BIFURCATING BLACK BOXES: OBSERVATION PERTURBATION ATTACKS AND DEFENCES ON REINFORCEMENT LEARNING AGENTS IN A CYBER PHYSICAL POWER SYSTEM

KIERNAN BRODA-MILIAN[1], RANWA AL-MALLAH[2], HANANE DAGHOUGUI[3]

[1]Department of Electrical and Computer Engineering, Royal Military College, Kingston, Ontario, Canada
[2]Department of Computer Engineering and Software Engineering, Polytechnic Montreal, Montreal, Quebec, Canada
[3]Department of Mathematical and Industrial Engineering, Polytechnique Montréal, Montreal, Quebec, Canada
E-MAIL: kbrodamilian@gmail.com, ranwa.al-mallah@polymtl.ca, hanane.dagdougui@polymtl.ca

Abstract:

Deep Reinforcement Learning (DRL) is an increasingly popular technique in the literature for control in stochastic environments, which are problematic to model, or have large state or action spaces. Distributed Energy Resources (DER) in smart grids are one such example, where the proliferation of renewables makes power systems more difficult to control. However, the existing literature remains limited regarding the robustness and security of these controllers intended for deployment in critical infrastructures. This paper extends our previous work on executing load altering attacks targeting DRL controllers by analyzing the impact of the proposed bifurcation attack in black-box settings, thereby providing a more realistic threat model. By monitoring the victim's observations and actions, an attacker can train a proxy model to generate bifurcation-enhanced adversarial observations. Our novel bifurcated black-box attack doubles the adversarial regret compared to methods proposed in prior works in our experiments. We adapt Alternating Training with Learned Adversary (ATLA) to both strengthen robustness against this more powerful attack, which is DRL algorithm agnostic. We demonstrate that ATLA is effective against strong attacks on the agent's function approximator, in addition to the attacks exploiting the its policy in previous works. Our findings are relevant for threat assessment and hardening for DRL in smart grids.

Keywords:

Deep Reinforcement Learning, Adversarial Examples, Smart Energy, Cyber Physical Systems, Load Altering Attacks, Bifurcation Method, ATLA

1. Introduction

Networks of Operations Technologies (OT) comprise the Cyber Physical Systems (CPS) controlling Critical Infrastructure (CI), such as smart grids. [1]. There are numerous vectors and precedents for cyber attacks on such CI [2]. Furthermore, the sparse nature of Distributed Energy Resources (DER) in modern smart grids presents a wide attack surface.

Deep Reinforcement Learning (DRL) has been extensively investigated as a control framework for complex systems in diverse domains, including energy, transportation, aviation, and maritime operations, with DER representing one such application [3, 4]. However, the robustness and security of DRL controllers in CI have not been similarly investigated, although these controllers are vulnerable to Artificial Neural Network (ANN)-specific exploits, such as adversarial examples [5].

In our previous work [6], agents with different architectures were trained in a smart building gym environment [7] to reduce energy usage by controlling the charging and discharging of a battery energy storage system. If executed on sufficient scale, such alterations to loads could disrupt a power grid [8,9]. These agents were used to test white-box attacks including a novel method called the bifurcation attack. Now we present *Bifurcating Black Boxes*, an observation perturbation attack that significantly increases power consumption without requiring access to the agent's parameters, and demonstrate the impact of the bifurcation method. This black-box attack is more feasible in practice, as the attacker only requires the historical observations of the agent. We then propose a robust training method that mitigates the effect of the black-box attack. Finally, we show that the choice of action space also has a significant impact on robustness to black-box attacks.

2. Related Work

2.1 Black-Box Attacks on DRL Agents

A properly constrained attack is imperceptible to human observers and threatens virtually any DRL agent much like a software vulnerability. Transfer attacks [10] are in DRL do not require knowledge of the victim agent or it's training environment, using the snooping attack [11]. The authors in [11] propose three threat models for the snooping attack where the attacker uses limited information on the victim's performance to train a proxy victim using supervised learning. All the attacker must do is observe the victim in its environment.

2.2. Robust RL

Adversarial training used in supervised learning [12] works by maximizing loss and gradient-based search methods provide reasonable solutions for maximum loss. Finding an analogous loss for DRL is less effective, as taking a sub-optimal action based on an adversarial example is not guaranteed to minimize the agent's reward. Simply adding adversarial perturbations during training can prevent the agent from learning or converging, while in many cases this training method does not provide significant protection [13, 14].

Other approaches summarized in [14] showed similarly mixed results or required new DRL algorithms. Instead, [15] proposes the Alternating Training with Learned Adversary (ATLA) method for improving model robustness to a hostile environment. This involves training a DRL Learned Adversary (LA) that adds optimal perturbations to the agent's observations, in a reward-based attack, as shown in Figure 1. Taking turns training the adversary and agent ensures that the newly learned policy is not easily exploitable by some new policy, where neither adversary nor agent has any incentive to change their policy lest the other exploit it.

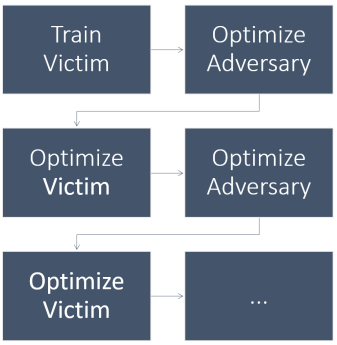

FIGURE 1. High level representation of ATLA.

To demonstrate the method's effect, ATLA was benchmarked against several gradient-based attacks that are unique to RL [14]. Of the five attacks, the Optimal/LA attack had the largest adversarial regret [16]. Adversarial regret was measured as the reduction in reward caused by the attacks during testing. ATLA was not tested against adversarial observations generated from the function approximator, such as those employed in [6]. ATLA aims to train a robust policy (π) rather than a robust function approximator or ANN. In this work, ATLA was used as a defence against LA attacks on an agent controlling multiple buildings in a CityLearn demand response environment [17].

2.3 The Bifurcation Method

We briefly summarize one of the contributions from [6]. Figure 2 shows how untargeted attacks do not significantly change the victim agent's behavior. Actor networks for discrete agents in CityLearn have symmetrically graduated actions for various levels of (dis)charge. These can be combined into (dis)charge groups by modifying the surrogate network used by the snooping attack to return only a pair of logits, corresponding to the maximal charge and discharge logits. We call this layer the bifurcation layer, which is described in Figure 3, and significantly increases the adversarial regret caused by the snooping attack.

3 Methodology

In this section, we will introduce the gym environment and discuss the attack and defence methods used in this study. We will then demonstrate how the snooping attack is improved with the Bifurcation method, generating *Bifurcating Black Boxes* and how ATLA was adapted to decrease the impact of this power attack.

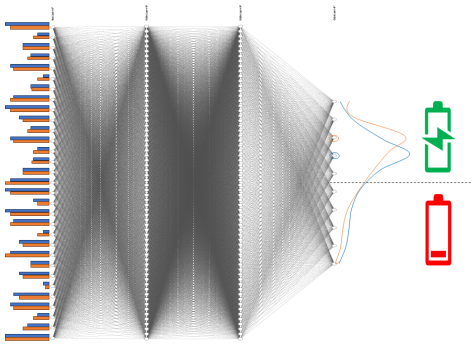

FIGURE 2. Example of an adversarial attack on a discrete actor network trained in CityLearn. Small changes to the original observation result in an adversarial action different from the original. But, the result is only slightly more charging than is optimal, so the impact on power consumption is limited.

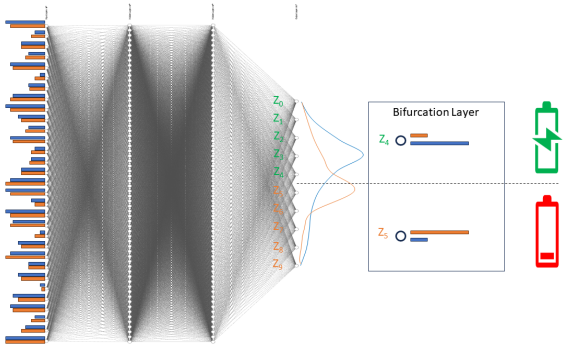

FIGURE 3. Example of an adversarial attack on a discrete actor network trained in CityLearn, using the bifurcation method. In this example, small changes to the original observation result in a discharge action instead of the original charge action. Inducing the victim agent to reverse its (dis)charge decisions increases electricity consumption.

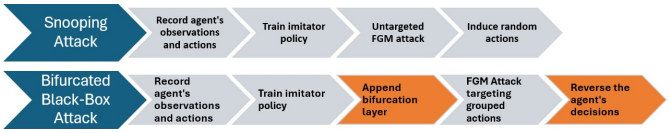

FIGURE 4. Comparison of previous work using the snooping attack to our novel approach. As the bifurcation method proposed in [6] improved white-box attacks, the adversarial regret of the snooping attack can also be increased.

3.1 Gym Environment

As per the methodology of [6], CityLearn [18] is used to simulate the victim controlling the energy storage of a smart building. The generation of this environment's observations from historical data makes it a realistic setting for this research. Our KPI is the building's electrical consumption.

3.2 Bifurcating Black Boxes

This section tests the impact of the black-box snooping attack on the same agents used to conduct white-box attacks, to show that a load altering attack targeting a DRL controller is possible without access to its ANN's parameters. The snooping attack [11] allows an attacker to leverage adversarial attacks with *no knowledge* of the victim agent's algorithm or architecture. The attacker is only able to observe the victim agent acting in its environment, which is necessary to craft adversarial perturbations. An attacker who has compromised internet of things sensors or improperly secured links between them and the controller could then observe and modify the observations.

The snooping attack [11] alone is relatively weak because proxy models only resemble the agent's policy, making targeted attacks and advanced iterative methods fail. In our control application in CityLearn a snooping attack is unlikely to cause a significant change to the agent's action, meaning many successful adversarial observations will fail to degrade its performance. We address these limitations using our bifurcation method, as shown in figure 4 to ensure that our adversarial observations significantly increase adversarial regret.

Using the State-Action threat model, an ANN proxy imitator was trained and used to enable Fast Gradient Method (FGM) [19] attacks. The bifurcation layer was appended to the trained proxy during inference, as it was to a trained agent's policy network in our prior work. The proxy was trained agnostic to the agent's architecture, instead hyperparameters were chosen during an Optuna [20] trial which maximized the mean time series split validation over one episode of the agent's observations and actions.

3.3 Robust Training

ATLA [15] is a training method that teaches an agent a robust policy (π), as described in Algorithm 1. ATLA does not require a unique or modified DRL algorithm to ensure robustness, because it is purely a training method, and the cost of adding it to an existing agent or training system is relatively small compared to using a different robust algorithm. Essentially, ATLA could be applied to any algorithm from existing libraries. Our review of the literature found no evaluation of resistance to adversarial attacks targeting a DRL agent's function approximator, so that is the goal of this experiment.

The environments for the adversary and victim use identical CityLearn parameters but have different action spaces and rewards. The adversary receives a clean observation from CityLearn and its action is adding a perturbation, then the victim's static policy in the adversary's environment selects an action based on the adversarial observation. Conventionally, the action space of the adversary is constrained by a function $A_{adv} \in B(s)$, where s is the current state. To avoid violating the Markov property and be consistent with the L_∞ constraint from previous attacks, $B(s)$ is a static boundary for each feature. This boundary was generated using $\sim 50\%$ of the difference of the maximum and mean variations between observations during a clean episode. $B(s)$ is a training hyperparameter and must be selected such that the adversary is capable of a significant adversarial regret without preventing the agent from learning.

4 Results

Snooping attacks were conducted using FGM, as per the methodology of [11]. Because FGM is a weaker attack compared to iterative methods, it requires a larger adversarial budget ϵ for a similar adversarial regret. The relationship between the adversarial budget and regrets for the snooping attack on the discrete PPO are shown in Figure 5. When random noise was compared to a snooping attack with the same ϵ as validation (absent from figures for brevity), the random attack requires almost 10 times the adversarial budget for a similar success rate. Random noise did not have a significant adversarial regret for any perturbation size tested, up to $\epsilon = 0.2$. The discrete PPO is robust in that it requires $\epsilon \approx 0.13$ before power consumption is equivalent to no controller, and is increased by over 10%. By comparing the slope of the direct and bifurcated attacks, the impact of the former becomes apparent. The bifurcation method significantly increases the cost and power consumption for the conventionally trained agent, though the ATLA trained agent sees a much smaller difference. Our *Bifurcating Black Box* attack produced a significantly higher adversarial regret for the conventional agent compared to the ATLA agent. This is the same to a smaller extent for the direct snooping attack, and this was partially counteracted by the ATLA agent's lower clean performance. Because black-box attacks require less knowledge, they are more likely to be encountered. These attacks are also easier to detect, as previous works have shown that adversarial samples crafted with FGM are detectable through statistical techniques [21]. Furthermore, [6] found that attacks with $\epsilon > 0.03$ are statistically detectable in CityLearn, while snooping attacks require four times that to undo the benefit of the DRL controller for a discrete agent.

On the other hand, the performance of the discrete and continuous PPO, and Soft Actor Critic (SAC) under a bifurcated black-box attack are compared in Figure 6. While $\epsilon \approx 0.13$ is required to remove the benefit of the discrete PPO in terms of power consumption, the same happens for the SAC for $\epsilon \approx 0.06$ and $\epsilon \approx 0.05$ for the continuous PPO. At this smaller value of ϵ the adversarial regret for the discrete PPO is only 0.03. Even without ATLA training, the discrete action space significantly improves the robustness of the PPO to the black-

Algorithm 1 Alternating Training with Learned Adversaries (ATLA) [15].

Inputs Transition function T(s,a) for the environment, number of alternations N_{alt}, number of pre-training iterations N_{pre}, agent training iterations N_π , adversary training iterations N_ν
Initialize θ_0 and ϕ_0 to parameterize π and ν
if $N_{pre} > 0$ **then** //optional agent pre-training
  for i = 0 to N_{pre} **do**
    $\theta_{i+1} \leftarrow train(T(\pi(s, \theta_i), s))$
  end for
  $\theta_0 \leftarrow \theta_{N_{pre}}$
end if
for i = 0 in N_{alt} **do**
  for j = 0 to N_ν **do**
    $\phi_{j+1} \leftarrow train(T((\pi(\nu(s, \phi_j), \theta_0)), s)$ //θ is constant
  end for
  $\phi_0 \leftarrow \phi_{N_\nu}$
  for j = 0 to N_π **do**
    $\theta_{j+1} \leftarrow train(T((\pi(\nu(s, \phi_0), \theta_j)), s)$
  end for
  $\theta_0 \leftarrow \theta_{N_\pi}$
end for

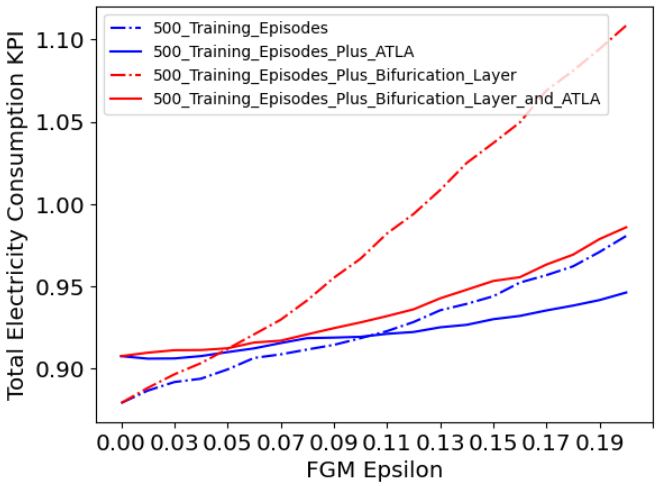

FIGURE 5. Comparison of electricity consumption KPI for snooping and our bifurcated black-box attacks with a range of ϵ, for discrete Proximal Policy Optimization (PPO) agents trained conventionally and with the ATLA method. While the robustness offered by ATLA is insignificant compared to its reduction in clean performance for energy consumption, the corresponding adversarial regret for bifurcated attacks is significantly reduced.

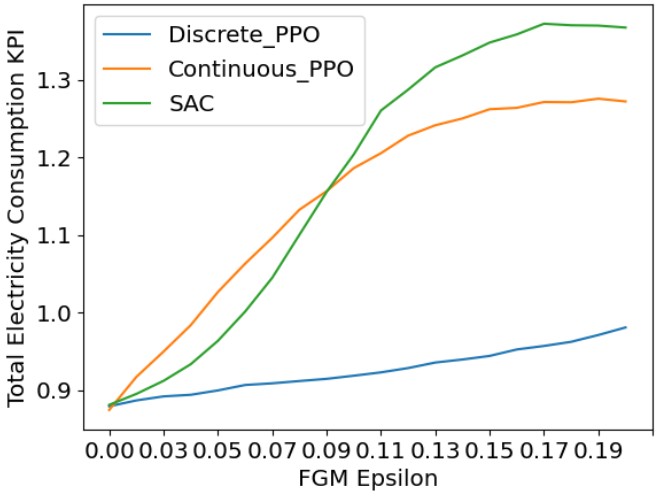

FIGURE 6. Comparison of electricity consumption KPI for bifurcated FGM snooping attacks with a range of ϵ, for discrete and continuous PPO, and SAC agents. This demonstrates that the discrete PPO is significantly more robust than either agent with a continuous action space.

box snooping attack.

In fact, an adversary can achieve a significant reduction in performance while only observing how the victim behaves in their environment and exploiting vulnerable sensor systems or links. Unlike the finding in [15] where the LA outperformed the snooping attack, we showed that combining the snooping attack with the bifurcation method outperforms the LA attack while using a smaller adversarial budget. The bifurcated black-box attack had a power consumption comparable to the LA attack with $\epsilon = 0.07$, which is much smaller than the $B(s)$ for the LA attack, for all except solar generation and net electricity consumption features.

5 Conclusion

Combining the bifurcation method with the snooping attack significantly increased the adversarial regret for the conventionally trained PPO, but the increase was less than half for the PPO trained using ATLA. Thus, ATLA is an effective defence against strong black-box attacks targeting the ANN function approximator. As with our previous work on white-box attacks, an algorithm with a discrete action space was significantly more robust to our black-box attack than an identical algorithm with continuous actions. This provides additional evidence that discrete action spaces improve robustness.

References

[1] C. C. f. C. Security, "Cyber threat bulletin: Cyber threat to operational technology," Dec 2021, last Modified: 2021-12-16. [Online]. Available: https://www.cyber.gc.ca/en/guidance/cyber-threat-bulletin-cyber-threat-operational-technology

[2] I. Zografopoulos, N. D. Hatziargyriou, and C. Konstantinou, "Distributed energy resources cybersecurity outlook: Vulnerabilities, attacks, impacts, and mitigations," *IEEE Systems Journal*, p. 1–15, 2023.

[3] D. Cao, W. Hu, J. Zhao, G. Zhang, B. Zhang, Z. Liu, Z. Chen, and F. Blaabjerg, "Reinforcement learning and its applications in modern power and energy systems: A review," *Journal of Modern Power Systems and Clean Energy*, vol. 8, no. 6, p. 1029–1042, Nov 2020.

[4] A. T. D. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110618, Mar 2021.

[5] I. Ilahi, M. Usama, J. Qadir, M. U. Janjua, A. Al-Fuqaha, D. T. Hoang, and D. Niyato, "Challenges and countermeasures for adversarial attacks on deep reinforcement learning," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 2, p. 90–109, Apr 2022.

[6] K. BRODA-MILIAN, H. DAGDOUGUI, and R. A. MALLAH, "The bifurcation method: White-box observation perturbation attacks on reinforcement learning agents on a cyber physical system," in *2024 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2024, pp. 571–578.

[7] K. E. Nweye, A. Wu, H. Park, Y. Almilaify, and Z. Nagy, "Citylearn: A tutorial on reinforcement learning control for grid-interactive efficient buildings and communities," in *ICLR 2023 Workshop on Tackling Climate Change with Machine Learning*, 2023. [Online]. Available: https://www.climatechange.ai/papers/iclr2023/2

[8] S. Soltan, P. Mittal, and H. V. Poor, "BlackIoT: IoT botnet of high wattage devices can disrupt the power grid," 2018, p. 15–32. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/soltan

[9] B. Huang, A. A. Cardenas, and R. Baldick, "Not everything is dark and gloomy: Power grid protections against IoT demand attacks," 2019, p. 1115–1132. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/huang

[10] F. Tramèr, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "The space of transferable adversarial examples," no. arXiv:1704.03453, May 2017, arXiv:1704.03453 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1704.03453

[11] M. Inkawhich, Y. Chen, and H. Li, "Snooping attacks on deep reinforcement learning," in *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, ser. AAMAS '20. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, May 2020, p. 557–565.

[12] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv*, 2018, accepted: 2021-11-05T14:56:05Z. [Online]. Available: https://dspace.mit.edu/handle/1721.1/137496

[13] V. Behzadan and A. Munir, "Whatever does not kill deep reinforcement learning, makes it stronger," no. arXiv:1712.09344, Dec. 2017, arXiv:1712.09344 [cs]. [Online]. Available: http://arxiv.org/abs/1712.09344

[14] H. Zhang, H. Chen, C. Xiao, B. Li, M. Liu, D. Boning, and C.-J. Hsieh, "Robust deep reinforcement learning against adversarial perturbations on state observations," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, p. 21024–21037. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/f0eb6568ea114ba6e293f903c34d7488-Abstract.html

[15] H. Zhang, H. Chen, D. Boning, and C.-J. Hsieh, "Robust reinforcement learning on state observations with learned optimal adversary," *International Conference on Learning Representation (ICLR)*, Jan. 2021. [Online]. Available: https://par.nsf.gov/biblio/10318889-robust-reinforcement-learning-state-observations-learned-optimal-adversary

[16] V. Behzadan and W. Hsu, "Rl-based method for benchmarking the adversarial resilience and robustness of deep reinforcement learning policies," in *Computer Safety, Reliability, and Security*, ser. Lecture Notes in Computer Science, A. Romanovsky, E. Troubitsyna, I. Gashi, E. Schoitsch, and F. Bitsch, Eds. Cham: Springer International Publishing, 2019, p. 314–325.

[17] L. Zeng, D. Qiu, and M. Sun, "Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks," *Applied Energy*, vol. 324, p. 119688, Oct 2022.

[18] J. R. Vazquez-Canteli, S. Dey, G. Henze, and Z. Nagy, "Citylearn: Standardizing research in multi-agent reinforcement learning for demand response and urban energy management," no. arXiv:2012.10504, Dec 2020, arXiv:2012.10504 [cs]. [Online]. Available: http://arxiv.org/abs/2012.10504

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," no. arXiv:1412.6572, Mar. 2015, arXiv:1412.6572 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1412.6572

[20] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," no. arXiv:1907.10902, Jul. 2019, arXiv:1907.10902 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1907.10902

[21] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," no. arXiv:1702.06280, Oct 2017, arXiv:1702.06280 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1702.06280