

JUXTALIGN: A FOUNDATIONAL ANALYSIS ON ALIGNMENT OF CERTIFIED REINFORCEMENT LEARNING

Anonymous authors

Paper under double-blind review

1 ANALYSIS ON THE STATE-ACTION VALUE FUNCTION

Below see the complete proof of Proposition 3.4. from the main body of the paper.

Proposition 1.1. *In the MDP \mathcal{M} let $\lambda > 0$ and suppose that $(1 - \lambda)\delta < (1 + \lambda)\eta < \delta$. Let $\theta = (\theta_1, \theta_2, \theta_3)$ be given by $\theta_1 = (1 + \lambda)\theta_1^*$, $\theta_2 = (1 + \lambda)\theta_2^*$ and $\theta_3 = (1 - \lambda)\theta_3^*$. Then $\mathcal{R}(\theta) < \mathcal{R}(\theta^*)$.*

Proof. By an identical argument to that in Proposition 3.4. we have that a_2 is always the action maximizing $\max_{a \neq a^*(s)} Q_\theta(\bar{s}, a) - Q_\theta(\bar{s}, a^*(s))$ whenever $(1 - \lambda)\delta < (1 + \lambda)\eta$. This condition is satisfied by assumption. Therefore, we conclude that for $s = s_1$, the optimal $\bar{s} \in D_\epsilon(s)$ for the scaled parameters θ is given by $\bar{s} = s + \frac{\epsilon}{\sqrt{2}(1+\lambda)}(\theta_2 - \theta_1)$. Therefore, the contribution to the sum defining $\mathcal{R}(\theta)$ from state s_1 is given by

$$\begin{aligned} \langle (\theta_2 - \theta_1), \bar{s} \rangle &= \langle (\theta_2 - \theta_1), s \rangle + \epsilon\sqrt{2}(1 + \lambda) \\ &= -(1 + \lambda) + (1 + \lambda)\eta + \epsilon\sqrt{2}(1 + \lambda) \end{aligned}$$

where the last step uses the fact that $s = \theta_1^* + \delta\theta_3^* + \eta\theta_2^*$ and that the vectors θ_i^* are orthonormal. Next using the fact that $(1 + \lambda)\eta < \delta$ by assumption we conclude

$$\begin{aligned} \langle (\theta_2 - \theta_1), \bar{s} \rangle &< -(1 + \lambda) + (1 + \lambda)\eta + \epsilon\sqrt{2} + \epsilon\lambda\sqrt{2} \\ &< -1 + \delta + \epsilon\sqrt{2}. \end{aligned} \tag{1}$$

The final inequality follows from the fact that $\epsilon < \frac{1}{\sqrt{2}}$ so $\epsilon\lambda\sqrt{2} - \lambda < 0$. Switching from state s_1 to state s_2 , an identical proof (with θ_1 replaced by θ_2) yields the same value for the contribution of state s_2 to the sum. By Proposition 3.4., the contribution of each type of state to the sum defining $\mathcal{R}(\theta^*)$ is

$$\langle (\theta_3^* - \theta_1^*), s + \frac{\epsilon}{\sqrt{2}}(\theta_3^* - \theta_1^*) \rangle = -1 + \delta + \epsilon\sqrt{2}. \tag{2}$$

Clearly the contribution of each state in 1 is strictly less than that in 2. Therefore $\mathcal{R}(\theta) < \mathcal{R}(\theta^*)$. \square

2 WHAT DOES IT ENTAIL TO LEARN INACCURATE, OVERESTIMATED AND INCONSONANT STATE-ACTION VALUES?

The fact that our paper explicitly theoretically and empirically demonstrates that certified adversarially trained policies learn inconsonant and inaccurate state-action values further implies significant concerns on the alignment with human decisions. This has been explained in Section 4. Note that humans conceptualize the values of the set actions they did not take, unlike the adversarially trained deep reinforcement learning policies. See more on the human cognitive decision making process and how humans can conceptualize the set of sub-optimal actions better than random here (Wunderlich et al., 2009; Hoeck et al., 2015; Phillips et al., 2019). Thus the results reported in our paper confirm that vanilla training is more aligned with the human cognitive decision making process.

Also further note that, as also initially described in the main body of our paper in Section 2.3, recent work demonstrated vulnerabilities of certified robust reinforcement learning policies from black-box

adversarial attacks (Korkmaz, 2022) to natural attacks that revealed the generalization problems of adversarially trained deep reinforcement learning policies when compared to straightforward reinforcement learning (Korkmaz, 2023). While these studies highlight the safety and security problems in certified adversarially trained policies, our paper dives into and explains the particular reasons why adversarial training experiences these safety problems. We believe it is crucial to understand the root causes of these problems regarding AI safety, because releasing models with guaranteed safety certifications with undiscovered non-robustness and vulnerabilities will in fact have serious consequences in the real world (The New York Times, 2024; The Washington Post, 2023; Guardian, 2022; The New York Times, 2023). These issues should be openly and transparently analyzed and discussed before these crucial consequences are faced in practice in real life.

3 SOCIETAL IMPACTS

We have described the potential detrimental effects of failing to provide safety while claiming robustness guarantees in the main body of the paper introduction of our paper, and further dedicated a section in the appendix (i.e. Section 2). In particular, these sections highlight the impacts of failing to deliver AI safety (The New York Times, 2024; The Washington Post, 2023; Guardian, 2022; The New York Times, 2023). Our paper discovers that the promises made in *certified-safety* in fact do not hold and furthermore we lay-out the theoretical foundations on why these promises made by *certified-safety* cannot hold. We believe it is crucial to study the exact issues arising and causing failures of machine learning systems both theoretically and empirically. Our paper discovers layers of detrimental issues with certified robust techniques. Our paper not only identifies these issues but further provides theoretical insights in to the fundamental trade-off between robustness and accuracy of the state-action value function.

4 OVERESTIMATION, INACCURACIES AND INCONSISTENCIES IN ADVERSARIAL TRAINING: RADIAL

The left and center column of Figure 1 demonstrate the performance drop $\mathcal{P}_2(p)$ with respect to action modification a_2 for the RADIAL adversarially trained deep reinforcement learning policy proposed by Oikarinen et al. (2021) and the vanilla trained deep reinforcement learning policy in BankHeist and RoadRunner respectively. The right column of the Figure 1 demonstrates the performance drop $\mathcal{P}_w(p)$ with respect to action modification a_w for the RADIAL adversarially trained deep reinforcement learning policy proposed by Oikarinen et al. (2021) and the vanilla trained deep reinforcement learning policy in RoadRunner. Again the results in Figure 1 demonstrate that the vanilla training technique has better estimates for state-action values compared to the adversarial training method RADIAL, quite recently proposed by Oikarinen et al. (2021).

In particular, the curve for $\mathcal{P}_2(p)$ for RADIAL in RoadRunner lies well above the corresponding vanilla training curve. This implies that, while taking the second best action has a relatively mild effect on the vanilla-trained policy, it causes a dramatic loss in performance for RADIAL. Similarly, the $\mathcal{P}_w(p)$ curve for RADIAL in RoadRunner lies above the corresponding curve for the vanilla-trained policy. This again implies that the vanilla-trained policy has a better estimate for which action will lead to lowest rewards than the RADIAL adversarially trained policy. The results reported in Figure 1 again demonstrate the loss of information in the state-action value function due to adversarial regulation of the temporal difference loss.

Figure 2 demonstrates that the overestimation bias discussed in the main body of our paper is again an issue for a newer adversarial training technique quite recently published in NeurIPS 2021. Furthermore, exactly as the previous adversarial training methods, RADIAL also learns inaccurate, inconsistent and overestimated state-action value functions. Hence, these results once more demonstrate the loss of information in the state-action value function as a novel fundamental trade-off intrinsic to adversarial training.

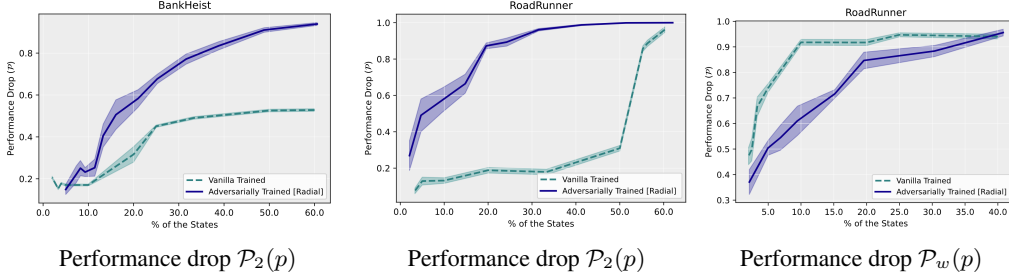


Figure 1: Left: Performance drop $\mathcal{P}_2(p)$ with respect to action modification a_2 for RADIAL adversarially trained deep neural policies Oikarinen et al. (2021) and vanilla trained policies for BankHeist. Center: Performance drop $\mathcal{P}_2(p)$ with respect to action modification a_2 for RADIAL adversarially trained deep neural policies Oikarinen et al. (2021) and vanilla trained policies for RoadRunner. Right: Performance drop $\mathcal{P}_w(p)$ with respect to action modification a_w for the RADIAL adversarially trained deep neural policy and the vanilla trained deep neural policy.

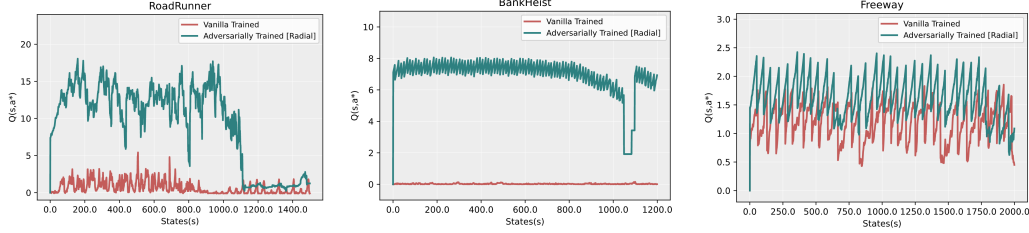


Figure 2: Q -value of the best action a^* over the states for the RADIAL adversarially trained deep neural policy proposed by Oikarinen et al. (2021) and vanilla trained deep neural policy.

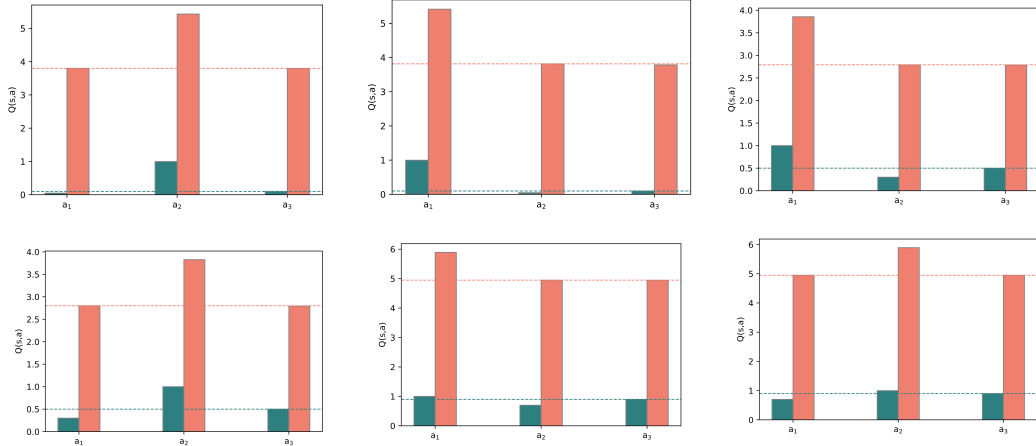


Figure 3: State-action values for the best action a^* , second best action a_2 and worst action a_w for the adversarially trained and vanilla trained deep neural policy loss function for the example MDP with linearly parameterized state-action values constructed in Section 3.

5 FURTHER EXPERIMENTS ON THE LINEARLY PARAMETRIZED MDP

To complement the theoretical results, we numerically optimized both the regularized and un-regularized loss function for the example MDP with linearly parameterized state-action values constructed in Section 3. Figure 3 demonstrates the state-action value function for each of the states the best action a^* , second best action a_2 and worst action a_w for the actions a_1, a_2, a_3 . Note that the numerical optimization of the un-regularized (i.e. vanilla training) loss converges to the true optimal state-action values computed analytically in Section 3. Thus, the results reported in Figure 3

further demonstrate that the addition of the certified training regularizer leads to overestimation of the optimal state-action value function, and re-ordering of the suboptimal actions.

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

As we mentioned in Section 6.1 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 4 in the \mathcal{P}_2 and \mathcal{P}_w results. Note that \mathcal{P}_2 represents the performance drop (Definition 4.1) with action modification a_2 , and \mathcal{P}_w (Definition 4.1) represents the action modification with a_w .

Thus, it can be observed from Figure 4 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 in BankHeist (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 6.1 this poses a problem for learning optimal state-action value functions Lin & Zhou (2020); Alshiekh et al. (2018).

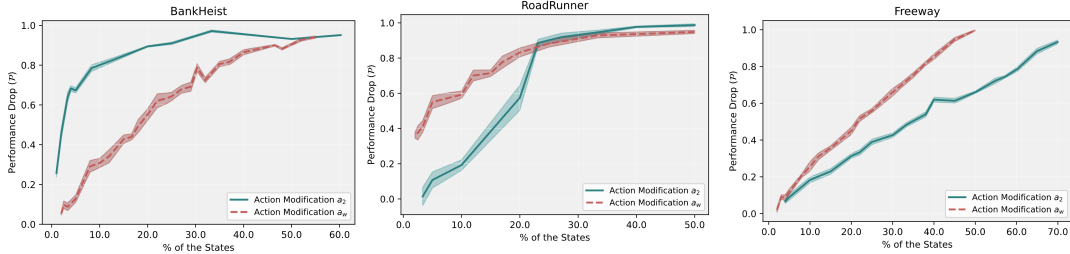


Figure 4: Consistency results for ranked actions via performance drop \mathcal{P}_2 and \mathcal{P}_w for the state-of-the-art adversarially trained deep neural policies.

7 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 5.3 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.

8 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 the SA-DDQN paper described in Section 3 and (Wang et al., 2016) respectively. In more detail for Double Deep Q-Network the batch size is 32, discount factor γ is 0.99, buffer size 50000, learning rate is 5×10^{-5}

Table 1: Average Q -values of the optimal action in state-of-the-art adversarially trained deep neural policies and vanilla trained deep neural policies.

Environments	BankHeist		RoadRunner		Freeway	
	Adversarial	Vanilla	Adversarial	Vanilla	Adversarial	Vanilla
$Q(s, a^*)$	5.903 ± 2.052	0.300 ± 0.434	8.806 ± 3.216	0.602 ± 0.781	1.667 ± 0.406	1.185 ± 0.348

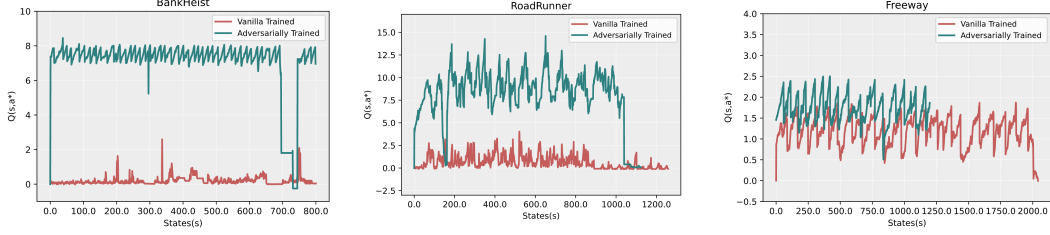


Figure 5: State-action values of the best action $Q(s, a^*)$ for vanilla trained deep neural policies and adversarially trained deep neural policies when p_{a_2} is 0.1.

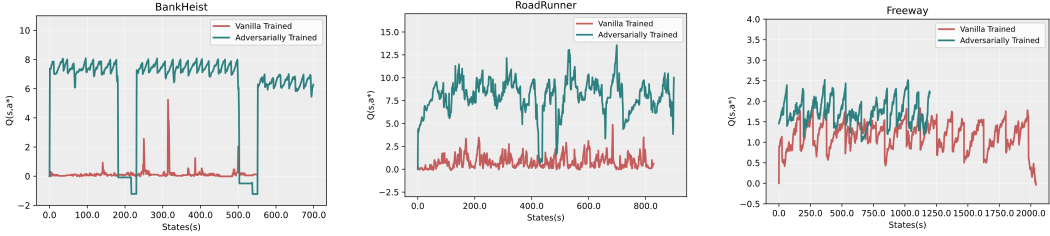


Figure 6: State-action values of the best action $Q(s, a^*)$ for vanilla trained deep neural policies and adversarially trained deep neural policies when p_{a_2} is 0.2.

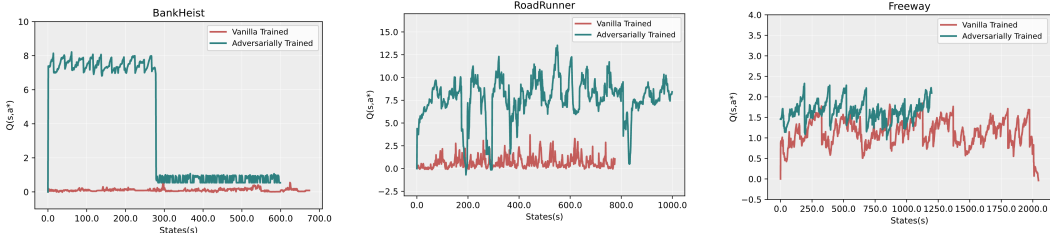


Figure 7: State-action values of the best action $Q(s, a^*)$ for vanilla trained deep neural policies and adversarially trained deep neural policies when p_{a_2} is 0.3.

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 in the SA-DDQN paper. 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 $\mathcal{R}(\theta) = \sum_s (\max_{\bar{s} \in D_\epsilon(s)} \max_{a \neq a^*(s)} Q_\theta(\bar{s}, a) - Q_\theta(\bar{s}, a^*(s)))$ 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.

9 FURTHER EXPERIMENTAL 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 $\kappa(Q, s) = \max_{a' \in A} Q(s, a') - \max_{a \in \arg \max_{a' \in A} Q(s, a')} Q(s, a)$. 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.

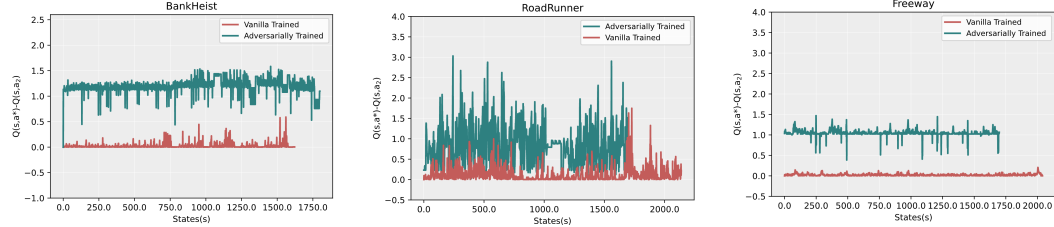


Figure 8: The action gap $Q(s, a^*) - Q(s, a_2)$ for the state-of-the-art adversarially trained deep neural policies and vanilla trained deep neural policies for p_{a_2} is 0.

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 8, Figure 9 and Figure 10 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 for vanilla trained deep neural policies.

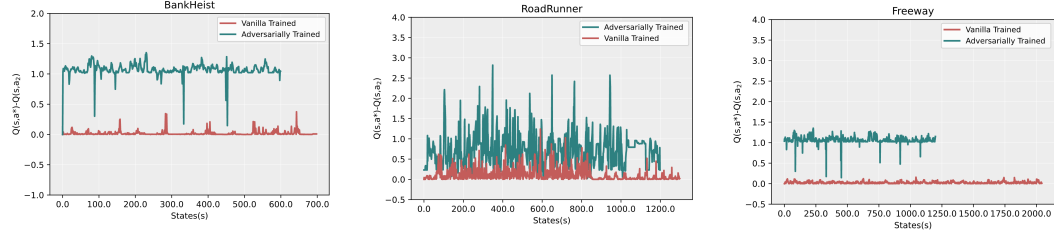


Figure 9: The action gap $Q(s, a^*) - Q(s, a_2)$ for state-of-the-art adversarially trained deep neural policies and vanilla trained deep neural policies for p_{a_2} is 0.1.

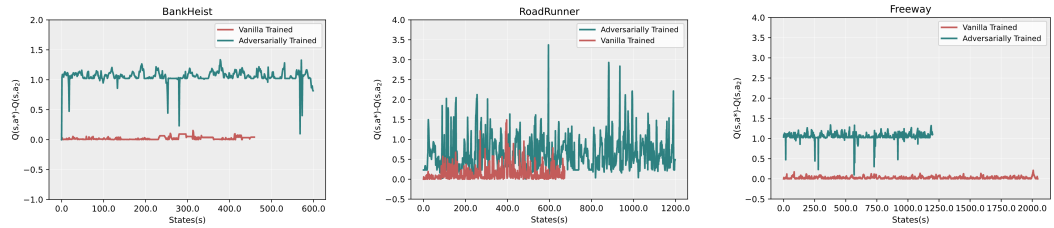


Figure 10: The action gap $Q(s, a^*) - Q(s, a_2)$ for state-of-the-art adversarially trained deep neural policies and vanilla trained deep neural policies for p_{a_2} is 0.2.

9.1 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 11 and Figure 12 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

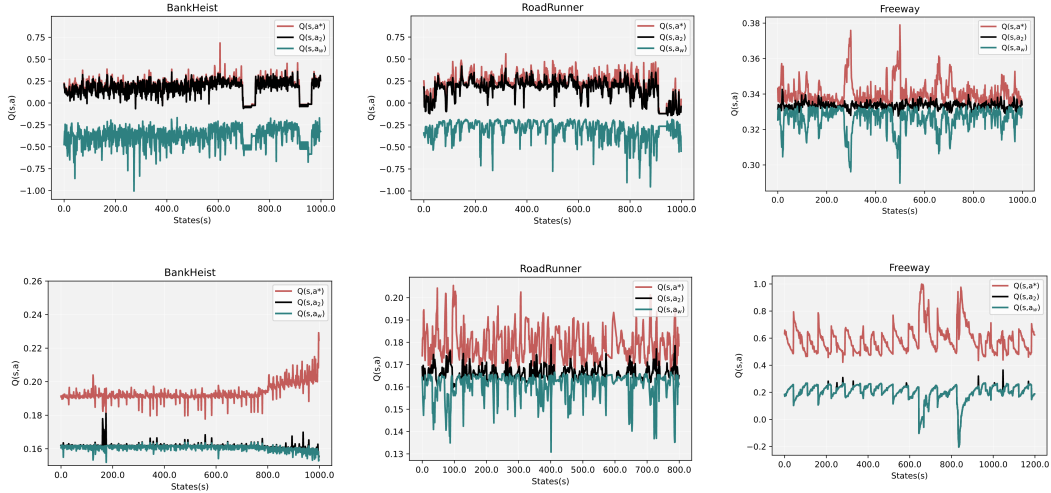


Figure 11: Normalized state-action values for the best action a^* , second best action a_2 and worst action a_w over states when p_{a_2} is 0.01. Row1: Vanilla trained deep neural policies. Row2: State-of-the-art adversarially trained deep neural policies.

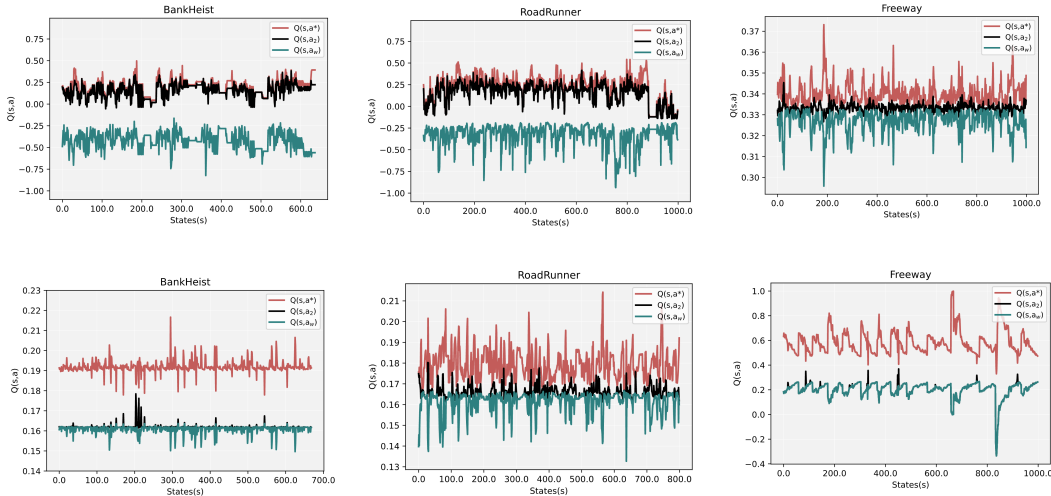


Figure 12: Normalized state-action values for the best action a^* , second best action a_2 and worst action a_w over states when p_{a_2} is 0.1. Row1: Vanilla trained deep neural policies. Row2: State-of-the-art adversarially trained deep neural policies.

trained deep neural policies when p_{a_2} is 0.01 and 0.1 respectively. Thus, Figure 11 and Figure 12 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 11 and Figure 12 are normalized Q -values (i.e. normalized via $Q(s, a) / \sum_a |Q(s, a)|$).

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.
- The Guardian. Tesla behind eight-vehicle crash was in ‘full self-driving’ mode, says driver. December 2022.
- Nicole Van Hoeck, Patrick D. Watson, and Aron K. Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience* 2015.
- Ezgi Korkmaz. Deep reinforcement learning policies learn shared adversarial features across mdps. *AAAI Conference on Artificial Intelligence*, 2022.
- Ezgi Korkmaz. Adversarial robust deep reinforcement learning requires redefining robustness. *AAAI Conference on Artificial Intelligence*, 2023.
- 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.
- Tuomas P. Oikarinen, Wang Zhang, Alexandre Megretski, Luca Daniel, and Tsui-Wei Weng. Robust deep reinforcement learning through adversarial loss. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 26156–26167, 2021.
- Jonathan Phillips, Adam Morris, and Fiery Cushman. How we know what not to think. *Trends in Cognitive Sciences* 2019.
- Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *International Conference on Learning Representations (ICLR)*, 2016.
- The New York Times. Driverless taxis blocked ambulance in fatal accident, san francisco fire department says. September 2023.
- The New York Times. Cruise says hostility to regulators led to grounding of its autonomous cars. 2024.
- The Washington Post. Cruise recalls all its driverless cars after pedestrian hit and dragged. November 2023.
- Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando. De Freitas. Dueling network architectures for deep reinforcement learning. *International Conference on Machine Learning ICML.*, pp. 1995–2003, 2016.
- Klaus Wunderlich, Antonio Rangel, and John P. O’Doherty. Neural computations underlying action-based decision making in the human brain. *Proceedings of the National Academy of Sciences (PNAS)* 2009.