TRAINING WITH WORST-CASE DISTRIBUTIONAL SHIFT CAUSES OVERESTIMATION AND INACCURACIES IN STATE-ACTION VALUE FUNCTIONS

Anonymous authors
Paper under double-blind review

1 SUPPLEMENTARY RESULTS ON INCONSISTENCIES IN ACTION RANKING IN ADVERSARIALLY TRAINED DEEP NEURAL POLICIES

As we mentioned in Section 6.2 of the main body of the paper the inaccuracies of the state-action value function reach a high enough level for the state-of-the-art adversarially trained deep neural policies such that the ranking of the sub-optimal actions is not correct anymore. This can be seen in Figure 1 in the $\mathcal{P}_2$ and $\mathcal{P}_w$ results. Note that $\mathcal{P}_2$ represents the performance drop (Definition 4.2) with action modification a_2, and $\mathcal{P}_w$ (Definition 4.2) represents the action modification with a_w. Thus, it can be observed from Figure 1 that the performance drop $\mathcal{P}_2$ with action modification a_2 is higher than the performance drop $\mathcal{P}_w$ with action modification a_w. In more detail $\mathcal{P}_2$ 0.18257-dominates $\mathcal{P}_w$ (Definition 4.3). This demonstrates that the state-of-the-art adversarially trained deep neural policies are not ranking the sub-optimal actions correctly. Note that as we discussed in the main body of the paper in Section 6.2 this poses a problem for learning optimal state-action value functions Lin & Zhou (2020); Alshiekh et al. (2018).

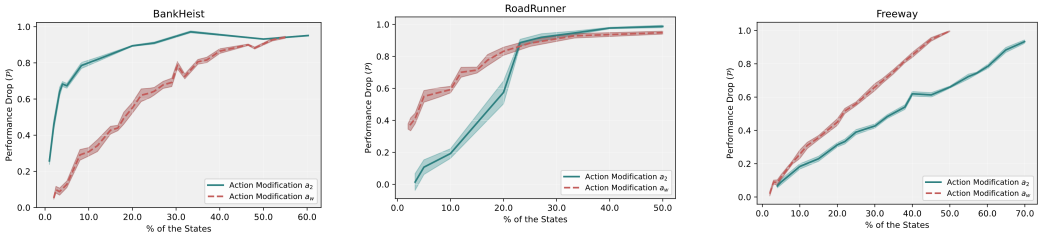

Figure 1: Consistency results for ranked actions via performance drop $\mathcal{P}_2$ and $\mathcal{P}_w$ for the state-of-the-art adversarially trained deep neural policies.

2 OVERESTIMATION OF STATE-ACTION VALUES

In this section we provide supplementary results for the overestimation bias caused by state-of-the-art adversarially trained deep neural policies. In particular, in Section 6.3 of the main body of the paper we explained the problem of overestimation of state-action values. Furthermore, in Section 6.2 we empirically demonstrate that state-of-the-art adversarially trained deep neural policies overestimate the state-action values. In this section we further provide results on state-action values of the optimal action for vanilla and adversarially trained deep neural policies when p_{a_2} is equal to 0.1, 0.2 and 0.3 respectively. Note that in the main body of the paper we claim that the reason for this overestimation lies in the fact that the state-of-the-art deep neural policy adversarial training is solely an extension of adversarial training in image classification tasks, which is based on penalizing the wrong "label". However, this approach does not directly correspond to deep neural policies. The correct label in image classification can be connected to the optimal action in deep neural policies in this analogy. However, the wrong label does not correspond to sub-optimal actions. An optimal Q-function represents the discounted expected cumulative rewards received when taking an action a in state s. Hence, the sub-optimal actions have much more meaning in collecting rewards than solely misclassifying an image.

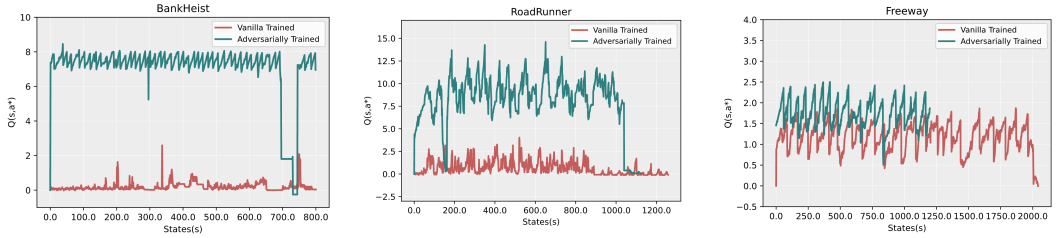

Figure 2: State-action values of the best action $Q(s, a^*)$ for vanilla trained deep neural policies and adversarially trained deep neural policies when p_{a_2} is 0.1.

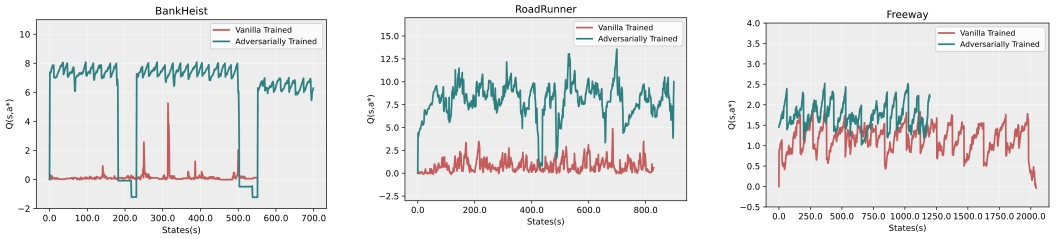

Figure 3: State-action values of the best action $Q(s, a^*)$ for vanilla trained deep neural policies and adversarially trained deep neural policies when p_{a_2} is 0.2.

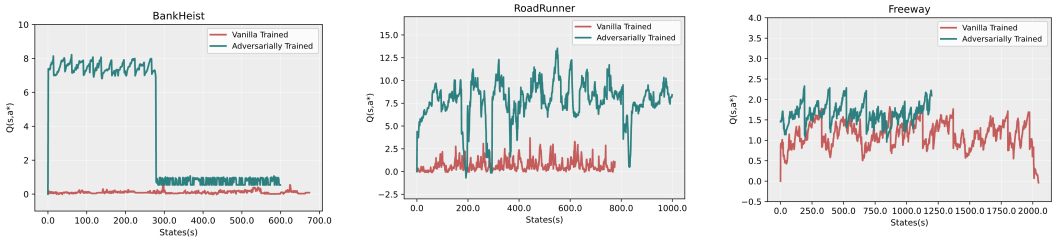

Figure 4: State-action values of the best action $Q(s, a^*)$ for vanilla trained deep neural policies and adversarially trained deep neural policies when p_{a_2} is 0.3.

3 Implementation Details

Note that to be able to provide a fair comparison State-Adversarial Double Deep Q-Network and Double Deep Q-Network are the exact same implementations described in Huan et al. (2020) and Wang et al. (2016) respectively. In more detail for Double Deep Q-Network the batch size is 32, discount factor γ is 1, buffer size 50000, learning rate is 5×10^{-5} for the Adam optimizer, and random action probability is 0.02. Note that experience replay Schaul et al. (2016) is utilized. More details can be found in Dhariwal et al. (2017) and Wang et al. (2016) on Double Deep Q-Networks. The state-of-the-art adversarial deep neural policy is the exact same implementation as Huan et al. (2020). Adversarial deep neural policies are trained via experience replay as well Schaul et al. (2016). Note that State-Adversarial Double Deep Q-Network is trained via the regularizer $\mathcal{R}(\theta) = \sum_s \left(\max_{\bar{s} \in D_\epsilon(s)} \max_{a \neq a^*(s)} Q_\theta(\bar{s}, a) - Q_\theta(\bar{s}, a^*(s))\right)$ where $a^*(s) = \arg\max_a Q(s, a)$ inside ϵ-ball $D_\epsilon(s) = \{\bar{s} : \|s - \bar{s}\|_\infty \leq \epsilon\}$. Hence, this ϵ is set to $1/255$. Note that the regularization is added to the temporal difference loss in the Q-update. The regularization parameter of state-adversarial is $\kappa \in \{0.005, 0.01, 0.02\}$. The initial 1.5×10^6 frames are trained without regularization. See more detail in Huan et al. (2020).

4 SUPPLEMENTARY RESULTS ON ACTION GAP

In Section 6.4 of the main body of our paper we discuss the action gap phenomenon introduced by Farahmand (2011). Note that the action gap is defined as $\kappa(Q, s) = \max_{a' \in A} Q(s, a') - \max_{a \notin \arg\max_{a' \in A} Q(s,a')} Q(s, a)$. Further, we argue that both the existence of overestimation of state action values and the higher action gap in state-of-the-art adversarially trained deep neural policies demonstrates that the hypothesis of Bellemare et al. (2016) cannot be true. In this section we provide supplementary results on the action gap without the normalization $Q(s, a)/\sum_a |Q(s, a)|$. In particular, Figure 5, Figure 6 and Figure 7 show the action gap for the vanilla trained deep neural policies and state-of-the-art adversarial deep neural policies when p_{a_2} is 0, 0.1 and 0.2 respectively. Hence, the action gap for adversarially trained deep neural policies is higher than the vanilla trained deep neural policies.

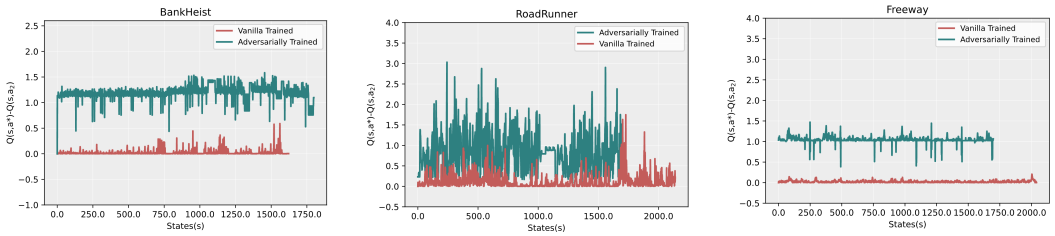

Figure 5: The action gap $Q(s, a^*) - Q(s, a_2)$ for the state-of-the-art adversarially trained deep neural policies and vanilla trained deep neural policies for p_{a_2} is 0.

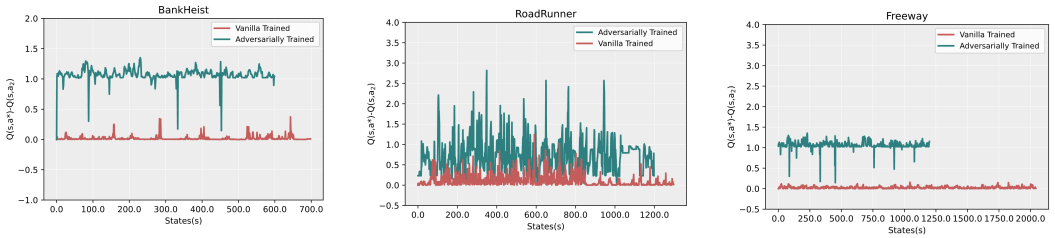

Figure 6: The action gap $Q(s, a^*) - Q(s, a_2)$ for state-of-the-art adversarially trained deep neural policies and vanilla trained deep neural policies for p_{a_2} is 0.1.

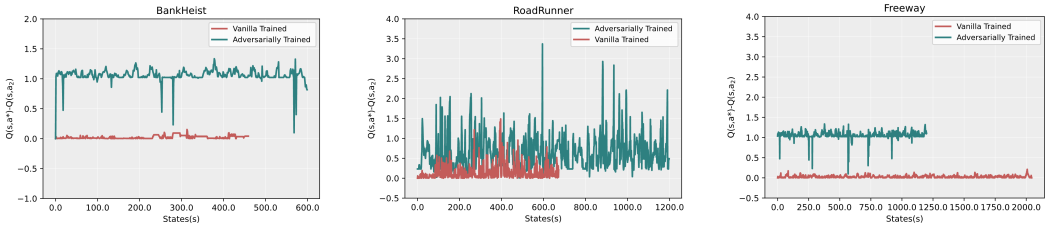

Figure 7: The action gap $Q(s, a^*) - Q(s, a_2)$ for state-of-the-art adversarially trained deep neural policies and vanilla trained deep neural policies for p_{a_2} is 0.2.

5 SUPPLEMENTARY RESULTS ON ACTION GAP WITH NORMALIZED STATE-ACTION VALUES

In the remainder of this section we provide additional results on normalized state-action values for adversarially trained and vanilla trained deep neural policies. In more detail, Figure 8 and Figure 9

show the normalized state-action values of the optimal action, second best action a_2 and worst action a_w for vanilla trained deep neural policies and adversarially trained deep neural policies when p_{a_2} is 0.01 and 0.1 respectively. Thus, Figure 8 and Figure 9 demonstrate that the action gap is higher for the state-of-the-art adversarially trained deep neural policies compared to vanilla trained deep neural policies. Note that the state-action values in Figure 8 and Figure 9 are normalized Q-values (i.e. normalized via $Q(s,a)/\sum_a |Q(s,a)|$).

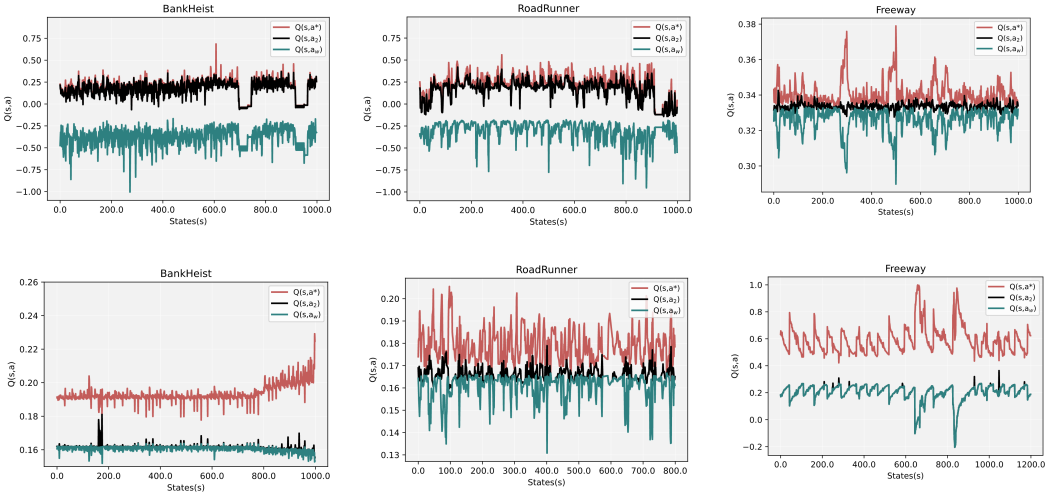

Figure 8: Normalized state-action values for the best action a^*, second best action a_2 and worst action a_w over states when p_{a_2} is 0.01. Row1: Vanilla trained deep neural policies. Row2: State-of-the-art adversarially trained deep neural policies.

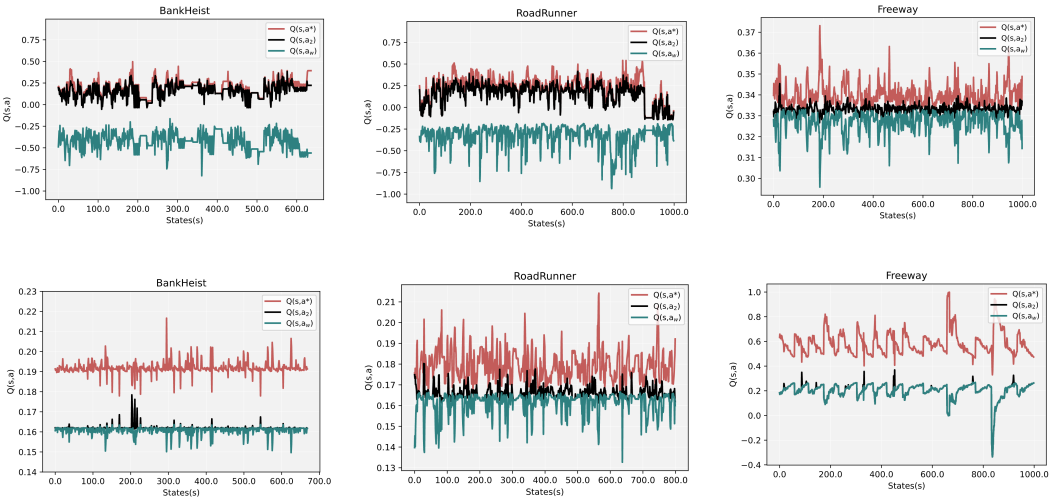

Figure 9: Normalized state-action values for the best action a^*, second best action a_2 and worst action a_w over states when p_{a_2} is 0.1. Row1: Vanilla trained deep neural policies. Row2: State-of-the-art adversarially trained deep neural policies.

REFERENCES

Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI*

Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pp. 2669–2678. AAAI Press, 2018.

Marc G. Bellemare, Georg Ostrovski, Arthur Guez, Philip S. Thomas, and Rémi Munos. Increasing the action gap: New operators for reinforcement learning. In Dale Schuurmans and Michael P. Wellman (eds.), *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 1476–1483. AAAI Press, 2016.

Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. Openai baselines. `https://github.com/openai/baselines`, 2017.

Amir Massoud Farahmand. Action-gap phenomenon in reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Zhang Huan, Chen Hongge, Xiao Chaowei, Bo Li, Mingyan Boning, Duane Liu, and ChoJui Hsiesh. Robust deep reinforcement learning against adversarial perturbations on state observatons. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Kaixiang Lin and Jiayu Zhou. Ranking policy gradient. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

Tom Schaul, John Quan, Ioannis Antonogloua, and David Silver. Prioritized experience replay. *International Conference on Learning Representations (ICLR)*, 2016.

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando. De Freitas. Dueling network architectures for deep reinforcement learning. *Internation Conference on Machine Learning ICML.*, pp. 1995–2003, 2016.