Adversarial poisoning attacks on reinforcement learning-driven energy pricing

Sam Gunn[*]
gunn@berkeley.edu

Doseok Jang[*]
djang@berkeley.edu

Orr Paradise[*]
orrp@eecs.berkeley.edu

Lucas Spangher[*]
lucas_spangher@berkeley.edu

Costas J. Spanos[*]
spanos@berkeley.edu

Abstract

Reinforcement learning (RL) has emerged as a strong candidate for implementing complex controls in energy systems, such as energy pricing in microgrids. But what happens when some of the microgrid controllers are compromised by a malicious entity? We demonstrate a novel attack in RL. Our attack perturbs each trajectory to reverse the direction of the estimated gradient. We demonstrate that if data from a small fraction of microgrid controllers is adversarially perturbed, the learning of the RL agent can be significantly slowed or (with larger perturbations) caused to operate at a loss. Prosumers also face higher energy costs, use their batteries less, and suffer from higher peak demand when the pricing aggregator is adversarially poisoned.

We address this vulnerability with a defense module; i.e., a "robustification" of RL algorithms against this attack. Our defense identifies the trajectories with the largest influence on the gradient and removes them from the training data.

1 Introduction

Artificial Intelligence (AI) heralds great benefits to power systems. In the future, AI-based controls could manage the use of passive appliances (Zhang et al., 2020; Chen et al., 2019), orchestrate demand response (Azuatalam et al., 2020), and optimize power flow (Chen et al., 2021; Dall'Anese et al., 2013). In the context of energy grids, local grid networks (i.e., microgrids) enable refined control at the cost of increased complexity, necessitating adoption of complex controls at scale.

At the same time, energy grids are known to be lucrative targets for cyberattacks (e.g., Kshetri and Voas, 2017). Our work investigates the robustness of an AI-based microgrid controller to malicious actors. We present a novel attack that enables a few compromised microgrid controllers to adversely affect the behavior of connected controllers by *poisoning the data* on which it is trained. We pair this finding with a gradient-based defense that eliminates the threat of this attack.

Concretely, we examine a setting in which a network of microgrid controllers collect supply and demand data that are continually aggregated by a central agent. The agent uses online reinforcement learning (RL) to optimize its profits. In our attack, a few microgrid controllers are compromised by a malicious adversary. The adversary applies a perturbation to the collected data, severely impacting the provider and *the entire network* of controllers. The provider is made to operate at a loss, and all prosumers are made to pay higher energy costs, use their batteries less, and increase peak demand.

Our work is set against a backdrop of developments in energy grid control that hold both promise and peril: RL-based controllers allow for sophisticated control in unprecedented granularity. Yet,

[*]University of California, Berkeley. Authors ordered alphabetically.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

we must be careful to minimize risk enabled by the opaque nature of deep learning. Our attack stands out in its subtlety and its scope. Other forms of large-scale interference such as blackouts and line disruptions are, by definition, easily detectable and local. Yet our attack causes harm by interfering with the agent's learning, and may not be detected until significant financial damage has been incurred. Furthermore, by interfering with the central agent's learning, our methods can damage systems that are physically disconnected from the compromised energy grid.

2 Background

Adversarial Attacks Adversarial attacks have seen great success in supervised learning. Fast gradient sign-based attacks (Goodfellow et al., 2014; Madry et al., 2018), decision boundary-based attacks (Moosavi-Dezfooli et al., 2016), and even attacks that learn an adversarial policy (Gleave et al., 2020) have been proposed to fool supervised learners. It has been shown that similar attacks can work on reinforcement learning agents (Huang et al., 2017), with the added nuance that these attacks can be strategically timed to maximize impact or move the agent into a desired state (Lin et al., 2017). However, these popular works have mostly been focused on so-called evasion attacks, which focus on generating adversarial examples *at test time*.

Our work focuses on data poisoning at *training* time. Usually, training phase attacks can be split into two categories (Chakraborty et al., 2018): label manipulation and input manipulation. In the context of RL, most work has been analogous to label manipulation: changing the recorded rewards (Ma et al., 2018; Liu and Shroff, 2019) rather than the actions or observations. Data poisoning has been examined in the context of supervised learning (Akhtar and Mian, 2018; Kloft and Laskov, 2010; Biggio et al., 2011), but not in deep reinforcement learning, to our knowledge.

RL for prosumer energy pricing RL has been applied to a number of demand response situations in prosumer microgrids; most work centers on agents that directly schedule resources (Vázquez-Canteli et al., 2019; Vázquez-Canteli and Nagy, 2019) or control appliances (Pinto et al., 2021; Zhou and Zheng, 2020). Recent works have used an RL controller as a price setter in a market: RL has been used to estimate dynamic prices in a multi agent environment of demand response assets (Jang et al., 2022; Agwan et al., 2021).

Demand response, an incentive mechanism geared towards moving consumption, is a no-material solution to variable wind and solar generation and is thus seen as an important technique in the energy transition. It has been demonstrated that learning local price controls is an effective demand response mechanism due to its generalizability and optimal local battery resource utilization (Spangher et al., 2020; Spangher, 2021).

The literature on adversarial attacks for RL in demand response focuses on *responding* to prices (Wan et al., 2021) rather than *setting* them. To our knowledge, there are no works on adversarial attacks on dynamic price setting for demand response.

3 Techniques

3.1 Threat model

In our setting, N controllers continuously collect data to be aggregated by a centralized agent. Learning takes place over multiple *iterations*; in each iteration, each controller collects a trajectory $\tau := (o_i, a_i, r_i)_i$ collected according to the agent policy π_θ. The agent's policy π_θ is described by a neural network. Nodes are required to feed observations through π_θ so as to collect policy-specified actions (pricing schemes), so we assume that the network parameters θ and architecture are shared with the controllers (and therefore the adversary).

The attacker's power is determined by a fraction of *corrupted controllers* $\varepsilon \in (0, 1)$, and a *perturbation bound* $\rho > 0$, as follows: An attacker controls $\varepsilon \cdot N$ of controllers. The attacker *perturbs* the trajectories collected by each compromised controller, causing it to report back a trajectory $\widetilde{\tau}$ instead of the collected trajectory τ. Crucially, these perturbations are of small norm, that is, $\|\widetilde{\tau} - \tau\|_\infty \leq \rho$, for some *perturbation bound* $\rho > 0$. Note that our attacker adheres to the suggested policy π_θ, but lies about the result to the agent.

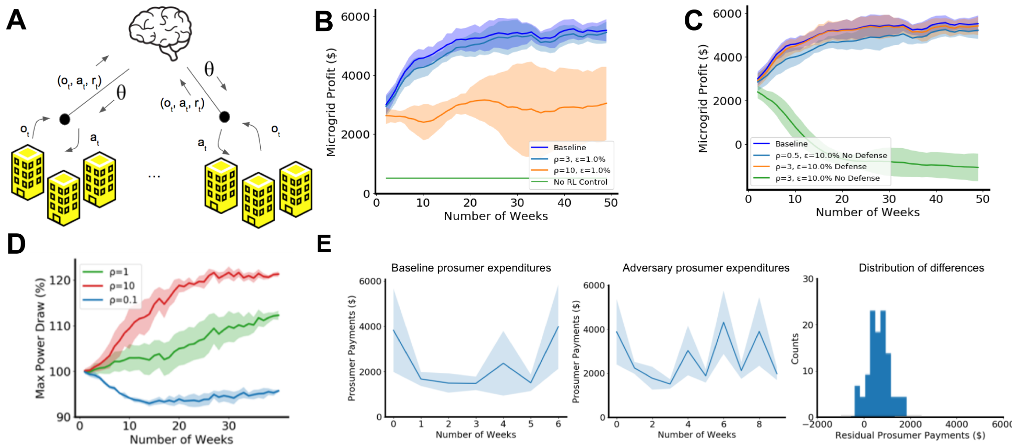

Figure 1: **A.** The microgrid environment; the brain is the RL agent, the black dot is the microgrid controller, and the adversary attacks the a_t that is sent back to the RL agent. **B.** Effect of the adversary on the agent's learning. Note that $\varepsilon = 1\%$ corresponds to only one adversarial microgrid. **C.** Effect of our defense in the presence of an adversary. **D.** Targeted attacks: the adversary is able to manipulate the RL agent's policy such that peak power consistently exceeds 120% of the grid's capacity, raising risk of transformer blowout. **E.** Prosumer costs in the baseline and adversarial scenarios: the prosumer consistently pays more in energy when the adversary interferes.

In our setting, the attacker perturbs the actions of each trajectory. Observations and rewards remain unperturbed, because such perturbations would be expensive or easily noticed. This is in contrast to previous work in RL poisoning in which only rewards are poisoned (Rakhsha et al., 2021).

3.2 Attacks and defense

At a high level, our **attack** aims to perturb each trajectory to reverse the direction of the estimated gradient $\nabla_\theta f(\tau_p)$. Let θ be the parameters of the agent's policy, τ_P be the unperturbed set of compromised trajectories (the trajectories collected by compromised controllers), $\tilde{\tau}_P$ be the set of perturbed adversarial trajectories (reported back to the agent), and τ_H be the set of honest trajectories (unaffected by the adversary). Our adversary minimizes the correlation of the gradient post-perturbation with the honest one by solving the following constrained optimization problem:

$$\min_{\tilde{\tau}_P} \quad \langle \nabla_\theta f_\theta\left(\tilde{\tau}_P\right), \nabla_\theta \left(f_\theta(\tau_P) + f_\theta(\tau_H)\right)\rangle \tag{1}$$
$$\text{such that} \quad ||\tilde{\tau}_P - \tau_P||_\infty \leq \rho.$$

Since compromised controllers report $\tilde{\tau}_P$ instead of τ_P, the agent will take gradient steps according to $\nabla_\theta \left(f_\theta(\tilde{\tau}_P) + f_\theta(\tau_H)\right)$. Therefore, choosing $\tilde{\tau}_P$ to minimize Equation (1) should maximally mislead the gradient towards a sub-optimal policy. Equation (1) is optimized by the adversary using the Fast Gradient Sign Method (FGSM, Goodfellow et al., 2014). Interestingly, we find that our adversaries can obtain nearly identical results by solving Equation (1) without the τ_H term, meaning that the adversary may not need any information about the honest (uncompromised) controllers.

The targeted attack. By tweaking the optimization objective, the adversary can cause the RL agent to learn optimize an auxiliary *target* "reward". Let $\tau' := (o_i, a_i, \tilde{r}_i)_i$, the set of all collected trajectories with rewards relabeled with (adversarially-chosen) "reward" $\tilde{r}$. Then we formulate our new constrained optimization problem as:

$$\max_{\tilde{\tau}_P} \quad \langle \nabla_\theta f_\theta\left(\tilde{\tau}_P\right), \nabla_\theta f_\theta\left(\tau'\right)\rangle \tag{2}$$
$$\text{such that} \quad ||\tilde{\tau}_P - \tau_P||_\infty \leq \rho.$$

By maximizing the correlation between $\nabla_\theta f_\theta\left(\tilde{\tau}_P\right)$ and $\nabla_\theta f_\theta\left(\tau'\right)$, we can maximally mislead the gradient towards a policy that maximizes the adversary's "reward" instead of the true reward.

Defense. Our defense works by identifying and removing the trajectories which have the largest influence on the gradient from the training data. Intuitively, this defense works because honest trajectories are not expected to have out-sized gradients. Note that the poisoned trajectories are not easily identifiable at first glance;while the adversarial perturbations significantly influence the gradient estimate, the perturbations themselves are small. More formally, if the RL agent suspects that some fraction $\hat{\varepsilon}$ of the microgrids are adversarially controlled, then, when estimating the gradient $\nabla_\theta f(\theta)$, it ignores the $\hat{\varepsilon}$-fraction of trajectories τ with largest $||\nabla_\theta f_\theta(\tau)||_2$ in each training batch.

4 Experimental setup

The Price-Setting Microgrid Problem Consider a setting of 100 microgrids. An Actor–Critic agent sets the policy parameters θ of all 100 microgrid controllers, which transact within each microgrid. Each microgrid consists of 7 prosumer office buildings, each of which has a battery, solar panel array, and baseline energy consumption; each wants to minimize their energy cost. Prosumers see both grid-set hourly energy buy and sell prices and local microgrid controller-set hourly energy buy and sell prices. Prosumers choose to transact with either the grid or the RL aggregator at each hour. Prosumers also decide when to discharge their battery according to both their demand and the energy prices. The microgrid does not produce or store energy, but sells energy straight from prosumers producing energy in a timestep to prosumers demanding energy in the same timestep. The aggregator balances the net load by purchasing from or selling to the energy utility under which they sit, usually at a loss. As the manager of the RL-aggregator, you see the grid's prices, and wish to learn a pricing policy such that you consistently turn a profit. See Figure 1.A for a graphical depiction of the environment. For a more precise description of the convex optimizations governing prosumer battery behavior and the reward function training the RL-aggregator, see (Agwan et al., 2021).

For testing the viability of a *targeted attack*, we define an auxiliary adversary objective as the maximization of peak power over the step period.

Hypothetical scenario of adversarial microgrid poisoning Suppose that Eastern Gas & Electric (EG&E) is piloting a dynamic, local pricing program. To do this, EG&E instantiates an RL agent to train across a sample of building clusters (i.e., microgrids grouped locally). Unfortunately, there is an attacker who wishes to disrupt the functioning of EG&E, and they intercept the outflow of data from one of the local microgrid controllers. In one attack strategy, the attacker wishes to minimize the extent to which the outgoing prices are perturbed so as to escape detection. In another attack strategy, the attacker considers high perturbations in order to maximally disrupt profitability.

5 Results

The attack. Figure 1.B shows our attacker can significantly hinder the RL agent's learning by co-opting a single microgrid controller. The maximal difference between successive actions taken by the true policy is around 6, so the strongest attack in the single-trajectory setting requires a relatively high perturbation budget $\rho = 10$. However in Figure 1.C, our attack utilizes a smaller perturbation budget of $\rho = 3$ with ten ($\varepsilon = 10\%$) compromised controllers to achieve significant damage.

The defense. We find that our defense recovers the original performance of the RL agent. In particular, the defense does not noticeably affect training time, even when $\varepsilon = 10\%$ of trajectories are removed. See Figure 1.C.

Characterizations of environmental response. We investigated several ways in which the environment responded to adversarial attack beyond sheer profit: individual prosumer energy costs (the sum of the building's energy expenditures with the adversary and without), battery utilization (the number of times batteries were charged and discharged) and peak power draw. Under all measures, the environment performed worse with an adversary, even those not directly targeted: prosumers paid on average *more* for energy, the battery was used *less*, and there was *more* peak demand. We present prosumer prices in Figure 1.E and omit the rest due to space constraints.

Targeted attacks. When we chose an adversarial reward of increasing peak power demanded by prosumers on the microgrid, we see that with increasing adversarial strength we were able to consistently exceed 120% of grid capacity. Exceeding thresholds of power consumption on the grid drastically increases risk of transformer power constraint violation. See Figure 1.D.

Acknowledgments and Disclosure of Funding

We are very thankful to Sergey Levine for thoughtful comments and guidance throughout all stages of this work. This research is supported by the *Simons Collaboration on the Theory of Algorithmic Fairness*, the *DARPA GRAIL project*, and by the *National Research Foundation, Prime Ministers Office, Singapore* under its *Campus for Research Excellence and Technological Enterprise (CREATE) programme*.

References

Utkarsha Agwan, Lucas Spangher, William Arnold, Tarang Srivastava, Kameshwar Poolla, and Costas J. Spanos. 2021. Pricing in Prosumer Aggregations using Reinforcement Learning. In *e-Energy '21: The Twelfth ACM International Conference on Future Energy Systems, Virtual Event, Torino, Italy, 28 June - 2 July, 2021*, Herman de Meer and Michela Meo (Eds.). ACM, 220–224.

Naveed Akhtar and Ajmal Mian. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access* 6 (2018), 14410–14430.

Donald Azuatalam, Wee-Lih Lee, Frits de Nijs, and Ariel Liebman. 2020. Reinforcement learning for whole-building HVAC control and demand response. *Energy and AI* 2 (2020), 100020.

Battista Biggio, Blaine Nelson, and Pavel Laskov. 2011. Support vector machines under adversarial label noise. In *Asian conference on machine learning*. PMLR, 97–112.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069* (2018).

Bingqing Chen, Zicheng Cai, and Mario Bergés. 2019. Gnu-rl: A precocial reinforcement learning solution for building hvac control using a differentiable mpc policy. In *Proceedings of the 6th ACM international conference on systems for energy-efficient buildings, cities, and transportation*. 316–325.

Bingqing Chen, Priya L Donti, Kyri Baker, J Zico Kolter, and Mario Bergés. 2021. Enforcing policy feasibility constraints through differentiable projection for energy optimization. In *Proceedings of the Twelfth ACM International Conference on Future Energy Systems*. 199–210.

Emiliano Dall'Anese, Hao Zhu, and Georgios B Giannakis. 2013. Distributed optimal power flow for smart microgrids. *IEEE Transactions on Smart Grid* 4, 3 (2013), 1464–1475.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. 2020. Adversarial Policies: Attacking Deep Reinforcement Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. `https://openreview.net/forum?id=HJgEMpVFwB`

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284* (2017).

Doseok Jang, Lucas Spangher, Selvaprabu Nadarajah, and Costas Spanos. 2022. Decarbonizing Buildings via Energy Demand Response and Deep Reinforcement Learning: The Deployment Value of Supervisory Planning and Guardrails. *Available at SSRN 4078206* (2022).

Marius Kloft and Pavel Laskov. 2010. Online anomaly detection under adversarial impact. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 405–412.

Nir Kshetri and Jeffrey M. Voas. 2017. Hacking Power Grids: A Current Problem. *Computer* 50, 12 (2017), 91–95.

Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. 2017. Tactics of adversarial attack on deep reinforcement learning agents. *arXiv preprint arXiv:1703.06748* (2017).

Fang Liu and Ness B. Shroff. 2019. Data Poisoning Attacks on Stochastic Bandits. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 4042–4050. `http://proceedings.mlr.press/v97/liu19e.html`

Yuzhe Ma, Kwang-Sung Jun, Lihong Li, and Xiaojin Zhu. 2018. Data Poisoning Attacks in Contextual Bandits. In *Decision and Game Theory for Security - 9th International Conference, GameSec 2018, Seattle, WA, USA, October 29-31, 2018, Proceedings (Lecture Notes in Computer Science, Vol. 11199)*, Linda Bushnell, Radha Poovendran, and Tamer Basar (Eds.). Springer, 186–204. `https://doi.org/10.1007/978-3-030-01554-1_11`

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. `https://openreview.net/forum?id=rJzIBfZAb`

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2574–2582.

Giuseppe Pinto, Marco Savino Piscitelli, José Ramón Vázquez-Canteli, Zoltán Nagy, and Alfonso Capozzoli. 2021. Coordinated energy management for a cluster of buildings through deep reinforcement learning. *Energy* 229 (2021), 120725.

Amin Rakhsha, Xuezhou Zhang, Xiaojin Zhu, and Adish Singla. 2021. Reward Poisoning in Reinforcement Learning: Attacks Against Unknown Learners in Unknown Environments. *CoRR* abs/2102.08492 (2021). arXiv:2102.08492

Lucas Spangher. 2021. Transactive multi-agent reinforcement learning for distributed energy price localization. In *BuildSys '21: The 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, Coimbra, Portugal, November 17 - 18, 2021*, Xiaofan Fred Jiang, Omprakash Gnawali, and Zoltan Nagy (Eds.). ACM, 244–245.

Lucas Spangher, Akash Gokul, Manan Khattar, Joseph Palakapilly, Akaash Tawade, Adam Bouyamourn, Alex Devonport, and Costas J. Spanos. 2020. Prospective Experiment for Reinforcement Learning on Demand Response in a Social Game Framework. In *e-Energy '20: The Eleventh ACM International Conference on Future Energy Systems, Virtual Event, Australia, June 22-26, 2020*. ACM, 438–444.

José R. Vázquez-Canteli, Jérôme Henri Kämpf, Gregor Henze, and Zoltán Nagy. 2019. CityLearn v1.0: An OpenAI Gym Environment for Demand Response with Deep Reinforcement Learning. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, BuildSys 2019, New York, NY, USA, November 13-14, 2019*. ACM, 356–357.

José R Vázquez-Canteli and Zoltán Nagy. 2019. Reinforcement learning for demand response: A review of algorithms and modeling techniques. *Applied energy* 235 (2019), 1072–1089.

Zhiqiang Wan, Hepeng Li, Hang Shuai, Yan Lindsay Sun, and Haibo He. 2021. Adversarial Attack for Deep Reinforcement Learning Based Demand Response. In *2021 IEEE Power & Energy Society General Meeting (PESGM)*. IEEE, 1–5.

Xiangyu Zhang, Xin Jin, Charles Tripp, David J Biagioni, Peter Graf, and Huaiguang Jiang. 2020. Transferable reinforcement learning for smart homes. In *Proceedings of the 1st International Workshop on Reinforcement Learning for Energy Management in Buildings & Cities*. 43–47.

Yuekuan Zhou and Siqian Zheng. 2020. Machine-learning based hybrid demand-side controller for high-rise office buildings with high energy flexibilities. *Applied Energy* 262 (2020), 114416.