Adaptive Discounting of Training Time Attacks

Extended Abstract

Ridhima Bector
Nanyang Technological University
Singapore
ridhima001@e.ntu.edu.sg

Abhay Aradhya
Nanyang Technological University
Singapore
abhayaradhya@ntu.edu.sg

Chai Quek
Nanyang Technological University
Singapore
ashcquek@ntu.edu.sg

Zinovi Rabinovich
Nanyang Technological University
Singapore
zinovi@ntu.edu.sg

ABSTRACT

Among the most insidious attacks on Reinforcement Learning (RL) solutions are training-time attacks (TTAs) that create loopholes and backdoors in the learned behaviour. Not limited to a simple disruption, *constructive* TTAs (C-TTAs) are now available, where the attacker forces a specific, target behaviour upon a training RL agent (victim). However, even state-of-the-art C-TTAs focus on target behaviours that could be naturally adopted by the victim if not for a particular feature of the environment dynamics, which C-TTAs exploit. In this work, we show that a C-TTA is possible even when the target behaviour is un-adoptable by the victim (in the default/un-attacked environment) due to *both* environment dynamics *as well as* due to the behaviour's non-optimality w.r.t. the victim's objective(s). To find efficient attacks in this context, we develop a specialised flavour of the DDPG algorithm, which we term γDDPG, that learns this stronger version of C-TTA. γDDPG dynamically alters the attack policy planning horizon based on the victim's current behaviour. This improves effort distribution throughout the attack timeline and reduces the effect of uncertainty that the attacker has about the victim. To demonstrate the features of our method and better relate the results to prior research, we borrow a 3D grid domain from a state-of-the-art C-TTA for our experiments. The full paper is available at "bit.ly/AdaptiveDiscountingofTTA".

KEYWORDS

Dynamic Discount; Adaptive Discount; Constructive Training-Time Attacks; Environment Poisoning; Reinforcement Learning

ACM Reference Format:
Ridhima Bector, Abhay Aradhya, Chai Quek, and Zinovi Rabinovich. 2024. Adaptive Discounting of Training Time Attacks: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

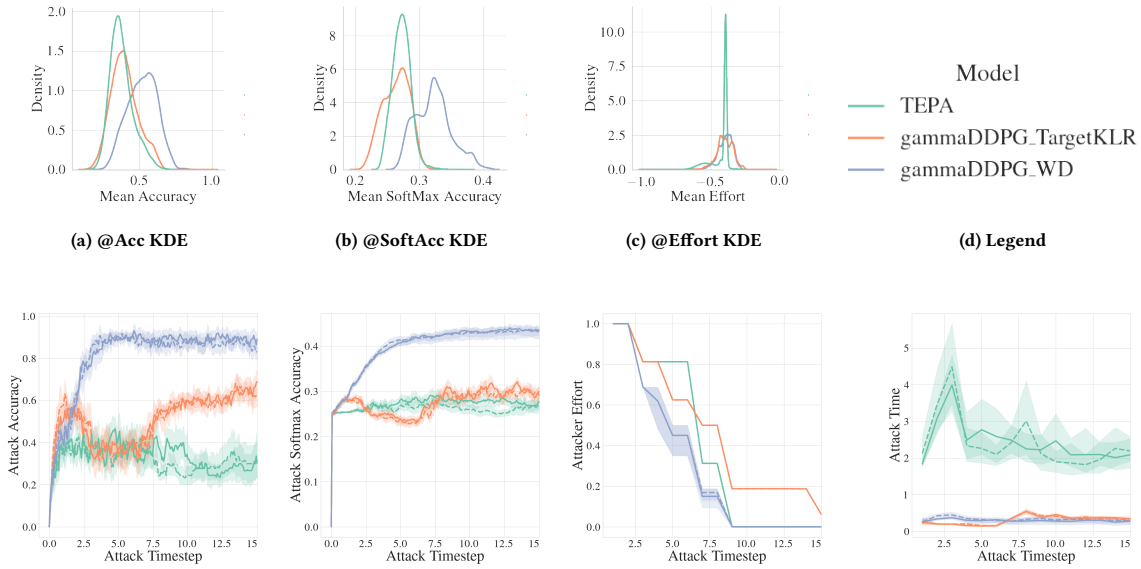
Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Success of RL stands threatened by a wide variety of attacks [2–4], most insidious of which are training-time attacks (TTAs) that "pre-program" back-doors and behavioural triggers into an RL strategy [1, 7, 9–12]. In TTAs, the attacker learns to optimally modify/poison a victim RL agent's internal aspects (i.e., sensor(s), processor(s), memory) and/or external influences (i.e., environment) while the victim agent trains to learn its task. The level of information access assumed by the adversary categorises the TTA as white-box [8, 10, 13] or black-box [7, 11, 12]. This work aims to develop and study an environment-poisoning black-box C-TTA which modifies/poisons the dynamics of the victim agent's environment without accessing any internal mechanism of the victim. Like prior works on environment-poisoning black-box C-TTAs [11, 12], the adversary in this research is an RL agent which learns the optimal C-TTA to be applied on the victim RL agent. However, unlike the prior works that enforce un-adoptable but optimal target behaviour on the victim agent and train the attack by infusing all attack objectives into the reward of the optimisation problem; our attack enforces a non-optimal target behaviour which is learned by distributing the attack objectives into the reward and the reward discounting factor of the attacker's optimisation problem. In addition to pushing the victim agent towards this non-optimal target behaviour, the attack must also preserve the environment as much as possible or, equivalently, reduce the effort expended to modify it. Attack actions are thus constrained by the magnitude of change a single attack action is permitted to make, as well as by treating environment modification effort as a second objective in the attacker's optimisation problem. The attacker, therefore, faces a multi-objective problem of finding an attack strategy that: a) generates the target behaviour in the victim with high accuracy, and b) has low-effort environment modifications.

2 METHODOLOGY : γ-VARIANT DDPG

Commonly, an RL agent's objectives are represented by a reward signal and the agent strives to find a behaviour/policy which, when executed in the given environment, maximises the produced cumulative reward. Likewise, in environment-poisoning C-TTAs, the attacker's reward typically inculcates both attack objectives: the accuracy with which the victim adopts the target behaviour, and the effort applied by the attacker, in terms of environment modifications, to achieve this accuracy. This can be done either by having several



(a) @Acc KDE (b) @SoftAcc KDE (c) @Effort KDE (d) Legend

(e) Test-Time @Acc (f) Test-Time @SoftAcc (g) Test-Time @Effort (h) Test-Time @Time

Figure 1: Training-Time statistics (a-c) and Test-Time performance (e-h) w.r.t. Accuracy (@Acc), Softmax Accuracy (@SoftAcc), Effort (@Effort), and Time (@Time) of baseline TEPA vs γDDPG with dynamic discounts TargetKLR and WD.

reward terms, allowing for prioritisation (through weights) of the attacker's objectives; or, by measuring the discrepancy between combined behaviour-environment pairs, as done in [11, 12]. More specifically, these works utilise the Kullback Leibler Divergence Rate (KLR) to provide a unified estimate of effort and effectiveness of an attack by measuring the discrepancy between the combination of the victim's current behaviour with the poisoned environment *and* the combination of the target behaviour with the default environment. However, both the aforementioned approaches have their shortcomings. Due to high symmetry, the KLR-based approach cannot properly distinguish between a high-accuracy, medium-effort behaviour-environment pair and a medium-accuracy, low-effort pair; while, weighted multiple terms of reward cannot address the fact that some behaviour-environment discrepancies cancel each other and, are thus, irrelevant.

In this work, we propose an alternative route. We avoid packing both attack effort and effectiveness into a single element of the attacker's problem. Rather, we use both the reward and the reward discounting factor to encode *and* prioritise these objectives. We propose a modification of DDPG [6] called γDDPG that supports dual-priority dual-objective optimisation with the aid of a dynamic discount function. Herein, the discount function, γ adapts in response to the current level of effort exerted by the attacker (and the current level of attack accuracy achieved with that effort) to create a bounded search space that bounds the lower priority objective (attacker effort), and enables the attacker to optimise the higher-priority objective (attack accuracy) within this bounded space. Furthermore, given that large discount factors lead to unreliable optimisation in uncertain environments [5], the bounded search space (created by the bounded discount function) in this work improves the optimisation capability of γDDPG by reducing the effect of uncertainty in the given black-box environment.

3 EXPERIMENTS

This work develops four adaptive discount functions based on Kullback Leibler Divergence Rate and Wasserstein Distance; two conditioned only on attacker effort (TargetKLR, TargetWD) while two conditioned on both attacker effort and accuracy (KLR, WD). Our results show that when TEPA, a SOTA baseline, is trained to enforce non-optimal target behaviour on a victim, it gets stuck in a local optima, unable to exit it even after ~20k training episodes. This is reflected in the KDE plots (a-c) in Figure 1 (and clearly shown via line-plots in the Full Paper) that compares the best-performing effort- and effort+accuracy-based dynamic discounts (TargetKLR and WD), with TEPA. In addition, the test-time plots (e-h) show that the best attack strategy found by TEPA performs worse than both WD and TargetKLR dynamic discounts w.r.t. @Acc, @SoftAcc and @Time while WD successfully finds strategies that maximise accuracy (higher-priority objective) and minimise effort (lower-priority objective) in low time.

4 FUTURE WORK

In this work, formulation of the attacker's state space, as well as (adaptive) discount, requires the underlying victim environment to be discrete in nature. Therefore, our next step entails extension of the proposed methodology to continuous victim environments. Furthermore, the proposed algorithm supports only dual-priority dual-objective optimisation. Future work constitutes expanding the developed methodology to multi-objective optimisation with more than two objectives and priority levels.

REFERENCES

[1] Kiarash Banihashem, Adish Singla, and Goran Radanovic. 2021. Defense Against Reward Poisoning Attacks in Reinforcement Learning. *arXiv preprint arXiv:2102.05776* (2021).

[2] Tong Chen, Jiqiang Liu, Yingxiao Xiang, Wenjia Niu, Endong Tong, and Zhen Han. 2019. Adversarial attack and defense in reinforcement learning-from AI security view. *Cybersecurity* 2, 1 (2019), 1–22.

[3] Ambra Demontis, Maura Pintor, Luca Demetrio, Kathrin Grosse, Hsiao-Ying Lin, Chengfang Fang, Battista Biggio, and Fabio Roli. 2022. A survey on reinforcement learning security with application to autonomous driving. *arXiv preprint arXiv:2212.06123* (2022).

[4] Inaam Ilahi, Muhammad Usama, Junaid Qadir, Muhammad Umar Janjua, Ala Al-Fuqaha, Dinh Thai Hoang, and Dusit Niyato. 2021. Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *IEEE Transactions on Artificial Intelligence* 3, 2 (2021), 90–109.

[5] MyeongSeop Kim, Jung-Su Kim, Myoung-Su Choi, and Jae-Han Park. 2022. Adaptive Discount Factor for Deep Reinforcement Learning in Continuing Tasks with Uncertainty. *Sensors* 22, 19 (2022), 7266.

[6] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016*, Yoshua Bengio and Yann LeCun (Eds.).

[7] Chris Lu, Timon Willi, Alistair Letcher, and Jakob Foerster. 2022. Adversarial Cheap Talk. *arXiv preprint arXiv:2211.11030* (2022).

[8] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. 2020. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning (ICML)*. PMLR, 7974–7984.

[9] Amin Rakhsha, Xuezhou Zhang, Xiaojin Zhu, and Adish Singla. 2021. Reward poisoning in reinforcement learning: Attacks against unknown learners in unknown environments. *arXiv preprint arXiv:2102.08492* (2021).

[10] Yanchao Sun, Da Huo, and Furong Huang. 2020. Vulnerability-aware poisoning mechanism for online RL with unknown dynamics. *arXiv preprint arXiv:2009.00774* (2020).

[11] Hang Xu, Xinghua Qu, and Zinovi Rabinovich. 2022. Spiking Pitch Black: Poisoning an Unknown Environment to Attack Unknown Reinforcement Learners. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 1409–1417.

[12] Hang Xu, Rundong Wang, Lev Raizman, and Zinovi Rabinovich. 2021. Transferable Environment Poisoning: Training-time Attack on Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. 1398–1406.

[13] Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu. 2020. Adaptive reward-poisoning attacks against reinforcement learning. In *International Conference on Machine Learning*. 11225–11234.