SCIENCE CHINA
Information Sciences

• **Supplementary File** •

Understanding Adversarial Attacks on Observations in Deep Reinforcement Learning

You Qiaoben[1], Chengyang Ying[1], Xinning Zhou[1], Hang Su[1,2], Jun Zhu[1,2*] & Bo Zhang[1]

[1]*Department of Computer Science and Technology,*
Beijing National Research Center for Information Science and Technology,
Tsinghua-Bosch Joint Center for Machine Learning, Institute for Artificial Intelligence,
Tsinghua University, Beijing 100084, China;
[2] *Peng Cheng Laboratory, Shenzhen, Guangdong, 518055, China*

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

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

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

$$\begin{aligned} R(\pi_h) - R(\pi^-) \leqslant & \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi^-},\ a\sim\pi_h}\left[A^{\pi^-}(s,a)\right] \\ & + \frac{2\gamma C}{(1-\gamma)^2}\mathbb{E}_{s\sim d^{\pi^-}}\left[D_{TV}(\pi^-(s)\|\pi_h(s))\right]. \end{aligned} \tag{A1}$$

By the definition of β_1 in Lemma 1:

$$\begin{aligned} & \mathbb{E}_{s\sim d^{\pi^-},\ a\sim\pi_h}\left[A^{\pi^-}(s,a)\right] \\ = & \mathbb{E}_{s\sim d^{\pi^-},\ a\sim\pi^-}\left[(\frac{\pi_h(a|s)}{\pi^-(a|s)} - 1)A^{\pi^-}(s,a)\right] \\ \leqslant & \beta_1\mathbb{E}_{s\sim d^{\pi^-},\ a\sim\pi^-}\left[A^{\pi^-}(s,a)\right] \leqslant \beta_1 C \end{aligned}$$

Combining this and the definition of C and β 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:

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

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):

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

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

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 $\hat{\alpha}$-adversary.

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

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

where C, β_0 and β_1 are defined in Lemma 1 and Lemma 2.

* Corresponding author (email: dcszj@mail.tsinghua.edu.cn)

Lemma 3 implies that the performance of the adversarial attack is bounded by the ability α of α-adversary and the distance from policy π_h and π^-.

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 β_1 satisfies:

$$\beta_1 < \frac{-\sqrt{2}\gamma C + \sqrt{2\gamma^2 C^2 + 4(\alpha_e - \hat{\alpha})(1-\gamma)^3}}{2(1-\gamma)C},$$

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:

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

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

$$\begin{aligned} R(\pi_h) &\leqslant \hat{\alpha} + \frac{C\beta_1}{1-\gamma} + \frac{2\gamma C\sqrt{\beta_0/2}}{(1-\gamma)^2} \\ &\leqslant \hat{\alpha} + \frac{C\beta_1}{1-\gamma} + \frac{2C\gamma\sqrt{\beta_1/2}}{(1-\gamma)^2}. \end{aligned} \tag{A2}$$

References

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