Vulnerability Assessment of Deep Reinforcement Learning Models for Power System Topology Optimization

Yan Zheng, Ziming Yan, *Graduate Student Member, IEEE*, Kangjie Chen, Jianwen Sun, *Member, IEEE*, Yan Xu, *Senior Member, IEEE*, and Yang Liu, *Senior Member, IEEE*

***Abstract*—This paper studies the vulnerability of deep reinforcement learning (DRL) models for power systems topology optimization under data perturbations and cyber-attack. DRL has recently solved many complex power system optimization problems. However, it has been practically proven that small perturbations of input data can lead to drastically different control decisions and induce danger. To evaluate and mitigate the security risks of DRL models in power systems, we propose a vulnerability assessment method for such DRL models under noisy data and cyber-attack. In specific, we assess the vulnerability of a DRL model in a way that perturbations are constructed to minimize the model's performance. Besides, several vulnerability indices are proposed to identify the characteristics of perturbations that may cause malfunction of DRL. Simulations on the 14-bus system and the IEEE 118-bus system for topology optimization are carried out to validate the effectiveness of the proposed vulnerability assessment method. The results show that the performance of DRL models for power systems can be significantly degraded under cyber-attack and data perturbations, especially when a proposed vulnerability index has abnormal values.**

***Index Terms*—Power system optimization applications, deep reinforcement learning, vulnerability assessment, false data injection attack.**

Nomenclature

Variables

S_{Li}, $S_{Lm,i}$	Apparent line flow and thermal limit of the i-th line.
P_{Di}, Q_{Di}	Active load and reactive load for the i-th node.
P_{Gi}, Q_{Gi}	Active generation and reactive generation for the i-th node.
δ_{ij}, V_j	The difference between angles of buses i and j. Voltage magnitude of the jth bus.
G_{ij}, B_{ij}	The (i, j) entry of conductance and susceptance in nodal admittance matrix.
S, A, P_t, r, γ	State space, action space, transition probability, reward function and discount factor.
$\pi(s_t)$, $Q^{\pi}(s, a)$	Agent's policy under state s_t. Action-value function under policy π.
π^*, a_t, s_t', λ, a_t'	Optimal policy, optimal action under the optimal policy, perturbed state, perturbation and the agent's action under the perturbed state.
$\mathcal{L}(\theta)$	Loss function under DNN parameters θ.
η, θ^-	The learning rate to control the training speed and the target network parameters for stabilizing the training procedure.
$p(s_t)$, $g_t(s_t)$	Attackers' preference to attack the DRL agent and the gradient of the loss function with respect to the state.
r_t, r_e	The instant reward at time t, penalty on power flow divergence
λ_t, ϵ, k_t	Perturbation at time t, the magnitude of perturbation, binary variable to decide attack/non-attack at time t.

Abbreviation

DRL	Deep reinforcement learning.
DQN, DNN	Deep Q network and deep neural network.
L2RPN	Learning to run a power network.
MDP Markov	decision process.
FDI, FGSM	False data injection and fast gradient sign method.

Manuscript received July 14, 2020; revised November 4, 2020 and January 22, 2021; accepted February 17, 2021. Date of publication March 1, 2021; date of current version June 21, 2021. This work was supported in part by the Singapore National Research Foundation, through the National Cybersecurity Research and Development Program under Grant NRF2018NCR-NCR005-0001; in part by the Singapore National Research Foundation under Grant NCR NRF2018NCR-NSOE003-0001; in part by NRF Investigatorship under Grant NRFI06-2020-0022 and Grant NRF-NRFI06-2020-0001; and in part by the Ministry of Education, Republic of Singapore, under Grant AcRF TIER 1 2017-T1-001-228 (RG92/17). The work of Yan Xu was supported by the Nanyang Assistant Professorship from Nanyang Technological University, Singapore. Paper no. TSG-01088-2020. *(Corresponding author: Jianwen Sun.)*

Yan Zheng is with the School of New Media and Communication, Tianjin University, Tianjin 300384, China, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: yanzheng@tju.edu.cn).

Ziming Yan and Yan Xu are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore.

Kangjie Chen and Yang Liu are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

Jianwen Sun is with the Trustworthy Software Engineering & Open Source Software Lab, Huawei Technologies, Shenzhen 518129, China, and also with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: buaasjw@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TSG.2021.3062700.

Digital Object Identifier 10.1109/TSG.2021.3062700

1949-3053 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

EPD, EPDR	Expected performance degradation, expected performance degradation rate.
GS	Gradient saliency.
CAR	Critical attack rate, the number of critical attacks divided by the total number of attacks.

I. Introduction

Many optimization problems of power systems are becoming more and more challenging with the increased integration of intermittent renewable energies and the deregulation of electricity markets [1]. In such a context, traditional model-based operation methods are becoming insufficient. Future smart grids call for more intelligent solutions with model-free and real-time computation capabilities [2], such as deep reinforcement learning (DRL). For instance, the Learning to Run a Power Network (L2RPN) competition [3] hosted by a French power transmission company uses DRL for optimizing network topology reconfiguration. The competition winner [4] utilizes a DRL algorithm named deep Q-networks (DQN) [9]. While many DRL models have reported satisfactory performance, few studies investigate the vulnerability (i.e., the ability to withstand hostile disturbances) of these models in power systems applications, especially under the risks of potential cyber-attack [5]–[6]. In the DRL research community, prior works have practically verified that adding small perturbations in the input data may lead to drastically different control actions for DQN [8]. Considering a DRL model's behavior can be intriguing and not always predictable [7], it is risky to widely deploy DRL models in real power grids without vulnerability assessment.

In practice, the advantages of DRL techniques have been recognized by the research community, and many attempts are made to leverage DRL models in various applications for power systems. For instance, a survey on reinforcement learning methods applied in the smart grid is summarized in [2]. Meanwhile, many problems in power systems can be tackled by leveraging DRL models, including electric vehicle scheduling [11]–[12], price-based demand response [13]–[14], load frequency control [15], voltage control [16], [17] and emergency control [18], [38]–[40], and network reconfiguration [19]–[20]. However, the DRL models leverage deep neural networks (DNNs) as function approximators, which might be misled by small perturbations in state space (i.e., input data) and give an incorrect control action (i.e., output) [25], [27]. If the DRL model is misled at a certain critical time, its incorrect control action may induce hazard power systems situations. Hence, it is necessary to assess the vulnerability of the DRL models in power systems, especially under noisy data and cyber-attack.

Very limited works have investigated the vulnerability of the DRL models in the power system. References [21]–[22] study the defense mechanism in power systems, and find that the state estimator may not be able to filter data perturbations in the presence of false data injection (FDI) cyber-attack. Reference [23] indicates that cyber-attack can cause massive power outages and cascading failures. Similar to these works, DRL models can also be vulnerable to such hostile cyber-attack and data perturbations. In general, a DRL model inherits the vulnerability from the DNNs that it adopts. To analyze the vulnerability of DNNs, the fast gradient sign method (FGSM) [25]–[27] and the Jacobian saliency map algorithm [28]–[30] have been proposed by constructing adversarial input data perturbations. These works have shown the importance of perturbations in vulnerability analysis: the malfunction of DRL might be a rare event and can hardly be identified without a proper perturbation method. Inspired by these works, it is possible to numerically find the common characteristics of data noises that induce the malfunction of DRL. In this paper, to assess the vulnerability of DRL models for power systems, a criticality-based perturbation model and several vulnerabilities indices are proposed. It's worth emphasizing that this paper aims to find the vulnerabilities of DRL models in advance to ensure the safe operation of power grids. Therefore, the target user of the proposed approach is the grid operators, who already know the parameters of DRL models and try to avoid potential failures inside the DRL model. Besides, we focus on a typical topology optimization problem (L2RPN challenge [3]) where incorrect decisions may induce branch overflows, voltage violations, and even blackouts [31].

This paper assesses the vulnerability of DRL models in the power system topology optimization problem by analyzing the characteristics of data perturbation (attack) that may cause malfunction of DRL. The main contributions of this paper are two-fold:

- To find the characteristics of perturbations that can induce DRL malfunction, a criticality-based adversarial perturbation model is proposed. This model constructs perturbations by minimizing expected DRL rewards (i.e., the objective function in power systems) but constrains the noise magnitude within a reasonable range. Besides, a criticality-based timing selection method is introduced to deploy the perturbations (proceed attack) at the moment when the power system can be most affected. After deploying the model, the post-perturbation power system status is evaluated to verify whether the DRL model has serious vulnerabilities (e.g., system crashing).
- Vulnerability indices based on the probability and gradient criteria are proposed. The indices aim to discover the scenarios and adversarial perturbations under which even small data perturbations could cause severe performance degradation of DRL models in power system topology optimizations (e.g., divergence of power flow solution). Furthermore, characteristics of perturbations and status of power systems under which DRL malfunction may happen are analyzed.

The rest of the paper is organized as follows. Section II describes the problem formulation, and the criticality-based perturbations model is presented in Section III. The vulnerability assessment indices are presented in Section IV. Simulation results and numerical analysis are shown in Section V, followed by the conclusions in Section VI.

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

II. Problem Descriptions

Many power system optimization problems are complex and non-linear and can hardly be computed in real-time. DRL has been successfully applied in many scenarios [46]–[49] and sheds a promising light to tackle these problems. However, DRL models may be vulnerable to data perturbation and cyber-attack. Hence, this paper studies the DRL model's vulnerabilities for power system topology optimization [3], which is of practical importance before deploying a DRL model in real systems.

A. Network Topology Optimization

For the power system topology optimization problem [3], transmission lines or power equipment will be disconnected if thermal limits are breached. After disconnection, the power flow will be rerouted to the new path (based on the resistance), but this rerouting may result in another line or lines being overloaded. Loads of each node are based on time-series data, and the generations have been pre-scheduled based on predicted loads. As set by the competition [3], contingency/hazards may randomly happen during operation.

The objective of network reconfiguration is to maximize the remaining transfer capabilities (sometimes after contingencies). Therefore, based on transmission line usage, the objective of topology reconfiguration is the summation of (percentage) remaining transmission capacity as [4, eq. (1a)], and the status of power systems after topology reconfiguration is solved by power flow [shown in Eq. (1b)-(1c)]:

$$Maximize \quad \sum_{i}^{N}\left[1-\left(S_{Li}/S_{Lm,i}\right)^{2}\right] \tag{1a}$$

$$\text{s.t.} \quad P_{Gi}-P_{Di}=V_{i}\sum_{j=1}^{n}V_{j}\left(G_{ij}cos\delta_{ij}+B_{ij}sin\delta_{ij}\right) \tag{1b}$$

$$Q_{Gi}-Q_{Di}=V_{i}\sum_{j=1}^{n}V_{j}\left(G_{ij}sin\delta_{ij}-B_{ij}cos\delta_{ij}\right) \tag{1c}$$

where N represents the total number of transmission lines. It is worth mentioning that the competition [20] does not require the satisfaction of inequality constraints and will disconnect the lines if thermal limits are violated.

B. DRL for Topology Optimization

To attain the optimization objective in the power system topology optimization problem, the decision-maker can reroute power flows by switching lines, splitting/coupling busbars at substations, and disconnecting the load. The sequential decision-making process of power flows rerouting can be modeled as the Markov decision process (MDP) [4], where the agent (i.e., topology controller) needs to interact with the environment (i.e., power systems) according to a certain policy π to achieve a certain goal (e.g., optimization objectives). The MDP is formulated as a tuple (S, A, P, r, γ), where the state space S includes active power outputs and voltage setpoints of generators, loads, line status, line flows, thermal limits and timestamps; the action space A includes different topological decisions (i.e., line switching, node splitting/rejoining). The transition probability P decides which state the agent will arrive after taking actions. r denotes the reward function and $\gamma \in [0, 1)$ is the discount factor.

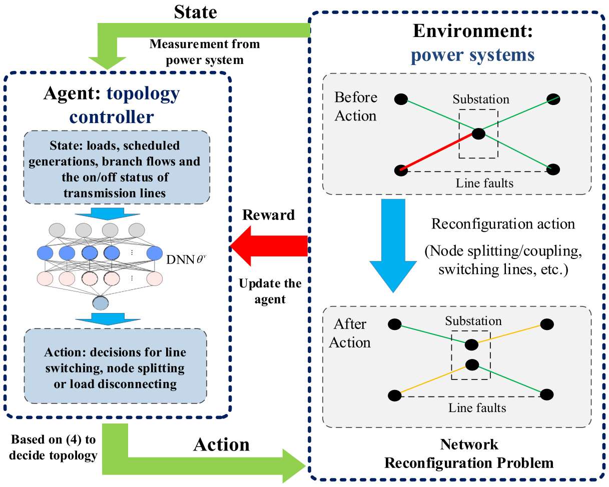

State
Measurement from power system
Agent: topology controller
State: loads, scheduled generations, branch flows and the on/off status of transmission lines
DNN $\theta^{\prime\prime}$
Action: decisions for line switching, node splitting or load disconnecting
Reward
Update the agent
Environment: power systems
Before Action
Substation
Line faults
Reconfiguration action
(Node splitting/coupling, switching lines, etc.)
After Action
Substation
Line faults
Network Reconfiguration Problem
Based on (4) to decide topology
Action

Fig. 1. Power system topology optimization problem for investigating the vulnerability of DRL models against cyber-attack.

The overall MDP process of the topology optimization problem is shown in Fig. 1. At each time slot t, the agent observes the state s_t. Based on s_t, the agent can select an action a_t to interact with the environment according to its policy $a_t \sim \pi(s_t)$, and receive an immediate reward r_t which is defined according to the optimization goal.

In order to penalize the divergence of power flow and satisfy transmission lines power limits, the reward function $r(s, a)$ is defined as follows:

$$r(s,a)=\begin{cases}\text{Penalty}(-r_e), & \text{power flow diverges}\\ \frac{1}{N}\sum_{i}^{N}max\left[0,1-\left(S_{Li}/S_{Lm,i}\right)^{2}\right], & \text{otherwise}\end{cases} \tag{2}$$

where the constant r_{e} is for penalizing power flow divergence (failure of topology reconfiguration), and r_t will be negative if power flow solution diverges.

The agent aims to select an optimal topology control action at each given state, such that the power network can stay safe under changing loads and contingencies. In general, the agent uses a policy to interact with the environment and received a cumulative reward $\sum_{t=0}^{T}\gamma^{t}r_{t}$. There exists an optimal policy π^*, by following which, the agent can achieve the maximum cumulative reward $\mathbb{E}_{a\sim\pi^*}\sum_{t=0}^{T}\gamma^{t}r_{t}$.

To obtain the optimal policy π^*(neural network parameterized by θ), we employ a standard DRL algorithm DQN [4], where the action-value (i.e., Q-value) function is defined as: $Q^{\pi}(s,a)=\mathbb{E}[\sum_{i=0}^{T-1}\gamma^{i}r_{t+i}]$. During the learning process, the Q-values is updated according to Eq. (3), and the network is trained with the experience replay by minimizing the mean squared error [shown in Eq. (4)]:

$$\begin{aligned}Q^{\pi}(s,a) \leftarrow\ & Q^{\pi}(s,a)+\eta\\ & \times\left[r+\gamma * \max Q^{\pi}\left(s',a'\right)-Q^{\pi}(s,a)\right]\end{aligned} \tag{3}$$

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

$$\mathcal{L}(\theta) = \sum_{i}\Big[\big(r_t + \gamma \max Q^{\pi}\left(s_{t+1}, a_{t+1}; \theta^{-}\right) - Q^{\pi}\left(s_t, a_t; \theta\right)\big)^2\Big] \tag{4}$$

Once the training is finished, the agent learns the optimal policy π^* and chooses actions a_t according to the π^* as:

$$a_t = argmax \;\; Q^{\pi^*}(s_t|\theta), \tag{5}$$

where the optimal action a_t can achieve the highest Q-values.

C. Perturbation-Based DRL Vulnerability Assessment

DRL has been shown as an effective solution for many problems in power systems [11]–[20], and the performance of DRL gives credit to the ability of DNNs. However, DNNs are known to be vulnerable to adversarial perturbations [25]–[30] (e.g., adding small but carefully crafted perturbations in the input may totally mislead the network to give an incorrect output with high confidence). Similar vulnerabilities may also exist in DRL models [8] that are built on DNNs. To spot these defects, adversarial attacks against the DRL are proposed, which, however, draw little attention.

For instance, [29] firstly attempts to attack DRL models by adding perturbations at every timestamp. Reference [36] generates adversarial perturbations to mislead agents to take the worst action. Another important facet of attacking DRL models is how to select appropriate attacking moments [41]. Instead of attacking at every step, [37] proposes to inject perturbations every N frames. Further, [33] proposes to compute the preference of an agent taking the optimal action for better-deploying perturbations. Apart from these, the transferability of adversarial perturbations across different DRL models are studied in [34]. The robustness and resilience of DRL models against attacks are investigated in [35].

Overall, the adversarial perturbations could be any kind of disturbances on the DRL state space s in power systems (e.g., due to cyber-attack, data noise or facilities malfunctions). Formally, the perturbations can be considered as additional noises on the input of DRL agents, and the post-perturbation states s'_t and DRL decisions a'_t can be denoted by:

$$s'_t = s_t + \lambda \tag{6}$$

$$a'_t = argmax \;\; Q^{\pi}\left(s'_t|\theta\right) \tag{7}$$

where the small perturbation λ is calculated based on a certain objective, and s'_t and a'_t represent the perturbed state and the corresponding DRL's action, respectively. It is possible to find λ such that $a' \neq a$ (i.e., the agent chooses a different action). Given a perturbed state s'_t, the DRL model may provide a wrong control action a' and induce danger in power systems. Intuitively, a DRL model is vulnerable if such perturbations can easily mislead it.

To recognize and assess the vulnerability of DRL models in power systems, we aim to find the conditions of DRL malfunction and find proper indices to analyze the conditions of DRL malfunction accordingly. In subsequent sections, we study: (*i*) constructing proper adversarial perturbation λ, and (*ii*) evaluating DRL vulnerability and analyzing DRL malfunction conditions. In short, in the former stage, the observation (i.e., measured power systems data) from the environment (i.e., power systems) is fed into the adversarial attack model, which launches a small perturbation on the observation based on the criticality of the power system status. After deploying perturbations, the DRL agent may change its control action, which might eventually lead to hazard situations of power systems operation. In the later stage, vulnerability indices are proposed to evaluate the post perturbation system performance. Moreover, according to the indices values, we analyze the entire system (e.g., the overall performance of DRL models, and the states s_t of power systems under which DRL can be easily misled).

III. CRITICALITY-BASED ADVERSARIAL PERTURBATION

This section firstly formalizes the vulnerability assessment problem of DRL-based controllers in the power system caused by the data perturbations. Then, we propose the criticality-based adversarial perturbation model to address the problem from two perspectives: 1) using a gradient-based method to construct the perturbations and 2) determining critical moments based on the agent's action preference values for deploying perturbations.

A. Perturbation Modelling

Three challenges are addressed to effectively construct the perturbation λ defined in E.q. (6). Firstly, the perturbation should be small enough. As indicated in [24], the AC state estimation cannot properly detect false data injection attacks with 5% to 10% perturbations on state variables. Consequently, the perturbation λ can be bounded to ±10% to find the DRL models' vulnerabilities more effectively. Since the state estimation and the measurement filter can eliminate abnormal data input under higher disturbances, it is unnecessary to analyze DRL vulnerabilities for a wide range of perturbations. Secondly, instead of merely misleading the decision such that $a' \neq a$, certain perturbation objective shall be satisfied, i.e., the performance of DRL can be degraded. Thirdly, the perturbation shall be rare, otherwise, frequent input distortions can be easily detected by the protection mechanism. Therefore, effective perturbations shall be constructed considering the objective function of DRL and a certain timing strategy.

Motivated by the above challenges, this paper proposes a perturbation model for degrading the DRL performance using only a few attacking steps. Specifically, the DRL model uses a policy $\pi_{\theta}(s)$, a non-linear DNN parameterized by θ, to control power systems. Crafting an adversarial perturbation and deploying it to mislead the DRL model to make a wrong action can be formalized as follows:

$$\|\lambda\|_p = \epsilon \quad s.t. \;\; \pi_{\theta}(s) \neq \pi_{\theta}(s+\lambda), \tag{8}$$

where $\|\cdot\|_p$ represents the p-norm, and $\|\lambda\|_p$ measures the difference between the original and the perturbed states. It is worth mentioning that, given a DRL model, we assume the model is fixed and its parameters θ can not be changed by the malicious attacker.

To interfere with the DRL model during the entire system controlling the process, the attackers need to find a perturbation strategy to perturb the agent's observation to minimize

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

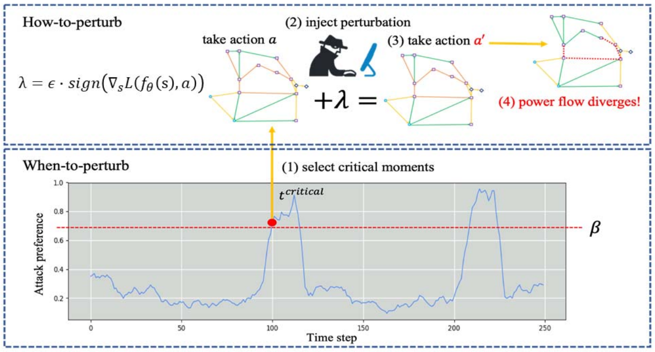


Fig. 2. Illustration of criticality-based adversarial perturbation.

the following accumulated reward R_1:

$$R_1 = \sum_{t=0}^{T-1} \mathbb{E}_{a \sim \pi(s_t + k_t \lambda_t)} \gamma^t r_t \tag{9}$$

where the immediate reward r_t is defined in Eq. (2). Note that, the DRL model controls the system according to its policy $\pi(s_t + k_t\lambda_t)$ even under data perturbations. At time step t, the attack strategy needs to decide whether deploying the perturbation $\lambda_t (k_t = 1)$ or not $(k_t = 0)$. To further improve attacking efficacy and avoid being detected, the attackers wish to achieve a stealthy attack by reducing the total number of deploying perturbations $\sum_{t=0}^{T-1} k_t$.

Therefore, finding an optimal attack strategy $\{k_0, \dots, k_{T-1}\}$ and corresponding perturbations $\{\lambda_0, \dots, \lambda_{T-1}\}$ can be regarded as an optimization problem formulated as follow:

$$\min_{k_0,\dots,k_{T-1},\lambda_0,\dots,\lambda_{T-1}} \quad \sum_{t=0}^{T-1} \mathbb{E}_{a \sim \pi(s_t + k_t \lambda_t)} \gamma^t r_t \tag{10a}$$

$$s.t. \quad \|\lambda_t\|_p = \epsilon \text{ for all } t = 0, \dots, T-1 \tag{10b}$$

$$\sum_{t=0}^{T-1} k_t \leq N \tag{10c}$$

where $\| \cdot \|_p$ represents the p-norm distance between original input s_t and perturbed input $s_t^{adv} = s_t + \lambda_t$.

It is difficult to directly solve the optimization problem in (10) as it involves numerous variables. Instead, we convert the problem by answering two sub-problems: ***when-to-perturb*** and ***how-to-perturb***. The when-to-perturb task aims at selecting an appropriate timing to inject perturbations on DRL input, i.e., deciding whether k_t equals 1 or 0. The how-to-perturb task focuses on finding an appropriate adversarial input s_t^{adv} to change a DRL model's behavior.

Fig. 2 depicts the process of employing criticality-based perturbation to compromise the DRL models in power systems. We calculate the preference of the trained agent in taking the most preferred action over the other actions at the current state. At the bottom of this figure, we plot the preference value of each step. The adversary first selects the critical moment t^{critical} by finding states with large preference values (larger than threshold β). Then, the adversarial perturbation λ is injected into the observation of the agent and the agent changes its action from a to a'. Eventually, the wrong action may lead to power flow divergence of the power gird. It's worth mentioning that β controls the attacking frequency, which should not be too high to be inefficient and unrealistic in real power grids, nor too low to be ineffective in discovering potential vulnerabilities. Therefore, we adopt a hyperparameter optimization approach, called grid search, to select an appropriate threshold.

B. When-to-Perturb: Criticality-Based Timing Selection

To solve the when-to-perturb problem, we introduce the attacker-preference value to measure each state's criticality [33]. If the DRL model is particularly inclined to select a specific action comparing to the others at time t, taking this action can lead to a higher future return. The attacker-preference value $p(s_t)$ is defined as:

$$p(s_t) = max_{a_t \in A} \pi(s_t|a_t) - max_{a_t \in A \setminus \{a^*\}} \pi(s_t|a_t) \tag{11}$$

where the policy network π maps state-action pair to a probability. A is the set of all possible actions, and $a^* = max_{a_t \in A} \pi(s_t|a_t)$ is the optimal action with the highest future return. The $p(s_t)$ value calculates the gap of the expected returns between the optimal action and the other actions. Intuitively, large p-value means that the agent strongly prefers the optimal action over the other and can yield large returns by choosing this action. To find the potentially dangerous perturbations, it is more appropriate and profitable to mislead the DQN agent not to take desired actions at moments with large p-values, so that the expected rewards in E.q. (3) will be greatly reduced.

The proposed method can be applied for standard DRL algorithms with a finite action space. For policy-based methods, the action probability in policy distribution can be used to indicate the preference when the agent chooses its action. If the action probability for a specific action is relatively high, it means that the agent strongly prefers this action and it is critical to take this action. Thus, $p(s_t)$ for policy-based methods is defined as (11). For value-based methods, such as DQN, we can convert the Q-values of actions into a probability distribution over actions using the softmax function. After that, the attacker-preference value $p(s_t)$ can be calculated following (11).

C. How-to-Perturb: Gradient-Based Perturbation

To address the how-to-perturb problem, the fast gradient sign method (FGSM) is leveraged to perturb the original input state s as follows:

$$s^{adv} = s + \epsilon \cdot sign\big(\nabla_s L\big(\pi_\theta(s), \overrightarrow{a}\big)\big) \tag{12}$$

where $\overrightarrow{a}$ is a one-hot vector representing the optimal action according to the policy $\pi_\theta(s)$. The cross-entropy function is adopted as the loss function L. $\nabla_s L(f_\theta(s), a)$ means the gradient with respect to s. Hence, the perturbation added to the original input s is obtained by multiplying ϵ with the *sign* $(+1$ or $-1)$ of the gradient $\nabla_s L(f_\theta(s), a)$. Intuitively, the original input is updated along the direction of the gradient's sign at each time step, which will maximize the loss function L and make a DRL model behave incorrectly (e.g., take wrong actions but $\overrightarrow{a}$).

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1 Criticality-Based Perturbation Algorithms for Assessing Vulnerability of DRL Models in Power Systems

```
Input: Trained DQN model parameterized f with $\theta$
       Perturbation magnitude $\epsilon$
       Magnitude of perturbation distance $p$
       The attack threshold $\beta$ for action preference value
Output: Adversarial perturbations $\{\lambda_1, \lambda_2, \dots, \lambda_L\}$.
1  Initialize: the random seed of the simulator
2  begin
3    for each step $t$ and the current episode not terminated do
4      # Calculate the action preference of the agent
5      $p(s_t) = max_{a_t \in A} \pi(s_t|a_t) - max_{a_t \in A \setminus \{a^*\}} \pi(s_t|a_t)$
6      # Decide whether to choose this moment to attack
7      if $p(s_t) > \beta$ then
8        # Generate the adversarial perturbation $\lambda_t$
9        Input $s_t$ and the gradient $\mathrm{g_t} = \nabla_{\mathrm{x}} L(f_\theta(s_t), a)$
10       $\lambda_t = \epsilon \cdot sign(\mathrm{g_t})$
11       Obtain the adversarial instances $s'_t = s_t + \lambda_t$
12       # Select action with the adversarial observation
13       $a_t = max\ \pi(s'_t|\theta)$
14     else
15       # Choose not to attack at step $t$
16       $a_t = max\ \pi(s_t|\theta)$
17     Perform $a_t$ and receive $r_{t+1}$ and $s_{t+1}$
18 end
```

D. Implementation Details

Algorithm 1 describes the pseudo-code of the criticality-based adversarial perturbation model to construct effective perturbations. The algorithm assumes that the parameters of DRL models are already known from the perspective of the power system operators. At each time t, the when-to-perturb problem is solved based on E.q. (12) (line 3) and thereby the adversarial example is crafted (lines 6 to 9) following E.q. (11). The algorithm first calculates the action preference value $p(s_t)$ to make the perturbation decision. If the action preference value is larger than the threshold β (line 5), the perturbation decision is made. Then, based on the gradient $\mathrm{g_t}$ with respect to the input s_t, the perturbation λ_t alongside the direction of the gradient is calculated. This method allows us to find perturbations that leverage a trained DRL controller's output to mislead it into the wrong action.

IV. Vulnerability Indices

The performance of DRL under data perturbations depends on many factors, including the defense mechanisms (e.g., state estimation), attack/perturbation severity, and state of power systems when attack/perturbation happens. To evaluate the vulnerability of a DRL model in power systems, the indices can be based on the characteristics of neural networks, including approximation errors, overfitting [32], lack of fitting (i.e., statistical measurement), and local optimum (i.e., due to solution of gradient-based learning algorithms). Hence, the vulnerability indices to evaluate the DRL-based power system controller can be investigated accordingly with probability-based criteria and gradient-based criteria.

The overall vulnerability of DRL in power systems can be assessed with probability-based criteria. Let N_v denote the number of vulnerable cases (divergence of power flow equations) and N_t denote the number of total test cases. It comes naturally that failure rate N_v/N_t, i.e., the percentage of diverged power flow solutions, provides a direct evaluation index of the global vulnerability for DRL-based control methods. Besides, a properly designed reward function can represent both economic performance and operation constraints compliance of the optimization model. For the generality of evaluation, this paper presents the expected performance decay (EPD) and expected performance decay rate (EPDR) for control against data perturbations:

$$EPD = \frac{1}{M}\sum_{i}^{M} \pi_i(s'_i)\left[R_i(s_i|\theta^v) - R'_i(s'_i|\theta^v)\right] \tag{13}$$

$$EPDR = \frac{1}{M}\sum_{i}^{M} \pi_i(s'_i)\left[1 - R'_i(s'_i|\theta^v)/R_i(s_i|\theta^v)\right] \tag{14}$$

where $\pi_i(s'_i)$ represents the probability of i-th abnormal state s'_i to happen, and $R_i(s_i|\theta^v)$ and $R'_i(s'_i|\theta^v)$ represent the control rewards for environment state before and after perturbations, respectively.

The operational vulnerability of DRL can be assessed with gradient-based criteria. Inspired by [25], [33], the gradient of objective function L with respect to state variable x_i can represent the sensitivity of neural networks to perturbations under certain states. This paper investigates the gradient saliency (GS) of perturbation to assess the operational vulnerability of DRL based control:

$$GS(s_t) = \frac{1}{N}\sum_{i}^{N}\left|\frac{\partial L(f_\theta(x), a)}{\partial x_i}\right| \tag{15}$$

where $L(f_\theta(x), a)$ represents the DRL training objective function, and x_i represents the i-th state variable.

Besides, inspired by [33], the importance of control actions can be quantified by the probability difference of two actions to be chosen. This paper employs a p-function to evaluate the importance of DQN control action and quantify the risks of being misled by DRL agent at a given state s_t:

$$p(s_t) = max_{a_t \in A}\pi(s_t|a_t) - max_{a_t \in A\setminus\{a^*\}}\pi(s_t|a_t) \tag{16}$$

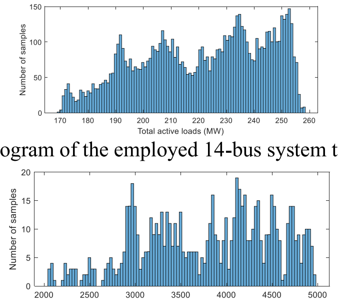
where A is the action set, and $a^* = max_{a_t \in A}\pi(s_t|a_t)$ is the optimal action with the highest future return. The larger p-function value represents that the action a_t is more critical to the power systems.

V. Simulation Results

A. Simulation Environment and System Information

The simulation is based on the simulation platform of Learning to Run a Power Network (L2RPN) competition [3] hosted by RTE, French power transmission company, and Chalearn, RTE's partner. The competition tests different reinforcement learning methods to control electricity transportation in power grids. The winner of the competition (Geirina team [4]) proposes a DRL-based grid control method with deep Q networks. The reward of DRL is computed for each scenario and control action based on economic indexes of transmission line usages [3], and the score will be 0 if

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.



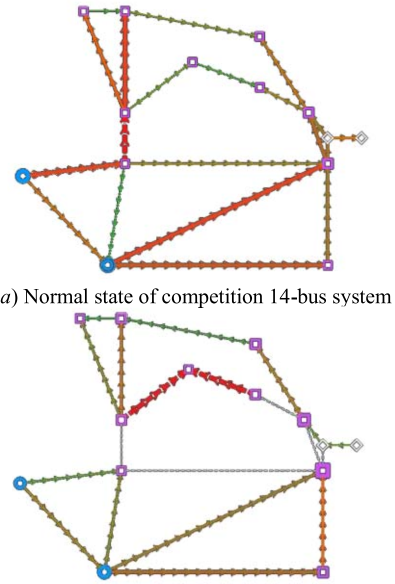
(*a*) Histogram of the employed 14-bus system total load

(*b*) Histogram of the employed IEEE 118-bus system total load

Fig. 3. Histogram of employed load profile in two case studies. (*a*) Histogram of the employed 14-bus system total load. (*b*) Histogram of the employed IEEE 118-bus system total load.

TABLE I
PERFORMANCE OF 14-BUS SYSTEM AGAINST FDI ATTACK AND PERTURBATIONS

Method	CAR = N_v/N_t	EPDR = EPD / A.S	EPDR (*C*/Non-*C*)	A.T
No attack	0% = 0/0	0% = 0 / 3347.1	0 / 0	0
Random Noise	0.021% = 5/23709	4.6% = 823 / 3157.7	4.6% / 0%	0.2ms
FGSM (every step)	0.101% = 20/19817	32.1% = 1073.2 / 2273.9	31.4% / 0.7%	2.2ms
Critical attack (chosen step)	0.386% = 18/4658	32.1% = 1072.6 / 2274.5	31.3% / 0.8%	3.1ms



(*a*) Normal state of competition 14-bus system

(*b*) State of competition 14-bus system under attack

Fig. 4. Topology decisions based on DRL (a) before and (b) after deploying perturbations (attacks) for the IEEE 14-bus system at time 18:55 of 3, Jan.

the solution of power flow diverges after the control decision (crash of control).

Two case studies, including the competition 14-bus system and the IEEE 118-bus system, are carried out to testify the effectiveness of the proposed vulnerability assessment method for DRL-based power system control models. For the 14-bus system and 118-bus system, the load profile including active load, reactive load, scheduled generation and voltage is provided in [3]. For statistical purposes, the histograms of the employed load data for both systems are provided in Fig. 3. Besides, the transmission line outages due to random hazards and maintenance are also considered for the test systems. The thermal limits to trigger transmission line protection are set as 1000MW. The inputs of the DRL agent have 538 variables for the competition 14-bus system and 4,967 variables for the IEEE 118-bus system. The output of the DRL agent is the power grids topology decisions as described in part B of Section II.

B. Simulation Results

Case 1 (Competition 14-Bus System): To testify the effectiveness of our vulnerability assessment method, the performance of DRL models in power systems is verified against random noise perturbations, FGSM adversarial attack [25] and the proposed criticality-based perturbations. The FGSM adversarial attack employs perturbations on every time steps of an episode. The criticality-based perturbation attack only launches attacks at specific states, which can mostly diminish the agent's cumulative score. The simulations numerically compare the performance of different attack methods (to find the vulnerability of DRL) with 500 scenarios in Table I. The magnitude of perturbation noise ϵ of attacks is set to 0.1. The random perturbation employs normal distribution noise with the same perturbation magnitude.

To visualize the impacts of DRL-based power grids control malfunction under the FDI attack, Fig. 4(a) shows the normal state on competition 14-bus system at dispatch time step 227 (3, Jan, 18:55). In Fig. 4, each square represents a substation, which has a double busbar layout. The width and color of lines in Fig. 4 represent the flows through the transmission lines, where red color represents overflow and white color represents disconnection. Without data perturbations, the system can normally run even though one bus is overloaded. However, in Fig. 4(b), there is one node isolated since two buses are overload simultaneously after the criticality-based perturbation attack. Therefore, the whole system cannot guarantee to provide power to every node and power flow solution diverges. Thus, the simulation environment will break and the game is over.

The numerical results and vulnerability indices are shown in Table I. As shown in Table I, all the perturbation methods may lead to crashes of power gird and power flow divergence. Under the same noise magnitude, the performance DRL-based power system control is obviously degraded under the existence of an FGSM attack (rewards reduced by 32.1%) and the criticality-based perturbation attack (rewards reduced by 32.1%). The critical attack rate (CAR) (N_v/N_t) that leads to power flow divergence is only 0.101% (20/19817) for FSGM attack but 0.386% (18/4658) for the criticality-based perturbation attack. The proposed vulnerability assessment model execute much fewer attacks (4658 vs. 19817) than FGSM, and can more effectively discover the vulnerabilities of DRL models under attacks. Besides, the vulnerability indices and power flow of transmission line #1 with respect to time are shown in Fig. 5 (10 scenarios with $\epsilon = 0.2$ and

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

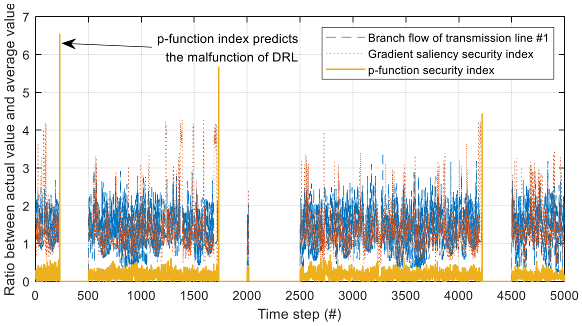
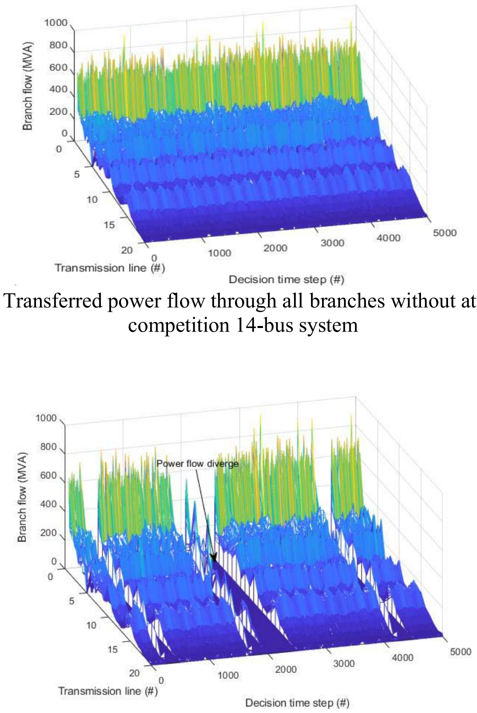


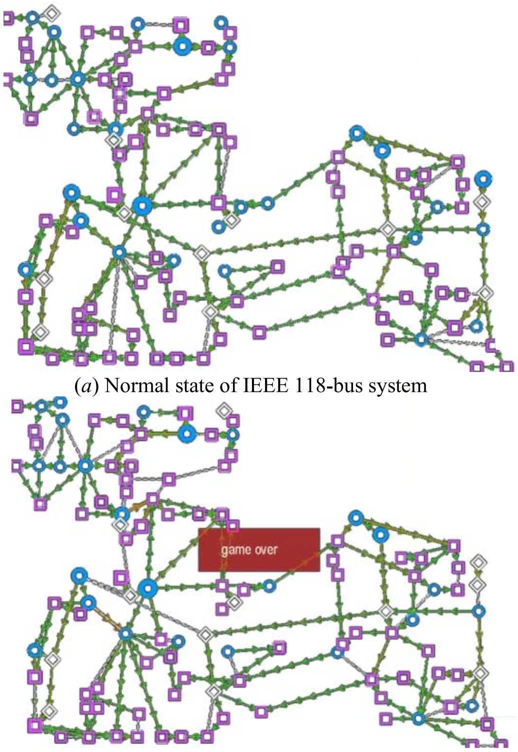
Fig. 5. Value of vulnerability indices with respect to time under FGSM attack (actual index value divided by average value for normalization).



(a) Transferred power flow through all branches without attack in competition 14-bus system

(b) Transferred power flow through all branches against FGSM-based timing false data injection in competition 14 bus system

Fig. 6. DRL-based control branch flows results with respect to time in competition 14-bus system against FGSM attack.



(*a*) Normal state of IEEE 118-bus system

(*b*) State of IEEE 118-bus system under attack

Fig. 7. Topology decisions made by the DRL model (a) before and (b) after deploying perturbations (attacks) for the IEEE 118-bus system.

$\beta = 0.2$) to investigate the characteristics of DRL malfunction. Fig. 6 provides the branch flows for all transmission lines during a simulation. As can be seen from Figs. 5 and 6, the preference function vulnerability indices successfully predict the malfunction of DRL. The value of action preference tends to be high before the failure of the DRL-based power system controller under data perturbation or attack. However, the gradient saliency index is not so distinguishing before the malfunction of DRL. To compare the severeness of the consequences of attacks, the average score (A.S) of operation, the EPD due to critical attack (C) that leads to power flow diverge and non-critical attacks (Non-C) is also compared. It is shown that the rewards degradation is much more severe if the attacks are critical (31.3% rewards degradation due to critical attacks and only 0.8% rewards degradation due to non-critical attack). The complexity of crafting adversarial perturbations is compared by the average time to craft attack signals. Since craft criticality-based perturbation only needs to calculate gradient and preference function, the average time consumption (A.T) of the proposed method is 3.1ms, similar to random noise FGSM attack.

Case 2 (IEEE 118-Bus System): For IEEE 118-bus systems, random line outages due to contingency are employed and the DRL-based controller cannot find a secure decision for some certain states. The action reduction is based on random selections of valid line switching and node splitting. Random perturbations with normal distribution, FGSM adversarial attack, and the criticality-based perturbation attack are employed to investigate the vulnerability of DRL-based control. Since a contingency is less likely to happen for this test system, the magnitude of perturbation noise ϵ for random state variation and FGSM attack is set as $\epsilon = 0.2$. Besides, we employ the grid search approach and set threshold $\beta = 0.15$. The comparison of normal operation state and operation state under data perturbations/attack on the IEEE 118-bus system is shown in Fig. 7(a) and Fig. 7(b). Similar to case 1, it can also be observed that loads might be isolated, and the power flow solution will diverge under attack with a small perturbation magnitude.

As shown in Table II, the performance of the DRL-based power system controller is significantly reduced under the existence of FGSM and the criticality-based perturbation attack under almost the same noise magnitude. Fortunately, the degradation of performance appears to be less compared with the smaller 14-bus system. Regarding the global

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

TABLE II
PERFORMANCE OF 118-BUS SYSTEM AGAINST FDI ATTACK AND PERTURBATIONS

Method	CAR = N_v/N_t	EPDR = EPD / A.S	EPDR (*C*/Non-*C*)	A.T
No attack	0% = 0/0	0% = 0 / 17900	0 / 0	0
Random Noise	0% = 0/900	4.6% = 823 / 17086	4.6% / 0%	1.3ms
FGSM (every step)	0.1% = 1/899	8.3% = 1479 / 16430	7.7% / 0.6%	4.8ms
Critical attack (chosen step)	0.2% = 1/507	8.9% = 1591 / 16318	7.1% / 1.8%	5.3ms

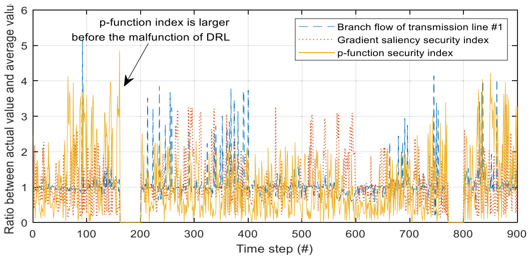


Fig. 8. Value of vulnerability indices with respect to time under FGSM attack (actual index value divided by average value for normalization).

index, compared to FGSM, the proposed method achieves a better-expected performance decay rate (EPDR = 8.9%) and the critical attack rate (CAR = 0.2%). If employing random noise as the data perturbation, the negative impacts on the IEEE 118-bus system controlled by the DRL model are nearly negligible (4.6% EPDR and no other power flow divergence of power systems). It is worth mentioning that, despite triggering the same number of system divergence, the proposed method uses only half of the total test cases that FGSM used (507 vs. 899), performing a better efficiency in discovering vulnerabilities.

Besides, as can be seen from Fig. 8 and Fig. 9, before a malfunction of DRL-based controllers occurs, the action preference index (*p*-function) has a higher value than normal states. The gradient saliency may not be effective enough to represent the operational vulnerability of DRL-based models in power systems control. This phenomenon can be explained by the fact that misleading DRL is sometimes useless because power systems at certain states can be robust enough even under wrong control action. Hence, it is shown that the value of action preference can be an effective index for assessing the vulnerability of DRL-based models in power systems operation and control, and the proposed vulnerability assessment model can more effectively discover critical attacks.

VI. CONCLUSION

This paper proposes a vulnerability assessment method for DRL models in power systems topology optimization. The presented approach aims to find the vulnerabilities of DRL models and enable the power grids operators to identify or reduce the security risks before practically applying DRL models. A criticality-based adversarial perturbation model is proposed to identify perturbation characteristics that may cause hazard situations in power systems. Adversarial perturbations are constructed by minimizing DRL rewards within minimum noise magnitude, and an attacker-preference value determines the timing of deploying perturbation. Several indices based on probability and gradient criteria are proposed to measure the DRL models' overall performance and operational vulnerability in power systems. With the proposed adversarial perturbation method and vulnerability indices, the conditions of DRL malfunction are found, and the consequences in power systems are analyzed. Simulations based on the L2RPN competition 14-bus system and IEEE 118-bus system show that the DRL-based power system control approaches can be vulnerable to false data injection attacks and uncertain data perturbation. Besides, it is shown that the *p*-function value can be an effective index to evaluate the operational risks of DRL models in power systems against the attack and data perturbation. In the future, we would like to study how to protect the power grid from dynamic attackers switching among multiples attacking policies [42]–[45].

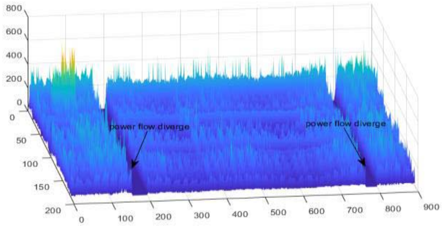


Fig. 9. Transferred power flow through all branches against FGSM-based timing false data injection attack in IEEE 118-bus system.

REFERENCES

[1] K. Moslehi and R. Kumar, "A reliability perspective of the smart grid," *IEEE Trans. Smart Grid*, vol. 1, no. 1, pp. 57–64, Jun. 2010.

[2] D. Zhang, X. Han, and C. Deng, "Review on the research and practice of deep learning and reinforcement learning in smart grids," *CSEE J. Power Energy Syst.*, vol. 4, no. 3, pp. 362–370, Sep. 2018.

[3] *Apogee Project Team, Learning to Run a Power Network Challenge, Réseau de Transport d'Électricité*. Accessed: Mar. 1, 2020. [Online]. Available: https://l2rpn.chalearn.org/

[4] *Geirina Team, Deep Reinforcement Learning to Run Power Networks With Deep Q Networks, GEIRI North America*. Accessed: Mar. 1, 2020. [Online]. Available: https://github.com/shidi1985/L2RPN

[5] Y. Xiang, L. Wang, and N. Liu, "A robustness-oriented power grid operation strategy considering attacks," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4248–4261, Sep. 2018.

[6] G. Liang, J. Zhao, F. Luo, S. R. Weller, and Z. Y. Dong, "A review of false data injection attacks against modern power systems," *IEEE Trans. Smart Grid*, vol. 8, no. 4, pp. 1630–1638, Jul. 2017.

[7] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013. [Online]. Available: arXiv:1312.6199.

[8] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," in *Proc. Int. Conf. Mach. Learn. Data Min. Pattern Recognit.*, 2017, pp. 262–275.

[9] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[10] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015. [Online]. Available: arXiv:1509.02971.

[11] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 10, no. 5, pp. 5246–5257, Sep. 2019.

[12] H. Li, Z. Wan, and H. He, "Constrained EV charging scheduling based on safe deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, May 2020.

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

[13] H. Xu, H. Sun, D. Nikovski, S. Kitamura, K. Mori, and H. Hashimoto, "Deep reinforcement learning for joint bidding and pricing of load serving entity," *IEEE Trans. Smart Grid*, vol. 10, no. 6, pp. 6366–6375, Nov. 2019.

[14] X. Xu, Y. Jia, Y. Xu, Z. Xu, S. Chai, and C. S. Lai, "A multi-agent reinforcement learning based data-driven method for home energy management," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3201–3211, Jul. 2020.

[15] Z. Yan and Y. Xu, "Data-driven load frequency control for stochastic power systems: A deep reinforcement learning method with continuous action search," *IEEE Trans. Power Syst.*, vol. 34, no. 2, pp. 1653–1656, Mar. 2019.

[16] J. Duan *et al.*, "Deep-reinforcement-learning-based autonomous voltage control for power grid operations," *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 814–817, Jan. 2020.

[17] W. Wang, N. Yu, Y. Gao, and J. Shi, "Safe off-policy deep reinforcement learning algorithm for volt-VAR control in power distribution systems," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3008–3018, Jul. 2020.

[18] Q. Huang, R. Huang, W. Hao, J. Tan, R. Fan, and Z. Huang, "Adaptive power system emergency control using deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1171–1182, Mar. 2020.

[19] Y. Gao, W. Wang, J. Shi, and N. Yu, "Batch-constrained reinforcement learning for dynamic distribution network reconfiguration," *IEEE Trans. Smart Grid*, vol. 11, no. 6, pp. 5357–5369, Nov. 2020.

[20] A. Marot *et al.*, "Learning to run a power network challenge for training topology controllers," 2019. [Online]. Available: arXiv: 1912.04211.

[21] H. He and J. Yan, "Cyber-physical attacks and defences in the smart grid: A survey," *IET Cyber Phys. Syst. Theory Appl.*, vol. 1, no. 1, pp. 13–27, 2016.

[22] G. Liang, S. R. Weller, J. Zhao, F. Luo, and Z. Y. Dong, "The 2015 Ukraine blackout: Implications for false data injection attacks," *IEEE Trans. Power Syst.*, vol. 32, no. 4, pp. 3317–3318, Jul. 2017.

[23] Q. Yang, D. Li, W. Yu, Y. Liu, D. An, X. Yang, and J. Lin, "Toward data integrity attacks against optimal power flow in smart grid," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1726–1738, Oct. 2017.

[24] G. Chaojun, P. Jirutitijaroen, and M. Motani, "Detecting false data injection attacks in AC state estimation," *IEEE Trans. Smart Grid*, vol. 6, no. 5, pp. 2476–2483, Sep. 2015.

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014. [Online]. Available: arXiv:1412.6572.

[26] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," presented at Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2574–2582.

[27] L. Hussenot, M. Geist, and Q. Pietquin, "CopyCAT: Taking control of neural policies with constant attacks," 2019. [Online]. Available: arXiv:1905.12282.

[28] N. Papernot, P. McDaniel, and S. Jha, "The limitations of deep learning in adversarial settings," presented at IEEE Eur. Symp. Security Privacy (EuroS&P), 2016, pp. 372–387.

[29] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," 2017. [Online]. Available: arXiv:1702.02284.

[30] J. Seo, J. Choe, J. Koo, S. Jeon, B. Kim, and T. Jeon, "Noise-adding methods of saliency map as series of higher order partial derivative," 2018. [Online]. Available: arXiv:1806.03000.

[31] G. Poyrazoglu and H. Oh, "Optimal topology control with physical power flow constraints and N-1 contingency criterion," *IEEE Trans. Power Syst.*, vol. 30, no. 6, pp. 3063–3071, Nov. 2015.

[32] C. Zhang, O. Vinyals, S. Bengio, and R. Munos, "A study on overfitting in deep reinforcement learning," 2018. [Online]. Available: arXiv:1804.06893.

[33] Y.-C. Lin, Z.-W. Hong, Y.-H. Liao, M.-L. Shih, M.-Y. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," 2017. [Online]. Available: arXiv:1703.06748.

[34] V. Behzadan and Arslan Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," presented at Int. Conf. Mach. Learn. Data Min. Pattern Recognit., Cham, Switzerland, 2017, p. 14.

[35] V. Behzadan and A. Munir, "Whatever does not kill deep reinforcement learning, makes it stronger," 2017. [Online]. Available: arXiv:1712.09344.

[36] A. Pattanaik, Z. Tang, S. Liu, G. Bommannan, and G. Chowdhary, "Robust deep reinforcement learning with adversarial attacks," 2017. [Online]. Available: arXiv:1712.03632.

[37] J. Kos and D. Song, "Delving into adversarial attacks on deep policies," 2017. [Online]. Available: arXiv:1705.06452.

[38] Y. Yang, J. Hao, Y. Zheng, X. Hao, and B. Fu, "Large-scale home energy management using entropy-based collective multiagent reinforcement learning framework," in *Proc. 18th Int. Conf. Auton. Agents Multiagent Syst.*, 2019, pp. 2285–2287.

[39] Y. Yang, J. Hao, Y. Zheng, and C. Yu, "Large-scale home energy management using entropy-based collective multiagent deep reinforcement learning framework," *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 630–636.

[40] J. Sun, Y. Zheng, J. Hao, Z. Meng, and Y. Liu, "Continuous multiagent control using collective behavior entropy for large-scale home energy management," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 922–929.

[41] J. Sun *et al.*, "Stealthy and efficient adversarial attacks against deep reinforcement learning," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 5883–5891.

[42] Y. Zheng, Z. Meng, J. Hao, Z. Zhang, T. Yang, and C. Fan, "A deep Bayesian policy reuse approach against non-stationary agents," in *Proc. 31st Annu. Conf. Neural Inf. Process. Syst.*, 2018, pp. 962–972.

[43] Y. Zheng *et al.* "Efficient policy detecting and reusing for non-stationarity in Markov games," *J. Syst. Softw.*, vol. 35, no. 1, pp. 1–29, 2020.

[44] T. Yang, J. Hao, Z. Meng, C. Zhang, Y. Zheng, and Z. Zheng, "Towards efficient detection and optimal response against sophisticated opponents," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 623–629.

[45] T. Yang, J. Hao, Z. Meng, Y. Zheng, C. Zhang, and Z. Zheng, "Bayes-ToMoP: A fast detection and best response algorithm towards sophisticated opponents," in *Proc. 18th Int. Conf. Auton. Agents Multiagent Syst.*, 2019, pp. 2282–2284.

[46] Y. Zheng *et al.*, "Wuji: Automatic online combat game testing using evolutionary deep reinforcement learning," in *Proc. 34th IEEE/ACM Int. Conf. Autom. Softw. Eng.*, 2019, pp. 772–784.

[47] Y. Zheng and Y. Liu, "Automatic Web testing using curiosity-driven reinforcement learning," in *Proc. 43rd Int. Conf. Softw. Eng.*, 2021, pp. 888–898.

[48] R. Shen *et al.*, "Generating behavior-diverse game ais with evolutionary multi-objectives deep reinforcement learning," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 3371–3377.

[49] Y. Zheng, J.-Y. Hao, Z.-Z. Zhang, Z.-P. Meng, and X.-T. Hao, "Efficient multiagent policy optimization based on weighted estimators in stochastic cooperative environments," *J. Comput. Sci. Technol.*, vol. 35, no. 2, pp. 268–280, 2020.

Yan Zheng received the Ph.D. degree from the College of Intelligent and Computing, Tianjin University. He was a Research Fellow with Nanyang Technological University. He is currently with the School of New Media and Communication, Tianjin University, where he is also a Co-Leader of the Deep Reinforcement Learning Laboratory. His research includes deep reinforcement learning and the multiagent system. He is the (senior) PC member of many top-tier conferences in computer science (e.g., AAAI, IJCAI, ECAI, and AAMAS).

Ziming Yan (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include data-driven power system operation and control.

Kangjie Chen received the B.E. degree from the University of Electronic Science and Technology of China in 2015, and the M.E. degree from Tianjin University in 2019. He is currently pursuing the Ph.D. degree (First Year) with the School of Computer Science and Engineering, Nanyang Technological University. His research interests include deep learning, deep reinforcement learning, and adversarial machine learning.

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.

Jianwen Sun (Member, IEEE) received the B.E. and Ph.D. degrees from Beihang University, Beijing, China, in 2011 and 2018, respectively. He is currently a Researcher with Huawei Technologies, China. His research interests include deep reinforcement learning, multiagent systems, and industrial application of intelligent algorithms.

Yan Xu (Senior Member, IEEE) received the B.E. and M.E. degrees from the South China University of Technology, Guangzhou, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Newcastle, Australia, in 2013. He did postdoctoral research with the University of Sydney Postdoctoral Fellowship in Australia, and then joined Nanyang Technological University (NTU) with the Nanyang Assistant Professorship. He is currently an Associate Professor with School of Electrical and Electronic Engineering, and a Cluster Director with Energy Research Institute@NTU (ERI@N), Singapore. His research interests include power system stability and control, microgrid, and data-analytics for smart grid applications. He is an Editor for IEEE TRANSACTIONS ON SMART GRID, IEEE TRANSACTIONS ON POWER SYSTEMS, *IET Generation, Transmission and Distribution*, *IET Energy Conversion and Economics*, and China's power engineering international journals *CSEE Journal of Power and Energy Systems* and *Journal of Modern Power Systems and Clean Energy*. He is also serving as the Chairman for IEEE Power and Energy Society Singapore Chapter.

Yang Liu (Senior Member, IEEE) received the Bachelor of Computing degree (Hons.) and the Ph.D. degree from the National University of Singapore (NUS) in 2005 and 2010, respectively. He started his postdoctoral work with NUS, MIT, and SUTD. In 2011, he was awarded the Temasek Research Fellowship at NUS to be the Principal Investigator in the area of cyber security. In 2012 fall, he joined Nanyang Technological University (NTU) as a Nanyang Assistant Professor, where he is currently a Full Professor and the Director of the Cybersecurity Lab. He specializes in software verification, security, and software engineering. His work led to the development of a state-of-the-art model checker, process analysis toolkit. He has more than 300 publications and six best paper awards in top tier conferences and journals. With more than 20 million Singapore dollar funding support, he leads a large research team working on state-of-the-art software engineering and cybersecurity problems. His research has bridged the gap between the theory and practical usage of formal methods and program analysis to evaluate the design and implementation of software for high assurance and security.

Authorized licensed use limited to: Illinois Institute of Technology. Downloaded on March 13,2024 at 00:24:40 UTC from IEEE Xplore. Restrictions apply.