product. We initialize $\widetilde{\phi}_1^{\mathcal{D}} = \phi_1^{\mathcal{D}} = \arg\min_\phi f(\phi)$. Here $\widetilde{\phi}_t^{\mathcal{D}}$ is an auxiliary update for $\phi_t^{\mathcal{D}}$ and the algorithm selects θ_t according to $\phi_t^{\mathcal{D}}$. Note that, since the learner observes the reward function $\overline{R}_t$ and the attacker's strategy $\phi_t^{\mathcal{A}}$ at the end of round t, the computation of $\phi_{t+1}^{\mathcal{D}}$ is possible. The pseudocode is given in the appendix.

Note that the proxy vector $\widehat{G}_t^T \phi_t^{\mathcal{A}}$ used by OMDUCB-DARP in the intermediate update $\widetilde{\phi}_{t+1}^{\mathcal{D}}$ depends on the UCB game matrix $\widehat{G}_t$, i.e. the $E \times D$-dimensional matrix composed of $\widehat{G}_t[\epsilon, \theta]$, for every $\epsilon \in \mathcal{E}$ and $\theta \in \Theta$, and the attacker's strategy. Then, the actual update $\phi_{t+1}^{\mathcal{D}}$ uses sample averages as proxies in order to compute the strategy of selection for the next round. This allows us to make use of the stochasticity of the environment and apply concentration bounds for the estimates, in order to shift the weight of the regret to quantities that we can control.

Moreover, note that the UCB term on the right-hand side of (5) depends on the actual round t, which is different from its traditional usage in bandit settings, where it depends on the number of times an action has previously been selected. In our case though, the feedback model we consider allows the algorithm to compute these terms in every round, hence the explicit dependence on t. It is important to emphasize that our feedback model is stronger than the usual *full information* feedback in static games, where only $\overline{G}_t^\top \phi_t^{\mathcal{A}}$ is observed at the end of round t. In our case, more information is needed, in order to account for the time-varying nature of the game. Now we state the main result of this section.

Theorem 3. *Let $\mathcal{A} \in \mathfrak{A}$, $\mathcal{D} \in \mathfrak{D}$ and $T \in \mathbb{N}$. Assume $\eta^{\mathcal{D}} \leq \eta^{\mathcal{A}} \in (0,1]$. Then, for any $\delta \in (0,1)$, the above algorithm incurs expected regret*

$$Reg_{\mathcal{D}}^*(T) \leq \eta^{\mathcal{A}} + \eta^{\mathcal{A}} \sum_{t=2}^{T} \left\| \phi_t^{\mathcal{A}} - \phi_{t-1}^{\mathcal{A}} \right\|_1^2 + \frac{1}{\eta^{\mathcal{D}}} f_{max}$$
$$+ \frac{\eta^{\mathcal{A}}}{2} \log(\pi^2 ED/(3\delta)) \log T ,$$

with probability at least $1 - \delta$, where $f_{max} = \max_\phi f(\phi)$.

Note that the regret bounds depend on the magnitude of change in the attacker's strategy through time. Next, we instantiate these bounds for different online learning methods.

Corollary 1. *Under the conditions of Theorem 3, we have*

- *$Reg_{\mathcal{D}}^*(T) \leq O(\log T)$, if the attack parameter is fixed.*
- *$Reg_{\mathcal{D}}^*(T) \leq \tilde{O}(T^{1/4})$, if the attacker plays Hedge.*
- *$Reg_{\mathcal{D}}^*(T) \leq \tilde{O}(T^{1/3})$, if the attacker plays an online learning method that satisfies the RVU property (Syrgkanis et al., 2015).*

Remark 3. *Under oblivious adversaries, that is, when the sequence of rewards is predictable (up to noise errors), we obtain order-optimal bounds, which degrade as the attacker's strategy becomes 'stronger' (in the sense of using better learning strategies). Note that OMDUCB-DARP's regret analysis is such that one can utilize the structure of the attacker's strategy. If we know nothing of the latter, then we can directly deploy Exp3-DARP, in which case we obtain the worst case $O(\sqrt{T})$ bounds.*

We have so far given sublinear regret bounds for our proposed methods, with respect to an optimal defense that knows the attack parameters beforehand. The defense is here assumed to utilize the poisoned reward and its knowledge of the attack structure (recall that one of the arguments of the defense function is the attack). However, it is not clear if such information is necessary for the learner to perform well. One can always choose to learn the true reward function solely based on delayed observations of the scores (in the bandit feedback case) or the true reward function (in the full information feedback case) and completely bypass the attack. Unfortunately, this is not optimal in our setting. Our results in the next section show that any defense model that does not utilize the poisoned reward and attack structure is hopeless when measured against our notion of the benchmark.

4 CHARACTERIZATION OF WEAK DEFENSES

In this section, we introduce different types of defense classes based on how much information they use about the attack and show that full utilization of the attack structure is, in fact, necessary in order to incur sublinear regret. All proofs of stated results can be found in the appendix.

First, let us introduce some additional notation that will help us quantify how well a given defense is doing compared to the optimal policy under the true MDP. Let $\mathcal{A} \in \mathfrak{A}$ and $\mathcal{D} \in \mathfrak{D}$ be given. For every $t \geq 1$, let π_t^* denote the optimal policy under the true MDP $\overline{M}_t$ and let $\pi_t^{\mathcal{D}} = \mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon_t), \theta_t)$ denote the policy that the defense commits to in round t. Further, let $\theta_{max} = \arg\max_\theta \sum_{t=1}^{T} \rho(\mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon_t), \theta), \overline{R}_t)$

[6] Note that the learner can compute both $\widetilde{G}_t[\epsilon, \theta]$ and $\widehat{G}_t[\epsilon, \theta]$ at the end of round t since it observes $\overline{R}_t$ and knows $\mathcal{A}$. We initialize their values to 0, for all $\epsilon \in \mathcal{E}$ and $\theta \in \Theta$.

	Regret*		Optimality gap	
	No attack	Adaptive	No attack	Adaptive
$\mathcal{D}_\emptyset$	No Reg.	Linear	0	$\Delta^{\pi_\dagger}$
$\mathcal{D}_1$	Linear	Linear	$\Omega(T)$	$\Delta^{opt} + \Omega(T)$
$\mathcal{D}_2$	No Reg.	Linear	0	$\Delta^{opt} + \Omega(T)$
Exp3	$\tilde{\Theta}(\sqrt{T})$	$\tilde{\Theta}(\sqrt{T})$	$\tilde{\Theta}(\sqrt{T})$	$\Delta^{opt} + \tilde{\Theta}(\sqrt{T})$
OMDUCB	$\tilde{O}(T^{1/3})$	$\tilde{O}(T^{1/3})$	$\tilde{O}(T^{1/3})$	$\Delta^{opt} + \tilde{O}(T^{1/3})$
$\mathcal{D}^{opt}$	No reg	No reg	0	Δ^{opt}
Oracle	NA	NA	0	0

Table 1: Summary of our results both in terms of regret and optimality gap. The Oracle always plays the optimal policy under the true reward. $\mathcal{D}_\emptyset$ represents a naive defense that is oblivious to the attack; $\mathcal{D}_1$ denotes a defense that does not utilize the poisoned reward and $\mathcal{D}_2$ is a defense that does not know the attack model. The detailed definitions are given in Section 4. Moreover, we only give the $\tilde{O}(T^{1/3})$ bounds for OMDUCB-DARP, omitting the other bounds, for brevity.
*The notion of regret corresponding to $\mathcal{D}_\emptyset$, $\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}^{opt}$ is the one given in (6); the regret of Exp3 is the one given in (1), while the regret of OMDUCB corresponds to the expected regret as defined in (2).

and let us denote by $\pi_t^{opt} = \mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon_t), \theta_{max})$ the policy that the optimal defense commits to at time t. We denote by $\Delta_t^{\mathcal{D}} = \rho(\pi_t^*, \overline{R}_t) - \mathbb{E}\rho(\pi_t^{\mathcal{D}}, \overline{R}_t)$ the gap between the optimal policy π_t^* and the one selected by the defense[7] at time t. Moreover, let $\Delta^{\mathcal{D}} = \sum_{t \le T} \Delta_t^{\mathcal{D}}$ denote the cumulative gap of the defense $\mathcal{D}$ in T rounds. Δ_t^{opt} and Δ^{opt} are analogously defined for the optimal defense that commits to policy π_t^{opt} in round t. Furthermore $\Delta_t^{\pi_\dagger}$ and $\Delta^{\pi_\dagger}$ correspond to the target policy $\pi_\dagger$. Note that, in general, we have $\Delta^{opt} \le \Delta^{\pi_\dagger}$. Banihashem et al. (2021) show that Δ^{opt} can be almost two times[8] smaller than $\Delta^{\pi_\dagger}$ in some examples.

We will use the notion of the gap to compare different types of defense. Table 1 depicts the contrast in performance (both in regret and in optimality gap) between different types of partial information defenses, which we will now introduce, and our defense.

Note that the regret defined in (1) changes when considering a defense model different from $\mathcal{D}$ (e.g. a non-parametric defense, or one that does not depend on the poisoned reward). Thus, for a given defense $\mathcal{D}'$, we define the *relative regret* of $\mathcal{D}'$ with respect to $\mathcal{D}$ as

$$Reg_{\mathcal{D}'}^{\mathcal{D}}(T) = \max_{\theta} \sum_{t=1}^{T} \rho(\mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon_t), \theta), \overline{R}_t)$$

$$- \mathbb{E} \sum_{t=1}^{T} \rho(\pi_t^{\mathcal{D}'}, \overline{R}_t) \, . \tag{6}$$

Note that the relative regret of a defense $\mathcal{D}$ with respect to itself is the one given in (1).

No Defense. Let $\mathcal{D}_\emptyset$ denote a learning agent that does not employ a defense strategy against attack model $\mathcal{A}$, but instead, commits to an optimal policy on the poisoned reward function, believing that there is no attack present, and therefore assumes $\widehat{R}_t$ to be the true reward, for every round t. Under no attack, we have $\widehat{R}_t = \overline{R}_t$, and thus $\pi_t^{\mathcal{D}_\emptyset} = \mathcal{D}(\overline{R}_t, \overline{\theta}_t) = \pi_t^{opt} = \pi_t^*$, for some $\overline{\theta}_t \in \Theta$. This is because, under no attack, the optimal defense would always select the parameter θ that represents no defense.[9] Thus, $\mathcal{D}_\emptyset$ incurs no regret in this case and $\Delta^{\mathcal{D}_\emptyset} = 0$. On the other hand, if there is an attack present in round t, and thus $\overline{R}_t \neq \widehat{R}_t$, the learner that employs no defense commits to policy $\pi_\dagger$, since $\pi_\dagger$ is optimal under $\widehat{R}_t$. Thus, we have $\Delta^{\mathcal{D}_\emptyset} = \sum_{t \le T} \Delta^{\pi_\dagger} = T\Delta^{\pi_\dagger}$. In this case, the optimal defense would incur a gap Δ^{opt}. The regret incurred by $\mathcal{D}_\emptyset$ with respect to an optimal defense that employs $\mathcal{D}$ with optimal parameter θ_{max} is $\sum_{t=1}^{T} \rho(\pi_t^{opt}, \overline{R}_t) - \rho(\pi_\dagger, \overline{R}_t)$, and there obviously exists a sequence of reward functions $\overline{R}_1, \ldots, \overline{R}_T$, under which $\pi_\dagger$ is never optimal. Thus, we obtain $Reg_{\mathcal{D}_\emptyset}^{\mathcal{D}}(T) = \Omega(T)$.

First Type Defense. Now let $\mathfrak{D}_1$ denote the class of defenses that bypass $\widehat{R}_t$ and only learn from past observations of reward functions. Formally, $\mathcal{D}_1 \in \mathfrak{D}_1$ is characterized by a sequence of mappings $\mathcal{D}_1^t : \mathcal{R}^{t-1} \rightarrow \Pi$, for all $t \ge 1$, so that

$$\pi_t = \mathcal{D}_1^t(\overline{R}_1, \ldots, \overline{R}_{t-1}).$$

When no attack is present, we have that $\widehat{R}_t = \overline{R}_t$, for all $t \ge 1$. In that case, the optimal defense would commit to policy $\pi_t^{opt} = \mathcal{D}(\overline{R}_t, \overline{\theta}_t) = \pi_t^*$. Any given defense $\mathcal{D}_1 \in \mathfrak{D}_1$, on the other hand, would have to compete against π_t^* based only on past observations. Whatever learning method $\mathcal{D}_1$ employs, it will only minimize the cumulative regret with respect to the optimal policy in hindsight but is not guaranteed to minimize our regret. This shows that our notion is stronger than the usual one. Our next result shows that $\mathcal{D}_1$ is hopeless against an optimal defender that employs $\mathcal{D}$, both under attack, and when no attack is present, even if it gets to observe the true reward function $\overline{R}_t$ at the end of round t.

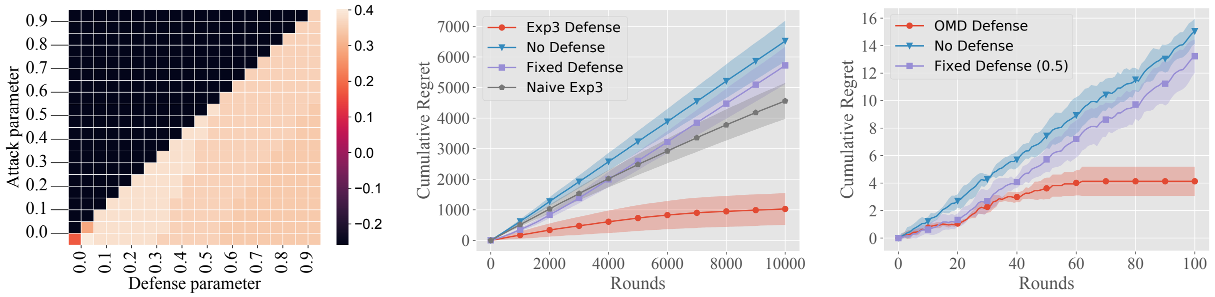
Proposition 1. *For every $\mathcal{D}_1 \in \mathfrak{D}_1$, there exists a sequence of MDPs $\overline{M}_1, \ldots, \overline{M}_T$ and a defense $\mathcal{D} \in \mathfrak{D}$, under which $\mathcal{D}_1$ incurs linear relative regret with respect to $\mathcal{D}$, under no attack. Moreover, for every $\mathcal{D}_1 \in \mathfrak{D}_1$, there exists a sequence of MDPs $\overline{M}_1, \ldots, \overline{M}_T$, a defense $\mathcal{D} \in \mathfrak{D}$, and an attack $\mathcal{A} \in \mathfrak{A}$, under which $\mathcal{D}_1$ incurs linear relative regret with respect to $\mathcal{D}$. For such a defense, we have $\Delta^{\mathcal{D}_1} \ge \Delta^{opt} + C_1 T$, for some $C_1 > 0$.*

Second Type Defense. Next, we consider $\mathfrak{D}_2$, a class of defenses that utilize the poisoned reward function and the

[7] The expectation in the score of $\pi_t^{\mathcal{D}}$ is with respect to any potential randomness in the parameter selection strategy of the defense.

[8] Explicit relation between them is less interpretable. We refer the reader to (Banihashem et al., 2021).

[9] We assume there exists $\theta \in \Theta$, such that $\mathcal{D}(\overline{R}_t, \theta) = \pi_t^*$.



(a) One-shot interaction

(b) Regrets in the adversarial setting.

(c) Regrets in the stochastic setting.

Figure 2: Results for the Navigation environment. (a) One shot attacker-defender interaction, with $\rho(\pi_\dagger, \overline{R}) = -0.26$ and $\rho(\pi^*, \overline{R}) = 0.45$. Note that the defense is able to recover a near-optimal policy in almost all scenarios. (b) Comparison of actual regrets in the adversarial setting for the given methods averaged over 5 runs, where we choose Fixed Defense parameter as 0.5, and No Defense learns $\pi_\dagger$ directly. (c) Comparison of actual regrets in the stochastic setting for the given methods averaged over 5 runs.

history of utilities, but are not aware of the attack model being employed by the attacker. A given $\mathcal{D}_2 \in \mathfrak{D}_2$ is characterized by the sequence $\mathcal{D}_2^t : (\mathbb{R}^{|S|\cdot|A|})^t \times \mathcal{R}^{t-1} \to \Pi$, for all $t \geq 1$, so that $\pi_t = \mathcal{D}_2^t(\widehat{R}_1, \dots, \widehat{R}_t, \overline{R}_1, \dots, \overline{R}_{t-1})$.

Note that the class $\mathfrak{D}_2$ is clearly a stronger defense class than $\mathfrak{D}_1$ since the poisoned reward is used as an argument. Under no attack, we similarly have $\pi_t^{opt} = \pi_t^*$, and, due to its observation of the true reward beforehand, there exists $\mathcal{D}_2 \in \mathfrak{D}_2$ that suffers no regret with respect to π^{opt}.

However, suppose the attack model that the attacker employs is not fixed,[10] i.e. $\mathcal{A}_t \in \{\mathcal{A}', \mathcal{A}''\} \subset \mathfrak{A}$. Furthermore, assume that the attacker adversarially selects which attack model to employ in round t. The $\mathcal{D}_2$ defense would not be able to tell whether $\widehat{R}_t$ is $\mathcal{A}(\overline{R}_t, \epsilon)$, for some $\epsilon \in \mathcal{E}$, or $\mathcal{A}'(\overline{R}_t, \epsilon')$, for some $\epsilon' \in \mathcal{E}$. Our next result shows that this can jeopardize the performance of the defense.

Proposition 2. *Assume that the attacker can change its attack model over time. Then, for every $\mathcal{D}_2 \in \mathfrak{D}_2$, there exists a sequence of MDPs $\overline{M}_1, \dots, \overline{M}_T$, attack models $\mathcal{A}_1, \dots, \mathcal{A}_T \in \{\mathcal{A}', \mathcal{A}''\} \in \mathfrak{A}$, and defense $\mathcal{D} \in \mathfrak{D}$, such that $\mathcal{D}_2$ incurs linear relative regret with respect to $\mathcal{D}$. Moreover, we have $\Delta^{\mathcal{D}_2} \geq \Delta^{opt} + C_2 T$, for some positive C_2.*

The previous results show that the utilization of the attack model, besides the poisoned reward and the history of utilities, is necessary if there is any hope of sublinear bounds on our notion of regret. We conclude that defense types $\mathcal{D}_\emptyset$, $\mathcal{D}_1 \in \mathfrak{D}_1$ and $\mathcal{D}_2 \in \mathfrak{D}_2$ are inefficient against strong adaptive attackers that utilize full information about the defense and modify their behavior accordingly.

5 EXPERIMENTS

We conduct experiments on three different environments, a bandit environment (Banihashem et al., 2021), a Navigation environment (Rakhsha et al., 2021a; Banihashem et al., 2021), and a Grid World environment (Ma et al., 2019; Banihashem et al., 2021). However, in this section, we only discuss the results on the Navigation environment. We refer the reader to Section B of the Appendix for the rest of the results.

The Attack and Defense Models. Let us instantiate an attack model following Rakhsha et al. (2021a). Given true reward vector $\overline{R}$, assume the attacker solves the following constrained optimization problem in order to obtain $\widehat{R}$, for a given $\epsilon \in \mathcal{E}$:

$$\mathcal{A}(\overline{R}, \epsilon) = \min_R \left\| R - \overline{R} \right\|_2 \quad \text{(P1)}$$
$$\text{s.t. } \rho(\pi_\dagger, R) \geq \rho(\pi_\dagger^{\{s,a\}}, R) + \epsilon, \ \forall s, a \neq \pi(s) \ ,$$

where $\pi^{\{s,a\}}$ denotes a policy that chooses $a \neq \pi(s)$ in state s and $\pi(\tilde{s})$ for all $\tilde{s} \neq s$. As shown by Rakhsha et al. (2021a), this optimization problem is feasible for ergodic MDPs and has a unique optimal solution. Given defense model $\mathcal{D}$ and parameter $\theta \in \Theta$, the learner commits to policy $\mathcal{D}(\widehat{R}, \theta)$. The cost of the attacker with respect to poisoned reward $\widehat{R}$ and policy π is defined as

$$\sigma(\pi, \widehat{R}) := C_{norm} \cdot \left(\|\pi - \pi_\dagger\|_1 + \lambda \left\| \widehat{R} - \overline{R} \right\|_2 \right) \ ,$$

where C_{norm} is a normalization factor which is chosen as the inverse of the maximum value that the quantity inside the brackets can have, for a given value of λ, where $\lambda \in [0, 1]$ is a regularization parameter. Lower values of λ imply a lower sensitivity of the attacker to the performed reward poisoning. Given this attack utility, and assuming that the attacker employs a no-regret algorithm, we define the regret of the attacker as follows.

$$Reg_{\mathcal{A}}(T) := \sum_{t=1}^{T} \sigma(\mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon_t), \theta_t), \mathcal{A}(\overline{R}_t, \epsilon_t))$$
$$- \min_\epsilon \sum_{t=1}^{T} \sigma(\mathcal{D}(\mathcal{A}(\overline{R}_t, \epsilon), \theta_t), \mathcal{A}(\overline{R}_t, \epsilon)) \ .$$

[10] Note that this does not violate our notion of regret, since the attack model is given as input to the defense, and thus, the optimization is only over the parameters. See Section 2.

On the other hand, similar to the defense strategy in (Banihashem et al., 2021), we instantiate the defense model as follows. Given poisoned reward vector $\widehat{R}$, the estimated reward with respect to $\theta \in \Theta$ is the solution of the following optimization problem:

$$\mathcal{D}(\widehat{R}, \theta) = \max_{\pi} \min_{R \in \mathcal{R}} \rho(\pi, R)$$
$$\text{s.t } \widehat{R} = \mathcal{A}(R, \theta) \ . \qquad \text{(P2)}$$

The optimal defense here solves (P2) with parameter $\theta = \epsilon$, where ϵ is such that $\widehat{R} = \mathcal{A}(\overline{R}, \epsilon)$, given that it knows the attack parameter (Banihashem et al., 2021). Thus, in this case, $\mathcal{E} = \Theta$. Note that the defender's optimization problem depends on the attack model $\mathcal{A}$. Assuming that the learner knows the attacker's chosen parameter, it is shown by Banihashem et al. (2021) that a closed-form solution of (P2) exists.

Setup and Results. In this attack-defense framework, the parameter set $\mathcal{E} = \Theta = \{0, 0.05, 0.1, 0.15, \dots, 0.95\}$ is the same for both players. Given the attack parameter ϵ, the optimal defense parameter is again ϵ since knowledge of the parameter, in this case, allows the defender to solve the inverse problem (P2).

In order to evaluate our online learning methods, we need to generate well-defined reward vectors that are meaningful for a given environment. In the adversarial setting, we generate the rewards in the following way. First, we compute the optimal policy π^* with respect to the clean environment using an MDP solver (in our case, Value Iteration). Next, we compute the set of the neighboring policies of π^*, i.e. the set $\mathcal{N}(\pi^*) = \{(\pi^*)^{\{s,a\}}\}_{s \in S, a \neq \pi^*(s)}$. Then, for each policy $\pi \in \mathcal{N}(\pi^*)$, we solve (P1) in order to generate a corresponding poisoned reward under which π is optimal. We use this set of reward vectors for the adversarial setting.

The Navigation environment (Rakhsha et al., 2021a; Banihashem et al., 2021), illustrated in Figure 3b (Section B of the Appendix), has 9 states and 2 actions. In this example we have $\overline{R}(s_0, \cdot) = \overline{R}(s_1, \cdot) = \overline{R}(s_2, \cdot) = \overline{R}(s_3, \cdot) = -2.5$, $\overline{R}(s_4, \cdot) = \overline{R}(s_5, \cdot) = 1$ and $\overline{R}(s_6, \cdot) = \overline{R}(s_7, \cdot) = \overline{R}(s_8, \cdot) = 0$. Also, in order to ensure ergodicity, we let the next state be sampled uniformly at random with probability 0.1 and such that follows the environment dynamics with probability 0.9. The discount factor is $\gamma = 0.99$ and the initial state is fixed as s_0. Figure 2a illustrates the one-shot interaction between the attacker and defender in the Navigation environment, that is, the score $\rho(\mathcal{D}(\mathcal{A}(\overline{R}, \epsilon), \theta), \overline{R})$ with respect to all possible (ϵ, θ).

The scores of target and optimal policies are $\rho(\pi_\dagger, \overline{R}) = -0.26$ and $\rho(\pi^*, \overline{R}) = 0.45$. Note that choosing $\theta \geq \epsilon$ is not the optimal defense, since overestimating the attack is suboptimal, as shown in Figure 2a. We set $\lambda = 0.01$ for the attack. We compare Exp3-DARP against Naive Exp3, No Defense, and Fixed Defense. Naive Exp3 is a defense of the first type, using only delayed feedback. Note that, if policies are seen as actions, then we have a total of $2^9 = 512$ actions. Fixed Defense always picks $\theta_t = 0.5$ (0.5 is selected as the mean value in Θ). As can be seen in Figure 2b the regret of No Defense and Fixed Defense is linear, but that of Naive Exp3 can be seen to show a sublinear tendency, only very late, due to the high number of actions. On the other hand, Exp3-DARP only has to find the best among 20 actions, and thus convergence happens much faster. This example clearly illustrates the need for defense in complex environments, when utilizing the defense is much more efficient than only observing delayed feedback.

We also run experiments on the stochastic setting, where we showcase the performances of OMDUCB-DARP, No Defense, and Fixed Defense. The results are shown in Figure 2c. We sample the reward vectors from the uniform distribution. Furthermore, we use the negative entropy function as our regularizer f, thus making the update rules of OMDUCB-DARP a variation of Multiplicative Weights, with proxies as defined in Section 3. The attack parameters are chosen similarly to the adversarial setting. As can be seen, the convergence of OMDUCB-DARP happens in less than 100 rounds. We argue that this is attributed to the full information feedback that is available to the learner and the small number of actions.

6 CONCLUSIONS AND FUTURE DIRECTIONS

We proposed a general game-theoretic framework for the problem of defense against adaptive reward-poisoning attacks in reinforcement learning. Our learning algorithms Exp3-DARP, and OMDUCB-DARP, designed for adversarial and stochastic environments, respectively, incur sublinear in time bounds on our notion of regret. We proved tight bounds for the adversarial case and improved the order of the time horizon in the stochastic case. Finally, we analyzed various defense strategies that utilize only partial information about the attack and showed that knowledge about its structure is necessary to obtain an efficient defense. This work initiates the discussion on the game-theoretic aspects of defenses against poisoning attacks in RL. We deem it as a starting point in the research direction that explores such aspects. While we are able to provide tighter than worst-case bounds on the expected regret for stochastic environments, a natural future direction is to investigate whether these bounds are order-optimal, for the given learning algorithms that the attack deploys. Furthermore, another natural direction is to consider the Stackelberg game formulation, where the sequential interaction between the attacker and defender is turn-based. Here, the defender would lead by choosing its parameter first, after which the attack would pick its parameter as the best response to the defense parameter, thus introducing a stronger attack that observes the defense before acting itself.

Acknowledgements

The authors thank anonymous reviewers for their valuable suggestions and comments. This research was, in part, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 467367360.

References

Scott Alfeld, Xiaojin Zhu, and Paul Barford. Data poisoning attacks against autoregressive models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002.

Kiarash Banihashem, Adish Singla, and Goran Radanovic. Defense against reward poisoning attacks in reinforcement learning. *arXiv preprint arXiv:2102.05776*, 2021.

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Adrian Rivera Cardoso, Jacob Abernethy, He Wang, and Huan Xu. Competing against equilibria in zero-sum games with evolving payoffs. *arXiv preprint arXiv:1907.07723*, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. *Advances in Neural Information Processing Systems*, 33:18990–18999, 2020.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, pages 6–1. JMLR Workshop and Conference Proceedings, 2012.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. *Advances in Neural Information Processing Systems*, 34, 2021.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606. PMLR, 2019.

Tom Everitt, Victoria Krakovna, Laurent Orseau, Marcus Hutter, and Shane Legg. Reinforcement learning with a corrupted reward channel. *arXiv preprint arXiv:1705.08417*, 2017.

Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55 (1):119–139, 1997.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *CoRR*, abs/1702.02284, 2017.

Yunhan Huang and Quanyan Zhu. Deceptive reinforcement learning under adversarial manipulations on cost signals. In *International Conference on Decision and Game Theory for Security*, pages 217–237. Springer, 2019.

Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

Aounon Kumar, Alexander Levine, and Soheil Feizi. Policy smoothing for provably robust reinforcement learning. *arXiv preprint arXiv:2106.11420*, 2021.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. Data poisoning attacks on factorization-based collaborative filtering. *Advances in neural information processing systems*, 29, 2016.

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 International Joint Conference on Artificial Intelligence*, pages 3756–3762, 2017.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108 (2):212–261, 1994.

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

Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu. Policy poisoning in batch reinforcement learning and control. In *NeurIPS*, pages 14543–14553, 2019.

Shike Mei and Xiaojin Zhu. Using machine teaching to identify optimal training-set attacks on machine learners. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, pages 427–436, 2015.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632*, 2017.

Andrea Paudice, Luis Muñoz-González, Andras Gyorgy, and Emil C Lupu. Detection of adversarial training examples in poisoning attacks through anomaly detection. *arXiv preprint arXiv:1802.03041*, 2018.

Chara Podimata and Alex Slivkins. Adaptive discretization for adversarial lipschitz bandits. In *Conference on Learning Theory*, pages 3788–3805. PMLR, 2021.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019. PMLR, 2013a.

Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning: Stochastic and constrained adversaries. *arXiv preprint arXiv:1104.5070*, 2011.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013b.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching in reinforcement learning via environment poisoning attacks. *Journal of Machine Learning Research*, 22(210):1–45, 2021a.

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

Anshuka Rangi, Haifeng Xu, Long Tran-Thanh, and Massimo Franceschetti. Understanding the limits of poisoning attacks in episodic reinforcement learning.

Anshuka Rangi, Long Tran-Thanh, Haifeng Xu, and Massimo Franceschetti. Saving stochastic bandits from poisoning attacks via limited data verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8054–8061, 2022a.

Anshuka Rangi, Haifeng Xu, Long Tran-Thanh, and Massimo Franceschetti. Understanding the limits of poisoning attacks in episodic reinforcement learning. *arXiv preprint arXiv:2208.13663*, 2022b.

Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pages 8230–8241. PMLR, 2020.

Pier Giuseppe Sessa, Ilija Bogunovic, Maryam Kamgarpour, and Andreas Krause. No-regret learning in unknown games with correlated payoffs. *Advances in Neural Information Processing Systems*, 32, 2019.

Jianwen Sun, Tianwei Zhang, Xiaofei Xie, Lei Ma, Yan Zheng, Kangjie Chen, and Yang Liu. Stealthy and efficient adversarial attacks against deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5883–5891, 2020a.

Yanchao Sun, Da Huo, and Furong Huang. Vulnerability-aware poisoning mechanism for online rl with unknown dynamics. In *International Conference on Learning Representations*, 2020b.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. *Advances in Neural Information Processing Systems*, 28, 2015.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Jingkang Wang, Yang Liu, and Bo Li. Reinforcement learning with perturbed rewards. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6202–6209, 2020.

Yizhen Wang and Kamalika Chaudhuri. Data poisoning attacks against online learning. *arXiv preprint arXiv:1808.08994*, 2018.

Fan Wu, Linyi Li, Chejian Xu, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li. Copa: Certifying robust policies for offline reinforcement learning against poisoning attacks. *arXiv preprint arXiv:2203.08398*, 2022.

Han Xiao, Huang Xiao, and Claudia Eckert. Adversarial label flips attack on support vector machines. In *ECAI*, pages 870–875, 2012.

Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *international conference on machine learning*, pages 1689–1698. PMLR, 2015.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33:21024–21037, 2020a.

Huan Zhang, Hongge Chen, Duane Boning, and Cho-Jui Hsieh. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*, 2021a.

Mengxiao Zhang, Peng Zhao, Haipeng Luo, and Zhi-Hua Zhou. No-regret learning in time-varying zero-sum games. *arXiv preprint arXiv:2201.12736*, 2022.

Xuezhou Zhang, Xiaojin Zhu, and Stephen Wright. Training set debugging using trusted items. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *International Conference on Machine Learning*, pages 11225–11234. PMLR, 2020b.

Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Robust policy gradient against strong data corruption. In *International Conference on Machine Learning*, pages 12391–12401. PMLR, 2021b.

Appendix

Table of Contents

A ALGORITHMS

First, we give an illustration of the interaction protocol of Algorithm 1.

$$\overline{R}_t \xrightarrow[\mathcal{A}(\overline{R}_t,\epsilon_t)]{} \widehat{R}_t \xrightarrow[\mathcal{D}(\widehat{R}_t,\theta_t)]{} \pi_t \longrightarrow \rho(\pi_t, \overline{R}_t) \tag{7}$$

Next, we give the pseudocodes for Exp3-DARP and OMDUCB-DARP. First, let us recall the update rule for Exp3. For any $t \geq 1$, we have

$$\phi^{\mathcal{D}}_{t+1}(\theta) := (1-\eta)\frac{w^{\mathcal{D}}_{t+1}(\theta)}{\sum_{\theta'} w^{\mathcal{D}}_{t+1}(\theta')} + \frac{\eta}{D} \ , \tag{8}$$

where, for any $\theta \in \Theta$

$$w^{\mathcal{D}}_{t+1}(\theta) = w^{\mathcal{D}}_{t}(\theta) \exp\left(\eta \tilde{\rho}_t(\theta)/D\right) \ ,$$

and

$$\tilde{\rho}_t = \begin{cases} \rho(\mathcal{D}(\widehat{R}_t, \theta), \overline{R}_t)/\phi^{\mathcal{D}}_t(\theta) & \text{if } \theta = \theta_t \\ 0 & \text{otherwise .} \end{cases}$$

The weights $w^{\mathcal{D}}_1(\theta)$ are initialized as 1, for all $\theta \in \Theta$.

B ADDITIONAL EXPERIMENTS

The environments we consider in this paper are illustrated in Figure 3. We provide experiments on the Bandit and Grid World environments, both in the adversarial and stochastic settings. We will use the same attack and defense models, as described in Section 5.

Algorithm 2 Exp3-DARP

```
Initialize: Attack model $\mathcal{A}$, defense model $\mathcal{D}$; attacker strategy $\phi_1^{\mathcal{A}}$; set $\phi_1^{\mathcal{D}}$ to the uniform distribution on $\Theta$.
for $t = 1, 2, 3, \ldots, T$ do:
    Attacker samples $\epsilon_t \sim \phi_t^{\mathcal{A}}$ and learner samples $\theta_t \sim \phi_t^{\mathcal{D}}$.
    $\overline{R}_t$ is chosen by the environment.
    $\widehat{R}_t = \mathcal{A}(\overline{R}_t, \epsilon_t)$ is computed and revealed to the learner.
    $\pi_t = \mathcal{D}(\widehat{R}_t, \theta_t)$ is computed.
    Attacker observes $\theta_t$, $\phi_t^{\mathcal{D}}$ and $\pi_t$, and incurs cost $\sigma(\pi_t, \widehat{R}_t)$.
    for $\epsilon \in \mathcal{E}$ do:
        Attacker updates its strategy $\phi_{t+1}^{\mathcal{A}}$ according to its learning method.
    end for
    Learner observes utility $\rho(\pi_t, \overline{R}_t)$.
    for $\theta \in \Theta$ do:
        Update $\phi_{t+1}^{\mathcal{D}}$ as in (8).
    end for
end for
```

Algorithm 3 Optimistic Mirror Descent for Stochastic Games

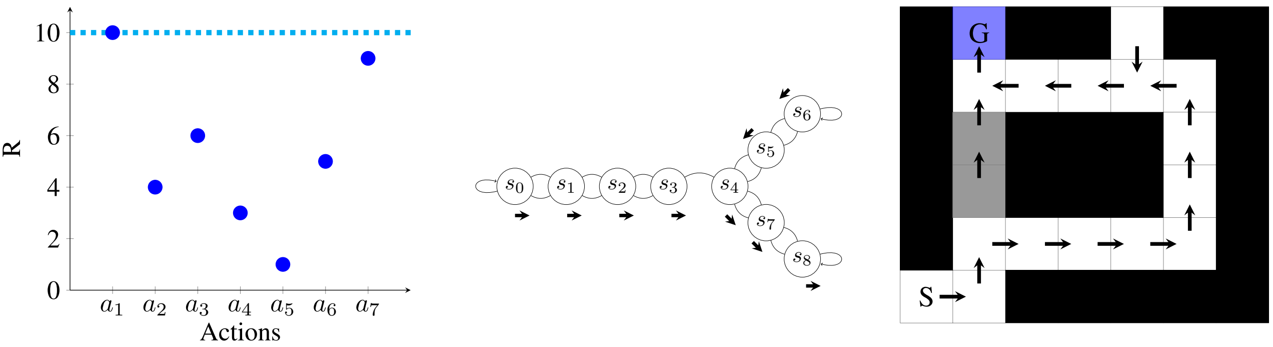
```
Initialize: Attack model $\mathcal{A}$, defense model $\mathcal{D}$; set $f$ to be 1-strongly convex with respect to $\|\cdot\|_1$ and let $\widetilde{\phi}_1^{\mathcal{D}} = \phi_1^{\mathcal{D}} = \arg\min_\phi f(\phi)$. Also, set $\widetilde{G}_1[\epsilon, \theta] = \widehat{G}_1[\epsilon, \theta] = 0$, for all $\epsilon \in \mathcal{E}$ and $\theta \in \Theta$.
for $t = 1, 2, 3, \ldots, T$ do:
    Attacker samples $\epsilon_t \sim \phi_t^{\mathcal{A}}$ and learner samples $\theta_t \sim \phi_t^{\mathcal{D}}$.
    $\overline{R}_t$ is chosen by the environment.
    $\widehat{R}_t = \mathcal{A}(\overline{R}_t, \epsilon_t)$ is computed and revealed to the learner.
    $\pi_t = \mathcal{D}(\widehat{R}_t, \theta_t)$ is computed.
    Attacker observes learner's strategy $\phi_t^{\mathcal{D}}$ and $\pi_t$, and incurs cost $\sigma(\pi_t, \widehat{R}_t)$.
    for $\epsilon \in \mathcal{E}$ do:
        Attacker updates its strategy $\phi_{t+1}^{\mathcal{A}}$ according to its learning method.
    end for
    Learner observes $\overline{R}_t$ and $\phi_t^{\mathcal{A}}$ and then computes $\widetilde{G}_{t+1}$ and $\widehat{G}_{t+1}$.
    for $\theta \in \Theta$ do:
        Update $\widetilde{\phi}_{t+1}^{\mathcal{D}}$ as in (3).
        Update $\phi_{t+1}^{\mathcal{D}}$ as in (4).
    end for
end for
```

B.1 Bandit Environment

In this section we consider a bandit environment, with 7 actions. Figure 4a illustrates the one shot interaction between the attacker and defender, that is, the score $\rho(\mathcal{D}(\mathcal{A}(\overline{R}, \epsilon), \theta), \overline{R})$ with respect to all possible (ϵ, θ). In this particular example, $\overline{R} = [10, 4, 6, 3, 1, 5, 9]$ and $\pi_\dagger = a_4$. Furthermore, we have $\rho(\pi_\dagger, \overline{R}) = 2.5$ and $\rho(\pi^*, \overline{R}) = 10$. As we can see, if the learner selects $\theta \geq \epsilon$, then the defense will output a policy which is near-optimal. Otherwise, the learner will inevitably adopt $\pi_\dagger$.

Next, we evaluate Exp3-DARP and OMDUCB-DARP in this environment. For the adversarial setting, we compare Exp3-DARP with No Defense and Fixed Defense, that is, a defender that does not use learning but employs only one defense parameter, thinking that the attacker is non-adaptive. We choose the fixed parameter 0.5, as the average of the parameter values. Similar to the Navigation case, the attack regularization parameter is $\lambda = 0.01$. The value of λ balances the trade-off between enforcing $\pi_\dagger$ and minimizing cost. A small value of λ means that the attacker is not sensitive to large amounts of poisoning. We did not compare with Naive Exp3 which only learns from feedback at the end of the round, since this particular case favours Naive Exp3, due to the low complexity of the environment (the state-action space is trivially small).

We run the algorithm for 10000 rounds, which takes approximately 10 minutes, since the environment is low-dimensional. As can be seen from Figure 4b, in both cases Exp3-DARP incurs sublinear regret, while both other methods incur linear regret.

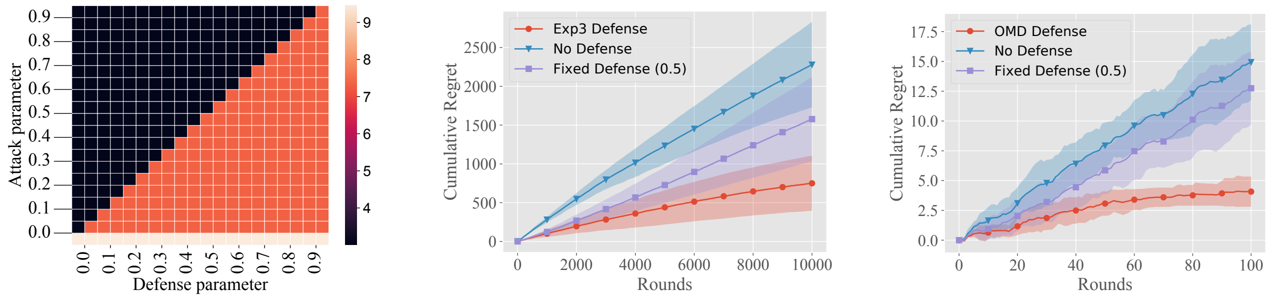

10
8
6
4
2
0
$\overline{R}$
a_1 a_2 a_3 a_4 a_5 a_6 a_7
Actions
s_0 s_1 s_2 s_3 s_4 s_5 s_6 s_7 s_8
G
S

(a) Bandit environment with 7 arms.

(b) Navigation environment with 9 states and 2 actions.

(c) Grid World environment with 18 states and 4 actions.

Figure 3: Visual illustration of the Bandit and Grid World environments. (a) A graphical representation of the Bandit environment with 7 arms, where $\pi^* = a_1$. (b) A visual representation of the Navigation environment with 9 states and 2 actions, where the arrows depict the optimal trajectory. (c) A visual representation of the Grid World environment with 18 states, initial state S and terminal state G. The idea is to avoid the grey states because the yield very low reward.



(a) One-shot interaction

(b) Regrets in the adversarial setting

(c) Regrets in the stochastic setting

Figure 4: Results for the Bandit environment. (a) One-shot attacker-defender interaction, with $\rho(\pi_\dagger, \overline{R}) = 2.5$ and $\rho(\pi^*, \overline{R}) = 10$. Here the defense is farther from the optimal policy than in the Navigation environment, possibly due to the large gap between $\rho(\pi_\dagger, \overline{R})$ and $\rho(\pi^*, \overline{R})$. (b) Comparison of actual regrets in the adversarial setting for the given methods, where we choose Fixed Defense parameter as 0.5, and No Defense learns $\pi_\dagger$ directly. (c) Comparison of actual regrets in the stochastic setting for the given methods.

For the stochastic case, we run the algorithm for only 100 rounds, and already observe convergence of OMDUCB-DARP, due to the stronger feedback model of the defense, and probably the low complexity of the distribution. Figure 4c illustrates the regret in the stochastic setting.

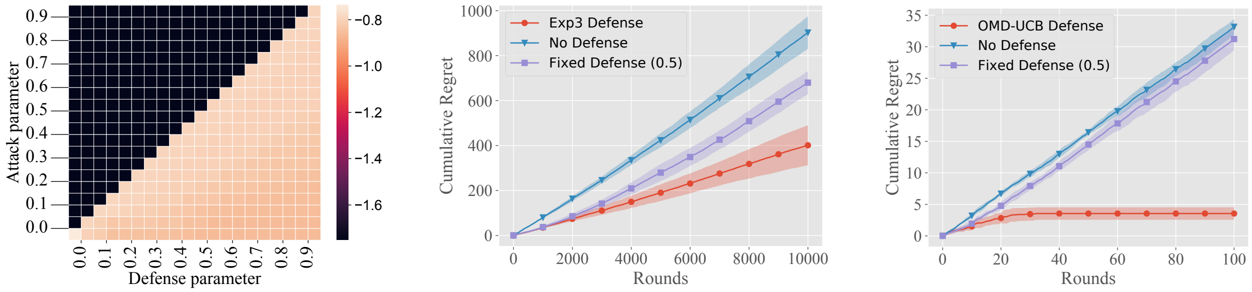
B.2 Grid World Environment

The Grid World environment (Ma et al., 2019; Banihashem et al., 2021) has 18 non-wall states and 4 actions, *up, down, left, right*. We have $\overline{R}(s_{14}, \cdot) = \overline{R}(s_{15}, \cdot) = -10$, $\overline{R}(s_{17}, \cdot) = 2$ and for all other states, the reward is -1. Again, ergodicity is ensured by letting the next state be sampled randomly with probability 0.1. Here we have $\gamma = 0.9$ and initial state s_0.

The optimal policy in the Grid World is to avoid the states s_{14} and s_{15} and go around from s_0 to s_{17}. The target policy of the attacker $\pi_\dagger$ is defined as following precisely these states to reach the terminal state s_{17}. In Figure 5a we can again see the interaction between both players for a fixed game. Here we have $\rho(\pi_\dagger, \overline{R}) = -1.75$ and $\rho(\pi^*, \overline{R}) = -0.7$. Again, it is obviously the case that the optimal defense parameter lies in the diagonal line of the plot, meaning that $\theta = \epsilon$.

Note that we observe similar results here for the adversarial setting, as can be seen in Figure 5b, where we again run the algorithm for 10000 rounds. We did not even bother to run Naive Exp3 in this environment since the number of arms in this case is $4^{18} = 68,719,476,736$, which makes learning from delayed feedback catastrophically non-efficient.

In the stochastic setting (see Figure 5c), on the other hand, linearity of No Defense and Fixed Defense is even more clear, due to the high complexity of the environment, while convergence of OMDUCB-DARP happens in around 30 rounds. This convergence rate, which is even faster than that in the simpler environments, can be attributed, besides the strong feedback model, also to a higher gap between the scores of the best and runner up policies.



(a) One-shot interaction.

(b) Regrets in the adversarial setting.

(c) Regrets in the stochastic setting.

Figure 5: Results for the Grid World environment. (a) One-shot attacker-defender interaction, with $\rho(\pi_\dagger, \overline{R}) = -1.75$ and $\rho(\pi^*, \overline{R}) = -0.7$. Note that here the defense is again able to recover a near-optimal policy, due to the proximity of $\rho(\pi_\dagger, \overline{R})$, to $\rho(\pi^*, \overline{R})$. (b) Comparison of actual regrets in the adversarial setting for the given methods, where we again let Fixed Defense parameter be 0.5, and No Defense learns $\pi_\dagger$ directly. Note that convergence in this setting is slower due to high complexity of the environment. (c) Comparison of actual regrets in the stochastic setting for the given methods.

C PROOFS OF RESULTS IN SECTION 3

In this section, we provide the proofs of Theorem 1, Theorem 2 and Theorem 3.

C.1 Proof Of Theorem 1

Statement. *Let $\mathcal{A} \in \mathfrak{A}$ and $\mathcal{D} \in \mathfrak{D}$. Moreover, given $T, D \in \mathbb{N}$, set*

$$\eta^{\mathcal{D}} = \min\{1, \sqrt{(D \ln D)/((e-1)T)}\}.$$

Then, we have $Reg_{\mathcal{D}}(T) \leq O(\sqrt{TD \ln D})$.

On the other hand, there exists an attack $\mathcal{A} \in \mathfrak{A}$ and a distribution β of rewards $\overline{R}_1, \dots, \overline{R}_T$, such that, for any defense $\mathcal{D} \in \mathfrak{D}$, the expected regret $\mathbb{E}_\beta Reg_{\mathcal{D}}(T)$ is at least $\Omega(\sqrt{TD})$.

Proof. For the upper bound, first note that in the adversarial scenario, we are not exploiting the structure in the data coming from the attacker's strategy. The results hold for any strategy of the attacker. We can thus view the problem as an adversarial bandit problem, where the arms are the parameters $\theta \in \Theta$ and the reward of pulling arm θ at time t is $\rho(\mathcal{D}(\widehat{R}_t, \theta), \overline{R}_t) \in [0, 1]$. The optimal arm, which serves as a benchmark for the adversarial bandit, is the optimal defense parameter that maximizes performance, i.e. θ_{max}. Thus, our notion of regret with respect to the optimal defense corresponds to the weak regret of Auer et al. (2002) and so Corollary 3.2 of (Auer et al., 2002) gives $O(\sqrt{TD \ln D})$ bounds on the regret of Exp3-DARP.

For the lower bound, we will consider the following construction. Let $M = (S, A, R, P, \gamma, \nu)$ with state space $S = \{s_0, s_{terminal}\}$, action space $A = \{a_1, a_2\}$, transitions $P(s_0, a_i, s_{terminal}) = 1$, for $i \in \{1, 2\}$, discount factor $\gamma = 1$ and initial state distribution $\nu(s_0) = 1$.

Now let us define the reward vector as $\overline{R}(s_0, a_1) = X$, where $X \sim \text{Ber}(\frac{1+\alpha}{2})$ and $\overline{R}(s_0, a_2) = Y$, where $Y \sim \text{Ber}(\frac{1-\alpha}{2})$, for some $\alpha \in (0, 1)$. In this case (since $\gamma = 1$) the score of a given policy π is $\rho(\pi, \overline{R}) = \overline{R}(s_o, \pi)$. Note that we have $\rho(a_1, \overline{R}) = X$ and $\rho(a_2, \overline{R}) = Y$.

Furthermore, let the action space of the attacker be $\mathcal{E} = \{\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4\}$, where we let

$$\mathcal{A}\left(\begin{bmatrix}1\\0\end{bmatrix}, \epsilon_1\right) = \begin{bmatrix}1\\0\end{bmatrix}, \quad \mathcal{A}\left(\begin{bmatrix}1\\1\end{bmatrix}, \epsilon_2\right) = \begin{bmatrix}1\\0\end{bmatrix} \quad \mathcal{A}\left(\begin{bmatrix}0\\1\end{bmatrix}, \epsilon_3\right) = \begin{bmatrix}1\\0\end{bmatrix} \quad \mathcal{A}\left(\begin{bmatrix}0\\0\end{bmatrix}, \epsilon_4\right) = \begin{bmatrix}1\\0\end{bmatrix}$$

Obviously, we have designed the attack so that $\pi_\dagger = a_1$. Note that $\pi_\dagger$ is always optimal under $\widehat{R}$ since, for any $\overline{R} \in \mathcal{R}$ and $\epsilon \in \mathcal{E}$, we have $\mathcal{A}(\overline{R}, \epsilon) = \widehat{R} = [1 \;\; 0]^T$.

On the other hand, let $\Theta = \{\theta_1, \dots, \theta_D\}$ denote the action space of the learner. Furthermore, let $\Theta_1 \subset \Theta$ be a subset of Θ such that $\mathcal{D}(\widehat{R}, \theta) = a_1$, for $\theta \in \Theta_1$, and let $\Theta_2 := \Theta \setminus \Theta_1$. We assume neither Θ_1 nor Θ_2 are empty, without loss of generality, since if any of these events were the case, we can always find a fixed attack that imposes $\pi_\dagger$ on the defense since there is no decision-making for the defense. Hence, we omit the trivial scenarios.

Assume that, in each round $t \geq 1$, we have $\overline{R}_t = \overline{R}$. Now if the defender selects $\theta_t \in \Theta_1$ at round t, we have

$$\rho(\mathcal{D}(\mathcal{A}(\overline{R}, \epsilon_t), \theta_t), \overline{R}) = \rho(\mathcal{D}(\widehat{R}, \theta_t), \overline{R}) = \rho(a_1, \overline{R}) = X ,$$

for any $\epsilon_t \in \mathcal{E}$. On the other hand, if $\theta_t \in \Theta_2$, we obtain

$$\rho(\mathcal{D}(\mathcal{A}(\overline{R}, \epsilon_t), \theta_t), \overline{R}) = \rho(\mathcal{D}(\widehat{R}, \theta_t), \overline{R}) = \rho(a_2, \overline{R}) = Y ,$$

for any $\epsilon_t \in \mathcal{E}$.

We will use the following result from (Bubeck and Cesa-Bianchi, 2012).

Lemma 1. *(Lemma 3.2 of (Bubeck and Cesa-Bianchi, 2012)) Let $\alpha \in (0, 1]$ and let $Y_{i,t}$ denote the reward received from playing action $i \leq D$ at time $t \geq 1$. Let $\mathbb{E}_i$ be the expectation with respect to the joint distribution of rewards where all actions are i.i.d. Bernoulli of parameter $\frac{1-\alpha}{2}$, but action i, which is i.i.d. Bernoulli of parameter $\frac{1+\alpha}{2}$. Then, for any randomized strategy A, if we denote by I_t the action played by A at time t, we have*

$$\max_{1 \leq i \leq D} \mathbb{E}_i \sum_{t=1}^{T} (Y_{i,t} - Y_{I_t,t}) \geq T\alpha \left(1 - \frac{1}{D} - \sqrt{\alpha \ln \frac{1+\alpha}{1-\alpha}} \sqrt{\frac{T}{2D}}\right) .$$

Note that all actions $\theta \in \Theta_1$ can be represented with one action, without loss of generality. Also, let us shorten notation and define

$$\rho(\epsilon, \theta, \overline{R}) = \rho(\mathcal{D}(\mathcal{A}(\overline{R}, \epsilon), \theta), \overline{R}) .$$

Now let us denote by θ_i^* any parameter from Θ_i. Lemma 1 implies

$$\max_{1 \leq i \leq D} \mathbb{E}_i \sum_{t=1}^{T} \left(\rho(\epsilon_t, \theta_i^*, \overline{R}) - \rho(\epsilon_t, \theta_t, \overline{R})\right) \geq T\alpha \left(1 - \frac{1}{D} - \sqrt{\alpha \ln \frac{1+\alpha}{1-\alpha}} \sqrt{\frac{T}{2D}}\right) ,$$

since

$$\rho(\epsilon_t, \theta_t, \overline{R}) = \begin{cases} \text{Ber}(\frac{1+\alpha}{2}) & \text{if } \theta_t = \theta_i^* \\ \text{Ber}(\frac{1-\alpha}{2}) & \text{if otherwise.} \end{cases}$$

Letting $\alpha = O(\sqrt{D/T})$, we obtain the $\Omega(\sqrt{TD})$ bounds. □

C.2 Proof Of Theorem 2

Statement. *Assume that, at the end of round t, the learner can observe the attacker's strategy and the true reward function $\overline{R}_t$. Then, there exists an attack $\mathcal{A} \in \mathfrak{A}$, such that, for any defense $\mathcal{D} \in \mathcal{D}$ and any sequence $\phi_1^{\mathcal{D}}, \dots, \phi_T^{\mathcal{D}}$, we have $Reg_{\mathcal{D}}(T) = \Omega(\sqrt{T \log D})$.*

Proof. For the full information feedback setting, we can use any algorithm that satisfies the *Regret bounded by Variation in Utilities* property (Equation (1) of (Syrgkanis et al., 2015)), such as any variation of Optimistic Mirror Descent procedure. For the usual adversarial case, when the structure of the other players cannot be exploited, we obtain $O(\sqrt{T \log D})$ bound on the regret.by an appropriate choice of the step-size $\eta^{\mathcal{D}}$.

In order to show the lower bound, we will consider the same construction as in the proof of Theorem 1. We use the same MDP and attack model. Thus, we again have

$$\rho(\mathcal{D}(\mathcal{A}(\overline{R}, \epsilon), \theta_t), \overline{R}) = \rho(\mathcal{D}(\widehat{R}, \theta_t), \overline{R}) = \rho(a_1, \overline{R}) = X ,$$

for all $\epsilon \in \mathcal{E}$, if $\theta_t \in \Theta_1$, and

$$\rho(\mathcal{D}(\mathcal{A}(\overline{R}, \epsilon), \theta_t), \overline{R}) = \rho(\mathcal{D}(\widehat{R}, \theta_t), \overline{R}) = \rho(a_2, \overline{R}) = Y ,$$

for any $\epsilon \in \mathcal{E}$, if $\theta_t \in \Theta_2$. However, instead of defining X and Y as Bernoulli random variables with different parameters, we instead let $X = 1/2$ and $Y = \text{Ber}(1/2)$.

Note that, given any strategy ϕ, we have $\mathbb{E}[\sum_{t=1}^{T} \langle \phi, \overline{\rho}_t \rangle] = \frac{1}{2}$, where the expectation is over the reward sequences. Let us denote by $Z_i = \sum_{t=1}^{T} \rho(\epsilon_t, \theta_i^*, \overline{R})$, where θ_i^* is defined as in the proof of Theorem 1. We have $Z_1 = T/2$ and $Z_2 \sim \text{Bin}(T, 1/2)$. We will use the following tail bounds for Binomial random variables.

Lemma 2. *Let Z be a Binomial random variable with T trials and success probability $1/2$. For $k \in [0, T/8]$, we have*

$$\mathbb{P}\left(Z \geq \frac{T}{2} + k\right) \geq \frac{1}{15} e^{-16k^2/T} \, .$$

Let $k = (1/4)\sqrt{T \log(D-1)}$. The result above implies

$$\mathbb{P}\left(\max_i Z_i \leq \frac{T}{2} + k\right) = \prod_{i=2}^{E} \mathbb{P}\left(Z_i \leq \frac{T}{2} + k\right) \leq \left(1 - \frac{1}{15} e^{-16k^2/T}\right)^{D-1} \leq 0.95$$

Thus, we have that $\mathbb{P}(\max_i Z_i \leq T/2 + k) \leq 0.95$ and $\mathbb{P}(\max_i Z_i \geq T/2 + k) \geq 0.05$. Since we always have that $\max_i Z_i \geq T/2$, we obtain

$$\begin{aligned}\mathbb{E}[\max_i Z_i] &\geq \mathbb{P}\left(\frac{T}{2} \leq \max_i Z_i \leq \frac{T}{2} + k\right)\frac{T}{2} + \mathbb{P}\left(\max_i Z_i \geq \frac{T}{2} + k\right)\left(\frac{T}{2} + k\right) \\ &= 0.95\frac{T}{2} + 0.05\left(\frac{T}{2} + k\right) \\ &= \frac{T}{2} + \frac{1}{80}\sqrt{T \log(D-1)} \, .\end{aligned}$$

Thus, for any sequence of randomized strategies of the learner $\phi_1^{\mathcal{D}}, \ldots, \phi_T^{\mathcal{D}}$, and for any sequence of attack parameters $\epsilon_1, \ldots, \epsilon_T$, we obtain

$$\mathbb{E}\left[\max_i \sum_{t=1}^{T} \rho(\epsilon_t, \theta_i^*, \overline{R}_t) - \sum_{t=1}^{T} \langle \phi_t^{\mathcal{D}}, \overline{\rho}_t \rangle\right] \geq \frac{1}{80}\sqrt{T \log(D-1)} \, ,$$

where the expectation is with respect to the distribution of rewards that we defined. Note that $Reg_{\mathcal{D}}(T) \leq T$. The reverse Markov inequality gives

$$\mathbb{P}\left(Reg_{\mathcal{D}}(T) \geq \frac{1}{160}\sqrt{T \ln(D-1)}\right) \geq \frac{\mathbb{E}Reg_{\mathcal{D}}(T) - \frac{1}{160}\sqrt{T \ln(D-1)}}{T - \frac{1}{160}\sqrt{T \ln(H-1)}} \geq \frac{1}{160}\sqrt{\frac{\ln(D-1)}{T}} \, .$$

Thus, using the probabilistic method, we can say that there exists a sequence $\overline{R}_1, \ldots, \overline{R}_T$, for which Optimistic Hedge incurs $\Omega(\sqrt{T \log D})$ regret. □

C.3 Proof Of Theorem 3

Statement. *Let $\mathcal{A} \in \mathfrak{A}$, $\mathcal{D} \in \mathfrak{D}$ and $T \in \mathbb{N}$. Assume $\eta^{\mathcal{D}} \leq \eta^{\mathcal{A}} \in (0, 1]$. Moreover, assume that, at every round $t \geq 1$, the attacker's randomized strategy $\phi_t^{\mathcal{A}}$ is updated via the Optimistic Hedge update rule. Then, for any $\delta \in (0, 1)$, the above algorithm incurs expected regret*

$$Reg_{\mathcal{D}}^*(T) \leq \eta^{\mathcal{A}} + O(\eta^{\mathcal{A}} \log E) + O\left((\eta^{\mathcal{A}})^2 + (\eta^{\mathcal{A}})^3\right) T + \frac{\eta^{\mathcal{A}}}{2} \log(\pi^2 ED/(3\delta)) \log T + \frac{1}{\eta^{\mathcal{D}}} f_{max} \, ,$$

with probability at least $1 - \delta$, where $f_{max} = \max_\phi f(\phi) - \min_\phi f(\phi)$.

Proof. First, we will prove an auxiliary result that gives tail bounds on the deviation of the sample average of the costs from their mean, for any action pair.

Lemma 3. *Given $\delta \in (0,1)$, for every $\epsilon \in \mathcal{E}$, $\theta \in \Theta$, and $t \geq 2$, we have*

$$\mathbb{P}\left(\left|\widetilde{G}_t[\epsilon,\theta] - G[\epsilon,\theta]\right| \leq \sqrt{\frac{\log\left(\pi^2 ED/(3\delta)\right)}{2(t-1)}}\right) \geq 1-\delta\ .$$

Proof. First, note that we have $\overline{G}_t[\epsilon,\theta] \in [0,1]$, $\forall \epsilon \in \mathcal{E}, \theta \in \Theta, t \geq 1$. Moreover, given any pair $(\epsilon,\theta) \in \mathcal{E} \times \Theta$, the sequence $\overline{G}_1[\epsilon,\theta], \overline{G}_2[\epsilon,\theta], \dots$ is i.i.d. with mean $G[\epsilon,\theta]$. Then, Hoeffding's inequality implies

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\epsilon\in\mathcal{E}}\bigcup_{\theta\in\Theta}\bigcup_{t\geq 2}\left\{\left|\sum_{k=1}^{t-1}\overline{G}_k[\epsilon,\theta] - \mathbb{E}\sum_{k=1}^{t-1}\overline{G}_k[\epsilon,\theta]\right| \geq \sqrt{\frac{(t-1)}{2}\log\left(\frac{\pi^2 ED}{3\delta}\right)}\right\}\right) \\
&\qquad \leq \sum_{\epsilon\in\mathcal{E}}\sum_{\theta\in\Theta}\sum_{t\geq 2}\mathbb{P}\left(\left|\sum_{k=1}^{t-1}\overline{G}_k[\epsilon,\theta] - (t-1)G[\epsilon,\theta]\right| \geq \sqrt{\frac{(t-1)}{2}\log\left(\frac{\pi^2 ED}{3\delta}\right)}\right) \\
&\qquad \leq \sum_{\epsilon\in\mathcal{E}}\sum_{\theta\in\Theta}\sum_{t\geq 2} 2\exp\left(-\frac{2\left(\sqrt{\frac{(t-1)}{2}\log(\frac{\pi^2 ED}{3\delta})}\right)^2}{t-1}\right) \\
&\qquad \leq ED\frac{\pi^2}{3}\exp\left(-\log\left(\frac{\pi^2 ED}{3\delta}\right)\right) \\
&\qquad = \delta\ .
\end{aligned}$$

Thus, for all $\theta \in \Theta$, $\epsilon \in \mathcal{E}$ and $t \geq 2$, the statement of the lemma follows. □

Now we are ready to prove the main result. Suppose the event of Lemma 3 holds. Note that we can write the regret as

$$\begin{aligned}
Reg^*_{\mathcal{D}}(T) &= \mathbb{E}_{\overline{G}_t\sim\mathcal{G},\epsilon_t\sim\phi^{\mathcal{A}}_t,\theta_t\sim\phi^{\mathcal{D}}_t}\sum_{t=1}^{T}\overline{G}_t[\epsilon_t,\theta_t] - \min_{\phi}\mathbb{E}_{\overline{G}_t\sim\mathcal{G},\epsilon_t\sim\phi^{\mathcal{A}}_t,\theta\sim\phi}\sum_{t=1}^{T}\overline{G}_t[\epsilon_t,\theta] \\
&= \sum_{t=1}^{T}\langle\phi^{\mathcal{D}}_t - \phi^{\mathcal{D}}_*, G^T\phi^{\mathcal{A}}_t\rangle\ ,
\end{aligned} \tag{9}$$

where $\phi^{\mathcal{D}}_* := \arg\min_\phi \sum_{t=1}^{T}\mathbb{E}_{\overline{G}_t\sim\mathcal{G},\epsilon_t\sim\phi^{\mathcal{A}}_t,\theta\sim\phi}\overline{G}_t[\epsilon_t,\theta]$. Given $t \geq 1$, Lemma 3 implies

$$\begin{aligned}
\langle\phi^{\mathcal{D}}_t - \phi^{\mathcal{D}}_*, G^T\phi^{\mathcal{A}}_t\rangle &= \langle\phi^{\mathcal{D}}_t - \widetilde{\phi}^{\mathcal{D}}_{t+1}, G^T\phi^{\mathcal{A}}_t - \widetilde{G}^T_t\phi^{\mathcal{A}}_{t-1}\rangle + \langle\phi^{\mathcal{D}}_t - \widetilde{\phi}^{\mathcal{D}}_{t+1}, \widetilde{G}^T_t\phi^{\mathcal{A}}_{t-1}\rangle + \langle\widetilde{\phi}^{\mathcal{D}}_{t+1} - \phi^{\mathcal{D}}_*, G^T\phi^{\mathcal{A}}_t\rangle \\
&\leq \langle\phi^{\mathcal{D}}_t - \widetilde{\phi}^{\mathcal{D}}_{t+1}, G^T\phi^{\mathcal{A}}_t - \widetilde{G}^T_t\phi^{\mathcal{A}}_{t-1}\rangle + \langle\phi^{\mathcal{D}}_t - \widetilde{\phi}^{\mathcal{D}}_{t+1}, \widetilde{G}^T_t\phi^{\mathcal{A}}_{t-1}\rangle + \langle\widetilde{\phi}^{\mathcal{D}}_{t+1} - \phi^{\mathcal{D}}_*, \widehat{G}^T_t\phi^{\mathcal{A}}_t\rangle
\end{aligned}$$

due to linearity of inner product and the fact that $G[\epsilon,\theta] \leq \widehat{G}_t[\epsilon,\theta]$, for any $\epsilon \in \mathcal{E}$, $\theta \in \Theta$ and $t \geq 2$. We will provide upper bounds for each term on the right-hand side. First, we give upper bounds for the last two terms. Any update of the form $a^* = \arg\min_a\langle a,x\rangle + \mathcal{B}_f(a,c)$ satisfies, for any $d \in \Theta$, (Equation 26 of (Rakhlin and Sridharan, 2013a)):

$$\langle a^* - d, x\rangle \leq \mathcal{B}_f(d,c) - \mathcal{B}_f(d,a^*) - \mathcal{B}_f(a^*,c)\ .$$

Thus, the update equations (3) and (4) imply

$$\langle\phi^{\mathcal{D}}_t - \widetilde{\phi}^{\mathcal{D}}_{t+1}, \widetilde{G}^T_t\phi^{\mathcal{A}}_{t-1}\rangle \leq \frac{1}{\eta^{\mathcal{D}}}\left(\mathcal{B}_f(\widetilde{\phi}^{\mathcal{D}}_{t+1}, \widetilde{\phi}^{\mathcal{D}}_t) - \mathcal{B}_f(\widetilde{\phi}^{\mathcal{D}}_{t+1}, \phi^{\mathcal{D}}_t) - \mathcal{B}_f(\phi^{\mathcal{D}}_t, \widetilde{\phi}^{\mathcal{D}}_t)\right)\ ,$$

and

$$\langle\widetilde{\phi}^{\mathcal{D}}_{t+1} - \phi^{\mathcal{D}}_*, \widehat{G}^T_t\phi^{\mathcal{A}}_t\rangle \leq \frac{1}{\eta^{\mathcal{D}}}\left(\mathcal{B}_f(\phi^{\mathcal{D}}_*, \widetilde{\phi}^{\mathcal{D}}_t) - \mathcal{B}_f(\phi^{\mathcal{D}}_*, \widetilde{\phi}^{\mathcal{D}}_{t+1}) - \mathcal{B}_f(\widetilde{\phi}^{\mathcal{D}}_{t+1}, \widetilde{\phi}^{\mathcal{D}}_t)\right)\ .$$

On the other hand, denoting by $\|\cdot\|_*$ the dual norm of $\|\cdot\|$, for the remaining term we have

$$\begin{aligned}\langle \phi_t^{\mathcal{D}} - \widetilde{\phi}_{t+1}^{\mathcal{D}}, G^T\phi_t^{\mathcal{A}} - \widetilde{G}_t^T\phi_{t-1}^{\mathcal{A}}\rangle &\le \left\|\phi_t^{\mathcal{D}} - \widetilde{\phi}_{t+1}^{\mathcal{D}}\right\| \cdot \left\|G^T\phi_t^{\mathcal{A}} - \widetilde{G}_t^T\phi_{t-1}^{\mathcal{A}}\right\|_* \\ &\le \frac{1}{2\eta^{\mathcal{A}}}\left\|\phi_t^{\mathcal{D}} - \widetilde{\phi}_{t+1}^{\mathcal{D}}\right\|_1^2 + \frac{\eta^{\mathcal{A}}}{2}\left\|G^T\phi_t^{\mathcal{A}} - \widetilde{G}_t^T\phi_{t-1}^{\mathcal{A}}\right\|_\infty^2 ,\end{aligned} \tag{10}$$

by Cauchy-Schwarz and the fact that the dual of $\|\cdot\|_1$ is $\|\cdot\|_\infty$. The second term of the right-hand side above can be bounded as

$$\begin{aligned}\frac{\eta^{\mathcal{A}}}{2}\left\|G^T\phi_t^{\mathcal{A}} - \widetilde{G}_t^T\phi_{t-1}^{\mathcal{A}}\right\|_\infty^2 &\le \eta^{\mathcal{A}}\left\|G^T(\phi_t^{\mathcal{A}} - \phi_{t-1}^{\mathcal{A}})\right\|_\infty^2 + \eta^{\mathcal{A}}\left\|(G-\widetilde{G}_t)^T\phi_{t-1}^{\mathcal{A}}\right\|_\infty^2 \\ &\le \eta^{\mathcal{A}}\left\|\phi_t^{\mathcal{A}} - \phi_{t-1}^{\mathcal{A}}\right\|_1^2 + \eta^{\mathcal{A}}\max_{\epsilon,\theta}\left|G[\epsilon,\theta] - \widetilde{G}_t[\epsilon,\theta]\right|^2 && (11) \\ &\le \eta^{\mathcal{A}}\left\|\phi_t^{\mathcal{A}} - \phi_{t-1}^{\mathcal{A}}\right\|_1^2 + \eta^{\mathcal{A}}\frac{\log(\pi^2 ED/(3\delta))}{2(t-1)} . && (12)\end{aligned}$$

where (11) follows from the fact that $\phi_t^{\mathcal{A}}$ is a probability simplex and $G[\epsilon,\theta] \le 1$, for all $\epsilon \in \mathcal{E}$, $\theta \in \Theta$, by assumption; (12) follows from Lemma 3. Putting everything together, we obtain

$$\begin{aligned}Reg_{\mathcal{D}}^*(T) &= \sum_{t=1}^T \langle \phi_t^{\mathcal{D}} - \phi_*^{\mathcal{D}}, G^T\phi_t^{\mathcal{A}}\rangle \\ &\le \sum_{t=1}^T \langle \phi_t^{\mathcal{D}} - \widetilde{\phi}_{t+1}^{\mathcal{D}}, G^T\phi_t^{\mathcal{A}} - \widetilde{G}_t^T\phi_{t-1}^{\mathcal{A}}\rangle + \langle \phi_t^{\mathcal{D}} - \widetilde{\phi}_{t+1}^{\mathcal{D}}, \widetilde{G}_t^T\phi_{t-1}^{\mathcal{A}}\rangle + \langle \widetilde{\phi}_{t+1}^{\mathcal{D}} - \phi_*^{\mathcal{D}}, \widehat{G}_t^T\phi_t^{\mathcal{A}}\rangle \\ &\le \sum_{t=1}^T \left(\frac{1}{2\eta^{\mathcal{A}}}\left\|\phi_t^{\mathcal{D}} - \widetilde{\phi}_{t+1}^{\mathcal{D}}\right\|_1^2 + \eta^{\mathcal{A}}\left\|\phi_t^{\mathcal{A}} - \phi_{t-1}^{\mathcal{A}}\right\|_1^2\right) + \eta^{\mathcal{A}} + \sum_{t=2}^T\left(\eta^{\mathcal{A}}\frac{\log(\pi^2 ED/(3\delta))}{2(t-1)}\right) \\ &\quad + \sum_{t=1}^T \frac{1}{\eta^{\mathcal{D}}}\left(\mathcal{B}_f(\widetilde{\phi}_{t+1}^{\mathcal{D}}, \widetilde{\phi}_t^{\mathcal{D}}) - \mathcal{B}_f(\widetilde{\phi}_{t+1}^{\mathcal{D}}, \phi_t^{\mathcal{D}}) - \mathcal{B}_f(\phi_t^{\mathcal{D}}, \widetilde{\phi}_t^{\mathcal{D}})\right) \\ &\qquad + \sum_{t=1}^T \frac{1}{\eta^{\mathcal{D}}}\left(\mathcal{B}_f(\phi_*^{\mathcal{D}}, \widetilde{\phi}_t^{\mathcal{D}}) - \mathcal{B}_f(\phi_*^{\mathcal{D}}, \widetilde{\phi}_{t+1}^{\mathcal{D}}) - \mathcal{B}_f(\widetilde{\phi}_{t+1}^{\mathcal{D}}, \widetilde{\phi}_t^{\mathcal{D}})\right) \\ &\le \eta^{\mathcal{A}} + \eta^{\mathcal{A}}\sum_{t=1}^T\left\|\phi_t^{\mathcal{A}} - \phi_{t-1}^{\mathcal{A}}\right\|_1^2 + \eta^{\mathcal{A}}\sum_{t=2}^T\frac{\log(\pi^2 ED/(3\delta))}{2(t-1)} \\ &\quad + \frac{1}{\eta^{\mathcal{D}}}\sum_{t=1}^T\left(\mathcal{B}_f(\phi_*^{\mathcal{D}}, \widetilde{\phi}_t^{\mathcal{D}}) - \mathcal{B}_f(\phi_*^{\mathcal{D}}, \widetilde{\phi}_{t+1}^{\mathcal{D}})\right) && (13) \\ &\le O(\eta^{\mathcal{A}}\log E) + O\left((\eta^{\mathcal{A}})^2 + (\eta^{\mathcal{A}})^3\right)T + \frac{\eta^{\mathcal{A}}}{2}\log(\pi^2 ED/(3\delta))\log T + \frac{1}{\eta^{\mathcal{D}}}f_{max} + \eta^{\mathcal{A}} && (14)\end{aligned}$$

where (13) follows from that fact that $\mathcal{B}_f(\widetilde{\phi}_{t+1}^{\mathcal{D}}, \phi_t^{\mathcal{D}}) \ge \frac{1}{2}\left\|\widetilde{\phi}_{t+1}^{\mathcal{D}} - \phi_t^{\mathcal{D}}\right\|_1^2$ by Pinsker's inequality and the assumption that $\eta^{\mathcal{D}} \le \eta^{\mathcal{A}}$; the first two terms of (14) follow from Lemma 3.2 of (Chen and Peng, 2020), the $\log T$ component of the third term is an upper bound on the $(T-1)$th Harmonic number, and the fourth term follows by definition of f_{max}. □

C.4 Proof Of Corollary 1

Before proving Corollary 1, we state and prove an auxiliary lemma that gives upper bounds on the magnitude of change in strategies, for common online learning methods.

Lemma 4. *Let $\phi_1(i) = 1/n$, for all $i \in [n]$, where $n \in \mathbb{N}$. Furthermore, let $T \in \mathbb{N}$ and fix learning rate $\eta \in (0,1]$. Then, assuming an online learning setting with cost vectors $\overline{\sigma}_t$ with entries in $[0,1]$, for all $t \in [T]$, we have the following:*

- $\sum_{t=1}^{T} \|\phi_t - \phi_{t-1}\|_1^2 \leq O(\log n/\eta + \eta\sqrt{T} + \eta^2 T)$*, if* ϕ_t *are updated using Optimistic Mirror Descent or Optimistic Follow the Regularized Leader update rules.*
- $\sum_{t=1}^{T} \|\phi_t - \phi_{t-1}\|_1^2 \leq O\left(\log n + (\eta + \eta^2)\right) T$ *if* ϕ_t *are updated using Optimistic Hedge update rule.*
- $\sum_{t=1}^{T} \|\phi_t - \phi_{t-1}\|_1^2 \leq O(\eta\sqrt{T})$*, if* ϕ_t *are updated using Hedge update rule.*

Proof. First, note that the regret is written as

$$Reg(T) = \sum_{t=1}^{T} \langle \phi_t, \overline{\sigma}_t \rangle - \min_{j\in[n]} \sum_{t=1}^{T} \overline{\sigma}_t(j) \ .$$

For the first point, we will use Proposition 5 and 7 of (Syrgkanis et al., 2015) and the RVU property of OMD and OFTRL. For both algorithms, we have

$$\begin{aligned}\sum_{t=1}^{T} \|\phi_t - \phi_{t-1}\|_1^2 &\leq O\left(\frac{\log n}{\eta} + \eta Reg(T) + \eta^2 \sum_{t=1}^{T} \|\overline{\sigma}_t - \overline{\sigma}_{t-1}\|_*^2\right) \\ &\leq O\left(\frac{\log n}{\eta} + \eta\sqrt{T} + \eta^2 T\right) \ ,\end{aligned}$$

where the first inequality follows from Propositions 5 and 7 of (Syrgkanis et al., 2015) and the second one follows from the reward assumption and the worst-case regret bounds for these methods. The second point follows from Lemma 3.2 of (Chen and Peng, 2020). For the third point, we use a similar argument as Lemma 3.2 of (Chen and Peng, 2020). We provide the proof here for completion. For every $2 \leq t \leq T$, we have

$$\begin{aligned}\frac{1}{2}\|\phi_t - \phi_{t-1}\|_1^2 &\leq \sum_{i=1}^{n} \phi_{t-1}(i) \log\left(\frac{\phi_{t-1}(i)}{\phi_t(i)}\right) \\ &= \sum_{i=1}^{n} \phi_{t-1}(i) \log\left(\sum_{j=1}^{n} \phi_{t-1}(j) \exp\left(-\eta\overline{\sigma}_{t-1}(j)\right)\right) + \eta \sum_{j=1}^{n} \phi_{t-1}(j)\overline{\sigma}_{t-1}(j) \\ &= \log\left(\sum_{j=1}^{n} \phi_{t-1}(j) \exp\left(-\eta\overline{\sigma}_{t-1}(j)\right)\right) + \eta\langle \phi_{t-1}, \overline{\sigma}_{t-1} \rangle \ ,\end{aligned}$$

where the first inequality follows from Pinsker's inequality; for the first equality we have used the Hedge update rule, given as

$$\phi_t(i) = \frac{\phi_{t-1}(i)\exp(-\eta\overline{\sigma}_{t-1}(i))}{\sum_{j\in[n]} \phi_{t-1}(j)\exp(-\eta\overline{\sigma}_{t-1}(j))} \ ,$$

for all $i \in [n]$; the second equality follows from the fact that $\sum_{i\in[n]} \phi_t(i) = 1$, for all $t \geq 1$.

Next, using induction, we will show that, for any $k \in \mathbb{N}$, we have

$$\sum_{t=1}^{k} \log\left(\sum_{j=1}^{n} \phi_{t-1}(j) \exp\left(-\eta\overline{\sigma}_{t-1}(j)\right)\right) = \log\left(\sum_{i\in[n]} \phi_1(i) \exp(-\eta \sum_{t=0}^{k-1} \overline{\sigma}_t(i))\right) \ .$$

For $k = 1$, the equality follows from the fact that $\overline{\sigma}_0$ is the 0 vector and the fact that $\phi_1(j) = 1/n$, for all $j \in [n]$. We assume the equality holds for k and prove it for $k+1$. We have

$$\begin{aligned}\sum_{t=1}^{k+1} \log\left(\sum_{j\in[n]} \phi_{t-1}(j) \exp\left(-\eta\overline{\sigma}_{t-1}(j)\right)\right) &= \sum_{t=1}^{k} \log\left(\sum_{j\in[n]} \phi_{t-1}(j) \exp\left(-\eta\overline{\sigma}_{t-1}(j)\right)\right) \\ &\quad + \log\left(\sum_{j\in[n]} \phi_k(j) \exp\left(-\eta\overline{\sigma}_k(j)\right)\right)\end{aligned}$$

$$
\begin{aligned}
&= \log\left(\sum_{i\in[n]} \phi_1(i)\exp(-\eta\sum_{t=0}^{k-1}\overline{\sigma}_t(i))\right) + \log\left(\sum_{j\in[n]}\phi_k(j)\exp\left(-\eta\overline{\sigma}_k(j)\right)\right)\\
&= \log\left(\left(\sum_{i\in[n]} \phi_1(i)\exp(-\eta\sum_{t=0}^{k-1}\overline{\sigma}_t(i))\right)\cdot\left(\sum_{j\in[n]}\phi_k(j)\exp\left(-\eta\overline{\sigma}_k(j)\right)\right)\right)\\
&= \log\left(\sum_{i\in[n]} \phi_1(i)\exp\left(-\eta\sum_{t=1}^{k}\overline{\sigma}_t(i)\right)\right)\ ,
\end{aligned}
$$

where the third inequality follows from the fact that, for all $i \in [n]$, we have

$$
\phi_t(i) = \frac{\phi_1(i)\exp(-\eta\sum_{\tau=1}^{t-1}\overline{\sigma}_\tau(i))}{\sum_{j\in[n]}\exp(-\eta\sum_{\tau=1}^{t-1}\overline{\sigma}_\tau(j))}\ .
$$

Finally, we have

$$
\begin{aligned}
\frac{1}{2\ln 2}\sum_{t=2}^{T}\|\phi_t-\phi_{t-1}\|_1^2 &\leq \sum_{t=2}^{T}\log\left(\sum_{j=1}^{n}\phi_{t-1}(j)\exp\left(-\eta\overline{\sigma}_{t-1}(j)\right)\right) + \eta\sum_{t=2}^{T}\langle\phi_{t-1},\overline{\sigma}_{t-1}\rangle\\
&= \log\left(\sum_{i\in[n]}\frac{1}{n}\exp\left(-\eta\sum_{t=1}^{T-1}\overline{\sigma}_t(i)\right)\right) + \eta\sum_{t=2}^{T}\langle\phi_{t-1},\overline{\sigma}_{t-1}\rangle\\
&\leq \eta\sum_{t=2}^{T}\langle\phi_{t-1},\overline{\sigma}_{t-1}\rangle - \eta\min_{i\in[n]}\sum_{t=1}^{T-1}\overline{\sigma}_t(i)\\
&= \eta Reg(T) \leq O(\eta\sqrt{T})\ ,
\end{aligned}
$$

where the second inequality follows from the fact that the $\min$ is less than the average, and the last inequality follows from the definition of regret and the worst-case bounds for Hedge.

□

Now we can plug in the results of the previous lemma and prove Corollary 1.

Statement. *Under the conditions of Theorem 3, we have*

- $Reg^*_{\mathcal{D}}(T) \leq O(\log T)$, *if the attacker plays a fixed strategy.*
- $Reg^*_{\mathcal{D}}(T) \leq O(T^{1/4})$, *if the attacker plays Hedge.*
- $Reg^*_{\mathcal{D}}(T) \leq O(T^{1/3})$, *if the attacker plays OMD, OH or OFTRL, or any other online learning method that satisfies the RVU property (Syrgkanis et al., 2015).*

Proof. The first point is obvious. For Hedge, we have from the previous lemma that $\sum_{t=1}^{T}\|\phi_t-\phi_{t-1}\|_1^2 \leq O(\eta\sqrt{T})$. Then, letting $\eta^{\mathcal{D}} = \eta^{\mathcal{A}} \leq O(\log T/T^{1/4})$, we obtain $Reg^*_{\mathcal{D}}(T) \leq O(T^{1/4})$.

For the rest, we let $\eta^{\mathcal{D}} = \eta^{\mathcal{A}} \leq O(\log T/T^{1/4})$, in order to obtain the desired bounds. □

D PROOFS OF RESULTS IN SECTION 4

In this section we provide the proofs of Proposition 1 and Proposition 2.

D.1 Proof Of Proposition 1

Statement. *For every $\mathcal{D}_1 \in \mathfrak{D}_1$, there exists a sequence of MDPs $\overline{M}_1, \dots, \overline{M}_T$ and a defense $\mathcal{D} \in \mathfrak{D}$, under which $\mathcal{D}_1$ incurs linear relative regret with respect to $\mathcal{D}$, under no attack. Moreover, for every $\mathcal{D}_1 \in \mathfrak{D}_1$, there exists a sequence of MDPs $\overline{M}_1, \dots, \overline{M}_T$, a defense $\mathcal{D} \in \mathfrak{D}$, and an attack $\mathcal{A} \in \mathfrak{A}$, under which $\mathcal{D}_1$ incurs linear relative regret with respect to $\mathcal{D}$. For such a defense, we have $\Delta^{\mathcal{D}_1} \geq \Delta^{opt} + C_1 T$, for some $C_1 > 0$.*

Proof. Let $\mathcal{D}_1 \in \mathfrak{D}_1$. We start by proving the first part of the statement. Let $M = (S, A, R, P, \gamma, \nu)$ with state space $S = \{s_0, s_{terminal}\}$, action space $A = \{a_1, a_2\}$, transitions $P(s_0, a_i, s_{terminal}) = 1$, for $i \in \{1, 2\}$, discount factor $\gamma = 1$ and initial state distribution $\nu(s_0) = 1$.

Let $\overline{R}$ be the reward vector that takes values $[1 \;\; 0]^T$ or $[0 \;\; 1]^T$, uniformly at random. Let $\mathcal{D} \in \mathfrak{D}$ with action set $\{\theta_1, \dots, \theta_H\}$ and let us pick $\theta^* \in \Theta$ such that $\mathcal{D}(\overline{R}, \theta^*) = \pi^*$, without loss of generality. Note that we have $\theta_{max} = \theta^*$, and thus, for every sample $\overline{R}_t$ of $\overline{R}$, it holds that $\pi_t^{opt} = \arg\max_i \overline{R}_t[i]$.

On the other hand, since $\mathcal{D}_1$ does not depend on $\widehat{R}_1, \dots, \widehat{R}_T$, the best strategy would be to pick each a_i with equal probability. Thus, we have

$$\begin{aligned} Reg_{\mathcal{D}_1}(T) &= \max_\theta \sum_{t=1}^T \rho(\mathcal{D}(\widehat{R}_t, \theta), \overline{R}_t) - \mathbb{E}\sum_{t=1}^T \rho(\mathcal{D}_1(\overline{R}_1, \dots, \overline{R}_{t-1}), \overline{R}_t) \\ &= \sum_{t=1}^T \rho(\pi_t^{opt}, \overline{R}_t) - \mathbb{E}\sum_{t=1}^T \rho(\pi_t^{\mathcal{D}_1}, \overline{R}_t) \\ &= \sum_{t=1}^T \max_{i=1,2} \overline{R}_t[i] - \sum_{t=1}^T \overline{R}_t[\pi_t^{\mathcal{D}_1}] \\ &= T - \frac{T}{2} = \frac{T}{2} \; . \end{aligned}$$

For the second part of the statement, note that an adaptive attack $\mathcal{A} \in \mathfrak{A}$ can always choose not to attack, and thus, the set of adaptive attacks contains the no attack case. Therefore, any $\mathcal{D}_1 \in \mathfrak{D}_1$ suffers linear regret under any $\mathcal{A} \in \mathfrak{A}$ by extension.

To give a characterization of the optimality gap, first recall that $\Delta^{\mathcal{D}_1} = \sum_{t=1}^T \rho(\pi_t^*, \overline{R}_t) - \mathbb{E}\rho(\pi_t^{\mathcal{D}_1}, \overline{R}_t)$. Now we can write

$$\begin{aligned} \Delta^{\mathcal{D}_1} &= \sum_{t=1}^T \rho(\pi_t^*, \overline{R}_t) - \mathbb{E}\rho(\pi_t^{\mathcal{D}_1}, \overline{R}_t) \\ &= \sum_{t=1}^T \left(\rho(\pi_t^*, \overline{R}_t) - \rho(\pi_t^{opt}, \overline{R}_t)\right) + \sum_{t=1}^T \Big(\rho(\pi_t^{opt}, \overline{R}_t) - \mathbb{E}\rho(\pi_t^{\mathcal{D}_1}, \overline{R}_t)\Big) \\ &= \Delta^{opt} + Reg_{\mathcal{D}_1}(T) \\ &\geq \Delta^{opt} + \Omega(T) \; , \end{aligned}$$

where the last inequality follows from above. □

D.2 Proof Of Proposition 2

Statement. *Assume that the attacker can change its attack model over time. Then, for every $\mathcal{D}_2 \in \mathfrak{D}_2$, there exists a sequence of MDPs $\overline{M}_1, \dots, \overline{M}_T$, attack models $\mathcal{A}_1, \dots, \mathcal{A}_T \in \{\mathcal{A}', \mathcal{A}''\} \in \mathfrak{A}$, and defense $\mathcal{D} \in \mathfrak{D}$, such that $\mathcal{D}_2$ incurs linear relative regret with respect to $\mathcal{D}$. Moreover, we have $\Delta^{\mathcal{D}_2} \geq \Delta^{opt} + C_2 T$, for some positive C_2.*

Proof. We will again construct an example to make the case. Let $M = (S, A, R, P, \gamma, \nu)$ with state space $S = \{s_0, s_{terminal}\}$, action space $A = \{a_1, a_2, a_3\}$, transitions $P(s_0, a_i, s_{terminal}) = 1$, for $i \in \{1, 2, 3\}$, discount factor $\gamma = 1$ and initial state distribution $\nu(s_0) = 1$.

Let us define the reward vector as a sample from the set $\mathfrak{R} = \{[0.7 \;\; 0.5 \;\; 0.3]^T, [0.5 \;\; 0.3 \;\; 0.7]^T\}$. Assume that, in every round t, the reward function $\overline{R}_t$ is a sample from $\mathfrak{R}$ uniformly at random.

Now let us define attack models $\mathcal{A}', \mathcal{A}'' \in \mathfrak{A}$, both with action set $\mathcal{E} = \{\epsilon_1, \epsilon_2\}$, such that

$$\mathcal{A}'([0.7 \;\; 0.5 \;\; 0.3]^T, \epsilon_1) = \mathcal{A}''([0.5 \;\; 0.3 \;\; 0.7]^T, \epsilon_2) = \widehat{R} = [0.5 \;\; 0.7 \;\; 0.3]^T \; .$$

Note that, in this example, $\pi_\dagger = a_2$. Assume that, given horizon T, the attacker employs $\mathcal{A}'$ in $T/2$ of the rounds and $\mathcal{A}''$ in the rest. Also, assume $\widehat{R} = \widehat{R}_t$, for all $t \leq T$.

On the other hand, let $\mathcal{D} \in \mathfrak{D}$ with action set $\Theta = \{\theta\}$, such that

$$\mathcal{D}(\mathcal{A}'([0.7 \;\; 0.5 \;\; 0.3]^T, \epsilon_1), \theta) = [0.7 \;\; 0.5 \;\; 0.3]^T \; ,$$
$$\mathcal{D}(\mathcal{A}''([0.5 \;\; 0.3 \;\; 0.7]^T, \epsilon_2), \theta) = [0.5 \;\; 0.3 \;\; 0.7]^T \; .$$

Now let $\mathcal{D}_2 \in \mathfrak{D}_2$. Note that $\mathcal{D}_2$ depends only on past observations of reward functions $\overline{R}_1, \ldots, \overline{R}_{t-1}$ and $\widehat{R}$. Furthermore, by assumption, $\mathcal{D}_2$ cannot tell which of the attack models is present at time t. Therefore, since the sequence of reward functions is selected by the environment uniformly at random and the sequence of attacks is adversarially selected by the attacker, the best $\mathcal{D}_2$ can do is to pick a_1 or a_3 uniformly at random.

The optimal defense, on the other hand, knowing the attack structure and which attack is employed in round t, can use the defense parameter θ in order to always recover the true reward function. Thus, using the original definition of the defense $\mathcal{D}$, where the dependence on $\mathcal{A}_t$ is explicit, we obtain

$$\begin{aligned}
Reg_{\mathcal{D}_2}(T) &= \max_{\theta} \sum_{t=1}^{T} \rho(\mathcal{D}(\widehat{R}_t, \theta, \mathcal{A}_t), \overline{R}_t) - \mathbb{E} \sum_{t=1}^{T} \rho(\mathcal{D}_2(\widehat{R}, \overline{R}_1, \ldots, \overline{R}_{t-1}), \overline{R}_t) \\
&= \sum_{t=1}^{T} \rho(\pi_t^{opt}, \overline{R}_t) - \mathbb{E} \sum_{t=1}^{T} \rho(\pi_t^{\mathcal{D}_2}, \overline{R}_t) \\
&= \sum_{t=1}^{T} \max_{i=1,2,3} \overline{R}_t[i] - \sum_{t=1}^{T} \overline{R}_t[\pi_t^{\mathcal{D}_2}] \\
&= 0.7T - \left(\frac{T}{2} 0.7 + \frac{T}{4} 0.3 + \frac{T}{4} 0.5 \right) \\
&= 0.55T \; .
\end{aligned}$$

The optimality gap is similarly computed. □