Appendix A  Proof of Theorem 3

In this part, we first begin with several lemmas and then provide a proof of Thm. 3. With the notations in Sec. 4, the following lemma connects the difference in discounted total reward between two arbitrary policies to an expected divergence between them.

**Lemma 1** (Upper bound for the performance gap between the attacked policy and the deceptive policy). Let $\beta = \mathbb{E}_{s \sim d^\pi} - [D_{TV}(\pi_h(|s|)\|\pi^-(|s|))]$, $C = \max_s \mathbb{E}_{s \sim d^\pi} \left[ A^\pi - (s, a) \right]$ and $\beta_1 = \max_s \mathbb{E}_{a \sim \pi} \left[ (\pi_h(a|s) - 1) \right]$. We have an upper bound on the performance gap between $\pi_h(s)$ and $\pi^-$ ($s$):

$$R(\pi_h) - R(\pi^-) \leq \frac{C\beta_1}{1 - \gamma} + \frac{2\gamma C\beta}{(1 - \gamma)^2}.$$  

**Proof.** Based on theorem 1 in [1], the performance of the attacked policy holds by the following bound:

$$R(\pi_h) - R(\pi^-) \leq \frac{1}{1 - \gamma} \mathbb{E}_{s \sim d^\pi^-} \left[ A^\pi^-(s, a) \right] + \frac{2\gamma C}{(1 - \gamma)^2} \mathbb{E}_{s \sim d^\pi^-} \left[ D_{TV}(\pi^- |s| \| \pi_h(s)) \right].$$  

(A1)

By the definition of $\beta_1$ in Lemma 1:

$$\mathbb{E}_{s \sim d^\pi^-} \left[ A^\pi^-(s, a) \right] = \mathbb{E}_{s \sim d^\pi^-} \left[ \pi_h(a|s) - 1 \right] A^\pi^-(s, a) \leq \beta_1 C.$$  

Combining this and the definition of $C$ and $\beta$ with inequality (A1), we get the bound in Lemma 1.

In [1], the authors prove the relation between the expected KL-divergence and the expected TV-divergence of the distribution $p$ and $q$ on state $s$ satisfies:

$$\mathbb{E}_{s \sim f(|s|)} D_{TV}(p(|s|)\|q(|s|)) \leq \mathbb{E}_{s \sim f(|s|)} \sqrt{D_{KL}(p(|s|)\|q(|s|))}/2,$$

where $f(s)$ is the distribution on state $s$. Therefore the expected TV-divergence can be bounded by KL-divergence.

**Lemma 2 (The adversary is stronger with a stronger adversarial optimizer).** We can bound the objective of the original problem (8):

$$\mathbb{E}_{s \sim d^{-}} \left[ D_{TV}(\pi_h(|s|)\|\pi^- (|s|)) \right] \leq \sqrt{\beta_0}/2,$$

here $\beta_0 = \max_{a \in \mathcal{A}} \left| D_{KL}(\pi_h(|s|)\|\pi^- (|s|)) \right|$.  

Lemma 2 shows that the bound of the objective in problem (9) is closely related to the optimization method solving problem (10). With Lemma 1 and Lemma 2, we further provide an upper bound of the performance after attack by $\hat{a}$-adversary.

**Lemma 3 (Upper bound of the $\hat{a}$-adversary’s performance).** Let the adversary be an $\hat{a}$-adversary. The performance of the perturbed policy $\pi_h$ satisfies:

$$R(\pi_h) \leq \hat{a} + \frac{C\beta_1}{1 - \gamma} + \frac{2\gamma C\sqrt{\beta_0}/2}{(1 - \gamma)^2} + R(\pi^-),$$

where $C$, $\beta_0$ and $\beta_1$ are defined in Lemma 1 and Lemma 2.

* Corresponding author (email: dcsj@mail.tsinghua.edu.cn)
Lemma 3 implies that the performance of the adversarial attack is bounded by the ability $\alpha$ of $\alpha$-adversary and the distance from policy $\pi_h$ and $\pi^-$. 

**Theorem 1 (\(\hat{\alpha}\)-adversary is stronger than other adversary under some conditions)**. Let $e$ be an arbitrary adversarial attack algorithm, set $\alpha_e = R(\pi_e) - R(\pi^-)$ and $\beta_1 = \max_s, a \| \frac{\pi_h(a|s)}{\pi^-(a|s)} - 1 \|$. If $\beta_1$ satisfies: 

$$
\beta_1 < -\sqrt{2 \gamma} C + \frac{\sqrt{2 \gamma}^2 C^2 + 4(\alpha_e - \hat{\alpha})(1 - \gamma)^2}{2(1 - \gamma) C},
$$

then the performance of the victim policy after our algorithm attack satisfies: $R(\pi_h) < R(\pi_e)$. In other words, our attack is stronger than adversarial attack $e$.

**Proof**. Let $p(a) = \pi_h(a|s)$, $q(a) = \pi^-(a|s)$. then:

$$
\sum_a p(a) \ln \frac{p(a)}{q(a)} \leq \sum_a p(a) \ln(1 + \beta_1) \leq \beta_1,
$$

with the inequality $\ln(1 + x) \leq x$ when $x \geq 0$. Therefore, $\beta_0 \leq \beta_1$, which bounds the performance of policy $\pi_h$:

$$
R(\pi_h) \leq \hat{\alpha} + \frac{C \beta_1}{1 - \gamma} + \frac{2\gamma C \sqrt{\beta_1 / 2}}{(1 - \gamma)^2}.
$$

(A2)

**References**

1. Achiam J, Held D, Tamar A, et al. Constrained Policy Optimization. In: International Conference on Machine Learning (ICML), 2017. 22–31.