Analysis and Improvement of Adversarial Training in DQN Agents With Adversarially-Guided Exploration (AGE)

Vahid Behzadan[1] and William Hsu[1]
{behzadan, bhsu}@ksu.edu

Kansas State University

Abstract. This paper investigates the effectiveness of adversarial training in enhancing the robustness of Deep Q-Network (DQN) policies to state-space perturbations. We first present a formal analysis of adversarial training in DQN agents and its performance with respect to the proportion of adversarial perturbations to nominal observations used for training. Next, we consider the sample-inefficiency of current adversarial training techniques, and propose a novel Adversarially-Guided Exploration (AGE) mechanism based on a modified hybrid of the ϵ-greedy algorithm and Boltzmann exploration. We verify the feasibility of this exploration mechanism through experimental evaluation of its performance in comparison with the parameter-space noise exploration algorithm.

Keywords: Deep Reinforcement Learning · State Perturbation · Policy Generalization · Resilience · robustness · Adversarial Exploration.

1 Introduction

Recent studies have established the brittleness of Deep Reinforcement Learning (DRL) policies to variations in the state space[10]. This can be attributed to failure in the generalization of the policy with respect to input features[13]. Consequently, many of the proposed techniques for enhancement of such brittleness are based on the idea of regularization. As a recent survey of literature on defensive techniques illustrates [3], a major emphasis in such techniques is on adversarial training [9], which is in effect a regularization technique based on data augmentation. In this paper, we first present an analysis of adversarial training in Deep Q-Network (DQN) agents[7], and its effectiveness with respect to the proportion of adversarial perturbations used for training. Next, we establish the sample-inefficiency of current adversarial training techniques, and develop a novel adversarially-guided exploration mechanism based on a modified hybrid of the ϵ-greedy and Boltzmann exploration techniques [12], and evaluate its performance in comparison with the parameter-space noise exploration[6] algorithm.

2 Limits of Adversarial Training

In this section, we analyze the effectiveness of training a DRL agent with experiences generated through an adversarial interaction. We consider an adversary constrained to a probabilistic budget $P(attack)$, which is the probability of perturbing any state $s'_t \leftarrow s_t + \delta$ such that the approximated policy at the ith iteration of training (π_i) produces an incorrect action, i.e., $\pi_i(s'_t) \neq \pi_i(s_t)$. We also consider two types of adversarial objectives, one is the *state-neutral* adversary, which imposes the perturbation so that the resulting s'_t induces any action other than $\pi_i(s_t)$. The second type type of adversary we consider is the *targeted* adversary, which crafts s'_t such that the induced action is the worst possible choice, i.e., $\pi_i(s'_t) = \arg\min_a Q_i(s, a)$. We assume that the adversary is always successful in crafting the desired perturbations.

We begin the analysis by noting the effect of such perturbations on the composition of the experience replay memory. For any state s_t, two types of experiences may be recorded. One represents the nominal (i.e., unperturbed) experiences, denoted by:

$$\langle s_t, a_t = \pi_i(s_t), s_{t+1}, r(s_t, a_t, s_{t+1}) \rangle \tag{1}$$

The second type are experiences in which s_t is the result of perturbing another state, i.e., $s_t \leftarrow s'_t + \delta$. Such adversarial experiences are denoted by:

$$\langle s_t, a_t = \pi_i(s_t), s'_{t+1}, r(s_t, a_t, s'_{t+1}) \rangle \tag{2}$$

Hence, the expected TD-error of state s_t in each iteration $i+1$ of training is given by:

$$\begin{aligned} \mathbb{E}[Err_{i+1}(s_t)] = {} & p_{i+1}(attack|s_t).[r(s_t, a_t, s'_{t+1}) + \gamma V^{\pi_i}(s'_{t+1}] \\ & + [p_{i+1}(s_t) - p_{i+1}(attack|s_t)].[r(s_t, a_t, s_{t+1)}) + \gamma V^{\pi_i}(s_{t+1})] \\ & - V^{\pi_i}(s_t) \end{aligned} \tag{3}$$

where $p_{i+1}(s_t)$ is the probability of choosing an experience beginning with either nominal or crafted state s_t from the experience memory in the $i+1$th iteration, and $p_{i+1}(attack|s_t) = p_{i+1}(s_t) - p_{i+1}^{nominal}(s_t)$ is the probability of choosing an experience sample beginning with an adversarially-crafted state s_t. It is noteworthy that adversarial perturbations add bias to the expected TD-error. It can be seen that, for the effect of this bias to be decreasing as i increases, the following condition must hold true:

$$p_{i+1}(s_t) - p_{i+1}(attack|s_t) > p_i(s_t) - p_i(attack|s_t) \tag{4}$$

That is, the probability of sampling nominal experiences starting with s_t from the experience memory must be increasing with i. In the case of a state-neutral adversary, and assuming the uniform sampling from experiences, this condition reduces to:

$$\forall s_t \in S : p_{i+1}^{nominal}(s_t) > p(attack) \tag{5}$$

Which can be interpreted as $p(attack) < 0.5$. This is in agreement with the results reported in [2] for non-contiguous, non-targeted adversarial example attacks against DQN agents.

2.1 Experimental Results

To evaluate the practical implications of the theoretical analyses of this section, we study the training performance of a CartPole DQN policy under non-targeted attacks with perturbation probabilities of $0.2, 0.4, 0.8,$ and 1.0. In these experiments, the attacks begin after the convergence of the policy to optimal performance.

The results are presented in figures 1 and 2. It can be seen that for $p(attack) = 0.2$ and $p(attack) = 0.4$, the training process recovers rather quickly. However, for $p(attack) = 0.8$ and $p(attack) = 1.0$, the recovery fails to realize within the observed training horizon. It is noteworthy that the early peaking observed in Figure 2 are due to residual unperturbed experiences still remaining in the replay memory, the impact of which immediately fades at around 50000 steps after the attack begins, which is equivalent to the number of experiences required to completely overwrite the memory.

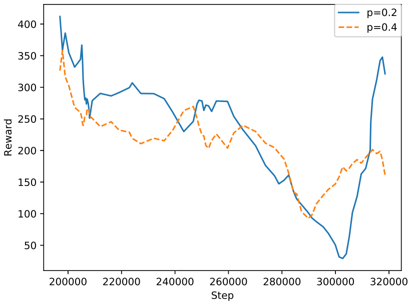

Fig. 1. Training Performance Under Non-Targeted Attack with p(attack)= 0.2 and p(attack) = 0.4

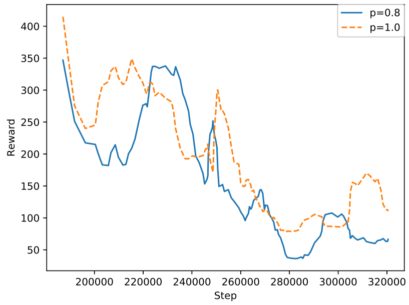

Fig. 2. Training Performance Under Non-Targeted Attack with p(attack)= 0.8 and p(attack) = 1.0

3 Adversarially-Guided Exploration Mechanism for Sample-Efficient Adversarial Training

There exists a noteworthy difference between the theoretical adversaries considered so far and one that crafts perturbations through adversarial examples. As reported in [9] and [2], training on adversarial examples enhances the resilience of the policy to perturbations crafted using the same technique. Similar to the case of adversarial training for deep learning classifiers [11], this phenomenon can be explained from the perspective of regularization: adversarial example perturbations of states provide the means for regularization of the policy (or value function) through data augmentation. Therefore, training the policy over adversarial examples of states generated with a certain attack mechanism results in the enhancement of resilience and robustness of the policy to perturbations crafted via that mechanism.

However, current procedures for training over adversarial examples (e.g., [9][8] are based on "blanket perturbation", in which all states have an equal probability of being perturbed during training, thus leading to the deterioration of sample efficiency in DRL training. To alleviate this adverse effect, we propose the Adversarially-Guided Exploration (AGE) mechanism, which efficiently reduces the number of perturbed observations required to produce similar or better improvements in robustness compared to the results achieved by previous techniques. The proposed mechanism is based on the fact that not all states are equal with respect to the total regret produced by their perturbation. To account for this fact, the proposed AGE mechanism extends the classical ϵ-greedy exploration mechanism by adjusting the probability of sampling actions for each state according to the *adversarial state-action significance*, defined as follows: In the $(i+1)$th training iteration, the adversarial significance of any action a in state s, denoted by $\zeta^{\pi_i}_{adv}(s,a)$, measures the maximum achievable adversarial gain, determined by the difference between maximum Q-value at state s and $Q^{\pi_i}(s,a)$ with respect to actions. We define ζ_{adv} as the ratio of this difference to the sum of this difference for all actions $a \in A$. Furthermore, to retain the GLIE (Greedy in the Limit with Infinite Exploration) criteria of the ϵ-greedy mechanism [12], we formulate ζ_{adv} in the form of the Boltzmann probability[5], with ϵ as the decaying temperature factor. Consequently, the formal definition of ζ_{adv} is as follows:

$$\zeta_{adv}^{\pi_i}(s,a) = \frac{\exp\left(\max_{a'} Q^{\pi_i}(s,a') - Q^{\pi_i}(s,a)/\epsilon\right)}{\sum_{\alpha \in A} \exp\left(\max_{a'} Q^{\pi_i}(s,a') - Q^{\pi_i}(s,\alpha)/\epsilon\right)} \tag{6}$$

Algorithm 1 presents the details of our proposed exploration mechanism:

Algorithm 1 Adversarially-Guided Exploration (AGE) for Adversarial Training

```
Require: $Q^{\pi_i}$, action space $A$
  function Adversarial_Exploration(Current state $s$, exploration probability $\epsilon$)
      for all $a \in A$ do
          $\zeta_{adv}^{\pi_i}(s,a) = \frac{\exp(\max_{a'} Q^{\pi_i}(s,a') - Q^{\pi_i}(s,a)/\epsilon)}{\sum_{\alpha \in A} \exp(\max_{a'} Q^{\pi_i}(s,a') - Q^{\pi_i}(s,\alpha)/\epsilon)}$
      end for
      if $rand() \leq \epsilon$ then
          Sample action according to $\zeta_{adv}^{\pi_i}$ to perform
      else
          Perform $\arg\max_a Q^{\pi_i}(s,a)$
      end if
```

4 Experiment Setup

Environment and Target Policies: To evaluate the performance of AGE in adversarial training, we study the training efficiency and adversarial resilience of a DQN policy in the CartPole environment in OpenAI Gym [4]. Table 1 presents the specifications of the CartPole environment, and Table 2 provides the parameter settings of each target policy.

Table 1. Specifications of the CartPole Environment

Observation Space	Cart Position [-4.8, +4.8] Cart Velocity [-inf, +inf] Pole Angle [-24 deg, +24 deg] Pole Velocity at Tip [-inf, +inf]
Action Space	0 : Push cart to the left 1 : Push cart to the right
Reward	+1 for every step taken
Termination	Pole Angle is more than 12 degrees Cart Position is more than 2.4 Episode length is greater than 500

Table 2. Parameters of DQN Policy

No. Timesteps	10^5
γ	0.99
Learning Rate	10^{-3}
Replay Buffer Size	50000
First Learning Step	1000
Target Network Update Freq.	500
Prioritized Replay	True
Exploration	Parameter-Space Noise
Exploration Fraction	0.1
Final Exploration Prob.	0.02
Max. Total Reward	500

Adversarial Agent: In these experiments, the adversarial agent is a DQN agent with the hyperparameters provided in Table 3. We consider a homogeneous perturbation cost function for all state perturbations, that is $\forall s, a' : c_{adv}(s, a') = c_{adv}$. For both the resilience and robustness measurements, we set $c_{adv} = 1$ (i.e., each perturbation incurs a cost of 1 to the adversary). The training process is terminated when the adversarial regret is maximized and the 100-episode average of the number of adversarial perturbations is quasi-stable for 200 episodes.

Table 3. Parameters of Adversarial DQN Agent

Max. Timesteps	10^5
γ	0.99
Learning Rate	10^{-3}
Replay Buffer Size	50000
First Learning Step	1000
Target Network Update Freq.	500
Experience Selection	Prioritized Replay
Exploration	Parameter-Space Noise
Exploration Fraction	0.1
Final Exploration Prob.	0.02

4.1 Results

Figure 3 illustrates the training performance of the DQN policy utilizing AGE for exploration. It can be seen that the training has successfully converged, and the progress is noticeably more stable than that of a DQN policy with NoisyNet exploration. Furthermore, Figure 5 depicts the training performance of a DQN-based adversarial resilience agent with the same configuration as presented in [1]. In comparison with the performance of the same agent against the same policy trained using NoisyNet exploration (Figure 4), two significant differences are observed: first, the adversarial agent targeting the AGE-trained policy achieves a lower regret and higher perturbation count in the same number of training iterations as its counter-part. Second, the training process targeting the AGE-trained policy fails to converge in 100000 iterations, whereas its counter-part converged at around 90000 iterations. These results indicate the superior resiliency of the AGE-trained policy over the nominal policy, thereby verifying the effectiveness of AGE in improving the adversarial resilience of policies.

Furthermore, in comparison with to the best-case scenario of adversarial training of the nominal DQN policy (as presented in Figure 1), it can be seen that the AGE-based training process requires significantly fewer samples for convergence. This comparison further verifies the efficiency of our proposed scheme with respect to sample complexity.

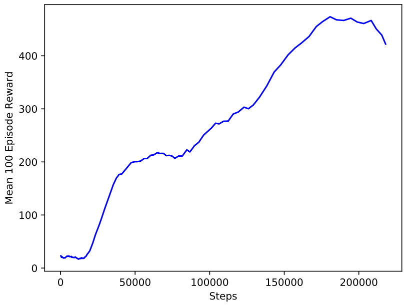

Fig. 3. Training Performance of a CartPole DQN policy with AGE exploration

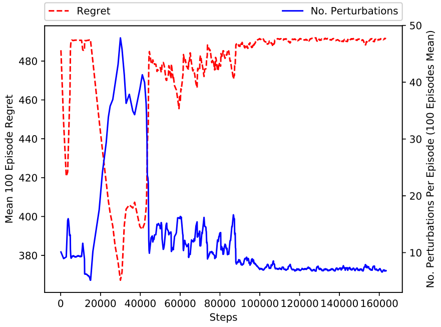

Fig. 4. Adversarial Training Progress for Resilience Benchmarking of the DQN Policy with NoisyNet exploration

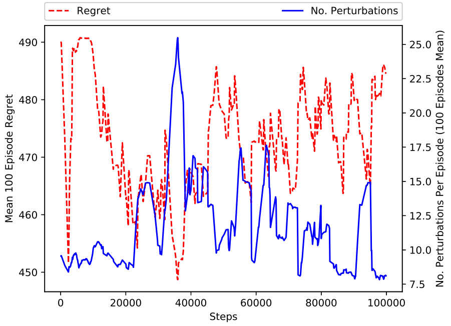

Fig. 5. Training Performance of an Adversarial Agent Targeting the AGE-Trained Policy

5 Conclusion

This paper formally establishes the limits of adversarial training in DQN agents with respect to the ratio of perturbed training experience to the nominal (i.e., unperturbed) experiences. We then address the sample-inefficiency of current adversarial training techniques, and present the Adversarially-Guided Exploration (AGE) mechanism to improve upon this shortcoming. The presented experimental results demonstrate the feasibility of this exploration mechanism in comparison with the parameter-space noise exploration algorithm.

References

1. Behzadan, V., Hsu, W.: Rl-based method for benchmarking the adversarial resilience and robustness of deep reinforcement learning policies. arXiv preprint arXiv:1906.01110 (2019)
2. Behzadan, V., Munir, A.: Whatever does not kill deep reinforcement learning, makes it stronger. arXiv preprint arXiv:1712.09344 (2017)
3. Behzadan, V., Munir, A.: The faults in our pi stars: Security issues and open challenges in deep reinforcement learning. arXiv preprint arXiv:1810.10369 (2018)
4. Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., Zaremba, W.: Openai gym. arXiv preprint arXiv:1606.01540 (2016)
5. Cesa-Bianchi, N., Gentile, C., Lugosi, G., Neu, G.: Boltzmann exploration done right. In: Advances in Neural Information Processing Systems. pp. 6284–6293 (2017)
6. Fortunato, M., Azar, M.G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., et al.: Noisy networks for exploration. arXiv preprint arXiv:1706.10295 (2017)
7. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529 (2015)

8. Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., Chowdhary, G.: Robust deep reinforcement learning with adversarial attacks. In: Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. pp. 2040–2042. International Foundation for Autonomous Agents and Multiagent Systems (2018)
9. Pinto, L., Davidson, J., Sukthankar, R., Gupta, A.: Robust adversarial reinforcement learning. arXiv preprint arXiv:1703.02702 (2017)
10. Rajeswaran, A., Lowrey, K., Todorov, E.V., Kakade, S.M.: Towards generalization and simplicity in continuous control. In: Advances in Neural Information Processing Systems. pp. 6550–6561 (2017)
11. Shaham, U., Yamada, Y., Negahban, S.: Understanding adversarial training: Increasing local stability of supervised models through robust optimization. Neurocomputing **307**, 195–204 (2018)
12. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction. MIT press (2018)
13. Zhang, A., Ballas, N., Pineau, J.: A dissection of overfitting and generalization in continuous reinforcement learning. arXiv preprint arXiv:1806.07937 (2018)