Admissible Policy Teaching through Reward Design

Kiarash Banihashem, Adish Singla, Jiarui Gan, Goran Radanovic

Max Planck Institute for Software Systems
{kbanihas, adishs, jrgan, gradanovic}@mpi-sws.org

Abstract

We study reward design strategies for incentivizing a reinforcement learning agent to adopt a policy from a set of admissible policies. The goal of the reward designer is to modify the underlying reward function cost-efficiently while ensuring that any approximately optimal deterministic policy under the new reward function is admissible and performs well under the original reward function. This problem can be viewed as a dual to the problem of optimal reward poisoning attacks: instead of forcing an agent to adopt a specific policy, the reward designer incentivizes an agent to avoid taking actions that are inadmissible in certain states. Perhaps surprisingly, and in contrast to the problem of optimal reward poisoning attacks, we first show that the reward design problem for admissible policy teaching is computationally challenging, and it is NP-hard to find an approximately optimal reward modification. We then proceed by formulating a surrogate problem whose optimal solution approximates the optimal solution to the reward design problem in our setting, but is more amenable to optimization techniques and analysis. For this surrogate problem, we present characterization results that provide bounds on the value of the optimal solution. Finally, we design a local search algorithm to solve the surrogate problem and showcase its utility using simulation-based experiments.

Introduction

Reinforcement learning (RL) (Sutton and Barto 2018) is a framework for deriving an agent's policy that maximizes its utility in sequential decision making tasks. In the standard formulation, the utility of an agent is defined via its reward function, which determines the decision making task of interest. Reward design plays a critical role in providing sound specifications of the task goals and supporting the agent's learning process (Singh, Lewis, and Barto 2009; Amodei et al. 2016).

There are different perspectives on reward design, which differ in the studied objectives. A notable example of reward design is *reward shaping* (Mataric 1994; Dorigo and Colombetti 1994; Ng, Harada, and Russell 1999) which modifies the reward function in order to accelerate the learning process of an agent. Reward transformations that are similar to or are based on reward shaping are not only used for accelerating learning. For example, reward penalties are often used in safe RL to penalize the agent whenever it violates safety constraints (Tessler, Mankowitz, and Mannor 2018). Similarly, reward penalties can be used in offline RL for ensuring robustness against model uncertainty (Yu et al. 2020), while exploration bonuses can be used as intrinsic motivation for an RL agent to reduce uncertainty (Bellemare et al. 2016).

In this paper, we consider a different perspective on reward design, and study it in the context of *policy teaching* and closely related (targeted) *reward poisoning attacks*. In this line of work (Zhang and Parkes 2008; Zhang, Parkes, and Chen 2009; Ma et al. 2019; Rakhsha et al. 2020b,a), the reward designer perturbs the original reward function to influence the choice of policy adopted by an optimal agent. For instance, (Zhang and Parkes 2008; Zhang, Parkes, and Chen 2009) studied policy teaching from a principal's perspective who provides incentives to an agent to influence its policy. In reward poisoning attacks (Ma et al. 2019; Rakhsha et al. 2020b,a), an attacker modifies the reward function with the goal of forcing a specific target policy of interest. Importantly, the reward modifications do not come for free, and the goal in this line of work is to alter the original reward function in a cost-efficient manner. The associated cost can, e.g., model the objective of minimizing additional incentives provided by the principal or ensuring the stealthiness of the attack.

The focus of this paper is on a dual problem to reward poisoning attacks. Instead of forcing a specific target policy, the reward designer's goal is to incentivize an agent to avoid taking actions that are inadmissible in certain states, while ensuring that the agent performs well under the original reward function. As in reward poisoning attacks, the reward designer cares about the cost of modifying the original reward function. Interestingly and perhaps surprisingly, the novel reward design problem leads to a considerably different characterization results, as we show in this paper. We call this problem *admissible policy teaching* since the reward designer aims to maximize the agent's utility w.r.t. the original reward function, but under constraints on admissibility of state-action pairs. These constraints could encode additional knowledge that the reward designer has about the safety and security of executing certain actions. Our key contributions are:

- We develop a novel optimization framework based on Markov Decision Processes (MDPs) for finding a minimal reward modifications which ensure that an optimal agent adopts a well-performing admissible policy.
- We show that finding an optimal solution to the reward design problem for admissible policy teaching is computationally challenging, in particular, that it is NP-hard to

find a solution that approximates the optimal solution.

- We provide characterization results for a surrogate problem whose optimal solution approximates the optimal solution to our reward design problem. For a specific class of MDPs, which we call *special* MDPs, we present an exact characterization of the optimal solution. For *general* MDPs, we provide bounds on the optimal solution value.
- We design a local search algorithm for solving the surrogate problem, and demonstrate its efficacy using simulation-based experiments.

Related Work

Reward design. A considerable number of works is related to designing reward functions that improve an agent's learning procedures. The optimal reward problem focuses on finding a reward function that can support computationally bounded agents (Sorg, Singh, and Lewis 2010; Sorg, Lewis, and Singh 2010). Reward shaping (Mataric 1994; Dorigo and Colombetti 1994), and in particular, potential-based reward shaping (Ng, Harada, and Russell 1999) and its extensions (e.g., (Devlin and Kudenko 2012; Grześ 2017; Zou et al. 2019)) densify the reward function so that the agent receives more immediate signals about its performance, and hence learns faster. As already mentioned, similar reward transformations, such as reward penalties or bonuses, are often used for reducing uncertainty or for ensuring safety constraints (Bellemare et al. 2016; Yu et al. 2020; Tessler, Mankowitz, and Mannor 2018). Related to safe and secure RL are works that study reward specification problem and negative side affects of reward misspecification (Amodei et al. 2016; Hadfield-Menell et al. 2017). The key difference between the above papers and our work is that we focus on policy teaching rather than on an agent's learning procedures.

Teaching and steering. As already explained, our work relates to prior work on policy teaching and targeted reward poisoning attacks (Zhang and Parkes 2008; Zhang, Parkes, and Chen 2009; Ma et al. 2019; Huang and Zhu 2019; Rakhsha et al. 2020b,a; Zhang et al. 2020b; Sun, Huo, and Huang 2021). Another line of related work is on designing steering strategies. For example, (Nikolaidis et al. 2017; Dimitrakakis et al. 2017; Radanovic et al. 2019) consider two-agent collaborative settings where a dominant agent can exert influence on the other agent, and the goal is to design a policy for the dominant agent that accounts for the imperfections of the other agent. Similar support mechanisms based on providing advice or helpful interventions have been studied by (Amir et al. 2016; Omidshafiei et al. 2019; Tylkin, Radanovic, and Parkes 2021). In contrast, we consider steering strategies based on reward design. When viewed as a framework for supporting an agent's decision making, this paper is also related to works on designing agents that are robust against adversaries (Pinto et al. 2017; Fischer et al. 2019; Lykouris et al. 2019; Zhang et al. 2020a, 2021a,b; Banihashem, Singla, and Radanovic 2021). These works focus on agent design and are complementary to our work on reward design.

Problem Setup

In this section we formally describe our problem setup.

Environment

The environment in our setting is described by a discrete-time Markov Decision Process (MDP) $M = (S, A, R, P, \gamma, \sigma)$, where S and A are the discrete finite state and action spaces respectively[1], $R : S \times A \to \mathbb{R}$ is a reward function, $P : S \times A \times S \to [0, 1]$ specifies the transition dynamics with $P(s, a, s')$ denoting the probability of transitioning to state s' from state s by taking action a, $\gamma \in [0, 1)$ is the discounted factor, and σ is the initial state distribution. A deterministic policy π is a mapping from states to actions, i.e., $\pi : S \to A$, and the set of all deterministic policies is denoted by Π_{det}.

Next, we define standard quantities in this MDP, which will be important for our analysis. First, we define the *score* of a policy π as the total expected return scaled by $1 - \gamma$, i.e., $\rho^{\pi,R} = \mathbb{E}\left[(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}R(s_t, a_t)|\pi, \sigma\right]$. Here states s_t and actions a_t are obtained by executing policy π starting from state s_1, which is sampled from the initial state distribution σ. Score $\rho^{\pi,R}$ can be obtained through state occupancy measure μ^π by using the equation $\rho^{\pi,R} = \sum_s \mu^\pi(s) \cdot R(s, \pi(s))$. Here, μ^π is the expected discounted state visitation frequency when π is executed, given by $\mu^\pi(s) = \mathbb{E}\left[(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{1}\left[s_t = s\right]|\pi, \sigma\right]$. Note that $\mu^\pi(s)$ can be equal to 0 for some states. Furthermore, we define $\mu^\pi_{\text{min}} = \min_{s|\mu^\pi(s)>0}\mu^\pi(s)$—the minimum always exists due to the finite state and action spaces. Similarly, we denote by μ_{min} the minimal value of μ^π_{min} across all deterministic policies, i.e., $\mu_{\text{min}} = \min_{\pi\in\Pi_{\text{det}}}\mu^\pi_{\text{min}}$.

We define the state-action value function, or Q values as $Q^{\pi,R}(s, a) = \mathbb{E}\left[\sum_{t=1}^{\infty}\gamma^{t-1}R(s_t, a_t)|\pi, s_1 = s, a_1 = a\right]$, where states s_t and actions a_t are obtained by executing policy π starting from state $s_1 = s$ in which action $a_1 = a$ is taken. State-action values $Q(s, a)$ relate to score ρ via the equation $\rho^{\pi,R} = \mathbb{E}_{s\sim\sigma}\left[(1-\gamma)\cdot Q^{\pi,R}(s, \pi(s))\right]$, where the expectation is taken over possible starting states.

Agent and Reward Functions

We consider a reinforcement learning agent whose behavior is specified by a deterministic policy π, derived offline using an MDP model given to the agent. We assume that the agent selects a deterministic policy that (approximately) maximizes the agent's expected utility under the MDP model specified by a reward designer. In other words, given an access to the MDP $M = (S, A, R, P, \gamma, \sigma)$, the agent chooses a policy from the set $\text{OPT}^\epsilon_{\text{det}}(R) = \{\pi \in \Pi_{\text{det}} : \rho^{\pi,R} > \max_{\pi'\in\Pi_{\text{det}}}\rho^{\pi',R} - \epsilon\}$, where ϵ is a strictly positive number. It is important to note that the MDP model given to the agent might be different from the true MDP model of the environment. In this paper, we focus on the case when only the reward functions of these two MDPs (possibly) differ.

Therefore, in our notation, we differentiate the reward function that the reward designer specifies to the agent, denoting it by $\widehat{R}$, from the original reward function of the environment, denoting it by $\overline{R}$. A generic reward function is

[1]This setting can encode the case where states have different number of actions (e.g., by adding actions to the states with smaller number of actions and setting the reward of newly added state-action pairs to $-\infty$).

denoted by R, and is often used as a variable in our optimization problems. We also denote by π^* a deterministic policy that is optimal with respect to $\overline{R}$ for any starting state, i.e., $Q^{\pi^*,\overline{R}}(s,\pi^*(s)) = \max_{\pi\in\Pi_{\text{det}}} Q^{\pi,\overline{R}}(s,\pi(s))$ for all states s.

Reward Designer and Problem Formulation

We take the perspective of a reward designer whose goal is to design a reward function, $\widehat{R}$, such that the agent adopts a policy from a class of admissible deterministic policies $\Pi^{\text{adm}}_{\text{det}} \subseteq \Pi_{\text{det}}$. Ideally, the new reward function $\widehat{R}$ would be close to the original reward function $\overline{R}$, thus reducing the cost of the reward design. At the same time, the adopted policy should perform well under the original reward function $\overline{R}$, since this is the performance that the reward designer wants to optimize and represents the objective of the underlying task. As considered in related works (Ma et al. 2019; Rakhsha et al. 2020b,a), we measure the *cost* of the reward design by L_2 distance between the designed $\widehat{R}$ and the original reward function $\overline{R}$. Moreover, we measure the agent's performance with the score $\rho^{\pi,\overline{R}}$, where the agent's policy π is obtained w.r.t. the designed reward function $\widehat{R}$. Given the model of the agent discussed in the previous subsection, and assuming the worst-case scenario (w.r.t. the tie-breaking in the policy selection), the following optimization problem specifies the reward design problem for *admissible policy teaching* (APT):

$$\begin{aligned} \min_{R}\max_{\pi}\ & \left\|\overline{R}-R\right\|_2 - \lambda\cdot\rho^{\pi,\overline{R}} && \text{(P1-APT)}\\ \text{s.t.}\quad & \textsc{Opt}^{\epsilon}_{\text{det}}(R)\subseteq\Pi^{\text{adm}}_{\text{det}} \\ & \pi\in\textsc{Opt}^{\epsilon}_{\text{det}}(R), \end{aligned}$$

where $\lambda \geq 0$ is a trade-off factor. While in this problem formulation $\Pi^{\text{adm}}_{\text{det}}$ can be any set of policies, we will primarily focus on admissible policies that can be described by a set of admissible actions per state. More concretely, we define sets of admissible actions per state denoted by $A^{\text{adm}}_s \subseteq A$. Given these sets A^{adm}_s, the set of admissible policies will be identified as $\Pi^{\text{adm}}_{\text{det}} = \{\pi|\pi(s)\in A^{\text{adm}}_s \vee \mu^\pi(s)=0 \text{ for } s\in S\}$.[2] In other words, these policies must take admissible actions for states that have non-zero state occupancy measure.

We conclude this section by validating the soundness of the optimization problem (P1-APT). The following proposition shows that the optimal solution to the optimization problem (P1-APT) is always attainable.

Proposition 1. *If $\Pi^{\text{adm}}_{\text{det}}$ is not empty, there always exists an optimal solution to the optimization problem* (P1-APT).

In the following sections, we analyze computational aspects of this optimization problem, showing that it is intractable in general and providing characterization results that bound the value of solutions. The proofs of our results are provided in the full version of the paper.

[2]In practice, we can instead put the constraint that $\mu^\pi(s)$ is greater than or equal to some threshold. For small enough threshold, our characterization results qualitatively remain the same.

Computational Challenges

We start by analyzing computational challenges behind the optimization problem (P1-APT). To provide some intuition, let us first analyze a special case of (P1-APT) where $\lambda = 0$, which reduces to the following optimization problem:

$$\begin{aligned} \min_{R}\ & \left\|\overline{R}-R\right\|_2 && \text{(P2-APT}_{\lambda=0}\text{)}\\ \text{s.t.}\quad & \textsc{Opt}^{\epsilon}_{\text{det}}(R)\subseteq\Pi^{\text{adm}}_{\text{det}}. \end{aligned}$$

This special case of the optimization problem with $\lambda = 0$ is a generalization of the reward poisoning attack from (Rakhsha et al. 2020b,a). In fact, the reward poisoning attack of (Rakhsha et al. 2020a) can be written as

$$\begin{aligned} \min_{R}\ & \left\|\overline{R}-R\right\|_2 && \text{(P3-ATK)}\\ \text{s.t.}\quad & \textsc{Opt}^{\epsilon}_{\text{det}}(R)\subseteq\{\pi|\pi(s)=\pi_\dagger(s) \text{ if } \mu^{\pi_\dagger}(s)>0\}, \end{aligned}$$

where $\pi_\dagger$ is the target policy that the attacker wants to force. However, while (P3-ATK) is tractable in the setting of (Rakhsha et al. 2020a), the same is not true for (P2-APT$_{\lambda=0}$); see Remark 1. Intuitively, the difficulty of solving the optimization problem (P2-APT$_{\lambda=0}$) lies in the fact that the policy set $\Pi^{\text{adm}}_{\text{det}}$ (in the constraints of (P2-APT$_{\lambda=0}$)) can contain exponentially many policies. Since the optimization problem (P2-APT$_{\lambda=0}$) is a specific instance of the optimization problem (P1-APT), the latter problem is also computationally intractable. We formalize this result in the following theorem.

Theorem 1. *For any constant $p\in(0,1)$, it is NP-hard to distinguish between instances of* (P2-APT$_{\lambda=0}$) *that have optimal values at most ξ and instances that have optimal values larger than $\xi\cdot\sqrt{(|S|\cdot|A|)^{1-p}}$. The result holds even when the parameters ϵ and γ in* (P2-APT$_{\lambda=0}$) *are fixed to arbitrary values subject to $\epsilon>0$ and $\gamma\in(0,1)$.*

The proof of the theorem is based on a classical NP-complete problem called EXACT-3-SET-COVER (X3C) (Karp 1972; Garey and Johnson 1979). The result implies that it is unlikely (assuming that P = NP is unlikely) that there exists a polynomial-time algorithm that always outputs an approximate solution whose cost is at most $\sqrt{(|S|\cdot|A|)^{1-p}}$ times that of the optimal solution for some $p>0$.

We proceed by introducing a surrogate problem (P4-APT), which is more amenable to optimization techniques and analysis since the focus is put on optimizing over policies rather than reward functions. In particular, the optimization problem takes the following form:

$$\begin{aligned} \min_{\pi\in\Pi^{\text{adm}}_{\text{det}},R}\ & \left\|\overline{R}-R\right\|_2 - \lambda\cdot\rho^{\pi,\overline{R}} && \text{(P4-APT)}\\ \text{s.t.}\quad & \textsc{Opt}^{\epsilon}_{\text{det}}(R)\subseteq\{\pi'|\pi'(s)=\pi(s) \text{ if } \mu^{\pi}(s)>0\}. \end{aligned}$$

Note that (P4-APT) differs from (P3-ATK) in that it optimizes over all admissible policies, and it includes performance considerations in its objective. The result in Theorem 1 extends to this case as well, so the main computational challenge remains the same.

The following proposition shows that the solution to the surrogate problem (P4-APT) is an approximate solution to the optimization problem (P1-APT), with an additive bound. More precisely:

Proposition 2. *Let $\widehat{R}_1$ and $\widehat{R}_2$ be the optimal solutions to* (P1-APT) *and* (P4-APT) *respectively and let $l(R)$ be a function that outputs the objective of the optimization problem* (P1-APT)*, i.e.,*

$$l(R) = \max_{\pi \in \text{OPT}^{\epsilon}_{\text{det}}(R)} \left\| \overline{R} - R \right\|_2 - \lambda \rho^{\pi, \overline{R}}. \quad (1)$$

Then $\widehat{R}_2$ satisfies the constraints of (P1-APT)*, i.e, $\text{OPT}^{\epsilon}_{\text{det}}(\widehat{R}_2) \subseteq \Pi^{\text{adm}}_{\text{det}}$, and*

$$l(\widehat{R}_1) \leq l(\widehat{R}_2) \leq l(\widehat{R}_1) + \frac{\epsilon}{\mu_{\min}} \cdot \sqrt{|S| \cdot |A|}.$$

Due to this result, in the following sections, we focus on the optimization problem (P4-APT), and provide characterization for it. Using Proposition 2 we can obtain analogous results for the optimization problem (P1-APT).

Remark 1. *The optimization problem* (P3-ATK) *is a strictly more general version of the optimization problem studied in (Rakhsha et al. 2020a) since* (P3-ATK) *does not require $\mu^{\pi}(s) > 0$ for all π and s. This fact also implies that the algorithmic approach presented in (Rakhsha et al. 2020a) is not applicable in our case. hm with provable guarantees. We provide an efficient algorithm for finding an approximate solution to* (P3-ATK) *with provable guarantees in the full version of the paper.*

Characterization Results for Special MDPs

In this section, we consider a family of MDPs where an agent's actions do not influence transition dynamics, or more precisely, all the actions influence transition probabilities in the same way. In other words, the transition probabilities satisfy $P(s, a, s') = P(s, a', s')$, for all s, a, a', s'. We call this family of MDPs *special* MDPs, in contrast to *general* MDPs that are studied in the next section. Since an agent's actions do not influence the future, the agent can reason myopically when deciding on its policy. Therefore, the reward designer can also treat each state separately when reasoning about the cost of the reward design. Importantly, the hardness result from the previous section does not apply for this instance of our setting, so we can efficiently solve the optimization problems (P1-APT) and (P4-APT).

Forcing Myopic Policies

We first analyze the cost of forcing a target policy $\pi_{\dagger}$ in special MDPs. The following lemma plays a critical role in our analysis.

Lemma 1. *Consider a* special *MDP with reward function $\overline{R}$, and let $\pi^*_{\text{adm}}(s) = \arg\max_{a \in \Pi^{\text{adm}}_{\text{det}}} \overline{R}(s, a)$. Then the cost of the optimal solution to the optimization problem* (P3-ATK) *with $\pi_{\dagger} = \pi^*_{\text{adm}}$ is less than or equal to the cost of the optimal solution to the optimization problem* (P3-ATK) *with $\pi_{\dagger} = \pi$ for any $\pi \in \Pi^{\text{adm}}_{\text{det}}$.*

In other words, Lemma 1 states that in special MDPs it is easier to force policies that are myopically optimal (i.e., optimize w.r.t. the immediate reward) than any other policy in the admissible set $\Pi^{\text{adm}}_{\text{det}}$. This property is important for the optimization problem (P4-APT) since its objective includes the cost of forcing an admissible policy.

Analysis of the Reward Design Problem

We now turn to the reward design problem (P4-APT) and provide characterization results for its optimal solution. Before stating the result, we note that for special MDPs $\mu^{\pi}(s)$ is independent of policy π, so we denote it by $\mu(s)$.

Theorem 2. *Consider a special MDP with reward function $\overline{R}$. Define $\widehat{R}(s, a) = \overline{R}(s, a)$ for $\mu(s) = 0$ and otherwise*

$$\widehat{R}(s, a) = \begin{cases} x_s + \frac{\epsilon}{\mu(s)} & \text{if } a = \pi^*_{\text{adm}}(s) \\ x_s & \text{if } a \neq \pi^*_{\text{adm}}(s) \wedge \overline{R}(s, a) \geq x_s \\ \overline{R}(s, a) & \text{otherwise} \end{cases},$$

where x_s is the solution to the equation

$$\sum_{a \neq \pi^*_{\text{adm}}(s)} \left[\overline{R}(s, a) - x\right]^+ = x - \overline{R}(s, \pi^*_{\text{adm}}(s)) + \frac{\epsilon}{\mu(s)}.$$

Then, $(\pi^*_{\text{adm}}, \widehat{R})$ is an optimal solution to (P4-APT)*.*

Theorem 2 provides an interpretable solution to (P4-APT): for each state-action pair $(s, a \neq \pi^*_{\text{adm}}(s))$ we reduce the corresponding reward $\overline{R}(s, a)$ if it exceeds a state dependent threshold. Likewise, we increase the rewards $\overline{R}(s, \pi^*_{\text{adm}}(s))$.

Characterization Results for General MDPs

In this section, we extend the characterization results from the previous section to general MDPs for which transition probabilities can depend on actions. In contrast to the previous section, the computational complexity result from Theorem 1 showcase the challenge of deriving characterization results for general MDPs that specify the form of an optimal solution. We instead focus on bounding the value of an optimal solution to (P4-APT) relative to the score of an optimal policy π^*. More specifically, we define the relative value Φ as

$$\Phi = \underbrace{\left\| \overline{R} - \widehat{R}_2 \right\|_2}_{\text{cost}} + \lambda \cdot \underbrace{[\rho^{\pi^*, \overline{R}} - \rho^{\pi_2, \overline{R}}]}_{\text{performance reduction}},$$

where $(\pi_2, \widehat{R}_2)$ is an optimal solution to the optimization problem (P4-APT). Intuitively, Φ expresses the optimal value of (P4-APT) in terms of the cost of the reward design and the agent's performance reduction.

The characterization results in this section provide bounds on Φ and are obtained by analyzing two specific policies: an optimal admissible policy $\pi^*_{\text{adm}} \in \arg\max_{\pi \in \Pi^{\text{adm}}_{\text{det}}} \rho^{\pi, \overline{R}}$ that optimizes for performance ρ, and a min-cost policy $\pi_{\min_c}$ that minimizes the cost of the reward design and is a solution to the optimization problem (P4-APT) with $\lambda = 0$. As we show in the next two subsections, bounding the cost of forcing π^*_{adm} and $\pi_{\min_c}$ can be used for deriving bounds on Φ. Next, we utilize the insights of the characterization results to devise a local search algorithm for solving the reward design problem, whose utility we showcase using experiments.

Perspective 1: Optimal Admissible Policy

Let us consider an optimal admissible policy $\pi^*_{\text{adm}} \in \arg\max_{\pi \in \Pi^{\text{adm}}_{\text{det}}} \rho^{\pi, \overline{R}}$. Following the approach presented in

the previous section, we can design $\widehat{R}$ by (approximately) solving the optimization problem (P3-ATK) (see Remark 1) with the target policy $\pi_\dagger = \pi^*_{\text{adm}}$. While this approach does not yield an optimal solution for general MDPs, the cost of its solution can be bounded by a quantity that depends on the gap between the scores of an optimal policy π^* and an optimal admissible policy π^*_{adm}.

In particular, for any policy π we can define the performance gap as $\Delta^\pi_\rho = \rho^{\pi^*,\overline{R}} - \rho^{\pi,\overline{R}}$. As we will show, the cost of forcing policy π can be upper and lower bounded by terms that linearly depend on Δ^π_ρ. Consequently, this means that one can also bound Φ with terms that linearly depend on $\Delta_\rho = \min_\pi \Delta^\pi_\rho$, which is nothing else but the performance gap of $\pi = \pi^*_{\text{adm}}$. Formally, we obtain the following result.

Theorem 3. *The relative value Φ is bounded by*

$$\alpha_\rho \cdot \Delta_\rho \leq \Phi \leq \beta_\rho \cdot \Delta_\rho + \frac{\epsilon}{\mu_{\min}} \cdot \sqrt{|S| \cdot |A|},$$

where $\alpha_\rho = \left(\lambda + \frac{1-\gamma}{2}\right)$ and $\beta_\rho = \left(\lambda + \frac{1}{\mu_{\min}}\right)$.

Note that the bounds in the theorem can be efficiently computed from the MDP parameters. Moreover, the reward design approach based on forcing π^*_{adm} yields a solution to (P4-APT) whose value (relative to the score of π^*) satisfies the bounds in Theorem 3. We use this approach as a baseline.

Perspective 2: Min-Cost Admissible Policy

We now take a different perspective, and compare Φ to the cost of the reward design obtained by forcing the min-cost policy $\pi_{\min_c}$. Ideally, we would relate Φ to the the smallest cost that the reward designer can achieve. However, this cost is not efficiently computable (due to Theorem 1), making such a bound uninformative.

Instead, we consider Q values: as Ma et al. (2019) showed, the cost of forcing a policy can be upper and lower bounded by a quantity that depends on Q values. We introduce a similar quantity, denoted by Δ_Q and defined as

$$\Delta_Q = \min_{\pi \in \Pi^{\text{adm}}_{\text{det}}} \max_{s \in S^\pi_{\text{pos}}} \left(Q^{\pi^*,\overline{R}}(s, \pi^*(s)) - Q^{\pi^*,\overline{R}}(s, \pi(s))\right),$$

where $S^\pi_{\text{pos}} = \{s | \mu^\pi(s) > 0\}$ contains the set of states that policy π visits with strictly positive probability. In the full version of the paper, we present an algorithm called QGREEDY that efficiently computes Δ_Q. The QGREEDY algorithm also outputs a policy π_{qg} that solves the corresponding min-max optimization problem. By approximately solving the optimization problem (P3-ATK) with $\pi_\dagger = \pi_{\text{qg}}$, we can obtain reward function $\widehat{R}$ as a solution to the reward design problem. We use this approach as a baseline in our experiments, and also for deriving the bounds on Φ relative to Δ_Q provided in the following theorem.

Theorem 4. *The relative value Φ is bounded by*

$$\alpha_Q \cdot \Delta_Q \leq \Phi \leq \beta_Q \cdot \Delta_Q + \frac{\epsilon}{\mu_{\min}} \sqrt{|S| \cdot |A|},$$

where $\alpha_Q = \left(\lambda \cdot \mu_{\min} + \frac{1-\gamma}{2}\right)$ and $\beta_Q = \left(\lambda + \sqrt{|S|}\right)$.

The bounds in Theorem 4 are obtained by analyzing the cost of forcing policy π_{qg} and the score difference $(\rho^{\pi^*,\overline{R}} - \rho^{\pi_{\text{qg}},\overline{R}})$. The well-known relationship between $\rho^{\pi,R} - \rho^{\pi',R}$ and $Q^{\pi,R}$ for any two policies π, π' (e.g., see (Schulman et al. 2015)) relates the score difference $(\rho^{\pi^*,\overline{R}} - \rho^{\pi_{\text{qg}},\overline{R}})$ to $Q^{\pi^*,\overline{R}}$, so the crux of the analysis lies in upper and lower bounding the cost of forcing policy π_{qg}. To obtain the corresponding bounds, we utilize similar proof techniques to those presented in (Ma et al. 2019) (see Theorem 2 in their paper). Since the analysis focuses on π_{qg}, the approach based on forcing π_{qg} outputs a solution to (P4-APT) whose value (relative to the score of π^*) satisfies the bounds in Theorem 4.

Practical Algorithm: CONSTRAIN&OPTIMIZE

In the previous two subsections, we discussed characterization results for the relative value Φ by considering two specific cases: optimizing performance and minimizing cost. We now utilize the insights from the previous two subsections to derive a practical algorithm for solving (P4-APT). The algorithm is depicted in Algorithm 1, and it searches for a well performing policy with a small cost of forcing it.

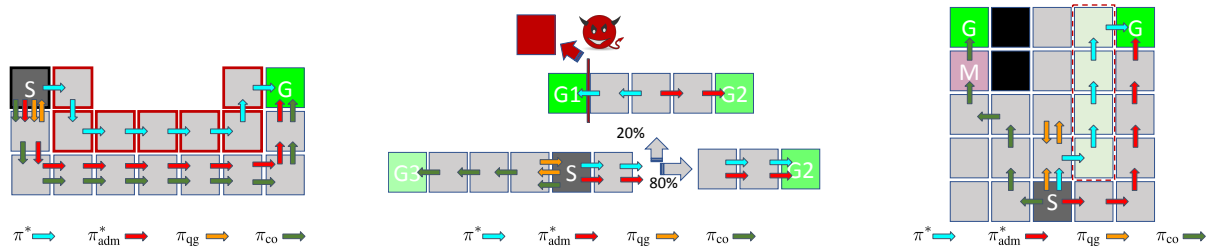
Algorithm 1: CONSTRAIN&OPTIMIZE

Input: MDP $\overline{M}$, admissible set $\Pi^{\text{adm}}_{\text{det}}$
Output: Reward function $\widehat{R}$, policy π_{co}

1: $\pi_{\text{co}} \leftarrow \arg\max_{\pi \in \Pi^{\text{adm}}_{\text{det}}} \rho^{\pi,\overline{R}}$
2: $cost_{\text{co}} \leftarrow$ approx. solve (P3-ATK) with $\pi_\dagger = \pi_{\text{co}}$
3: set $\Pi_{\text{co}} \leftarrow \Pi^{\text{adm}}_{\text{det}}$
4: **repeat**
5: output$_{\text{new}} \leftarrow$false
6: **for** s in *priority-queue*$(S^{\pi_{\text{co}}}_{\text{pos}})$ **do**
7: $\Pi' \leftarrow \{\pi | \pi \in \Pi_{\text{co}} \wedge \pi(s) \neq \pi_{\text{co}}(s)\}$
8: $\pi' \leftarrow \arg\max_{\pi \in \Pi'} \rho^{\pi,\overline{R}}$
9: $cost' \leftarrow$ approx. solve (P3-ATK) with $\pi_\dagger = \pi'$
10: **if** $cost' - \lambda\rho^{\pi',\overline{R}} < cost_{\text{co}} - \lambda\rho^{\pi_{\text{co}},\overline{R}}$ **then**
11: set $\pi_{\text{co}} \leftarrow \pi'$, $cost_{\text{co}} \leftarrow cost'$, and $\Pi_{\text{co}} \leftarrow \Pi'$
12: set output$_{\text{new}} \leftarrow$true and **break**
13: **end if**
14: **end for**
15: **until** output$_{\text{new}}$ = false
16: $\widehat{R} \leftarrow$ approx. solve (P3-ATK) with $\pi_\dagger = \pi_{\text{co}}$

The main blocks of the algorithm are as follows:

- **Initialization (lines 1-2).** The algorithm selects π^*_{adm} as its initial solution, i.e., $\pi_{\text{co}} = \pi^*_{\text{adm}}$, and evaluates its cost by approximately solving (P3-ATK).
- **Local search (lines 4-15).** Since the initial policy π_{co} is not necessarily cost effective, the algorithm proceeds with a local search in order to find a policy that has a lower value of the objective of (P4-APT). In each iteration of the local search procedure, it iterates over all states that are visited by the current π_{co} (i.e., $S^{\pi_{\text{co}}}_{\text{pos}}$), prioritizing those that have a higher value of $Q^{\pi^*,\overline{R}}(s, \pi^*(s)) - Q^{\pi^*,\overline{R}}(s, \pi_{\text{co}}(s))$ (obtained via *priority-queue*). The intuition behind this prioritization is that this Q value difference is reflective



(a) Cliff environment (b) Action hacking environment (c) Grass and mud environment

Figure 1: **Qualitative assessment.** (**a**) π_{co} is the same as π^*_{adm}, taking the longer path to the goal state than π^* in order to avoid the cliff edge, while π_{qg} simply alternates between the starting state and the state below it; (**b**) π_{co} takes the path to the goal state G3—this behavior is less costly to incentivize than navigating to G2 since in the former case the reward designer mainly needs to compensate for the difference in rewards between G3 and G2 (G1 is reachable only 20% of the time from S); (**c**) π^* takes a path through the grass states, π^*_{adm} takes a different but admissible path towards the same goal, while π_{co} navigates the agent towards the other goal (with a lower cumulative reward, but the corresponding policy is less costly to force). **Quantitative assessment.** The objective values of (P4-APT) for the approaches based on forcing π^*, π^*_{adm}, π_{qg}, and π_{co} are respectively: (**a**) 0.27, 1.59, 3.93, and 1.59; (**b**) -2.04, 14.96, 5.00, and 3.82; and (**c**) -9.54, 9.46, 17.26, and 7.92.

of the cost of forcing action $\pi_{\text{co}}(s)$ (as can be seen by setting $\lambda = 0$ in the upper bound of Theorem 4). Hence, deviations from π_{co} that are considered first are deviations from those actions that are expected to induce high cost.

- **Evaluating a neighbor solution (lines 7-12).** Each visited state s defines a neighbor solution in the local search. To find this neighbor, the algorithm first defines a new admissible set of policies Π' (line 7), obtained from the current one by making action $\pi_{\text{co}}(s)$ inadmissible. The neighbor solution is then identified as $\pi' \in \arg\max_{\pi \in \Pi'} \rho^{\pi, \overline{R}}$ (line 8) and the costs of forcing it is calculated by approximately solving (P3-ATK) with $\pi_\dagger = \pi'$ (line 9). If π' yields a better value of the objective of (P4-APT) than π_{co} does (line 10), we have a new candidate policy and the set of admissible policies is updated to Π' (lines 11-12).
- **Returning solution (line 16).** Once the local search finishes, the algorithm outputs π_{co} and the reward function $\widehat{R}$ found by approximately solving (P3-ATK) with $\pi_\dagger = \pi_{\text{co}}$.

In each iteration of the local search (lines 5-14), the algorithm either finds a new candidate (output$_{\text{new}}$=true) or the search finishes with that iteration (output$_{\text{new}}$=false). Notice that the former cannot go indefinitely since the admissible set reduces between two iterations. This means that the algorithm is guaranteed to halt. Since the local search only accepts new policy candidates if they are better than the current π_{co} (line 10), the output of CONSTRAIN&OPTIMIZE is guaranteed to be better than forcing an optimal admissible policy (i.e., approx. solving (P3-ATK) with $\pi_\dagger = \pi^*_{\text{adm}}$).

Numerical Simulations

We analyze the efficacy of CONSTRAIN&OPTIMIZE in solving the optimization problem (P4-APT) and the policy it incentivizes, π_{co}. We consider three baselines, all based on approximately solving the optimization problem (P3-ATK), but with different target policies $\pi_\dagger$: a) forcing an optimal policy, i.e., $\pi_\dagger = \pi^*$, b) forcing an optimal admissible policy, i.e., $\pi_\dagger = \pi^*_{\text{adm}}$, c) forcing the policy obtained by the QGREEDY algorithm, i.e., $\pi_\dagger = \pi_{\text{qg}}$.[3] We compare these approaches by measuring their performance w.r.t. the objective value of (P4-APT)—lower value is better. By default, we set the parameters $\gamma = 0.9$, $\lambda = 1.0$ and $\epsilon = 0.1$.

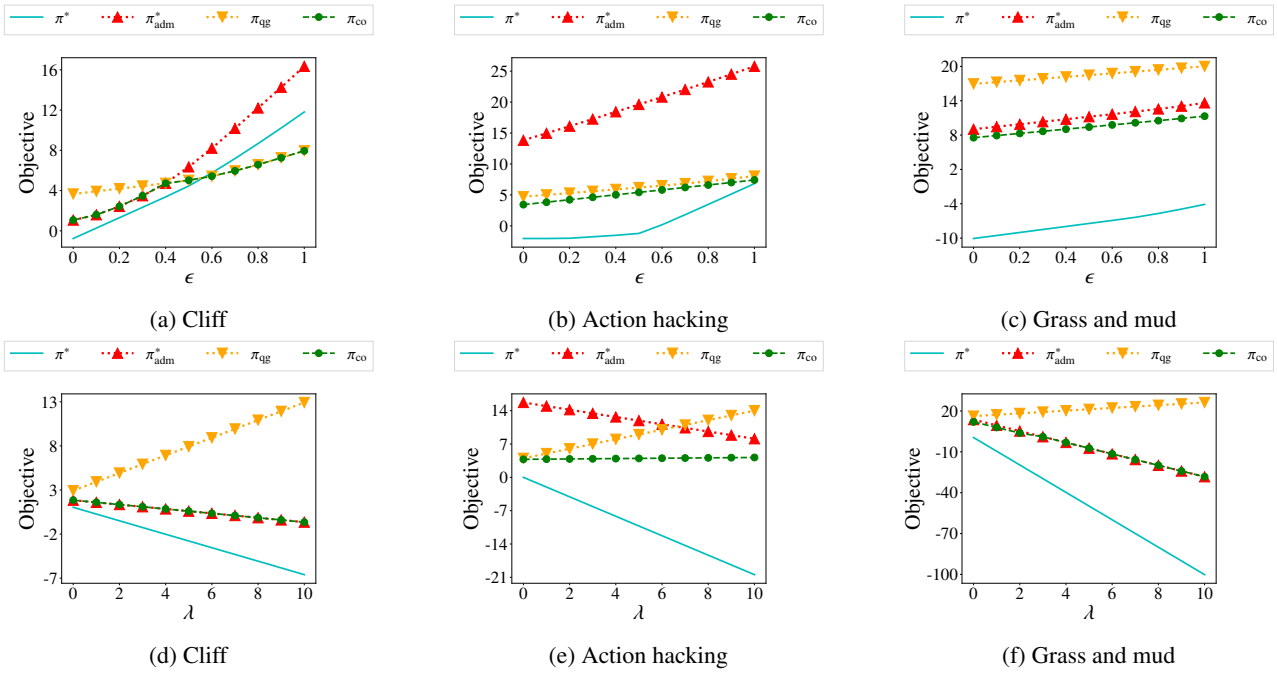
Experimental Testbeds

As an experimental testbed, we consider three simple navigation environments, shown in Figure 1. Each environment contains a start state S and goal state(s) G. Unless otherwise specified, in a non-goal state, the agent can navigate in the left, right, down, and up directions, provided there is a state in that direction. In goal states, the agent has a single action which transports it back to the start state.

Cliff environment (Figure 1a). This environment depicts a scenario where some of the states are potentially unsafe due to model uncertainty that the reward designer is aware of. More concretely, the states with "red" cell boundaries in Figure 1a represent the edges of a cliff and are unsafe; as such, all actions leading to these states are considered inadmissible. In this environment, the action in the goal state yields a reward $\overline{R}$ of 20 while all other actions yield a reward $\overline{R}$ of -1.

Action hacking environment (Figure 1b). This environment depicts a scenario when some of the agent's actions could be hacked at the deployment phase, taking the agent to a bad state. The reward designer is aware of this potential hacking and seeks to design a reward function so that these actions are inadmissible. More concretely, the action leading the agent to G1 is considered inadmissible. In this environment, we consider the reward function $\overline{R}$ and dynamics P as follows. Whenever an agent reaches any of the goal states (G1, G2, or G3), it has a single action that transports it back to the starting state and yields a reward of 50, 10, and 5 for G1, G2, and G3 respectively. In all other states, the agent can take either the left or right action and navigate in the corresponding direction, receiving a reward of -1. With a

[3] π^* might not be admissible; also, even though π^* is an optimal policy, there is still a cost of forcing it to create the required gap.



(a) Cliff (b) Action hacking (c) Grass and mud

(d) Cliff (e) Action hacking (f) Grass and mud

Figure 2: Effect of λ and ϵ on the objective value of (P4-APT) for different approaches. **(a, b, c)** vary ϵ with $\lambda = 1.0$; **(d, e, f)** vary λ with $\epsilon = 0.1$. Lower values on the y-axis denote better performance. Note that π^* is not an admissible policy in these environments; importantly, the objective value for π_{co} is consistently better (lower) than for π^*_{adm} and π_{qg}, highlighting its efficacy.

small probability of 0.20, taking the right action in the state next to S results in the agent moving up instead of right.

Grass and mud environment (Figure 1c). This environment depicts a policy teaching scenario where the reward designer and the agent do not have perfectly aligned preferences (e.g., the agent prefers to walk on grass, which the reward designer wants to preserve). The reward designer wants to incentivize the agent not to step on the grass states, so actions leading to them are considered inadmissible. In addition to the starting state and two goal states, the environment contains four grass states, one mud state, and 16 ordinary states, shown by "light green", "light pink", and "light gray" cells respectively. The "black" cells in the figure represent inaccessible blocks. The reward function $\overline{R}$ is as follows: the action in the goal states yields a reward of 50; the actions in the grass and mud states yield rewards of 10 and -2 respectively; all other actions have a reward of -1.

Results

Figure 1 provides an assessment of different approaches by visualizing the agent's policies obtained from the designed reward functions $\widehat{R}$. For these results, we set the parameters $\lambda = 1.0$ and $\epsilon = 0.1$. In order to better understand the effect of the parameters λ and ϵ, we vary these parameters and solve (P4-APT) with the considered approaches. The results are shown in Figure 2 for each environment separately. We make the following observations based on the experiments. First, the approaches based on forcing π^* and π^*_{adm} benefit more from increasing λ. This is expected as these two policies have the highest scores under $\overline{R}$; the scores of π_{qg} and π_{co} is less than or equal to the score of π^*_{adm}. Second, the approaches based on forcing π_{qg} and π_{co} are less susceptible to increasing ϵ. This effect is less obvious, and we attribute it to the fact that QGREEDY and CONSTRAIN&OPTIMIZE output π_{qg} and π_{co} respectively by accounting for the cost of forcing these policies. Since this cost clearly increases with ϵ—intuitively, forcing a larger optimality gap in (P3-ATK) requires larger reward modifications—we can expect that increasing ϵ deteriorates more the approaches based on forcing π^* and π^*_{adm}. Third, the objective value of (P4-APT) is consistently better (lower) for π_{co} than for π^*_{adm} and π_{qg}, highlighting the relevance of CONSTRAIN&OPTIMIZE.

Conclusion

The characterization results in this paper showcase the computational challenges of optimal reward design for admissible policy teaching. In particular, we showed that it is computationally challenging to find minimal reward perturbations that would incentivize an optimal agent into adopting a well-performing admissible policy. To address this challenge, we derived a local search algorithm that outperforms baselines which either account for only the agent's performance or for only the cost of the reward design. On the flip side, this algorithm is only applicable to tabular settings, so one of the most interesting research directions for future work would be to consider its extensions based on function approximation. In turn, this would also make the optimization framework of this paper more applicable to practical applications of interest, such as those related to safe and secure RL.

References

Amir, O.; Kamar, E.; Kolobov, A.; and Grosz, B. 2016. Interactive teaching strategies for agent training. In *IJCAI*, 804–811.

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *CoRR*, abs/1606.06565.

Banihashem, K.; Singla, A.; and Radanovic, G. 2021. Defense Against Reward Poisoning Attacks in Reinforcement Learning. *CoRR*, abs/2102.05776.

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016. Unifying count-based exploration and intrinsic motivation. *NeurIPS*, 29: 1471–1479.

Devlin, S. M.; and Kudenko, D. 2012. Dynamic potential-based reward shaping. In *AAMAS*, 433–440.

Dimitrakakis, C.; Parkes, D. C.; Radanovic, G.; and Tylkin, P. 2017. Multi-View Decision Processes: The Helper-AI Problem. In *NeurIPS*, 5443–5452.

Dorigo, M.; and Colombetti, M. 1994. Robot shaping: Developing autonomous agents through learning. *Artificial intelligence*, 71(2): 321–370.

Fischer, M.; Mirman, M.; Stalder, S.; and Vechev, M. 2019. Online robustness training for deep reinforcement learning. *CoRR*, abs/1911.00887.

Garey, M. R.; and Johnson, D. S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman.

Grześ, M. 2017. Reward Shaping in Episodic Reinforcement Learning. In *AAMAS*, 565–573.

Hadfield-Menell, D.; Milli, S.; Abbeel, P.; Russell, S.; and Dragan, A. D. 2017. Inverse reward design. In *NeurIPS*, 6768–6777.

Huang, Y.; and Zhu, Q. 2019. Deceptive Reinforcement Learning Under Adversarial Manipulations on Cost Signals. In *GameSec*, 217–237.

Karp, R. M. 1972. Reducibility among combinatorial problems. In *Complexity of computer computations*, 85–103. Springer.

Lykouris, T.; Simchowitz, M.; Slivkins, A.; and Sun, W. 2019. Corruption robust exploration in episodic reinforcement learning. *CoRR*, abs/1911.08689.

Ma, Y.; Zhang, X.; Sun, W.; and Zhu, J. 2019. Policy poisoning in batch reinforcement learning and control. In *NeurIPS*, 14543–14553.

Mataric, M. J. 1994. Reward functions for accelerated learning. In *ICML*, 181–189.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, 278–287.

Nikolaidis, S.; Nath, S.; Procaccia, A. D.; and Srinivasa, S. 2017. Game-theoretic modeling of human adaptation in human-robot collaboration. In *HRI*, 323–331.

Omidshafiei, S.; Kim, D.-K.; Liu, M.; Tesauro, G.; Riemer, M.; Amato, C.; Campbell, M.; and How, J. P. 2019. Learning to teach in cooperative multiagent reinforcement learning. In *AAAI*, 6128–6136.

Pinto, L.; Davidson, J.; Sukthankar, R.; and Gupta, A. 2017. Robust adversarial reinforcement learning. In *ICML*, 2817–2826.

Radanovic, G.; Devidze, R.; Parkes, D.; and Singla, A. 2019. Learning to collaborate in markov decision processes. In *ICML*, 5261–5270.

Rakhsha, A.; Radanovic, G.; Devidze, R.; Zhu, X.; and Singla, A. 2020a. Policy Teaching in Reinforcement Learning via Environment Poisoning Attacks. *CoRR*, abs/2011.10824.

Rakhsha, A.; Radanovic, G.; Devidze, R.; Zhu, X.; and Singla, A. 2020b. Policy Teaching via Environment Poisoning: Training-time Adversarial Attacks against Reinforcement Learning. In *ICML*, 7974–7984.

Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; and Moritz, P. 2015. Trust region policy optimization. In *ICML*, 1889–1897.

Singh, S.; Lewis, R. L.; and Barto, A. G. 2009. Where do rewards come from. In *the Annual Conference of the Cognitive Science Society*, 2601–2606.

Sorg, J.; Lewis, R. L.; and Singh, S. 2010. Reward design via online gradient ascent. *NeurIPS*, 2190–2198.

Sorg, J.; Singh, S.; and Lewis, R. 2010. Internal rewards mitigate agent boundedness. In *ICML*, 1007–1014.

Sun, Y.; Huo, D.; and Huang, F. 2021. Vulnerability-Aware Poisoning Mechanism for Online RL with Unknown Dynamics. In *ICLR*.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tessler, C.; Mankowitz, D. J.; and Mannor, S. 2018. Reward Constrained Policy Optimization. In *ICLR*.

Tylkin, P.; Radanovic, G.; and Parkes, D. C. 2021. Learning robust helpful behaviors in two-player cooperative Atari environments. In *AAMAS*, 1686–1688.

Yu, T.; Thomas, G.; Yu, L.; Ermon, S.; Zou, J. Y.; Levine, S.; Finn, C.; and Ma, T. 2020. MOPO: Model-based Offline Policy Optimization. In *NeurIPS*, 14129–14142.

Zhang, H.; Chen, H.; Boning, D.; and Hsieh, C.-J. 2021a. Robust reinforcement learning on state observations with learned optimal adversary. *CoRR*, abs/2101.08452.

Zhang, H.; Chen, H.; Xiao, C.; Li, B.; Boning, D.; and Hsieh, C.-J. 2020a. Robust deep reinforcement learning against adversarial perturbations on observations. *CoRR*, abs/2003.08938.

Zhang, H.; and Parkes, D. C. 2008. Value-Based Policy Teaching with Active Indirect Elicitation. In *AAAI*, 208–214.

Zhang, H.; Parkes, D. C.; and Chen, Y. 2009. Policy teaching through reward function learning. In *EC*, 295–304.

Zhang, X.; Chen, Y.; Zhu, X.; and Sun, W. 2021b. Robust policy gradient against strong data corruption. *CoRR*, abs/2102.05800.

Zhang, X.; Ma, Y.; Singla, A.; and Zhu, X. 2020b. Adaptive Reward-Poisoning Attacks against Reinforcement Learning. In *ICML*, 11225–11234.

Zou, H.; Ren, T.; Yan, D.; Su, H.; and Zhu, J. 2019. Reward shaping via meta-learning. *CoRR*, abs/1901.09330.

Appendix: Table of Contents

Appendix is structured according to the following sections:

- Section Background introduces additional quantities and lemmas relevant for the formal proofs.
- An approach for solving (P3-ATK), which is used in the experiments, is provided in Section Approximately Solving the Optimization Problem (P3-ATK). This section also analyzes this approach and provides provable guarantees.
- Section QGREEDY Algorithm provides the description of QGREEDY and shows that it is sound.
- The proof or Proposition 1 is given in section Proofs of the Results in Section Problem Setup.
- Theorem 1 and Proposition 2 are proven in section Proofs of the Results in Section Computational Challenges and Additional Results. The same section contains an additional hardness result, which proves that the optimization problem (P4-APT) is computationally hard.
- Proofs of Lemma 1 and Theorem 2 are in given in section Proofs of the Results from Section Characterization Results for Special MDPs.
- Proofs of Theorem 3 and Theorem 4 are provided in section Proofs of the Results in Section Characterization Results for General MDPs. The same section includes additional results relevant for proving the statements.

Background

As explained in the main paper, for a policy π and reward function R, we define its state-action value function $Q^{\pi,R}$ as

$$Q^{\pi,R}(s,a) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t) | \pi, s_1 = s, a_1 = a\right],$$

where states s_t and actions a_t are obtained by executing policy π starting from state $s_1 = s$ in which action $a_1 = a$ is taken. The state value function $V^{\pi,R}$ is similarly defined as

$$V^{\pi,R}(s) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t) | \pi, s_1 = s\right] = Q^{\pi,R}(s, \pi(s))$$

We define $Q^{*,R}$ and $V^{*,R}$ as the maximum of these values over all policies, i.e.,

$$Q^{*,R}(s,a) = \max_{\pi \in \Pi_{\text{det}}} Q^{\pi,R}(s,a)$$
$$V^{*,R}(s) = \max_{\pi \in \Pi_{\text{det}}} V^{\pi,R}(s)$$

The optimal policy in an MDP can be calculated by setting $\pi(s) \in \arg\max_a Q^{*,R}(s,a)$ and satisfies $Q^{\pi,R} = Q^{*,R}$. For $R = \overline{R}$, we denote this policy with π^*.
We define the state occupancy measure μ^π as

$$\mu^\pi(s) = \mathbb{E}\left[(1-\gamma)\sum_{t=1}^{\infty} \gamma^{t-1} \mathbb{1}\left[s_t = s\right] | \pi, \sigma\right].$$

μ^π can be efficiently calculated as it is the unique solution to the Bellman flow constraint

$$\mu^\pi(s) = (1-\gamma)\cdot\sigma(s) + \gamma \sum_{s'} P(s', \pi(s'), s)\mu^\pi(s'). \tag{2}$$

An important result that we utilize repeatedly in our proofs, is the following lemma that relates the score difference $\rho^{\pi_1,R} - \rho^{\pi_2,R}$ for two policies π_1, π_2 to their Q-values through the state occupancy measure μ.

Lemma 2. *(Schulman et al. 2015) Any two deterministic policies, π_1 and π_2, and reward function R satisfy:*

$$\rho^{\pi_1,R} - \rho^{\pi_2,R} = \sum_{s \in S} \mu^{\pi_1}(s)\big(Q^{\pi_2,R}(s, \pi_1(s)) - Q^{\pi_2,R}(s, \pi_2(s))\big).$$

For a policy π, we define S^π_{pos} as

$$S^\pi_{\text{pos}} = \{s | \mu^\pi(s) > 0\}. \tag{3}$$

We also prove the following lemma, which we use in several results in the following sections.

Lemma 3. *Let π, π' be deterministic policies such that $\pi(s) = \pi'(s)$ for all $s \in S^\pi_{\text{pos}} \cap S^{\pi'}_{\text{pos}}$. Then $\mu^\pi = \mu^{\pi'}$.*

Proof. **Part 1:** We first prove a simpler version of the Lemma; we assume that $\pi(s) = \pi'(s)$ for all $s \in S^{\pi}_{\text{pos}}$. We then show how to extend the result to the general case.

We prove by induction on t that for all states s,

$$\mathbb{P}\left[s_t = s|\pi\right] = \mathbb{P}\left[s_t = s|\pi'\right]$$

where s_t denotes the state visited at time t.

The claim holds for $t = 1$ as the initial probabilities are sampled from σ. Assuming the claim holds for t,

$$\begin{aligned}
\mathbb{P}\left[s_t = s|\pi\right] &= \sum_{s'} \mathbb{P}\left[s_{t-1} = s'|\pi\right] \mathbb{P}\left[s', \pi(s'), s\right] \\
&= \sum_{s':\mathbb{P}[s_{t-1}=s'|\pi]>0} \mathbb{P}\left[s_{t-1} = s'|\pi\right] \mathbb{P}\left[s', \pi(s'), s\right] \\
&\overset{(i)}{=} \sum_{s':\mathbb{P}[s_{t-1}=s'|\pi]>0} \mathbb{P}\left[s_{t-1} = s'|\pi'\right] \mathbb{P}\left[s', \pi(s'), s\right] \\
&\overset{(ii)}{=} \sum_{s':\mathbb{P}[s_{t-1}=s'|\pi]>0} \mathbb{P}\left[s_{t-1} = s'|\pi'\right] \mathbb{P}\left[s', \pi'(s'), s\right] \\
&\leq \sum_{s'} \mathbb{P}\left[s_{t-1} = s'|\pi'\right] \mathbb{P}\left[s', \pi'(s'), s\right] \\
&= \mathbb{P}\left[s_t = s|\pi'\right],
\end{aligned}$$

where (i) follows from the induction hypotheses and (ii) follows from the fact that if $\mathbb{P}\left[s_{t-1} = s'|\pi\right] > 0$, then $\mu^{\pi}(s') > 0$ and therefore $\pi(s) = \pi'(s)$. Since

$$\sum_{s} \mathbb{P}\left[s_t = s|\pi\right] = \sum_{s} \mathbb{P}\left[s_t = s|\pi'\right] = 1,$$

it follows that

$$\mathbb{P}\left[s_t = s|\pi\right] = \mathbb{P}\left[s_t = s|\pi'\right].$$

Therefore,

$$\mu^{\pi}(s) = \sum_{t=1}^{\infty} \mathbb{P}\left[s_t = s|\pi\right] = \sum_{t=1}^{\infty} \mathbb{P}\left[s_t = s|\pi'\right] = \mu^{\pi'}(s).$$

Part 2: Now, in order to obtain the general case, define $\tilde{\pi}$ as follows.

$$\tilde{\pi}(s) := \begin{cases} \pi(s) & \text{if} \quad s \in S^{\pi}_{\text{pos}} \\ \pi'(s) & \text{otherwise} \end{cases},$$

By the simpler version just proved, since $\tilde{\pi}(s) = \pi(s)$ for all $s \in S^{\pi}_{\text{pos}}$, $\mu^{\tilde{\pi}} = \mu^{\pi}$.

Furthermore, for all $s \in S^{\pi'}_{\text{pos}}$, either $s \in S^{\pi}_{\text{pos}}$, in which case $\pi(s) = \pi'(s)$ by assumption and therefore by definition of $\tilde{\pi} = \pi(s) = \pi'(s)$, or $s \notin S^{\pi}_{\text{pos}}$, in which case, $\tilde{\pi}(s) = \pi'(s)$. Since $\tilde{\pi}(s) = \pi'(s)$ in both cases, it follows that by the simpler version just proved, $\mu^{\tilde{\pi}} = \mu^{\pi'}$.

Therefore $\mu^{\pi} = \mu^{\tilde{\pi}} = \mu^{\pi'}$ as claimed. □

Approximately Solving the Optimization Problem (P3-ATK)

In this section, we show to efficiently approximate the optimization problem (P3-ATK). In order to obtain the approximate solution, we will consider the following optimization problem

$$\min_{R} \quad \left\|R - \overline{R}\right\|_2 \tag{P5-ATK}$$

$$\text{s.t.} \quad \forall s, a : Q(s,a) = R(s,a) + \gamma \sum_{s'} P(s,a,s')V(s') \tag{4}$$

$$\forall s \in S^{\pi_\dagger}_{\text{pos}}, a \neq \pi_\dagger(s) : Q(s, \pi_\dagger(s)) \geq Q(s,a) + \epsilon'(s,a) \tag{5}$$

$$\forall s \in S_{\text{pos}}^{\pi_\dagger} : V(s) = Q(s, \pi_\dagger(s)) \tag{6}$$
$$\forall s \notin S_{\text{pos}}^{\pi_\dagger}, a : V(s) \geq Q(s, a), \tag{7}$$

where $\epsilon'(s, a) \geq 0$ are arbitrary non-negative values that will be specified later. We first show that the constraints of the optimization problem effectively ensure that V and Q can be thought of as the $V^{*,R}$ and $Q^{*,R}$ vectors respectively. Note that this is not trivial since for $s \notin S_{\text{pos}}^{\pi_\dagger}$, the constraint $V(s) = \max_a Q(s, a)$ is not explicitly enforced. Formally, we have the following lemma.

Lemma 4. *Let $\epsilon'(s, a)$ be a non-negative vector. If the vectors (R, Q, V) satisfy the constraints of the optimization problem* (P5-ATK)*, the vectors $(R, Q^{*,R}, V^{*,R})$ satisfy the constraints as well.*

Proof. Starting with Q, V, we run the standard value iteration algorithm for finding $Q^{*,R}, V^{*,R}$ and claim that at the end of each step, all the constraints would still be satisfied. Concretely, We set $V^0 = V$ and $Q^0 = Q$ and for all $t \geq 0$:

$$V^{t+1}(s, a) = \max_a Q^t(s, a)$$
$$Q^{t+1}(s, a) = R(s, a) + \gamma \sum_{s'} P(s, a, s') V^{t+1}(s').$$

We claim that for all $t \geq 0$, the vectors (R, Q^t, V^t) satisfty the constraints (4) to (7). We prove that claim by induction on t.

For $t = 0$, the claim holds by assumption. Assume that the claim holds for $t - 1$; we will show that it holds for t as well by proving the constraints (4), (6), (5) and (7) repsectively. Constraint (4) holds by definition of $Q^t(s, a)$. For constraint (6), observe that for all $s \in S_{\text{pos}}^{\pi_\dagger}$,

$$\begin{aligned} V^t(s) &\overset{(i)}{=} \max_a Q^{t-1}(s, a) \\ &\overset{(ii)}{=} Q^{t-1}(s, \pi_\dagger(s)) \\ &\overset{(iii)}{=} V^{t-1}(s) \end{aligned} \tag{8}$$

where (i) follows from the definition of V^t, (ii) follows from (5) and (iii) follows from (6) for V^{t-1} and Q^{t-1}.

Now observe that if $s \in S_{\text{pos}}^{\pi_\dagger}$ and $s' \notin S_{\text{pos}}^{\pi_\dagger}$, then $P(s, \pi_\dagger(s), s') = 0$ as otherwise given (2), $\mu^{\pi_\dagger}(s')$ would be lower bounded by $\mu^{\pi_\dagger}(s) \cdot P(s, \pi_\dagger(s), s') > 0$, contradicting the assumption $s' \notin S_{\text{pos}}^{\pi_\dagger}$. Therefore, for all $s \in S_{\text{pos}}^{\pi_\dagger}$,

$$\begin{aligned} Q^t(s, \pi_\dagger(s)) &= R(s, \pi_\dagger(s)) + \gamma \sum_{s'} P(s, \pi_\dagger(s), s') V^t(s'). \\ &= R(s, \pi_\dagger(s)) + \gamma \sum_{s' \in S_{\text{pos}}^{\pi_\dagger}} P(s, \pi_\dagger(s), s') V^t(s'). \\ &\overset{(i)}{=} R(s, \pi_\dagger(s)) + \gamma \sum_{s' \in S_{\text{pos}}^{\pi_\dagger}} P(s, \pi_\dagger(s), s') V^{t-1}(s'). \\ &= R(s, \pi_\dagger(s)) + \gamma \sum_{s'} P(s, \pi_\dagger(s), s') V^{t-1}(s'). \\ &= Q^{t-1}(s, \pi_\dagger(s)), \end{aligned} \tag{9}$$

where (i) follows from (8). Together with (8), (9) implies that the constraint (6) still holds.

Now observe that for $s \notin S_{\text{pos}}^{\pi_\dagger}$,

$$V^t(s) = \max_a Q^{t-1}(s, a) \leq V^{t-1}(s),$$

where the inequality follows from the induction hypothesis; namely, constraint (7) for V^{t-1} and Q^{t-1}. This means that for all s (both when $s \notin S_{\text{pos}}^{\pi_\dagger}$ and when $s \in S_{\text{pos}}^{\pi_\dagger}$), $V^t(s) \leq V^{t-1}(s)$ and therefore given (4),

$$Q^t(s, a) \leq Q^{t-1}(s, a).$$

for all s, a. This implies that the constraint (5) still holds because the LHS has stayed the same and RHS hasn't increased. Finally, constraint (7) holds as well because it holds with V^t, Q^{t-1} by definition of V^t and $Q^t(s, a) \leq Q^{t-1}(s, a)$.

Since (Q^t, V^t) converge to $(Q^{*,R}, V^{*,R})$ and the constraints characterize a closed set, $(R, Q^{*,R}, V^{*,R})$ also satisfy the constraints. □

While the value of ϵ' can be arbitrary in (5), in our analysis we will mainly consider $\epsilon'_{\pi_\dagger}$, which for a non-negative number $\epsilon \geq 0$ and policy $\pi_\dagger$ we define as

$$\epsilon'_{\pi_\dagger}(\tilde{s},\tilde{a}) := \begin{cases} \dfrac{\epsilon}{\min_{\pi \in D(\pi_\dagger,\tilde{s},\tilde{a})} \mu^\pi(\tilde{s})} & \text{if} \quad \tilde{s} \in S^{\pi_\dagger}_{\text{pos}} \text{ and } \tilde{a} \neq \pi(s) \\ 0 & \text{otherwise.} \end{cases}, \tag{10}$$

where

$$D(\pi_\dagger,\tilde{s},\tilde{a}) = \big\{\pi : \pi(\tilde{s}) = \tilde{a} \text{ and } \pi(s) = \pi_\dagger(s) \text{ for all } s \in S^{\pi_\dagger}_{\text{pos}} \backslash \{\tilde{s}\}\big\}.$$

Of course, in order for the above definition to be valid, we need to ensure that the denominator is non-zero, i.e. $\min_{\pi \in D(\pi_\dagger,\tilde{s},\tilde{a})} \mu^\pi(\tilde{s}) > 0$. The following lemma ensures that this is the case.

Lemma 5. *Let $\pi_\dagger \in \Pi_{det}$ be a deterministic policy. Define $S^{\pi_\dagger}_{\text{pos}}$ as in* (3)*. Let $\tilde{s}$ be an arbitrary state in $S^{\pi_\dagger}_{\text{pos}}$ and π be a deterministic policy such that*

$$\pi(s) = \pi_\dagger(s) \quad \textit{for all} \quad s \in S^{\pi_\dagger}_{\text{pos}} \backslash \{\tilde{s}\}.$$

Then

$$\mu^\pi(\tilde{s}) > 0.$$

Proof. Assume that this is not the case and $\mu^\pi(\tilde{s}) = 0$. Then $s \notin S^\pi_{\text{pos}}$ and therefore $\pi(s) = \pi_\dagger(s)$ for all $S^{\pi_\dagger}_{\text{pos}} \cap S^\pi_{\text{pos}}$. Lemma 3 (from section Background) implies that $\mu^\pi = \mu^{\pi_\dagger}$ which is a contradiction since $\mu^{\pi_\dagger}(s) > 0 = \mu^\pi(\tilde{s})$. Therefore the initial assumption was wrong and $\mu^\pi(s) > 0$. □

Proposition 3. *Let ϵ be a non-negative number*[4]*. Denote by $\widehat{R}^{\pi_\dagger}$ the solution of the optimization problem* (P3-ATK) *and let $\widehat{R}'$ be the solution to* (P5-ATK) *with $\epsilon' = \epsilon'_{\pi_\dagger}$ where $\epsilon'_{\pi_\dagger}$ is defined as in Equation* (10)*. Then $\widehat{R}'$ satisfies the constraints of* (P3-ATK) *and*

$$0 \leq \left\|\widehat{R}' - \overline{R}\right\|_2 - \left\|\widehat{R}^{\pi_\dagger} - \overline{R}\right\|_2 \leq \|\epsilon'\|_2$$

Proof. Before we proceed with the proof, note that $\min_{\pi \in D(\pi_\dagger,\tilde{s},\tilde{a})} \mu^\pi(\tilde{s}) > 0$ because of Lemma 5.

Part 1: We first prove that $\widehat{R}'$ satisfies the constraints of (P3-ATK), which automatically proves the left inequality by optimality of $\widehat{R}^{\pi_\dagger}$. Set $Q = Q^{*,\widehat{R}'}$ and $V = V^{*,\widehat{R}'}$. Given Lemma 4, (R, Q, V) satisfy the constraints of (P5-ATK). Define $\tilde{\pi}_\dagger \in \Pi_{\text{det}}$ as

$$\tilde{\pi}_\dagger(s) = \arg\max Q(s,a). \tag{11}$$

It is clear that $Q^{\tilde{\pi}_\dagger,\widehat{R}'} = Q$. Now note that since $\arg\max Q(s,a) = \pi_\dagger(s)$ for all states $s \in S^{\pi_\dagger}_{\text{pos}}$, Lemma 2 implies that

$$\begin{aligned} \rho^{\pi_\dagger,\widehat{R}'} - \rho^{\tilde{\pi}_\dagger,\widehat{R}'} &= \sum_s \mu^{\pi_\dagger}(s)\big(Q(s,\pi_\dagger(s)) - Q(s,\tilde{\pi}_\dagger(s))\big) \\ &= \sum_{s \in S^{\pi_\dagger}_{\text{pos}}} \mu^{\pi_\dagger}(s)\big(Q(s,\pi_\dagger(s)) - Q(s,\tilde{\pi}_\dagger(s))\big) \\ &= 0. \end{aligned}$$

Given this, it suffices to prove that

$$\rho^{\tilde{\pi}_\dagger} \geq \rho^\pi + \epsilon \quad \text{if} \quad \exists s \in S^{\pi_\dagger}_{\text{pos}} : \pi(s) \neq \pi_\dagger(s),$$

Our proof now proceeds in a similar fashion to the proof of Lemma 1 in (Rakhsha et al. 2020b): We start by proving the claim policies that effectively, differ from $\pi_\dagger$ in only a single state. We then generalise the claim for other policies via induction.

Concretely, we first claim that if $\pi \in D(\pi_\dagger,\tilde{s},\tilde{a})$ for some $\tilde{s} \in S^{\pi_\dagger}_{\text{pos}}$ and $\tilde{a}$, then $\rho^{\tilde{\pi}_\dagger,\widehat{R}'} - \rho^{\pi,\widehat{R}'} \geq \epsilon$. To see why, note that

$$\begin{aligned} \rho^{\tilde{\pi}_\dagger,\widehat{R}'} - \rho^{\pi,\widehat{R}'} &= \sum_s \mu^\pi(s)\big(Q(s,\tilde{\pi}_\dagger(s)) - Q(s,\pi(s)\big) \\ &= \mu^\pi(\tilde{s})\big(Q(\tilde{s},\tilde{\pi}_\dagger(\tilde{s})) - Q(\tilde{s},\pi(\tilde{s})\big) + \sum_{s \neq \tilde{s}} \mu^\pi(s)\big(Q(s,\tilde{\pi}_\dagger(s)) - Q(s,\pi(s)\big) \end{aligned}$$

[4]The case of $\epsilon = 0$ is also covered by the lemma.

$$\overset{(i)}{\geq} \mu^{\pi}(\tilde{s})\big(Q(\tilde{s},\tilde{\pi}_{\dagger}(\tilde{s})) - Q(\tilde{s},\pi(\tilde{s}))\big)$$
$$\overset{(ii)}{\geq} \mu^{\pi}(\tilde{s})\epsilon'(\tilde{s},\pi(\tilde{s}))$$
$$\overset{(iii)}{\geq} \epsilon.$$

where (i) follows from the definition of $\tilde{\pi}_{\dagger}$, (ii) follows from (5) since $\tilde{s} \in S_{\text{pos}}^{\pi_{\dagger}}$ and (iii) follows from the definition of $\epsilon'_{\pi_{\dagger}}$ in (10).

We now generalize the above result by showing that if π is a policy such that there exists state $\tilde{s} \in S_{\text{pos}}^{\pi_{\dagger}}$ satisfying $\pi(\tilde{s}) \neq \pi_{\dagger}(\tilde{s})$, then $\rho^{\tilde{\pi}_{\dagger},\widehat{R}'} - \rho^{\pi,\widehat{R}'} \geq \epsilon$. We do this by induction on $d_{S_{\text{pos}}^{\pi_{\dagger}}}(\pi,\pi_{\dagger})$ where we define $d_{\tilde{S}}(\pi,\pi')$ for $\tilde{S} \subseteq S$ as

$$d_{\tilde{S}}(\pi,\pi') := \left|\left\{s \in \tilde{S} : \pi(s) \neq \pi'(s)\right\}\right|.$$

For $d_{S_{\text{pos}}^{\pi_{\dagger}}}(\pi,\pi_{\dagger}) = 1$, the claim is already proved since this is equivalent to $\pi \in D(\pi_{\dagger},\tilde{s},\tilde{a})$ for some $\tilde{s} \in S_{\text{pos}}^{\pi_{\dagger}}$ and $\tilde{a}$. Suppose the claim holds for all π satisfying $d_{S_{\text{pos}}^{\pi_{\dagger}}}(\pi,\pi_{\dagger}) \leq k$ where $k \geq 1$. We prove it holds for all π satisfying $d_{S_{\text{pos}}^{\pi_{\dagger}}}(\pi,\pi_{\dagger}) = k+1$. Let π be one such policy and note that by Lemma 2,

$$\rho^{\tilde{\pi}_{\dagger},\widehat{R}'} - \rho^{\pi,\widehat{R}'} = \sum_{s} \mu^{\pi}(s)\big(Q(s,\tilde{\pi}_{\dagger}(s)) - Q(s,\pi(s))\big) \geq 0.$$

On the other hand, again by Lemma 2,

$$\begin{aligned}
\rho^{\tilde{\pi}_{\dagger},\widehat{R}'} - \rho^{\pi,\widehat{R}'} &= \sum_{s} \mu^{\tilde{\pi}_{\dagger}}(s)\big(Q^{\pi,\widehat{R}'}(s,\tilde{\pi}_{\dagger}(s)) - Q^{\pi,\widehat{R}'}(s,\pi(s))\big) \\
&= \sum_{s \in S_{\text{pos}}^{\tilde{\pi}_{\dagger}}} \mu^{\tilde{\pi}_{\dagger}}(s)\big(Q^{\pi,\widehat{R}'}(s,\tilde{\pi}_{\dagger}(s)) - Q^{\pi,\widehat{R}'}(s,\pi(s))\big). \\
&\overset{(i)}{=} \sum_{s \in S_{\text{pos}}^{\pi_{\dagger}}} \mu^{\tilde{\pi}_{\dagger}}(s)\big(Q^{\pi,\widehat{R}'}(s,\tilde{\pi}_{\dagger}(s)) - Q^{\pi,\widehat{R}'}(s,\pi(s))\big).
\end{aligned}$$

Where (i) follows from the fact that $S_{\text{pos}}^{\pi_{\dagger}} = S_{\text{pos}}^{\tilde{\pi}_{\dagger}}$ by Lemma 3. Therefore, there exists a state $s \in S_{\text{pos}}^{\pi_{\dagger}}$ such that

$$Q^{\pi,\widehat{R}'}(s,\tilde{\pi}_{\dagger}(s)) - Q^{\pi,\widehat{R}'}(s,\pi(s)) \geq 0.$$

Define the policy $\tilde{\pi}$ as

$$\tilde{\pi}(\tilde{s}) = \begin{cases} \pi_{\dagger}(\tilde{s}) & \text{if} \quad \tilde{s} = s \\ \pi(\tilde{s}) & \text{otherwise} \end{cases}.$$

It follows that

$$\rho^{\tilde{\pi},\widehat{R}'} - \rho^{\pi,\widehat{R}'} = \mu^{\tilde{\pi}}(s)\big(Q^{\pi,\widehat{R}'}(s,\tilde{\pi}_{\dagger}(s)) - Q^{\pi,\widehat{R}'}(s,\pi(s))\big) \geq 0.$$

However, $d_{S_{\text{pos}}^{\pi_{\dagger}}}(\pi_{\dagger},\tilde{\pi}) \leq k-1$ and therefore

$$\rho^{\tilde{\pi}_{\dagger},\widehat{R}'} - \rho^{\tilde{\pi},\widehat{R}'} \geq \epsilon.$$

Which proves the claim.

Part 2: For the right inequality, define $\widehat{R}$" as $\widehat{R}$" $:= \widehat{R}^{\pi_{\dagger}} - \epsilon'_{\pi_{\dagger}}$. We claim that

$$\pi_{\dagger}(s) \in \arg\max_{a} Q^{*,\widehat{R}^{\pi_{\dagger}}}(s,a) \quad \text{for all} \quad s \in S_{\text{pos}}^{\pi_{\dagger}}. \tag{12}$$

To prove this, define $\tilde{\pi}_{\dagger} \in \Pi_{\text{det}}$ as

$$\tilde{\pi}_{\dagger}(s) = \arg\max_{a} Q^{*,\widehat{R}^{\pi_{\dagger}}}(s,a).$$

Now note that

$$0 \overset{(i)}{=} \rho^{\pi_{\dagger},\widehat{R}^{\pi_{\dagger}}} - \rho^{\tilde{\pi}_{\dagger},\widehat{R}^{\pi_{\dagger}}}$$

$$= \sum_s \mu^{\pi_\dagger}(s) \cdot \left(Q^{\tilde{\pi}_\dagger, \widehat{R}^{\pi_\dagger}}(s, \pi_\dagger(s)) - Q^{\tilde{\pi}_\dagger, \widehat{R}^{\pi_\dagger}}(s, \tilde{\pi}_\dagger(s)) \right)$$
$$\overset{(ii)}{=} \sum_s \mu^{\pi_\dagger}(s) \cdot \left(Q^{*, \widehat{R}^{\pi_\dagger}}(s, \pi_\dagger(s)) - Q^{*, \widehat{R}^{\pi_\dagger}}(s, \tilde{\pi}_\dagger(s)) \right)$$

where (i) follows from optimality of $\pi_\dagger$ in $\widehat{R}^{\pi_\dagger}$ and (ii) follows from the fact that $Q^{*, \widehat{R}^{\pi_\dagger}} = Q^{\tilde{\pi}_\dagger, \widehat{R}^{\pi_\dagger}}$. Therefore $Q^{*, \widehat{R}^{\pi_\dagger}}(s, \pi_\dagger(s)) = Q^{*, \widehat{R}^{\pi_\dagger}}(s, \tilde{\pi}_\dagger(s))$ for all $s \in S^{\pi_\dagger}_{\text{pos}}$ which is equivalent to (12).

Given this result, we further assume that $\tilde{\pi}_\dagger(s) = \pi_\dagger(s)$ for all $s \in S^{\pi_\dagger}_{\text{pos}}$ since we did not originially specify how to break ties in the definition of $\tilde{\pi}_\dagger$. This also implies $S^{\pi_\dagger}_{\text{pos}} = S^{\tilde{\pi}_\dagger}_{\text{pos}}$ since $\mu^{\pi_\dagger} = \mu^{\tilde{\pi}_\dagger}$ by Lemma 3.

Now note that given the construction of $\widehat{R}$",

$$Q^{\tilde{\pi}_\dagger, \widehat{R}^{\pi_\dagger}}(s, \tilde{\pi}_\dagger(s)) = Q^{\tilde{\pi}_\dagger, \widehat{R}"}(s, \tilde{\pi}_\dagger(s))$$

for all s. This is because the rewards for the state-action pairs $(s, \tilde{\pi}_\dagger(s))$ were not modified. Since no reward has increased, this further implies that $V^{\tilde{\pi}_\dagger, \widehat{R}"} = V^{\tilde{\pi}_\dagger, \widehat{R}^{\pi_\dagger}}$. Now note that for all s, a,

$$Q^{\tilde{\pi}_\dagger, \widehat{R}"}(s, a) - Q^{\tilde{\pi}_\dagger, \widehat{R}^{\pi_\dagger}}(s, a) = \widehat{R}"(s, a) - \widehat{R}^{\pi_\dagger}(s, a) \leq -\epsilon'(s, a) \cdot \mathbb{1}\left[s \in S^{\pi_\dagger}_{\text{pos}}\right]$$

Recall however that by definition of $\tilde{\pi}_\dagger$,

$$Q^{\tilde{\pi}_\dagger, \widehat{R}^{\pi_\dagger}}(s, \tilde{\pi}_\dagger(s)) \geq Q^{\tilde{\pi}_\dagger, \widehat{R}^{\pi_\dagger}}(s, a).$$

We can therefore conclude that

$$Q^{\tilde{\pi}_\dagger, \widehat{R}"}(s, \tilde{\pi}_\dagger(s)) \geq Q^{\tilde{\pi}_\dagger, \widehat{R}"}(s, a) + \epsilon'(s, a) \cdot \mathbb{1}\left[s \in S^{\pi_\dagger}_{\text{pos}}\right].$$

This means that $\left(R = \widehat{R}", Q = Q^{\tilde{\pi}_\dagger, \widehat{R}"}, V = V^{\tilde{\pi}_\dagger, \widehat{R}"}, \pi_\dagger = \tilde{\pi}_\dagger\right)$ satisfy all of the constraints of (P5-ATK). Since $\pi_\dagger(s) = \tilde{\pi}_\dagger(s)$ for all $s \in S^{\pi_\dagger}_{\text{pos}}$, this means that $\left(R = \widehat{R}", Q = Q^{\tilde{\pi}_\dagger, \widehat{R}"}, V = V^{\tilde{\pi}_\dagger, \widehat{R}"}, \pi_\dagger = \pi_\dagger\right)$ satisfy the constraints of (P5-ATK) as well. Therefore, by optimality of $\widehat{R}'$,

$$\begin{aligned}
\left\| \overline{R} - \widehat{R}' \right\|_2 &\leq \left\| \overline{R} - \widehat{R}" \right\|_2 \\
&\leq \left\| \overline{R} - \widehat{R}^{\pi_\dagger} \right\|_2 + \left\| \widehat{R}^{\pi_\dagger} - \widehat{R}" \right\|_2 \\
&= \left\| \overline{R} - \widehat{R}^{\pi_\dagger} \right\|_2 + \left\| \epsilon' \right\|_2 .
\end{aligned}$$

□

QGREEDY Algorithm

In this section, we present the QGREEDY algorithm and prove its correctness. Recall that this algorithm finds a solution to the optimization problem:

$$\Delta_Q = \min_{\pi \in \Pi^{\text{adm}}_{\text{det}}} \max_{s \in S^\pi_{\text{pos}}} \left(Q^{*, \overline{R}}(s, \pi^*(s)) - Q^{*, \overline{R}}(s, \pi(s)) \right).$$

The intuition behind the QGREEDY algorithm as follows. It starts by finding the state s which has the highest value of $\delta(s) = \min_{a \in A^{\text{adm}}} Q^{*, \overline{R}}(s, \pi^*(s)) - Q^{*, \overline{R}}(s, a)$ among admissible state-action pairs—this value provides an upper bound on Δ_Q and is tight if s is reachable by any admissible policy. If this is not the case, then there might exists a policy π which results in lower value of Δ^π_Q by not reaching s (making $\mu^\pi(s) = 0$). Therefore, after finding s, the algorithm proceeds by finding the set of state-action pairs that are "connected" to s in that policies defined on these pairs reach s with strictly positive probability. These state action pairs are removed from the admissible set of state-action pairs, and the algorithm proceeds with the next iteration. The output is defined by the minimum of all of the gaps δ found in each iteration, and the policy can be reconstructed from the set of state-action pairs that are "connected" to the state that defines this gap.

The pseudo-code of QGREEDY can be found in Algorithm 2 and provides a more detailed description of the algorithm. The following result formally shows that QGREEDY outputs a correct result.

Lemma 6. *Let Δ_A, π_A denote the output of the algorithm 2. Then it holds that*

$$\Delta_Q = \Delta_A = \Delta^{\pi_A}_Q,$$

where Δ^π_Q is defined as

$$\Delta^\pi_Q = \max_{s \in S^\pi_{\text{pos}}} \left(Q^{*, \overline{R}}(s, \pi^*(s)) - Q^{*, \overline{R}}(s, \pi(s)) \right). \tag{13}$$

Algorithm 2: QGREEDY

Input: MDP $\overline{M}$, admissible action set A_s^{adm} for each state s.
Output: Δ_Q, Policy $\pi \in \arg\min_\pi \Delta_Q^\pi$.

1: Calculate Q-values $Q^{*,\overline{R}}$.
2: Let $\pi^*(s) = \max_a Q^{*,\overline{R}}(s,a)$.
3: Let $\text{ADM}^{(0)} = \{(s,a) | a \in A_s^{\text{adm}}\}$.
4: Let $\tilde{S}^{(0)} = S$.
5: Let $S_\sigma = \{s | \sigma(s) \neq 0\}$.
6: Let $t = 0$.
7: **while** $S_\sigma \subseteq \tilde{S}^{(t)}$ **do**
8: Let $\text{ADM}_s^{(t)} = \{a | (s,a) \in \text{ADM}^{(t)}\}$ for all $s \in \tilde{S}^{(t)}$.
9: Let $\delta^{(t)}(s) = \min_{a \in \text{ADM}_s^{(t)}} Q^{*,\overline{R}}(s,\pi^*(s)) - Q^{*,\overline{R}}(s,a)$ for all $s \in \tilde{S}^{(t)}$.
10: Let $s_t = \arg\max_{s \in \tilde{S}^{(t)}} \delta^{(t)}(s)$ and $\Delta_t = \delta^{(t)}(s_t)$.
11: Let $\pi_t(s) = \arg\min_{a \in \text{ADM}_s^{(t)}} Q^{*,\overline{R}}(s,\pi^*(s)) - Q^{*,\overline{R}}(s,a)$ for all $s \in \tilde{S}^{(t)}$ and choose $\pi_t(s)$ arbitrarily otherwise.
12: Let $\tilde{S}^{(t+1)} = \tilde{S}^{(t)}$ and $\text{ADM}^{(t+1)} = \text{ADM}^{(t)}$.
13: Let $S_\delta = \{s_t\}$
14: **while** $S_\delta \neq \emptyset$ **do**
15: $\tilde{S}^{(t+1)} = \tilde{S}^{(t+1)} \backslash S_\delta$.
16: $\text{ADM}^{(t+1)} = \text{ADM}^{(t+1)} \backslash \{(s,a) | P(s,a,\tilde{s}) > 0 \text{ for some } \tilde{s} \notin \tilde{S}^{(t+1)}\}$.
17: Let $S_\delta = \{s \in \tilde{S}^{(t+1)} | (s,a) \notin \text{ADM}^{(t+1)} \text{ for all } a\}$
18: **end while**
19: $t = t + 1$.
20: **end while**
21: Let $t^* = \arg\min_t \Delta_t$
22: **return** Δ_{t^*}, π_{t^*}

Proof. Before we discuss the algorithm's correctness, observe that it guaranteed to terminate since S is finite, S_σ is nonempty, and $|\tilde{S}^{(t)}|$ strictly decreases in each iteration. We now prove the correctness of the algorithm. We divide the proof into three parts.

Part 1: We show that $\Delta_Q^{\pi_A} \leq \Delta_A$. In order to understand why this is the case, for each iteration t, consider the policy π_t. We first claim that $S_{\text{pos}}^{\pi_t} \subseteq \tilde{S}^{(t)}$. Note that this is the reason we have defined π_t only for the states $s \in \tilde{S}^{(t)}$ in the algorithm, the value of $\pi_t(s)$ for $s \notin \tilde{S}^{(t)}$ is not important and can be chosen arbitrarily.

In order to prove the claim, note that it is obviously true for $t = 0$ because $\tilde{S}^{(0)} = S$. As for $t \geq 1$, observe that an agent following π_t initially starts in a state in $S_\sigma \subseteq \tilde{S}^{(t)}$ and therefore the agent is in $\tilde{S}^{(t)}$ in the beginning. Therefore, in order to reach a state in $S \backslash \tilde{S}^t$, at some point the agent would need to take an action that could lead to, i.e., with strictly positive probability would transition to, $S \backslash \tilde{S}^t$. This is not possible however as $\pi_t(s) \in \text{ADM}_s^{(t)}$ and all state-action pairs in $\tilde{S}^{(t)}$ that could lead to $S \backslash \tilde{S}^{(t)}$ were removed from $\text{ADM}^{(t)}$ during its construction in Line 16.

Given this result, it is clear that

$$\begin{aligned}
\Delta_t &= \delta^{(t)}(s_t) \\
&= \max_{s \in \tilde{S}^{(t)}} \delta^{(t)}(s) \\
&\geq \max_{s \in S_{\text{pos}}^{\pi_t}} \delta^{(t)}(s) \\
&= \max_{s \in S_{\text{pos}}^{\pi_t}} Q^{*,\overline{R}}(s,\pi^*(s)) - Q^{*,\overline{R}}(s,\pi_t(s)) \\
&= \Delta_Q^{\pi_t}.
\end{aligned}$$

Therefore, $\Delta_Q^{\pi_t} \leq \Delta_t$. Since this holds for all t, the claim is proved.

Part 2: We show that $\Delta_Q \geq \Delta_A$. We use proof by contradiction. Assume that this is not the case and $\Delta_Q < \Delta_A$. This means that $\Delta_Q^\pi < \Delta_A$ for some $\pi \in \Pi_{\text{det}}^{\text{adm}}$. Now, observe that since the loop terminates for some t, there exists a t such that $S_\sigma \not\subseteq \tilde{S}^{(t)}$. Since $S_\sigma \subseteq S_{\text{pos}}^\pi$, this means that there exists a t such that $S_{\text{pos}}^\pi \not\subseteq \tilde{S}^{(t)}$. We claim that this contradicts the assumption $\Delta_Q^\pi < \Delta_A$.

Concretely, we will use the assumption $\Delta_Q^\pi < \Delta_A$ and show by induction on t that $S_{\text{pos}}^\pi \subseteq \tilde{S}^{(t)}$ and $(s, \pi(s)) \in \text{ADM}^{(t)}$ for all $s \in S_{\text{pos}}^\pi$.

For $t = 0$, the claim holds because since $\pi \in \Pi_{\text{det}}^{\text{adm}}$. Assume the claim holds for t, we will show that it holds for $t+1$ as well. We first claim that $s_t \notin S_{\text{pos}}^\pi$. Concretely, $(s_t, \pi(s_t)) \in \text{ADM}^{(t)}$ by the induction hypotheses. If $s_t \in S_{\text{pos}}^\pi$, this would imply that

$$\begin{aligned}
\Delta_Q^\pi &\geq Q^{*,\overline{R}}(s_t, \pi^*(s_t)) - Q^{*,\overline{R}}(s_t, \pi(s_t)) \\
&\geq \min_{a \in \text{ADM}_{s_t}^t} Q^{*,\overline{R}}(s_t, \pi^*(s_t)) - Q^{*,\overline{R}}(s_t, a) \\
&= \delta^{(t)}(s_t) \\
&= \Delta_t \\
&\geq \Delta_A,
\end{aligned}$$

which contradicts the assumption $\Delta_Q^\pi < \Delta_A$. Therefore, $s_t \notin S_{\text{pos}}^\pi$. Now consider the inner loop in lines 14-18. We claim by induction that the loop does not remove any states $s \in S_{\text{pos}}^\pi$ from $\tilde{S}^{(t+1)}$ and does not remove any state-action pairs $(s, \pi(s))$ such that $s \in S_{\text{pos}}^\pi$ from $\text{ADM}^{(t+1)}$. Since $s_t \notin S_{\text{pos}}^\pi$, this is true the first time line 15 is executed. In each execution of the loop, assuming the constraint is not violated in line 15, then it will not be violated in line 16 either. This is because if $(s, \pi(s))$ is removed from $\text{ADM}^{(t+1)}$ for some $s \in S_{\text{pos}}^\pi$, then $P(s, a, s') > 0$ for some $s' \notin \tilde{S}^{(t+1)}$. However, $P(s, a, s') > 0$ implies that $s' \in S_{\text{pos}}^\pi$ and therefore $s' \in S_{\text{pos}}^\pi$ was already removed from $\tilde{S}^{(t+1)}$. Likewise, if a state $s \in S_{\text{pos}}^\pi$ is removed in later executions of 15, then it must be the case that $(s, a) \notin \text{ADM}^{(t+1)}$ for all a. This means that at some point, the state-action pair $(s, \pi(s))$ must have been removed from $\text{ADM}^{(t+1)}$ which means the constraint must have already been violated. Therefore, the constraint is not violated at any point. This means that the induction is complete and we have reached a contradiction using the assumption $\Delta_Q < \Delta_A$. Therefore, the assumption was wrong and $\Delta_Q \geq \Delta_A$.

Part 3: Putting both parts together, note that

$$\Delta_A \leq \Delta_Q \leq \Delta_Q^{\pi_A} \leq \Delta_A,$$

where the first inequality follows from Part 2, the second inequality follows from the definition of Δ_Q and the final inequality follows from Part 1. Therefore the proof is complete. □

Proofs of the Results in Section Problem Setup

In this section, we provide proofs of our results in Section Problem Setup, namely, Proposition 1.

Proof of Proposition 1

Statement: *If $\Pi_{\text{det}}^{\text{adm}}$ is not empty, there always exists an optimal solution to the optimization problem* (P1-APT).

Proof. to prove the statement, we first show that the following three claims hold.

Claim 1. *Consider a function that evaluates the objective of the optimization problem* (P1-APT) *for a given R:*

$$l(R) = \max_{\pi \in \text{OPT}_{\text{det}}^\epsilon(R)} \left\|\overline{R} - R\right\|_2 - \lambda \rho^{\pi, \overline{R}}.$$

This function, $l(R)$, is lower-semi continuous.

Proof. First note that the function is real-valued (i.e., $l(R) \notin \{\infty, -\infty\}$) as the set of all deterministic policies is finite. To prove the claim, we need to show that for all R:

$$\forall \delta : \exists \alpha : \forall \tilde{R} : \left\|\tilde{R} - R\right\|_2 \leq \alpha \implies l(\tilde{R}) \geq l(R) - \delta.$$

Assume to the contrary that there exists an R such that

$$\exists \delta : \forall \alpha : \exists \tilde{R}_\alpha : \left\|\tilde{R}_\alpha - R\right\|_2 \leq \alpha \wedge l(\tilde{R}_\alpha) < l(R) - \delta.$$

By setting $\alpha_i = \frac{1}{2^i}$, we obtain a series $\{R_i\}_{i=1}^\infty$ such that R_i tends to R and

$$\forall i : l(R_i) < l(R) - \delta.$$

Take π to be an arbitrary policy in

$$\arg\max_{\pi \in \text{OPT}_{\text{det}}^\epsilon(R)} \left\|\overline{R} - R\right\|_2 - \lambda \rho^{\pi, \overline{R}}.$$

We claim that there exists N such that

$$\forall i \geq N : \pi \in \textsc{Opt}^{\epsilon}_{\text{det}}(R_i).$$

If this would not be the case, then there would exist an infinite sub-sequence of R_is such that for all of them $\pi \notin \textsc{Opt}^{\epsilon}_{\text{det}}(R_i)$ and therefore there exists a deterministic $\tilde{\pi}_i$ such that $\rho^{\tilde{\pi}_i,R_i} \geq \rho^{\pi,R_i} + \epsilon$. Since the number of deterministic policies is finite, at least one of these deterministic policies would occur infinitely often. This would mean that there is a policy $\tilde{\pi}$ and an infinite subsequence of R_js that tends to R, and for all of the js in the subsequence

$$\rho^{\tilde{\pi},R_j} \geq \rho^{\pi,R_j} + \epsilon.$$

Since $\rho(R, \tilde{\pi})$ is continuous in R for fixed $\tilde{\pi}$, this would imply that

$$\rho^{\tilde{\pi},R} \geq \rho^{\pi,R} + \epsilon,$$

which contradicts the assumption $\pi \in \textsc{Opt}^{\epsilon}_{\text{det}}(R)$. Therefore, as we stated above, there exists N such that

$$\forall i \geq N : \pi \in \textsc{Opt}^{\epsilon}_{\text{det}}(R_i).$$

Now, note that

$$\forall i \geq N : l(R_i) = \max_{\pi \in \textsc{Opt}^{\epsilon}_{\text{det}}(R_i)} \left\|\overline{R} - R_i\right\|_2 - \lambda\rho^{\pi,\overline{R}} \geq \left\|\overline{R} - R_i\right\|_2 - \lambda\rho^{\pi,\overline{R}}.$$

Therefore, we have that

$$\begin{aligned} -\delta &> l(R_i) - l(R) \\ &\geq \left\|\overline{R} - R_i\right\|_2 - \lambda\rho^{\pi,\overline{R}} - \left\|\overline{R} - R\right\|_2 + \lambda\rho^{\pi,\overline{R}} = \left\|\overline{R} - R_i\right\|_2 - \left\|\overline{R} - R\right\|_2, \end{aligned}$$

which is a contradiction, since $\left\|\overline{R} - R_i\right\|_2 - \left\|\overline{R} - R\right\|_2$ goes to 0 as $i \to \infty$. This proves the claim. □

Claim 2. *The set* $\{R : \textsc{Opt}^{\epsilon}_{\text{det}}(R) \subseteq \Pi^{\text{adm}}_{\text{det}}\}$ *is closed.*

Proof. To see why, note that R is in this set if and only if

$$\exists \pi \in \Pi^{\text{adm}}_{\text{det}} : \forall \tilde{\pi} \in \Pi_{\text{det}} \backslash \Pi^{\text{adm}}_{\text{det}} : \rho^{\pi,R} \geq \rho^{\tilde{\pi},R} + \epsilon$$

For a fixed $\pi, \tilde{\pi}$, the set $\{R : \rho^{\pi,R} \geq \rho^{\tilde{\pi},R} + \epsilon\}$ is closed. Since the set $\{R : \textsc{Opt}^{\epsilon}_{\text{det}}(R) \subseteq \Pi^{\text{adm}}_{\text{det}}\}$ is a finite union of a finite intersection of such sets, $\{R : \textsc{Opt}^{\epsilon}_{\text{det}}(R) \subseteq \Pi^{\text{adm}}_{\text{det}}\}$ is closed as well. □

Claim 3. *If* $\Pi^{\text{adm}}_{\text{det}}$ *is not empty,* (P1-APT) *is feasible.*

Proof. Assume $\pi \in \Pi^{\text{adm}}_{\text{det}}$. Let $t \geq 0$ be an arbitrary positive number. Define reward function R as

$$R(s,a) = t \cdot \mathbb{1}\left[\mu^{\pi}(s) > 0 \wedge a = \pi(s)\right].$$

Note that this implies

$$\begin{aligned} \rho^{\pi,R} &= \sum_s \mu^{\pi}(s) R(s, \pi(s)) \\ &= \sum_{s \in S^{\pi}_{\text{pos}}} \mu^{\pi}(s) R(s, \pi(s)) \\ &= \sum_{s \in S^{\pi}_{\text{pos}}} \mu^{\pi}(s) \cdot t \\ &= t. \end{aligned}$$

We show that if t is large enough, R is feasible, i.e., for all $\tilde{\pi} \notin \Pi^{\text{adm}}_{\text{det}}$, $\tilde{\pi} \notin \textsc{Opt}^{\epsilon}_{\text{det}}(R)$. Since the number of deterministic policies is finite, it suffices to show for a fixed that $\tilde{\pi} \notin \Pi^{\text{adm}}_{\text{det}}$, $\tilde{\pi} \notin \textsc{Opt}^{\epsilon}_{\text{det}}(R)$ for large enough t. To prove this, note that since $\tilde{\pi} \notin \Pi^{\text{adm}}_{\text{det}}$, there exists a state $\tilde{s}$ such that $\mu^{\tilde{\pi}}(\tilde{s}) > 0$ and $\tilde{\pi}(\tilde{s}) \notin A^{\text{adm}}_{\tilde{s}}$. Since π was admissible, this means that either $\pi(\tilde{s}) \neq \tilde{\pi}(\tilde{s})$ or $\mu^{\pi}(\tilde{s}) = 0$. Either way, $R(s, \tilde{\pi}(\tilde{s})) = 0$. Therefore,

$$\begin{aligned} \rho^{\pi,R} &= \sum_s \mu^{\tilde{\pi}}(s) \cdot R(s, \tilde{\pi}(s)) \\ &= \sum_{s \neq \tilde{s}} \mu^{\tilde{\pi}}(s) \cdot R(s, \tilde{\pi}(s)) \\ &\leq t - \mu^{\tilde{\pi}}(\tilde{s}) \cdot t. \end{aligned}$$

Setting $t > \frac{\epsilon}{\mu^{\tilde{\pi}}(\tilde{s})}$ proves the claim. □

Let us now prove the statement of the proposition. Since the optimization problem (P1-APT) has no limits on R, in other words the set of all feasible R in the optimization problem is not bounded, we cannot claim that the feasible set is compact. Note however that since the optimization problem is feasible for any fixed $\overline{R}$, there is an upper bound on its value. Furthermore, the second term in the objective, i.e, $\rho^{\pi,\overline{R}}$ is bounded for any fixed $\overline{R}$. This means that for every fixed $\overline{R}$, there exists a number Θ such that the optimization problem (P1-APT) is equivalent to

$$\begin{aligned} \min_{R} \max_{\pi} \ & \left\|\overline{R} - R\right\|_2 - \lambda \rho^{\pi,\overline{R}} \\ \text{s.t.} \quad & \textsc{Opt}^{\epsilon}_{\text{det}}(R) \subseteq \Pi^{\text{adm}}_{\text{det}} \\ & \pi \in \textsc{Opt}^{\epsilon}_{\text{det}}(R) \\ & \left\|R - \overline{R}\right\|_2 \leq \Theta, \end{aligned}$$

This turns the problem into minimizing a lower-semi-continous function over a compact set which has an optimal solution (i.e., the infimum is attainable). □

Proofs of the Results in Section Computational Challenges and Additional Results

In this section, we provide proofs of our results in Section Computational Challenges, namely, Theorem 1, and Proposition 2. We also provide an additional computational complexity result for the optimization problem (P4-APT).

Proof of Theorem 1

Statement: *For any constant $p \in (0,1)$, it is NP-hard to distinguish between instances of* (P2-APT$_{\lambda=0}$) *that have optimal values at most ξ and instances that have optimal values larger than $\xi \cdot \sqrt{(|S| \cdot |A|)^{1-p}}$. The result holds even when the parameters ϵ and γ in* (P2-APT$_{\lambda=0}$) *are fixed to arbitrary values subject to $\epsilon > 0$ and $\gamma \in (0,1)$.*

Proof. We show a reduction from the NP-complete problem EXACT-3-SET-COVER (X3C) (Karp 1972; Garey and Johnson 1979). An instance of X3C is given by a set $E = \{e_1, \ldots, e_{3k}\}$ of $3k$ elements and a collection $\mathcal{S}$ of 3-element subsets of E. It is a yes-instance if there exists a sub-collection $\mathcal{Q} \subseteq \mathcal{S}$ of size k such that $\cup_{S \in \mathcal{Q}} S = E$, and a no-instance otherwise.

Given an X3C instance, we construct the following instance of (P2-APT$_{\lambda=0}$). The underlying MDP is illustrated in Figure 3 with the following specifications, where we let $N = \left\lceil 3k \cdot \varphi^{1-p} \cdot \left(\frac{9l}{\gamma}\right)^2 \right\rceil$, $\delta = \frac{\gamma^2}{8 \cdot m \cdot l}$, and $m = (3k+1)N$; the value of φ will be defined shortly. Intuitively, we need N to be sufficiently large and $\delta > 0$ to be sufficiently close to 0.

- s_0 is the starting state, in which taking the only action available leads to a non-deterministic transition to each of the states $s_1, \ldots, s_{3k}$ and s_* with probability $\frac{1}{m}$.
- In each state s_i, $i = 1, \ldots, 3k$, taking action $a_\dagger$ leads to the state transitioning to $\tilde{s}_0$, yielding a reward $z_i := 0$. Action $a_\dagger$ is *not* admissible in any of these states.
- Suppose there are l subsets $S_1, \ldots, S_l$ in the collection $\mathcal{S}$. For each subset S_j, we create a state t_j. We also create l actions $a_1, \ldots, a_l$. If $e_i \in S_j$, then taking action a_j in state s_i leads to the state transitioning to t_j and yields a reward $x_{ij} := \frac{m}{\gamma}(\frac{\epsilon}{1-\gamma} + \delta) - \gamma$. From each state t_j, the only action available leads to state $\tilde{s}_f$, yielding a reward $\omega_j := 0$.
- In state s_*, taking action $a_\dagger$ leads to the state transitioning to each t_j with probability $1/l$, yielding a reward $z_* := 0$; this action is *not* admissible. Taking the other available action — let it be a_1 — leads to $\tilde{s}_1$ and a reward $y := \gamma \cdot \frac{k}{l} + \frac{m}{\gamma}\left(\frac{\epsilon}{1-\gamma} + \delta\right)$ is yielded.
- In states $\tilde{s}_0$ there is a single action yielding a reward of $u = 0$ and transitioning to the state $\tilde{s}_f$. Similarly, in $\tilde{s}_1$ there is a single action yielding a reward of $r = 0$ and transitioning to $\tilde{s}_f$.
- In state $\tilde{s}_f$, there is a single action, yielding a reward 0 and transitioning back to s_0

After the above construction is done, we start to create copies of some of the states, which will be essential for the reduction. We repeat the following step $N-1$ times to create $N-1$ sets of copies:

- Create a copy of each state $s \in \{s_1, \ldots, s_{3k}\} \cup \{s_*, \tilde{s}_0, \tilde{s}_1\}$. Connect these new copies in the following way: Connect the copy $\bar{s}$ of each state s to the copy $\bar{s}'$ of every other state s' (only those created in this step) the same way s and s' are connected; In addition, connect $\bar{s}$ to every other state s' which does not have a copy the same way s and s' are connected.

Finally, we let

$$\varphi = \left(6k \cdot \left(\frac{9l}{\gamma}\right)^2 \cdot (3k + l + 5) \cdot (l+1)\right)^{1/p}.$$

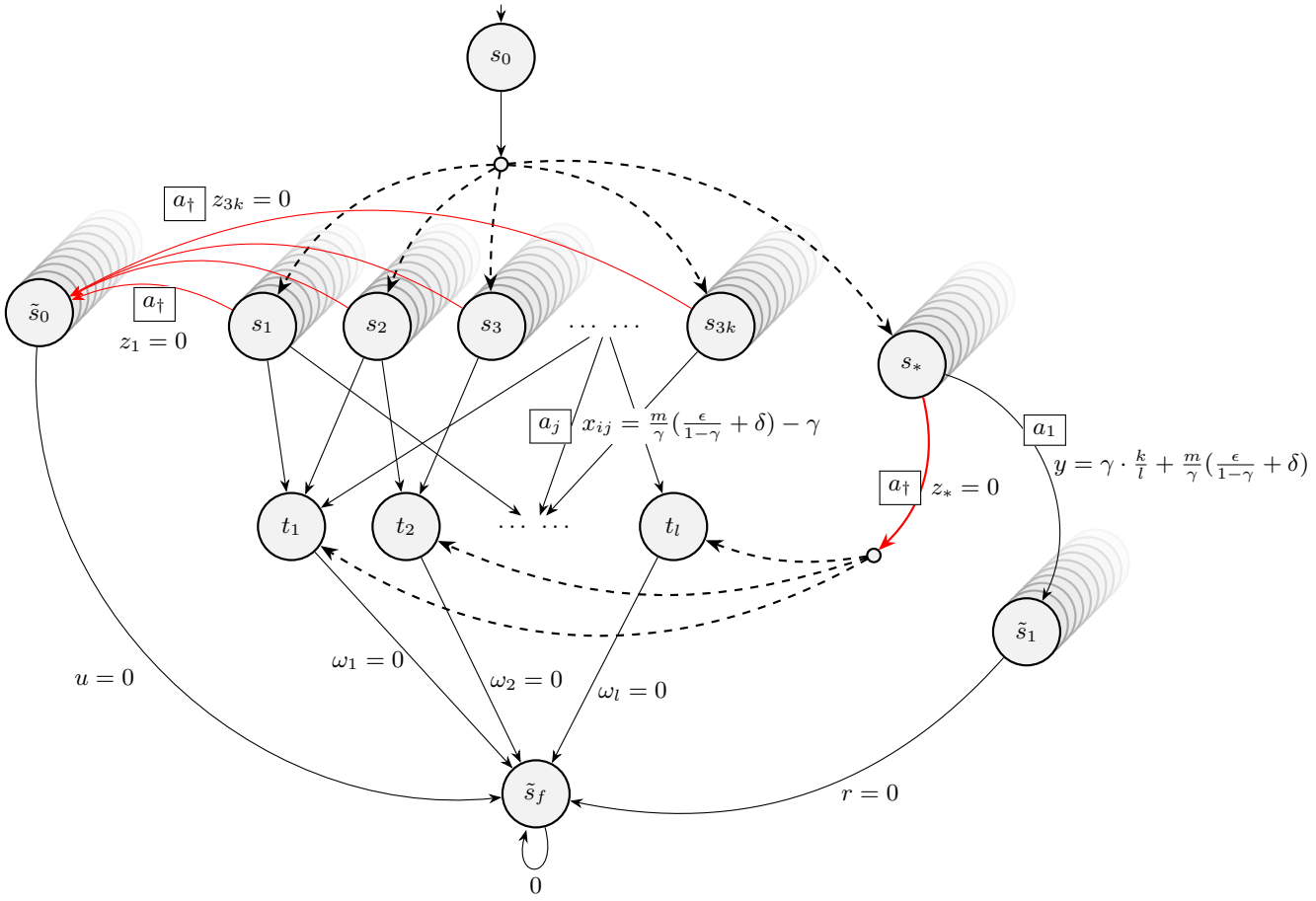


Figure 3: Reduction. Solid edges represent actions and dashed edges represent non-deterministic transitions. Red edges are *not* admissible. Labels in boxes are names of some important actions in the reduction, of the corresponding edges; other values on the edges denote rewards for the corresponding actions. Each of the states $s_1, \ldots, s_{3k}$ and s_* has N copies, and copies of each state are connected to other states in the same way.

Now note that $\varphi \geq 1$ and therefore $N \geq 3$ and $N \leq 2(N-1) \leq 2 \cdot \left(3k \cdot \varphi^{1-p} \cdot \left(\frac{9l}{\gamma}\right)^2\right)$, which implies

$$|S| \cdot |A| \leq N \cdot (3k + l + 5) \cdot (l + 1) \leq \varphi^{1-p} \cdot 2 \cdot \left(3k \cdot \left(\frac{9l}{\gamma}\right)^2\right) \cdot (3k + l + 5) \cdot (l + 1) \leq \varphi. \tag{14}$$

Without loss of generality, we can also assume that $k \leq |\mathcal{S}|$; the X3C problem is always a no-instance when $|\mathcal{S}| < k$ since this implies $|E| > 3|\mathcal{S}|$. Therefore the X3C problem remains NP-hard with this restriction.[5] The restriction implies that $k + 1 \leq 2|\mathcal{S}| = 2l$, which will be useful in the sequel.

Observe that by the above construction, regardless of what policy is chosen, the agent will end up in $\tilde{s}_f$ in exactly three steps and will stay there forever. Therefore, the reward of this state does not matter as it cancels out when considering $\rho^{\pi,R} - \rho^{\tilde{\pi},R}$ for all $\pi, \tilde{\pi}, R$. We now proceed with the proof.

Correctness of the Reduction Let $\xi = \sqrt{k}$. We will show next that if the X3C instance is a yes-instance, then this (P2-APT$_{\lambda=0}$) instance admits an optimal solution R with $\left\|\overline{R} - R\right\|_2 \leq \xi$; otherwise, any feasible solution R of (P2-APT$_{\lambda=0}$) is such that $\left\|\overline{R} - R\right\|_2 > \xi \cdot \sqrt{(|S| \cdot |A|)^{1-p}}$.

[5] We can further assume that $k \geq 1/\epsilon$ without loss of generality given that ϵ is fixed, in which case we have $x_{ij} = \frac{m}{\gamma}\left(\frac{\epsilon}{1-\gamma} + \delta\right) - \gamma > 0$, so the reduction would not rely on negative rewards.

First, suppose that the X3C instance if a yes-instance. By definition, there exists a size-k set $Q \subseteq \{1, \ldots, l\}$, such that $\cup_{j \in Q} S_j = E$. Consider a solution R obtained by increasing each ω_j to 1 if $j \in Q$. Since $|Q| = k$, we have $\left\|\overline{R} - R\right\|_2 = \sqrt{k} = \xi$. We verify that R is a feasible solution, with $\text{OPT}^{\epsilon}_{\text{det}}(R) \subseteq \Pi^{\text{adm}}_{\text{det}}$. In particular, in each state s_i, $i = 1, \ldots, 3k$, we verify that taking action $a_\dagger$ would results in a loss of greater than ϵ in the policy score, as compared with taking an action a_j such that $e_i \in S_j$ and $j \in Q$; we know such an a_j exists because Q is an exact set cover. Similarly, we show that in state s_*, taking action a_1 is at least ϵ better than taking action $a_\dagger$. Note this proves the claim that $\text{OPT}^{\epsilon}_{\text{det}}(R) \subseteq \Pi^{\text{adm}}_{\text{det}}$: if a policy takes an inadmissible action in one of the states, changing that action to a_j or a_1 will cause an increase of at least ϵ in its score. This means that the inadmissible policy could not have been in $\text{OPT}^{\epsilon}_{\text{det}}(R)$. Note that since there may be many copies of some states in the MDP, when referring to a state s_i or a state s_*, we are technically referring to one of these copies. For convenience however, we will not consider this dependence in our notation.

Formally, let $\hat{\pi}$ be a policy that chooses that action $a_\dagger$ in some s_i and denote by π a policy such that $\pi(s) = \hat{\pi}(s)$ for all $s \neq s_i$ and $\pi(s) = a_j$ for $s = s_i$ where j is chosen such that $e_i \in S_j$ and $j \in Q$. It follows that

$$\begin{aligned}
\frac{\rho^{\pi,R} - \rho^{\hat{\pi},R}}{1-\gamma} &= \frac{1}{m} \cdot \gamma \left(V^{\pi,R}(s_i) - V^{\hat{\pi},R}(s_i) \right) \\
&= \frac{1}{m} \cdot \gamma \left(R(s_i, a_j) + \gamma V^{\pi,R}(t_j) - R(s_i, a_\dagger) - \gamma R(\tilde{s}_0) \right) \\
&= \frac{1}{m} \cdot \gamma \left(x_{ij} + \gamma \cdot 1 - 0 - \gamma \cdot 0 \right) \\
&= \frac{1}{m} \cdot \gamma \left(\frac{m}{\gamma} \left(\frac{\epsilon}{1-\gamma} + \delta \right) - \gamma + \gamma - 0 \right) \\
&= \frac{\epsilon}{1-\gamma} + \delta \\
&\geq \frac{\epsilon}{1-\gamma}
\end{aligned}$$

Similarly, if $\hat{\pi}(s_*) = a_\dagger$, by defining π such that $\pi(s) = \hat{\pi}(s)$ for $s \neq s_*$ and $\pi(s) = a_1$ for $s = s_*$,

$$\begin{aligned}
\frac{\rho^{\pi,R} - \rho^{\hat{\pi},R}}{1-\gamma} &= \frac{1}{m} \cdot \gamma \left(V^{\pi,R}(s_*) - V^{\hat{\pi},R}(s_*) \right) \\
&= \frac{1}{m} \cdot \gamma \left(R(s_*, a_1) + \gamma R(\tilde{s}_1) - R(a_\dagger) - \frac{\gamma}{l} \sum_j V^{\pi,R}(t_j) \right) \\
&= \frac{1}{m} \cdot \gamma \left(\gamma \cdot \frac{k}{l} + \frac{m}{\gamma} \left(\frac{\epsilon}{1-\gamma} + \delta \right) - \frac{\gamma k}{l} \right) \\
&= \frac{1}{m} \cdot \gamma \left(\frac{m}{\gamma} \left(\frac{\epsilon}{1-\gamma} + \delta \right) \right) \\
&= \frac{\epsilon}{1-\gamma} + \delta \geq \frac{\epsilon}{1-\gamma}
\end{aligned}$$

It is therefore clear that in both cases, $\hat{\pi}$ cannot be in $\text{OPT}^{\epsilon}_{\text{det}}(R)$ as $\rho^{\pi,R} - \rho^{\hat{\pi},R} \geq \epsilon$.

Conversely, suppose that the X3C instance is a no-instance. Consider an arbitrary feasible solution R, i.e., $\text{OPT}^{\epsilon}_{\text{det}}(R) \subseteq \Pi^{\text{adm}}_{\text{det}}$. Suppose that the parameters ω_j, x_{ij}, y, z_j, u and r are modified in R to $\tilde{\omega}_j$, $\tilde{x}_{ij}$, $\tilde{y}$, $\tilde{z}_j$, $\tilde{u}$ and $\tilde{r}$ respectively. Note that while technically these values may be modified differently for different copies of the copied states, our results focus on one copy and obtain bounds on these parameters. Since our choice of copy is arbitrary, the bound holds for any copy.
We will consider two cases and we will show that in both cases it holds that

$$(\tilde{r} - r)^2 + (\tilde{y} - y)^2 + (\tilde{z}_* - z_*)^2 + (\tilde{u} - u)^2 + \sum_{ij} (\tilde{x}_{ij} - x_{ij})^2 + \sum_i (\tilde{z}_i - z_i)^2 \geq \frac{1}{3} \cdot (\frac{\gamma}{8l})^2.$$

Since there are N copies of each of these values, it follows that

$$\begin{aligned}
\left\|\overline{R} - R\right\|_2 &\geq \sqrt{\frac{N}{3} \cdot \left(\frac{\gamma}{8l}\right)^2} \\
&\geq \sqrt{k \cdot \varphi^{1-p} \cdot \left(\frac{9l}{\gamma}\right)^2 \cdot \left(\frac{\gamma}{8l}\right)^2}
\end{aligned}$$

$$
\begin{aligned}
&> \sqrt{k \cdot \varphi^{1-p}} \\
&\overset{(14)}{\geq} \sqrt{k \cdot (|S| \cdot |A|)^{1-p}} \\
&= \xi \cdot \sqrt{(|S| \cdot |A|)^{1-p}}
\end{aligned}
$$

which completes the proof.

Case 1. $\sum_{j=1}^{l} \tilde{\omega}_j \geq k + 1/2$. Let $\hat{\pi}$ be an optimal policy under R. Hence, $\hat{\pi} \in \textsc{Opt}^{\epsilon}_{\text{det}}(R) \subseteq \Pi^{\text{adm}}_{\text{det}}$, which means that $\hat{\pi}(s_*) \neq a_\dagger$ as $a_\dagger$ is not admissible. Now, consider an alternative policy π, such that $\pi(s) = \hat{\pi}(s)$ for all $s \neq s_*$, and $\pi(s_*) = a_\dagger$. Since π is not admissible, we have $\pi \notin \textsc{Opt}^{\epsilon}_{\text{det}}(R)$, which means $\rho^{\pi,R} \leq \rho^{\hat{\pi},R} - \epsilon$. Note that

$$
\begin{aligned}
\frac{\rho^{\pi,R} - \rho^{\hat{\pi},R}}{1-\gamma} &= \frac{1}{m} \cdot \gamma \left(V^{\pi,R}(s_*) - V^{\hat{\pi},R}(s_*) \right) \\
&= \frac{\gamma}{m} \left(R\left(s_*, a_\dagger\right) + \frac{\gamma}{l} \sum_{j=1}^{l} V^{\pi,R}(t_j) - R\left(s_*, a_1\right) - \gamma R(\tilde{s}_1) \right) \\
&= \frac{\gamma}{m} \left(\tilde{z}_* + \frac{\gamma}{l} \sum_{j=1}^{l} \tilde{\omega}_j - \tilde{y} - \gamma \tilde{r} \right) \\
&= \frac{\gamma}{m} \left((\tilde{z}_* - z_*) + \frac{\gamma}{l} \sum_{j=1}^{l} \tilde{\omega}_j - (\tilde{y} - y) - \gamma(\tilde{r} - r) + (z_* - y - \gamma \cdot r) \right) \\
&= \frac{\gamma}{m} \left((\tilde{z}_* - z_*) + \frac{\gamma}{l} \sum_{j=1}^{l} \tilde{\omega}_j - (\tilde{y} - y) - \gamma(\tilde{r} - r) + \left(0 - \gamma \cdot \frac{k}{l} - \frac{m}{\gamma} \cdot (\frac{\epsilon}{1-\gamma} + \delta) + \gamma \cdot 0 \right) \right) \\
&= \frac{\gamma}{m} \left(\frac{\gamma}{l} \left(\sum_{j=1}^{l} \tilde{\omega}_j - k \right) + (\tilde{z}_* - z_*) - (\tilde{y} - y) - \gamma \left(\tilde{r} - r \right) \right) - (\frac{\epsilon}{1-\gamma} + \delta).
\end{aligned}
$$

Now that $\rho^{\pi,R} \leq \rho^{\hat{\pi},R} - \epsilon$, plugging this in the above equation and rearranging the terms leads us to the following result:

$$
\left(\sum_{j=1}^{l} \tilde{\omega}_j - k \right) + \frac{l}{\gamma} \left(\tilde{z}_* - z_* \right) - \frac{l}{\gamma} \left(\tilde{y} - y \right) - l \cdot (\tilde{r} - r) \leq \delta \cdot \frac{m \cdot l}{\gamma^2} \leq 1/4.
$$

Given the assumption that $\sum_{j=1}^{l} \tilde{\omega}_j \geq k + 1/2$ with this case, we then have

$$
\gamma \left(\tilde{r} - r \right) + (\tilde{y} - y) - (\tilde{z}_* - z_*) \geq \frac{\gamma}{4l}.
$$

Now note that for any three real numbers a, b, c, it holds by Cauchy–Schwarz that

$$
a^2 + b^2 + c^2 \geq \frac{(\gamma a + b - c)^2}{1 + 1 + \gamma} \geq \frac{(\gamma a + b - c)^2}{3}.
$$

Applying this result, we have

$$
\begin{aligned}
(\tilde{r} - r)^2 + (\tilde{y} - y)^2 + (\tilde{z}_* - z_*)^2 &\geq \frac{1}{3} \cdot \left(\frac{\gamma}{4l} \right)^2 \\
&\geq \frac{1}{3} \cdot \left(\frac{\gamma}{8l} \right)^2.
\end{aligned} \tag{15}
$$

Case 2. $\sum_{j=1}^{l} \tilde{\omega}_j < k + 1/2$. Let $Q = \left\{ j : \tilde{\omega}_j \geq \frac{k+1/2}{k+1} \right\}$. Hence, the size of Q is at most k: otherwise, there are at least $k+1$ numbers in $\tilde{\omega}_1, \ldots, \tilde{\omega}_l$ bounded by $\frac{k+1/2}{k+1}$ from below, which would imply that $\sum_{j=1}^{l} \tilde{\omega}_j \geq k + 1/2$.

By assumption, the X3C instance is a no-instance, so by definition, Q cannot be an exact cover, which means that there exists an element $e_\ell \in E$ not in any subset S_j, $j \in Q$. Accordingly, in the MDP constructed, besides $\tilde{s}_0$, state s_ℓ only connects subsequently to states t_η, with $\eta \notin Q$ (and hence, $\tilde{\omega}_\eta < \frac{k+1/2}{k+1}$).

Similarly to the analysis of Case 1, let $\hat{\pi}$ be an optimal policy under R, so $\hat{\pi}(s_\ell) \neq a_\dagger$ as $a_\dagger$ is not admissible. Hence, $\hat{\pi}(s_\ell) = a_\eta$ for some $\eta \in \{1, \ldots, l\}$ such that $e_\ell \in S_\eta$; in this case, $\eta \notin Q$.

Consider an alternative policy π, such that $\pi(s) = \hat{\pi}(s)$ for all $s \neq s_\ell$, and $\pi(s_\ell) = a_\dagger$, so π is not admissible. By assumption R is a feasible solution, which means $\pi \notin \textsc{Opt}^{\epsilon}_{\text{det}}(R)$ and hence, $\rho^{\pi,R} \leq \rho^{\hat{\pi},R} - \epsilon$. We have

$$\begin{aligned}
-\frac{\epsilon}{1-\gamma} &\geq \frac{\rho^{\pi,R} - \rho^{\hat{\pi},R}}{1-\gamma} \\
&= \frac{1}{m} \cdot \gamma \left(V^{\pi,R}(s_\ell) - V^{\hat{\pi},R}(s_\ell) \right) \\
&= \frac{\gamma}{m} \Big(R\left(s_\ell, a_\dagger\right) + \gamma R\left(\tilde{s}_0\right) - R\left(s_\ell, a_\eta\right) - \gamma \cdot \tilde{\omega}_\eta \Big) \\
&= \frac{\gamma}{m} \left(\tilde{z}_\ell + \gamma \tilde{u} - \tilde{x}_{\ell,\eta} - \gamma \cdot \tilde{\omega}_\eta \right) \\
&= \frac{\gamma}{m} \left(\tilde{z}_\ell + \gamma \tilde{u} - \tilde{x}_{\ell,\eta} - \gamma \cdot \tilde{\omega}_\eta \right) + \frac{\gamma}{m} \cdot \left(z_\ell + \gamma \cdot u - x_{\ell,\eta} \right) \\
&= \frac{\gamma}{m} \left(\left(\tilde{z}_\ell - z_\ell \right) + \gamma \left(\tilde{u} - u \right) - \left(\tilde{x}_{\ell,\eta} - x_{\ell,\eta} \right) \right) + \frac{\gamma^2}{m} (1 - \tilde{\omega}_\eta) - \left(\frac{\epsilon}{1-\gamma} + \delta \right),
\end{aligned}$$

which means

$$\begin{aligned}
\left(\tilde{x}_{\ell,\eta} - x_{\ell,\eta} \right) - \left(\tilde{z}_\ell - z_\ell \right) - \gamma \left(\tilde{u} - u \right) &\geq \gamma \left(1 - \tilde{\omega}_\eta \right) - \frac{m}{\gamma} \cdot \delta \\
&> \frac{1/2}{k+1} \cdot \gamma - \frac{m}{\gamma} \cdot \delta \\
&= \gamma \left(\frac{1}{2(k+1)} - \frac{1}{8l} \right) \\
&\geq \gamma \cdot \frac{1}{8l},
\end{aligned}$$

where we use the inequality $k + 1 \leq 2l$, which is implied by the assumption that $k \leq |\mathcal{S}|$ as we mentioned previously. In the same way we derived (15), we can obtain the following lower bound:

$$\left(\tilde{u} - u \right)^2 + \left(\tilde{x}_{\ell,\eta} - x_{\ell,\eta} \right)^2 + \left(\tilde{z}_\ell - z_\ell \right)^2 \geq \frac{1}{3} \cdot \left(\frac{\gamma}{8l} \right)^2,$$

which completes the proof. □

We now show that the hardness result in Theorem 1, holds for the optimization problem (P4-APT) as well.

Theorem 5. *For any constant $p \in (0,1)$, it is NP-hard to distinguish between instances of* (P4-APT) *that have optimal values at most ξ and instances that have optimal values larger than $\xi \cdot \sqrt{(|S| \cdot |A|)^{1-p}}$. The result holds even when the parameters ϵ and γ in* (P4-APT) *are fixed to arbitrary values subject to $\epsilon > 0$ and $\gamma \in (0,1)$.*

Proof. Throughout the proof, we assume that $\lambda = 0$. To prove the theorem, we use the same reduction as the one used in the proof of Theorem 1 with slight modification. Fomrally, given an instance of the X3C problem $(E, \mathcal{S})$ and parameter ϵ, let $\overline{R}_\epsilon$ be the reward function $\overline{R}$ as defined in the proof of Theorem 5 with the same underlying MDP. We have made the dependence on ϵ explicit here for reasons that will be clear shortly. As before, the underlying MDP is shown in Figure 3. Similarly, for instances of the X3C problem where an exact cover Q exists, let R_ϵ be the reward function defined in Part 1 of the same proof. Note that R_ϵ was constructed using Q and therefore implicitly depends on it. Define $\xi = \sqrt{k}$ as before.

Given these definitions, recall that our proof showed that $(\overline{R}_\epsilon, R_\epsilon, \xi)$ satisfied the following two properties.

- For instances of the problem where exact cover is possible, $\left\| R_\epsilon - \overline{R}_\epsilon \right\|_2 \leq \xi$ and R_ϵ satisfied the constraints of (P1-APT) with parameter ϵ, i.e,

  $$\textsc{Opt}^{\epsilon}_{\text{det}}(R_\epsilon) \subseteq \Pi^{\text{adm}}_{\text{det}},$$

  which is equivalent to

  $$\forall \pi \notin \Pi^{\text{adm}}_{\text{det}} : \rho^{\pi,R_\epsilon} \leq \max_{\pi'} \rho^{\pi',R_\epsilon} - \epsilon.$$

- For instances of the X3C problem where an exact cover does not exist, for any reward function $\widetilde{R}$ satisfying

  $$\forall \pi \notin \Pi^{\text{adm}}_{\text{det}} : \rho^{\pi,\widetilde{R}} \leq \max_{\pi'} \rho^{\pi',\widetilde{R}} - \epsilon,$$

  it follows that $\left\| \widetilde{R} - \overline{R}_\epsilon \right\|_2 > \sqrt{(|S| \cdot |A|)^{1-p}} \cdot \xi$.

Now, for $\eta > 0$, define the reward functions $\overline{R'}_{\epsilon,\eta} := \eta \cdot \overline{R}_{\frac{\epsilon}{\eta}}$ and $R'_{\epsilon,\eta} := \eta \cdot R_{\frac{\epsilon}{\eta}}$ and set $\xi'_\eta := \eta \cdot \xi$. Since the score function $\rho^{\pi,R}$ is linear in R, it follows that $(\overline{R'}_{\epsilon,\eta}, R'_{\epsilon,\eta}, \xi'_\eta)$ satisfy the same two properties listed above.

In order to prove the theorem statement, for a given instance of X3C and a given parameter ϵ, we will need to provide an MDP with reward function $\overline{R'}$ and a parameter ξ' such that:

- If exact cover of the X3C instance is possible, the cost of the optimization problem (P4-APT) with parameter ϵ is less than or equal to ξ'.
- If exact cover of the X3C instance is not possible, the cost of the optimization problem (P4-APT) with parameter ϵ is more than $\sqrt{(|S| \cdot |A|)^{1-p}} \cdot \xi'$.

Set $\eta = \frac{\epsilon \cdot m}{\gamma^2 \cdot (1-\gamma)}$. We claim that $\overline{R'} = \overline{R'}_{\epsilon,\eta}$ and $\xi' = \xi'_\eta$ satisfy the above properties. The second property is easy to check. Formally, if the cost of (P4-APT) is less than equal to $\sqrt{(|S| \cdot |A|)^{1-p}} \cdot \xi'$, so is the cost of (P1-APT). By the second property of $(\overline{R'}_{\epsilon,\eta}, R'_{\epsilon,\eta}, \xi'_\eta)$ discussed before, it follows that the X3C instance has an exact cover.

For the first property, assume that the X3C instance has an exact cover Q and consider $R' = R'_{\epsilon,\eta}$. By the first property of $(\overline{R'}_{\epsilon,\eta}, R'_{\epsilon,\eta}, \xi'_\eta)$, it follows that $\left\|R' - \overline{R'}\right\|_2 \leq \xi'$. Furthermore,

$$\forall \pi \notin \Pi^{\text{adm}}_{\text{det}} : \rho^{\pi,R'} \leq \max_{\pi'} \rho^{\pi',R'} - \epsilon. \tag{16}$$

Now, consider the policy $\hat{\pi}$ as follows: for each $i = 1, \ldots, 3k$, $\hat{\pi}(s_i) = a_j$ such that $e_i \in S_j$ and $j \in Q$ (such j exist and is unique given that Q is an exact set cover); and $\hat{\pi}(s_*) = a_1$. We claim that $(\overline{R'}, \hat{\pi})$ are feasible for (P4-APT) which would prove the claim.

Formally, let $\pi \neq \hat{\pi}$ be a deterministic policy. We need to show that $\rho^{\pi,R'} \leq \rho^{\hat{\pi},R'} - \epsilon$. We assume without loss of generality that $\pi \in \Pi^{\text{adm}}_{\text{det}}$. If we prove this, then since the optimal policy in R' is admissible by (16), it follows that $\hat{\pi}$ is the optimal policy and therefore the case of $\pi \notin \Pi^{\text{adm}}_{\text{det}}$ is covered by (16).

With this assumption in mind, note that since π and $\hat{\pi}$ are admissible, there for each $i \in \{1, \ldots, 3k\}$, there exist $j(i)$ and $\hat{j}(i)$ such that $\pi(s_i) = t_{j(i)}$ and $\hat{\pi}(s_i) = t_{\hat{j}(i)}$. Furthermore, $\pi(s_*) = \hat{\pi}(s_*)$ since both policies are admissible. It therefore follows that

$$\begin{aligned}
\frac{\rho^{\hat{\pi},R'} - \rho^{\pi,R'}}{1-\gamma} &= \frac{\gamma}{m} \cdot \left(\sum_{i=1}^{3k} \left(V^{\hat{\pi},R'}(s_i) - V^{\pi,R'}(s_i) \right) \right) \\
&= \frac{\gamma}{m} \cdot \left(\sum_{i=1}^{3k} \left(x_{ij} + \gamma \cdot V^{\hat{\pi},R'}(t_{\hat{j}(i)}) - x_{ij} - \gamma \cdot V^{\pi,R'}(t_{j(i)}) \right) \right) \\
&= \frac{\gamma^2}{m} \cdot \left(\sum_{i=1}^{3k} \left(V^{\hat{\pi},R'}(t_{\hat{j}(i)}) - V^{\pi,R'}(t_{j(i)}) \right) \right)
\end{aligned}$$

Note however that by construction of R', for any $1 \leq j \leq l$,

$$V^{\hat{\pi},R'}(t_j) = R'_{\epsilon,\eta}(t_j) = \eta \cdot R_{\frac{\epsilon}{\eta}}(t_j) = \eta \cdot \mathbb{1}\left[S_j \in Q\right].$$

It therefore follows that $V^{\hat{\pi},R'}(t_{\hat{j}(i)}) = \eta$ since Q is a cover. Furthermore, since Q is an *exact* cover, e_i is only covered by $S_{\hat{j}(i)}$ which means $V^{\pi,R'}(t_{j(i)}) = \eta \cdot \mathbb{1}\left[j(i) = \hat{j}(i)\right]$. Threfore,

$$\begin{aligned}
\frac{\rho^{\hat{\pi},R'} - \rho^{\pi,R'}}{1-\gamma} &= \frac{\gamma^2}{m} \cdot \left(\sum_{i=1}^{3k} \left(V^{\hat{\pi},R'}(t_{\hat{j}(i)}) - V^{\pi,R'}(t_{j(i)}) \right) \right) \\
&= \frac{\gamma^2}{m} \cdot \eta \cdot \left(\sum_{i=1}^{3k} \mathbb{1}\left[\pi(s_i) \neq \hat{\pi}(s_i)\right] \right)
\end{aligned}$$

Note however that $\pi(s_i) \neq \hat{\pi}(s_i)$ for some $1 \leq i \leq 3k$. This is because $\pi \neq \hat{\pi}$ and since both are admissible policies, they can only disagree on some state s_i. Therefore, $\sum_{i=1}^{3k} \mathbb{1}\left[\pi(s_i) \neq \hat{\pi}(s_i)\right] \geq 1$ which implies

$$\rho^{\hat{\pi},R'} - \rho^{\pi,R'} \geq (1-\gamma) \cdot \frac{\gamma^2}{m} \cdot \eta = \epsilon,$$

which completes the proof. □

Proof of Proposition 2

Statement: *Let $\widehat{R}_1$ and $\widehat{R}_2$ be the optimal solutions to* (P1-APT) *and* (P4-APT) *respectively and let $l(R)$ be a function that outputs the objective of the optimization problem* (P1-APT)*, i.e.,*

$$l(R) = \max_{\pi \in \text{OPT}^{\epsilon}_{\text{det}}(R)} \left\| \overline{R} - R \right\|_2 - \lambda \rho^{\pi, \overline{R}}.$$

Then $\widehat{R}_2$ satisfies the constraints of (P1-APT)*, i.e,* $\text{OPT}^{\epsilon}_{\text{det}}(\widehat{R}_2) \subseteq \Pi^{\text{adm}}_{\text{det}}$*, and*

$$l(\widehat{R}_1) \leq l(\widehat{R}_2) \leq l(\widehat{R}_1) + \frac{\epsilon}{\mu_{\min}} \cdot \sqrt{|S| \cdot |A|}.$$

Proof. We first prove that $\widehat{R}_2$ is feasible for (P1-APT). This would prove the left inequlity given the optimality of $\widehat{R}_1$. In order to prove this, we just need to show that for all $\pi \in \Pi^{\text{adm}}_{\text{det}}$,

$$\{\pi' : \pi'(s) = \pi(s) \text{ if } \mu^{\pi}(s) > 0\} \subseteq \Pi^{\text{adm}}_{\text{det}}$$

Given Lemma 3, if $\pi'(s) = \pi(s)$ for all s such that $\mu^{\pi}(s) > 0$, then $\mu^{\pi} = \mu^{\pi'}$. Therefore, for all states s, either $\mu^{\pi}(s) > 0$, in which case $\pi'(s) = \pi(s) \in A^{\text{adm}}_s$, or $\mu^{\pi}(s) = 0$, in which case $\mu^{\pi'}(s) = 0$. Therefore, $\pi' \in \Pi^{\text{adm}}_{\text{det}}$ and the claim is proved.

As for the right inequality, let π_1 be an optimal policy under $\widehat{R}_1$. Given the constraints of (P1-APT), $\pi_1 \in \Pi^{\text{adm}}_{\text{det}}$. Define R' as

$$R'(s,a) = \begin{cases} \widehat{R}_1(s,a) - \frac{\epsilon}{\mu_{\min}} & \text{if} \quad a \neq \pi_1(s) \wedge \mu^{\pi_1}(s) > 0 \\ \widehat{R}_1(s,a) & \text{o.w.} \end{cases}.$$

We first show that R', π_1 are feasible for (P4-APT). Let π be a policy such that $\pi(\tilde{s}) \neq \pi_1(\tilde{s})$ for some $\tilde{s}$ that satisfies $\mu^{\pi_1}(\tilde{s}) > 0$. We claim that this means there exists a state $s \in S^{\pi_1}_{\text{pos}} \cap S^{\pi}_{\text{pos}}$ such that $\pi(s) \neq \pi_1(s)$. If this is not the case, then Lemma 3 (from section Background) implies that $\mu^{\pi_1} = \mu^{\pi}$. This further implies that $S^{\pi}_{\text{pos}} = S^{\pi_1}_{\text{pos}}$ and therefore since $\tilde{s} \in S^{\pi_1}_{\text{pos}}$ and $\pi(\tilde{s}) \neq \pi_1(\tilde{s})$, we have reached a contradiction. Therefore, there exists a state state $s \in S^{\pi_1}_{\text{pos}} \cap S^{\pi}_{\text{pos}}$ such that $\pi(s) \neq \pi_1(s)$. Without loss of generality, assume that $\tilde{s}$ is this state.
It follows that

$$\rho^{\pi_1, R'} - \rho^{\pi, R'} = (\rho^{\pi_1, \widehat{R}_1} - \rho^{\pi, \widehat{R}_1}) + (\rho^{\pi_1, R'} - \rho^{\pi_1, \widehat{R}_1}) + (\rho^{\pi, \widehat{R}_1} - \rho^{\pi, R'})$$

The first term is non-negative since π_1 was assumed to be optimal under $\widehat{R}_1$. The second term equals zero given the definition of R'. As for the last term,

$$\begin{aligned}
\rho^{\pi, \widehat{R}_1} - \rho^{\pi, R'} &= \sum_s \mu^{\pi}(s)(\widehat{R}_1(s, \pi(s)) - R'(s, \pi(s))) \\
&= \mu^{\pi}(\tilde{s})(\widehat{R}_1(\tilde{s}, \pi(\tilde{s})) - R'(\tilde{s}, \pi(\tilde{s})) + \sum_{s \neq \tilde{s}} \mu^{\pi}(s)(\widehat{R}_1(s, \pi(s)) - R'(s, \pi(s))) \\
&= \mu^{\pi}(\tilde{s}) \cdot \frac{\epsilon}{\mu_{\min}} + \sum_{s \neq \tilde{s}} \mu^{\pi}(s)(\widehat{R}_1(s, \pi(s)) - R'(s, \pi(s))) \\
&\overset{(i)}{\geq} \mu^{\pi}(\tilde{s}) \cdot \frac{\epsilon}{\mu_{\min}} \\
&\overset{(ii)}{\geq} \epsilon
\end{aligned}$$

where (i) follows from the fact that $\widehat{R}_1 - R'$ is non-negative and (ii) follows from the definition of $\mu_{\min}$ and the fact $\tilde{s} \in S^{\pi}_{\text{pos}}$. Therefore, R', π_1 are feasible for (P4-APT).

Now, note that

$$\begin{aligned}
\left\| \overline{R} - R' \right\|_2 &\leq \left\| \overline{R} - \widehat{R}_1 \right\|_2 + \left\| \widehat{R}_1 - R' \right\|_2 \\
&\leq \left\| \overline{R} - \widehat{R}_1 \right\|_2 + \sqrt{\sum_{s, a \neq \pi'(s)} \left(\frac{\epsilon}{\mu_{\min}} \right)^2} \\
&\leq \left\| \overline{R} - \widehat{R}_1 \right\|_2 + \frac{\epsilon}{\mu_{\min}} \sqrt{|S| \cdot |A|}.
\end{aligned}$$

This means that

$$\begin{aligned}\left\|\overline{R}-R'\right\|_2-\lambda\rho^{\pi_1,\overline{R}} &\le \left\|\overline{R}-\widehat{R}_1\right\|_2-\lambda\rho^{\pi_1,\overline{R}}+\frac{\epsilon}{\mu_{\min}}\sqrt{|S|\cdot|A|}\\ &\le \max_{\pi\in \textsc{Opt}^{\epsilon}_{\text{det}}(\widehat{R}_1)}\left\|\overline{R}-\widehat{R}_1\right\|_2-\lambda\cdot\rho^{\pi,\overline{R}}+\frac{\epsilon}{\mu_{\min}}\sqrt{|S|\cdot|A|}\\ &= l(\widehat{R}_1)+\frac{\epsilon}{\mu_{\min}}\sqrt{|S|\cdot|A|},\end{aligned}$$

where the second inequality is due to the fact that π_1 is optimal under $\widehat{R}_1$, so it belongs to the set $\textsc{Opt}^{\epsilon}_{\text{det}}(\widehat{R}_1)$.

Now, let π_2 be a deterministic optimal policy under $\widehat{R}_2$. Since $\widehat{R}_2$ was the solution to (P4-APT), for any $\pi\in\textsc{Opt}^{\epsilon}_{\text{det}}(\widehat{R}_2)$, it holds that $\pi(s)=\pi_2(s)$ for all $s\in S^{\pi_2}_{\text{pos}}$ and therefore by Lemma 3, $\rho^{\pi,\overline{R}}=\rho^{\pi_2,\overline{R}}$. Denote the solution to the optimization problem (P3-ATK) (reward poisoning attack) for the target policy $\pi_\dagger=\pi_1$ by $\widehat{R}_{\mathcal{A}}$. We obtain

$$\begin{aligned}l(\widehat{R}_2) &= \max_{\pi\in\textsc{Opt}^{\epsilon}_{\text{det}}(\widehat{R}_2)}\left\|\overline{R}-\widehat{R}_2\right\|_2-\lambda\cdot\rho^{\pi,\overline{R}}\\ &= \left\|\overline{R}-\widehat{R}_2\right\|_2-\lambda\rho^{\pi_2,\overline{R}}\\ &\overset{(i)}{\le} \left\|\overline{R}-\widehat{R}_{\mathcal{A}}\right\|_2-\lambda\cdot\rho^{\pi_1,\overline{R}}\\ &\overset{(ii)}{\le} \left\|\overline{R}-R'\right\|_2-\lambda\cdot\rho^{\pi_1,\overline{R}}\\ &\le l(\widehat{R}_1)+\frac{\epsilon}{\mu_{\min}}\cdot\sqrt{|S|\cdot|A|}.\end{aligned}$$

where (i) follows from the optimality of $\widehat{R}_2$ for (P4-APT) and (ii) follows from the definition of the reward poisoning attack (P3-ATK). We have therefore shown the right inequality of the Lemma's statement holds and the proof is complete. □

Proofs of the Results from Section Characterization Results for Special MDPs

In this section, we provide proofs of our results in Section Characterization Results for Special MDPs: Lemma 1 and Theorem 2. Recall that for special MDPs since the transition probabilities are independent of the agent's policy, so is the state occupancy measure. Concretely, since the Bellman flow constraint (2) characterizing μ^π is independent of policy, so is μ^π. We therefore use μ instead of μ^π to denote the state occupancy measure for special MDPs. Similarly, since S^π_{pos} depends on π only through μ^π, we use S_{pos} instead of S^π_{pos}.

Before we prove the main results, we present and prove the following two lemmas.

Lemma 7. *Consider a special MDP with a reward function $\overline{R}$ and a policy of interest $\pi_\dagger$. For all $s\in S$ s.t. $\mu(s)>0$, there exists a unique x such that:*

$$\sum_{a\neq\pi_\dagger(s)}\left[\overline{R}(s,a)-x\right]^+ = x-\overline{R}(s,\pi_\dagger(s))+\frac{\epsilon}{\mu(s)}. \tag{17}$$

Proof. Fix the state s. Consider the following function

$$f:\mathbb{R}\to\mathbb{R},\quad f(x):=\sum_{a\neq\pi_\dagger(s)}\left[\overline{R}(s,a)-x\right]^+-x.$$

This function is strictly decreasing because $x\to -x$ is strictly decreasing and $x\to\left[\overline{R}(s,a)-x\right]^+$ is decreasing. Furthermore, $\lim_{x\to\infty}f(x)=-\infty$ and $\lim_{x\to-\infty}f(x)=\infty$. Therefore given the intermediate value theorem, there exists a unique number x such that

$$f(x)=-\overline{R}(s,\pi_\dagger(s))+\frac{\epsilon}{\mu(s)}.$$

□

Lemma 8. *Consider a special MDP with reward function $\overline{R}$. Let $\pi_\dagger\in\Pi_{det}$ be an arbitrary deterministic policy. Define $\widehat{R}^{\pi_\dagger}$ as*

$$\widehat{R}^{\pi_\dagger}(s,a)=\begin{cases}x_s+\frac{\epsilon}{\mu(s)} & \text{if } \mu(s)>0\wedge a=\pi_\dagger(s)\\ x_s & \text{if } \mu(s)>0\wedge a\neq\pi_\dagger(s)\wedge\overline{R}(s,a)\ge x_s\\ \overline{R}(a,s) & \text{otherwise}\end{cases},$$

where x_s is the solution to the Eq. (17)*. $\widehat{R}^{\pi_\dagger}$ is an optimal solution to the optimization problem* (P3-ATK)*.*

Proof. In order to show feasibility, let $\tilde{\pi} \in \Pi_{\text{det}}$ be a deterministic policy such that $\tilde{\pi}(\tilde{s}) \neq \pi_\dagger(s)$ for some $\tilde{s}$ such that $\mu(\tilde{s}) > 0$. It is clear that

$$\begin{aligned}
\rho^{\pi_\dagger, \widehat{R}^{\pi_\dagger}} - \rho^{\tilde{\pi}, \widehat{R}^{\pi_\dagger}} &= \sum_s \mu(s)\big(\widehat{R}^{\pi_\dagger}(s, \pi_\dagger(s)) - \widehat{R}^{\pi_\dagger}(s, \tilde{\pi}(s))\big) \\
&= \mu(\tilde{s})\big(\widehat{R}^{\pi_\dagger}(\tilde{s}, \pi_\dagger(\tilde{s})) - \widehat{R}^{\pi_\dagger}(\tilde{s}, \tilde{\pi}(\tilde{s}))\big) + \sum_{s \neq \tilde{s}} \mu(s)\big(\widehat{R}^{\pi_\dagger}(s, \pi_\dagger(s)) - \widehat{R}^{\pi_\dagger}(s, \tilde{\pi}(s))\big) \\
&\overset{(i)}{\geq} \mu(\tilde{s})\big(\widehat{R}^{\pi_\dagger}(\tilde{s}, \pi_\dagger(\tilde{s})) - \widehat{R}^{\pi_\dagger}(\tilde{s}, \tilde{\pi}(\tilde{s}))\big) \\
&\overset{(ii)}{\geq} \mu(\tilde{s})\frac{\epsilon}{\mu(\tilde{s})} \geq \epsilon
\end{aligned}$$

where (i) and (ii) both follow from the definition of $\widehat{R}^{\pi_\dagger}$; (i) follows from the fact that $\widehat{R}^{\pi_\dagger}(s, \pi_\dagger(s)) \geq \widehat{R}^{\pi_\dagger}(s, a)$ for all s, a such that $\mu(s) > 0$ and (ii) follows from the fact that $\tilde{\pi}(\tilde{s}) \neq \pi_\dagger(\tilde{s})$.

We now show that if R is also feasible for the optimization problem (P3-ATK), then $\left\|\overline{R} - R\right\|_2 \geq \left\|\overline{R} - \widehat{R}^{\pi_\dagger}\right\|_2$. The key point about $\widehat{R}^{\pi_\dagger}$, is the definition of x_s in Equation (17). Concretely, it is clear that for $s \in S_{\text{pos}}$:

$$\begin{aligned}
\overline{R}(s, \pi_\dagger(s)) - \widehat{R}^{\pi_\dagger}(s, \pi_\dagger(s)) &= \overline{R}(s, \pi_\dagger(s)) - x_s - \frac{\epsilon}{\mu(s)} \\
&= -\big(x_s + \frac{\epsilon}{\mu(s)} - \overline{R}(s, \pi_\dagger(s))\big) \\
&\overset{(17)}{=} -\sum_{a \neq \pi_\dagger(s)} \big[\overline{R}(s, a) - x_s\big]^+ \\
&= -\sum_{a \neq \pi_\dagger(s)} \mathbb{1}\left[\overline{R}(s, a) > x_s\right] \cdot (\overline{R}(s, a) - x_s) \\
&= -\sum_{a \neq \pi_\dagger(s)} (\overline{R}(s, a) - \widehat{R}^{\pi_\dagger}(s, a)) \\
&= \sum_{a \neq \pi_\dagger(s)} (\widehat{R}^{\pi_\dagger}(s, a) - \overline{R}(s, a)) \qquad (18)
\end{aligned}$$

Now note that

$$\begin{aligned}
\left\|\overline{R} - R\right\|_2^2 - \left\|\overline{R} - \widehat{R}^{\pi_\dagger}\right\|_2^2 &= \left\langle \overline{R} - R - \overline{R} + \widehat{R}^{\pi_\dagger}, \overline{R} - R + \overline{R} - \widehat{R}^{\pi_\dagger} \right\rangle \\
&= \left\langle \widehat{R}^{\pi_\dagger} - R, \overline{R} - R + \overline{R} - \widehat{R}^{\pi_\dagger} \right\rangle \\
&= \left\langle \widehat{R}^{\pi_\dagger} - R, \overline{R} - R + \overline{R} + \widehat{R}^{\pi_\dagger} - 2\widehat{R}^{\pi_\dagger} \right\rangle \\
&= \left\|\widehat{R}^{\pi_\dagger} - R\right\|_2^2 + 2\left\langle \widehat{R}^{\pi_\dagger} - R, \overline{R} - \widehat{R}^{\pi_\dagger} \right\rangle \\
&\geq 2\left\langle \widehat{R}^{\pi_\dagger} - R, \overline{R} - \widehat{R}^{\pi_\dagger} \right\rangle \\
&= 2\Big(\left\langle \overline{R} - \widehat{R}^{\pi_\dagger}, \widehat{R}^{\pi_\dagger} \right\rangle - \left\langle \overline{R} - \widehat{R}^{\pi_\dagger}, R \right\rangle \Big)
\end{aligned}$$

It therefore suffices to show that the above quantity is non-negative. Note however,

$$\begin{aligned}
\left\langle \overline{R} - \widehat{R}^{\pi_\dagger}, R \right\rangle &= \sum_{(s,a)} (\overline{R}(s, a) - \widehat{R}^{\pi_\dagger}(s, a)) \cdot R(s, a) \\
&= \sum_{s \in S_{\text{pos}}, a} (\overline{R}(s, a) - \widehat{R}^{\pi_\dagger}(s, a)) \cdot R(s, a) + \sum_{s \notin S_{\text{pos}}, a} (\overline{R}(s, a) - \widehat{R}^{\pi_\dagger}(s, a)) \cdot R(s, a) \\
&= \sum_{s \in S_{\text{pos}}, a} (\overline{R}(s, a) - \widehat{R}^{\pi_\dagger}(s, a)) \cdot R(s, a) + \sum_{s \notin S_{\text{pos}}, a} 0 \cdot R(s, a) \\
&= \sum_{s \in S_{\text{pos}}, a} (\overline{R}(s, a) - \widehat{R}^{\pi_\dagger}(s, a)) \cdot R(s, a)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{s\in S_{\text{pos}}} (\overline{R}(s,\pi_\dagger(s)) - \widehat{R}^{\pi_\dagger}(s,\pi_\dagger(s))) \cdot R(s,\pi_\dagger(s)) + \sum_{s\in S_{\text{pos}}, a\neq\pi_\dagger(s)} (\overline{R}(s,a) - \widehat{R}^{\pi_\dagger}(s,a)) \cdot R(s,a) \\
&\overset{(i)}{=} \sum_{s\in S_{\text{pos}}} R(s,\pi_\dagger(s)) \cdot \sum_{a\neq\pi_\dagger(s)} (\widehat{R}^{\pi_\dagger}(s,a) - \overline{R}(s,a)) + \sum_{s\in S_{\text{pos}}, a\neq\pi_\dagger(s)} (\overline{R}(s,a) - \widehat{R}^{\pi_\dagger}(s,a)) \cdot R(s,a) \\
&= \sum_{s\in S_{\text{pos}}, a\neq\pi_\dagger(s)} \big(\overline{R}(s,a) - \widehat{R}^{\pi_\dagger}(s,a)\big) \cdot \big(R(s,a) - R(s,\pi_\dagger(s))\big). \qquad (19)
\end{aligned}
$$

where (i) follows from (18). Now note that since R was assumed to be feasible, if $s \in S_{\text{pos}}$ and $a \neq \pi_\dagger(s)$, then by defining $\tilde{\pi}$ as the policy that chooses $\pi_\dagger(\tilde{s})$ in states $\tilde{s} \neq s$ and chooses a in state s, it follows that

$$
\begin{aligned}
\frac{\epsilon}{\mu(s)} &\leq \frac{\rho^{\pi_\dagger,R} - \rho^{\tilde{\pi},R}}{\mu(s)} \\
&= \sum_{\tilde{s}} \frac{\mu(\tilde{s})\big(R(\tilde{s},\pi_\dagger(\tilde{s})) - R(\tilde{s},\tilde{\pi}(\tilde{s}))\big)}{\mu(s)} \\
&= \frac{\mu(s)\big(R(s,\pi_\dagger(s)) - R(s,\tilde{\pi}(s))\big)}{\mu(s)} \\
&= \frac{\mu(s)\big(R(s,\pi_\dagger(s)) - R(s,a)\big)}{\mu(s)} \\
&= R(s,\pi_\dagger(s)) - R(s,a).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\Big\langle \overline{R} - \widehat{R}^{\pi_\dagger}, R \Big\rangle &= \sum_{s\in S_{\text{pos}}, a\neq\pi_\dagger(s)} \big(\overline{R}(s,a) - \widehat{R}^{\pi_\dagger}(s,a)\big) \cdot \big(R(s,a) - R(s,\pi_\dagger(s))\big) \\
&\leq - \sum_{s\in S_{\text{pos}}, a\neq\pi_\dagger(s)} \big(\overline{R}(s,a) - \widehat{R}^{\pi_\dagger}(s,a)\big) \cdot \big(\frac{\epsilon}{\mu(s)}\big)
\end{aligned}
$$

Using (19) for $\widehat{R}^{\pi_\dagger}$ (note that since no assumptions on R were made in deriving the identity, it is valid for $R = \widehat{R}^{\pi_\dagger}$),

$$
\begin{aligned}
\Big\langle \overline{R} - \widehat{R}^{\pi_\dagger}, \widehat{R}^{\pi_\dagger} \Big\rangle &= \sum_{s\in S_{\text{pos}}, a\neq\pi_\dagger(s)} \big(\overline{R}(s,a) - \widehat{R}^{\pi_\dagger}(s,a)\big) \cdot \big(\widehat{R}^{\pi_\dagger}(s,a) - \widehat{R}^{\pi_\dagger}(s,\pi_\dagger(s))\big) \\
&\overset{(i)}{=} - \sum_{s\in S_{\text{pos}}, a\neq\pi_\dagger(s)} \big(\overline{R}(s,a) - \widehat{R}^{\pi_\dagger}(s,a)\big) \cdot \big(\frac{\epsilon}{\mu(s)}\big)
\end{aligned}
$$

where (i) follows form the definition of $\widehat{R}^{\pi_\dagger}$. Concretely, for state action pairs $s \in S_{\text{pos}}, a \neq \pi_\dagger(s)$ such that $\widehat{R}^{\pi_\dagger}(s,a) \neq \overline{R}(s,a)$, $\overline{R}(s,a) \geq x_s$ and therefore

$$
\widehat{R}^{\pi_\dagger}(s,a) = x_s = \widehat{R}^{\pi_\dagger}(s,\pi_\dagger(s)) - \frac{\epsilon}{\mu(s)}
$$

We therefore obtain $\Big\langle \overline{R} - \widehat{R}^{\pi_\dagger}, \widehat{R}^{\pi_\dagger} \Big\rangle \geq \Big\langle \overline{R} - \widehat{R}^{\pi_\dagger}, R \Big\rangle$ and the proof is concluded. □

Proof of Lemma 1

Statement: *Consider a* special *MDP with reward function $\overline{R}$, and let $\pi^*_{\text{adm}}(s) = \arg\max_{a\in A_s^{\text{adm}}} \overline{R}(s,a)$. Then the cost of the optimal solution to the optimization problem* (P3-ATK) *with $\pi_\dagger = \pi^*_{\text{adm}}$ is less than or equal to the cost of the optimal solution to the optimization problem* (P3-ATK) *with $\pi_\dagger = \pi$ for any $\pi \in \Pi^{\text{adm}}_{\text{det}}$.*

Proof.
Part 1: We first prove the claim for single-state MDPs which are equivalent to multi-arm bandits. Since there is a single state, for ease of notation we drop the dependence on the state s when referring to quantities that would normally depend on s such as $\overline{R}(s,a)$ and $\pi(s)$.
Let a_1, a_2 be two actions such that $\overline{R}(a_1) \geq \overline{R}(a_2)$. Denote by π_1 and π_2 the policies that deterministically choose a_1 and

a_2 respectively and and denote by $\widehat{R}^{\pi_1}, \widehat{R}^{\pi_2}$ the solutions to the optimization problem (P3-ATK) with $\pi_\dagger = \pi_1$ and $\pi_\dagger = \pi_2$ respectively. We will show that

$$\left\|\overline{R} - \widehat{R}^{\pi_2}\right\|_2 \geq \left\|\overline{R} - \widehat{R}^{\pi_1}\right\|_2.$$

Consider the reward function R' defined as

$$R'(a) = \begin{cases} \widehat{R}^{\pi_2}(a_1) & \text{if} \quad a = a_2 \\ \widehat{R}^{\pi_2}(a_2) & \text{if} \quad a = a_1 \\ \widehat{R}^{\pi_2}(a) & \text{o.w.} \end{cases};$$

In other words, we have swtiched the reward for a_1 and a_2 in $\widehat{R}^{\pi_2}$.

We claim that

$$||\overline{R} - \widehat{R}^{\pi_2}||_2 \geq ||\overline{R} - R'||_2 \geq ||\overline{R} - \widehat{R}^{\pi_1}||_2.$$

To prove the second inequality, note that a_1 is ϵ-robust optimal in R' since a_2 was ϵ-robust optimal in R' and R' was obtained from switching a_1, a_2 in $\widehat{R}^{\pi_2}$. Therefore the inequality follows from the optimality of $\widehat{R}^{\pi_1}$.

As for the first inequality, note that it can be rewritten as

$$\begin{aligned}
&||\overline{R} - \widehat{R}^{\pi_2}||_2^2 \geq ||\overline{R} - R'||_2^2 \\
\iff &\sum_i (\overline{R}(a_i) - \widehat{R}^{\pi_2}(a_i))^2 \geq \sum_i (\overline{R}(a_i) - R'(a_i))^2 \\
\overset{(i)}{\iff} &(\overline{R}(a_1) - \widehat{R}^{\pi_2}(a_1))^2 + (\overline{R}(a_2) - \widehat{R}^{\pi_2}(a_2))^2 \geq (\overline{R}(a_1) - \widehat{R}^{\pi_2}(a_2))^2 + (\overline{R}(a_2) - \widehat{R}^{\pi_2}(a_1))^2 \\
\iff &-2\overline{R}(a_1)\widehat{R}^{\pi_2}(a_1) - 2\overline{R}(a_2)\widehat{R}^{\pi_2}(a_2) \geq -2\overline{R}(a_1)\widehat{R}^{\pi_2}(a_2) - 2\overline{R}(a_2)\widehat{R}^{\pi_2}(a_1) \\
\iff &\overline{R}(a_1)\widehat{R}^{\pi_2}(a_2) + \overline{R}(a_2)\widehat{R}^{\pi_2}(a_1) - \overline{R}(a_1)\widehat{R}^{\pi_2}(a_1) - \overline{R}(a_2)\widehat{R}^{\pi_2}(a_2) \geq 0 \\
\iff &(\overline{R}(a_1) - \overline{R}(a_2))(\widehat{R}^{\pi_2}(a_2) - \widehat{R}^{\pi_2}(a_1)) \geq 0,
\end{aligned}$$

where (i) follows from the definition of $R'(a_i)$. Note however that the last equation holds trivially since $\overline{R}(a_1) \geq \overline{R}(a_2)$ by assumption and $\widehat{R}^{\pi_2}(a_2) > \widehat{R}^{\pi_2}(a_1)$ by definition of $\widehat{R}^{\pi_2}$. Note further that if $\overline{R}(a_1) > \overline{R}(a_2)$, then the inequality is strict.

The statement of the lemma now follows by setting $a_1 = \arg\max_{a \in A_s^{\text{adm}}} \overline{R}(a)$ and a_2 to be any $a \in A_s^{\text{adm}}$.

Part 2: We now extend this result to multi-state special MDPs. Since the MDP is special, given Lemma 8 we can view the attack as separate single-state attacks for $s \in S_{\text{pos}}$ with parameters $\frac{\epsilon}{\mu(s)}$. Since $\pi^*_{\text{adm}}(s)$ equals $\arg\max_{a \in A_s^{\text{adm}}} \overline{R}(s,a)$ by definition, Part 1 implies that the cost of the optimization problem (P3-ATK) with $\pi_\dagger = \pi^*_{\text{adm}}$ is not more than the cost of the optimization problem with $\pi_\dagger = \pi$ for all $\pi \in \Pi^{\text{adm}}_{\text{det}}$. □

Proof of Theorem 2

Statement: *Consider a special MDP with reward function $\overline{R}$. Define $\widehat{R}(s,a) = \overline{R}(s,a)$ for $\mu(s) = 0$ and otherwise*

$$\widehat{R}(s,a) = \begin{cases} x_s + \frac{\epsilon}{\mu(s)} & \textit{if } a = \pi^*_{\text{adm}}(s) \\ x_s & \textit{if } a \neq \pi^*_{\text{adm}}(s) \wedge \overline{R}(s,a) \geq x_s \\ \overline{R}(s,a) & \textit{otherwise} \end{cases},$$

where x_s is the solution to the equation

$$\sum_{a \neq \pi^*_{\text{adm}}(s)} \left[\overline{R}(s,a) - x\right]^+ = x - \overline{R}(s, \pi^*_{\text{adm}}(s)) + \frac{\epsilon}{\mu(s)}.$$

Then, $(\pi^*_{\text{adm}}, \widehat{R})$ is an optimal solution to (P4-APT).

Proof. The claim follows from Lemmas 1 and 8. Concretely, given Lemma 1, the solution to the optimization problem (P4-APT) is π^*_{adm} and $\widehat{R}^{\pi_\dagger}$ where $\widehat{R}^{\pi_\dagger}$ is the solution to (P3-ATK) with $\pi_\dagger = \pi^*_{\text{adm}}$. The claim now follows from Lemma 8 which characterizes $\widehat{R}^{\pi_\dagger}$.

□

Proofs of the Results in Section Characterization Results for General MDPs

In this section, we provide proofs of our results in Section Characterization Results for General MDPs, namely Theorem 3 and Theorem 4. Before we present the proofs, we introduce and prove two auxiliary lemmas.

Lemma 9. *For an arbitrary policy π, define R', V, Q as*

$$R'(s,a) = \begin{cases} \overline{R}(s,a) + \delta^\pi(s) & \text{if} \quad \mu^\pi(s) > 0 \text{ and } a = \pi(s) \\ \overline{R}(s,a) - \epsilon'_\pi(s,a) & \text{if} \quad \mu^\pi(s) > 0 \text{ and } a \neq \pi(s) \\ \overline{R}(s,a) & \text{o.w.} \end{cases}$$

$$V(s) = V^{*,\overline{R}}(s)$$

$$Q(s,a) = R'(s,a) + \gamma \sum_{s'} P(s,a,s')V(s'),$$

where

$$\delta^\pi(s) = \begin{cases} Q^{*,\overline{R}}(s,\pi^*(s)) - Q^{*,\overline{R}}(s,\pi(s)) & \text{if} \quad \mu^\pi(s) > 0 \\ 0 & \text{otherwise} \end{cases},$$

and ϵ'_π is defined as in (10)*. The vectors R', V, Q are feasible in* (P5-ATK) *with ϵ' set to ϵ'_π and $\pi_\dagger = \pi$. Furthermore, given Proposition 3, R' is feasible for* (P3-ATK) *with $\pi_\dagger = \pi$ and therefore (R', π) are feasible for* (P4-APT)*.*

Proof. We check all of the conditions. (4) holds by definition of Q. Now note that since $V = V^{*,\overline{R}}$ and $R'(s,a) = \overline{R}(s,a)$ for all $s \notin S^\pi_{\text{pos}}$, we conclude that $Q(s,a) = Q^{*,\overline{R}}(s,a)$ for all $s \notin S^\pi_{\text{pos}}$. Therefore, since $V^{*,\overline{R}}(s) \geq Q^{*,\overline{R}}(s,a)$ for all s, a, the constraint (7) holds as well. The constraint (6) holds by definition of δ^π as

$$\begin{aligned} Q(s,\pi(s)) - V(s) &= R'(s,\pi(s)) + \gamma \sum_{s'} P(s,\pi(s),s')V^{*,\overline{R}}(s') - V^{*,\overline{R}}(s) \\ &= \overline{R}(s,\pi(s)) + \delta^\pi(s) + \gamma \sum_{s'} P(s,\pi(s),s')V^{*,\overline{R}}(s') - V^{*,\overline{R}}(s) \\ &= \overline{R}(s,\pi(s)) + V^{*,\overline{R}}(s) - Q^{*,\overline{R}}(s,\pi(s)) + \gamma \sum_{s'} P(s,\pi(s),s')V^{*,\overline{R}}(s') - V^{*,\overline{R}}(s) \\ &= \left(\overline{R}(s,\pi(s)) + \gamma \sum_{s'} P(s,\pi(s),s')V^{*,\overline{R}}(s') - Q^{*,\overline{R}}(s,\pi(s))\right) + \left(V^{*,\overline{R}}(s) - V^{*,\overline{R}}(s)\right) \\ &= 0. \end{aligned}$$

Furthermore, (5) holds because for all $s \in S^\pi_{\text{pos}}, a \neq \pi(s)$,

$$\begin{aligned} Q(s,a) &= \overline{R}(s,a) - \epsilon'(s,a) + \gamma \sum_{s'} P(s,a,s')V^{*,\overline{R}}(s') \\ &= Q^{*,\overline{R}}(s,a) - \epsilon'(s,a) \\ &\leq V^{*,\overline{R}}(s) - \epsilon'(s,a) \\ &\overset{(6)}{=} Q(s,\pi(s)) - \epsilon'(s,a). \end{aligned}$$

Therefore, all 4 sets of constraints are satisfied which proves feasibility. Finally, given Proposition 3, R' is feasible for (P3-ATK) with $\pi_\dagger = \pi$ and therefore R', π are feasible for (P4-APT). □

Lemma 10. *Lemma 7 in (Ma et al. 2019) For arbitrary reward functions R_1 and R_2,*

$$(1-\gamma) \cdot \left\| Q^{*,R_1} - Q^{*,R_2} \right\|_\infty \leq ||R_1 - R_2||_\infty$$

Corollary 1. *Let $\widehat{R}^\pi$ be the solution to the optimization problem* (P3-ATK) *with $\pi_\dagger = \pi$. Define Δ^π_Q as in* (13)*. The following holds.*

$$\left\| \overline{R} - \widehat{R}^\pi \right\|_2 \geq \frac{1-\gamma}{2} \cdot \Delta^\pi_Q$$

Proof. Define $s_{\max}$ as

$$\arg\max_{s\in S^{\pi}_{\text{pos}}} \left(Q^{*,\overline{R}}(s,\pi^*(s)) - Q^{*,\overline{R}}(s,\pi(s))\right).$$

Since $s_{\max} \in S^{\pi}_{\text{pos}}$ and π is optimal in $\widehat{R}^{\pi}$, it is clear that $Q^{*,\widehat{R}^{\pi}}(s_{\max},\pi(s_{\max})) \geq Q^{*,\widehat{R}^{\pi}}(s_{\max},\pi^*(s_{\max}))$. However, $Q^{*,\overline{R}}(s_{\max},\pi(s_{\max})) \leq Q^{*,\overline{R}}(s_{\max},\pi^*(s_{\max})) - \Delta^{\pi}_{Q}$. Summing up the two inequalities,

$$\begin{aligned}
\Delta^{\pi}_{Q} &\leq Q^{*,\overline{R}}(s_{\max},\pi^*(s_{\max})) - Q^{*,\overline{R}}(s_{\max},\pi(s_{\max})) + Q^{*,\widehat{R}^{\pi}}(s_{\max},\pi(s_{\max})) - Q^{*,\widehat{R}^{\pi}}(s_{\max},\pi^*(s_{\max})) \\
&= \left[Q^{*,\overline{R}}(s_{\max},\pi^*(s_{\max})) - Q^{*,\widehat{R}^{\pi}}(s_{\max},\pi^*(s_{\max}))\right] + \left[Q^{*,\widehat{R}^{\pi}}(s_{\max},\pi(s_{\max})) - Q^{*,\overline{R}}(s_{\max},\pi(s_{\max}))\right] \\
&\leq 2\cdot \left\| Q^{*,\overline{R}} - Q^{*,\widehat{R}^{\pi_\dagger}} \right\|_{\infty} \\
&\overset{(i)}{\leq} \frac{2}{1-\gamma}\cdot \left\| \overline{R} - \widehat{R}^{\pi_\dagger} \right\|_{\infty} \\
&\leq \frac{2}{1-\gamma}\cdot \left\| \overline{R} - \widehat{R}^{\pi_\dagger} \right\|_{2}.
\end{aligned}$$

where (i) follows from Lemma 10. □

Proof of Theorem 3

To prove Theorem 3 we will utilize the following lemma.

Lemma 11. *Let $\pi \in \Pi_{det}$. be an arbitrary deterministic policy. Define Δ^{π}_{ρ} as $\Delta^{\pi}_{\rho} = \rho^{\pi^*,\overline{R}} - \rho^{\pi,\overline{R}}$ and let $\widehat{R}$ be the solution to the optimization problem* (P3-ATK) *with $\pi_\dagger = \pi$. The following holds:*

$$\frac{1-\gamma}{2}\Delta^{\pi}_{\rho} \leq \left\| \widehat{R} - \overline{R} \right\|_2 \leq \frac{1}{\mu^{\pi}_{\min}}\cdot \Delta^{\pi}_{\rho} + \frac{\epsilon}{\mu_{\min}}\sqrt{|S|\cdot|A|}.$$

Proof.
Upper bound: We prove the upper bound in a constructive manner, using the reward vector R' as defined in Lemma 9. Setting δ^{π} as in Lemma 9, the cost of modifying $\overline{R}$ to R' is bounded by:

$$\begin{aligned}
\left\| R' - \overline{R} \right\|_2 &\leq \left\| \delta^{\pi} \right\|_2 + \left\| \epsilon'_{\pi} \right\|_2 \\
&\leq \left\| \delta^{\pi} \right\|_2 + \frac{\epsilon}{\mu_{min}}\cdot\sqrt{|S|\cdot|A|}
\end{aligned}$$

It remains to bound the term $\left\| \delta^{\pi} \right\|_2$. From Lemma 2 and the definition of π, we have:

$$\begin{aligned}
\rho^{\pi,\overline{R}} - \rho^{\pi^*,\overline{R}} &= \sum_{s} \mu^{\pi}(s)\cdot\left(Q^{\pi^*,\overline{R}}(s,\pi'(s)) - Q^{\pi^*,\overline{R}}(s,\pi^*(s))\right) \\
&= \sum_{s\in S^{\pi}_{\text{pos}}} \mu^{\pi}(s)\left(Q^{\pi^*,\overline{R}}(s,\pi'(s)) - Q^{\pi^*,\overline{R}}(s,\pi^*(s))\right) \\
&= \sum_{s\in S^{\pi}_{\text{pos}}} \mu^{\pi}(s)\cdot\delta^{\pi}(s) \\
&\geq \sum_{s\in S^{\pi}_{\text{pos}}} \mu^{\pi}_{\min}\cdot\delta^{\pi}(s) \\
&\overset{(i)}{=} \sum_{s} \mu^{\pi}_{\min}\cdot\delta^{\pi}(s) \\
&= \mu^{\pi}_{\min}\cdot\left\|\delta^{\pi}\right\|_1 \\
&\geq \mu^{\pi}_{min}\cdot\left\|\delta^{\pi}\right\|_2,
\end{aligned}$$

where (i) follows from the fact that $\delta^{\pi}(s) = 0$ for $s \notin S^{\pi}_{\text{pos}}$. We can therefore conclude that

$$\left\| \widehat{R} - \overline{R} \right\|_2 \overset{(i)}{\leq} \left\| R' - \overline{R} \right\|_2 \leq \frac{1}{\mu^{\pi}_{min}}\cdot\left[\rho^{\pi^*,\overline{R}} - \rho^{\pi,\overline{R}}\right] + \frac{\epsilon}{\mu_{min}}\cdot\sqrt{|S|\cdot|A|}$$

$$= \frac{1}{\mu^{\pi}_{\min}} \cdot \Delta^{\pi}_{\rho} + \frac{\epsilon}{\mu_{\min}} \sqrt{|S| \cdot |A|}.$$

where (i) follows from the fact that R' is feasible in the optimization problem (P3-ATK) with $\pi_\dagger = \pi$ by Lemma 2 while $\widehat{R}$ is optimal for the problem.

Lower Bound: Given Corollary 1,

$$\left\| \overline{R} - \widehat{R} \right\|_2 \geq \frac{1-\gamma}{2} \cdot \Delta^{\pi}_{Q}.$$

Note however that Lemma 2 implies

$$\begin{aligned}
\rho^{\pi^*,\overline{R}} - \rho^{\pi,\overline{R}} &= \sum_s \mu^{\pi}(s) \big(Q^{\pi^*,\overline{R}}(s, \pi^*(s)) - Q^{\pi^*,\overline{R}}(s, \pi(s)) \big) \\
&= \sum_{s \in S^{\pi}_{\text{pos}}} \mu^{\pi}(s) \big(Q^{*,\overline{R}}(s, \pi^*(s)) - Q^{*,\overline{R}}(s, \pi(s)) \big) \\
&\leq \sum_{s \in S^{\pi}_{\text{pos}}} \mu^{\pi}(s) \Delta^{\pi}_{Q} = \Delta^{\pi}_{Q},
\end{aligned}$$

which proves the claim. □

We can now prove Theorem 3.

Statement: *The relative value* Φ *is bounded by*

$$\alpha_\rho \cdot \Delta_\rho \leq \Phi \leq \beta_\rho \cdot \Delta_\rho + \frac{\epsilon}{\mu_{\min}} \cdot \sqrt{|S| \cdot |A|},$$

where $\alpha_\rho = \left(\lambda + \frac{1-\gamma}{2}\right)$ *and* $\beta_\rho = \left(\lambda + \frac{1}{\mu_{\min}}\right)$.

Proof.
Given the Lemma 11, it is clear that since $\min_\pi \Delta^{\pi}_{\rho} = \Delta_\rho$, setting $(\widehat{R}_2, \pi_2)$ as the solution to (P4-APT),

$$\frac{1-\gamma}{2} \cdot \Delta_\rho \leq \left\| \overline{R} - \widehat{R}_2 \right\|_2 \leq \frac{1}{\mu_{\min}} \cdot \Delta_\rho + \frac{\epsilon}{\mu_{\min}} \cdot \sqrt{|S| \cdot |A|}$$

Now note that for any $\pi \in \Pi^{\text{adm}}_{\text{det}}$, including π_2, $\rho^{\pi^*,\overline{R}} - \rho^{\pi,\overline{R}} \geq \Delta_\rho$, which proves the lower bound. As for the upper bound, setting $\widehat{R}^{\pi^*_{\text{adm}}}$ as the solution to (P3-ATK) with $\pi_\dagger = \pi^*_{\text{adm}}$,

$$\begin{aligned}
\Phi &= \left\| \overline{R} - \widehat{R}_2 \right\|_2 + \lambda \cdot [\rho^{\pi^*,\overline{R}} - \rho^{\pi_2,\overline{R}}] \\
&\overset{(i)}{\leq} \left\| \overline{R} - \widehat{R}^{\pi^*_{\text{adm}}} \right\|_2 + \lambda \cdot [\rho^{\pi^*,\overline{R}} - \rho^{\pi^*_{\text{adm}},\overline{R}}] \\
&\overset{(ii)}{\leq} \beta_\rho \cdot \Delta_\rho + \frac{\epsilon}{\mu_{\min}} \cdot \sqrt{|S| \cdot |A|},
\end{aligned}$$

where (i) follows from the optimality of $(\widehat{R}_2, \pi_2)$ and (ii) follows from Lemma 11 and the fact that $\rho^{\pi^*,\overline{R}} - \rho^{\pi^*_{\text{adm}},\overline{R}} = \Delta_\rho$. □

Proof of Theorem 4

To prove Theorem 4, we utilize the following result.

Lemma 12. *Let* $\widehat{R}$ *be the solution to* (P3-ATK) *with* $\pi_\dagger = \pi$. *The following holds:*

$$\frac{1-\gamma}{2} \cdot \Delta^{\pi}_{Q} \leq \left\| \overline{R} - \widehat{R} \right\| \leq \sqrt{|S|} \cdot \Delta^{\pi}_{Q} + \frac{\epsilon}{\mu_{min}} \cdot \sqrt{|S| \cdot |A|}.$$

Proof. The lower bound follows from Corollary 1. As for the upper bound, define R' as in Lemma 9 and note that

$$\left\| \widehat{R} - \overline{R} \right\|_2 \overset{(i)}{\leq} \left\| R' - \overline{R} \right\|_2 \leq \|\delta^{\pi}\|_2 + \|\epsilon'_{\pi}\|_2 \leq \sqrt{|S|} \cdot \Delta^{\pi}_{Q} + \frac{\epsilon}{\mu_{min}} \cdot \sqrt{|S| \cdot |A|}.$$

where (i) follows from the fact that R' is feasible in the optimization problem (P3-ATK) with $\pi_\dagger = \pi$ by Lemma 2 while $\widehat{R}$ is optimal for the problem. □

We can now prove Theorem 4.

Statement: *The relative value* Φ *is bounded by*

$$\alpha_Q \cdot \Delta_Q \leq \Phi \leq \beta_Q \cdot \Delta_Q + \frac{\epsilon}{\mu_{\min}} \sqrt{|S| \cdot |A|},$$

where $\alpha_Q = \left(\lambda \cdot \mu_{\min} + \frac{1-\gamma}{2}\right)$ *and* $\beta_Q = \left(\lambda + \sqrt{|S|}\right)$.

Proof.
Lower bound: Note that for any admissible π, by Lemma 2,

$$\begin{aligned}\rho^{\pi^*,\overline{R}} - \rho^{\pi,\overline{R}} &= \sum_s \mu^{\pi}(s)\big(Q^{*,\overline{R}}(s,\pi^*(s)) - Q^{*,\overline{R}}(s,\pi(s))\big) \\ &\geq \mu_{\min} \cdot \Delta_Q^{\pi} \geq \mu_{\min} \cdot \Delta_Q.\end{aligned}$$

the claim now follows from Lemma 12 and the definition of Φ.
Upper bound: Recall that $\pi_{\text{qg}} = \arg\min_{\pi \in \Pi_{\text{det}}^{\text{adm}}} \Delta_Q^{\pi}$. Using Lemma 2,

$$\begin{aligned}\rho^{\pi^*,\overline{R}} - \rho^{\pi_{\text{qg}},\overline{R}} &= \sum_s \mu^{\pi_{\text{qg}}}(s)\big(Q^{*,\overline{R}}(s,\pi^*(s)) - Q^{*,\overline{R}}(s,\pi_{\text{qg}}(s))\big) \\ &= \sum_{s \in S_{\text{pos}}^{\pi_{\text{qg}}}} \mu^{\pi_{\text{qg}}}(s)\big(Q^{*,\overline{R}}(s,\pi^*(s)) - Q^{*,\overline{R}}(s,\pi_{\text{qg}}(s))\big) \\ &\leq \sum_{s \in S_{\text{pos}}^{\pi_{\text{qg}}}} \mu^{\pi_{\text{qg}}}(s) \cdot \Delta_Q \\ &= \Delta_Q.\end{aligned}$$

Similiar to the proof of Theorem 3, the claim now follows form Lemma 12, the definition of Φ and optimality of π_2. Concretely, setting $\widehat{R}^{\pi_{\text{qg}}}$ as the solution to (P3-ATK) with $\pi_\dagger = \pi_{\text{qg}}$,

$$\begin{aligned}\Phi &= \left\|\overline{R} - \widehat{R}_2\right\|_2 + \lambda \cdot [\rho^{\pi^*,\overline{R}} - \rho^{\pi_2,\overline{R}}] \\ &\leq \left\|\overline{R} - \widehat{R}^{\pi_{\text{qg}}}\right\|_2 + \lambda \cdot [\rho^{\pi^*,\overline{R}} - \rho^{\pi_{\text{qg}},\overline{R}}] \\ &\leq \beta_Q \cdot \Delta_Q + \frac{\epsilon}{\mu_{\min}} \cdot \sqrt{|S| \cdot |A|}.\end{aligned}$$

□