Exploring the Vulnerability of Deep Reinforcement Learning-based Emergency Control for Low Carbon Power Systems

Xu Wan , Lanting Zeng , Mingyang Sun*
Zhejiang University
{wanxu, ltzeng, mingyangsun}@zju.edu.cn

Abstract

Decarbonization of global power systems significantly increases the operational uncertainty and modeling complexity that drive the necessity of widely exploiting cutting-edge Deep Reinforcement Learning (DRL) technologies to realize adaptive and real-time emergency control, which is the last resort for system stability and resiliency. The vulnerability of the DRL-based emergency control scheme may lead to severe real-world security issues if it can not be fully explored before implementing it practically. To this end, this is the first work that comprehensively investigates adversarial attacks and defense mechanisms for DRL-based power system emergency control. In particular, recovery-targeted (RT) adversarial attacks are designed for gradient-based approaches, aiming to dramatically degrade the effectiveness of the conducted emergency control actions to prevent the system from restoring to a stable state. Furthermore, the corresponding robust defense (RD) mechanisms are proposed to actively modify the observations based on the distances of sequential states. Experiments are conducted based on the standard IEEE reliability test system, and the results show that security risks indeed exist in the state-of-the-art DRL-based power system emergency control models. The effectiveness, stealthiness, instantaneity, and transferability of the proposed attacks and defense mechanisms are demonstrated with both white-box and black-box settings.

1 Introduction

To facilitate the transition to a secure, smart, and low carbon power system, the increasing penetration of renewable energy sources, the widespread deployment of power electronic devices, and the integration of cyber and physical spaces bring about unprecedented challenges for Transmission System Operators to maintain the reliability and resilience of power systems. For example, the UK suffered its most significant power outage in over a decade in 2019 because of the generator faults, lasting more than 1.5 hours and causing widespread disruption to the traffic signal network, affecting about 1 million people. As a last resort, power system emergency control is of great importance for real-time operation to diminish the occurrence frequency and severe impact of power outages or blackouts. In particular, emergency control includes Under Voltage Load Shedding (UVLS), Under Frequency Load Shedding (UFLS), generation redispatch or tripping, dynamic braking, and controlled system separation [Kundur *et al.*, 2000], which are highly dependent on power system physical models and their abilities are limited to deal with complex and rapidly changing dynamic conditions [Zhang *et al.*, 2018]. For example, although there are some adaptive schemes of UFLS [Banijamali and Amraee, 2018] based on system behavior that have been investigated, these approaches exhibit limited efficiency and effectiveness in guaranteeing the stability for the unknown failures as the parameters for calculating integral are almost off-line designed based on the specific failure set. To this end, it is imperative to develop a new paradigm of adaptive and real-time emergency control schemes by employing cutting-edge Deep Reinforcement Learning (DRL) technologies.

In the literature, the DRL algorithm, which makes use of the advantages of Reinforcement Learning (RL) in sequential decision-making problems and combines with the idea of Deep Learning (DL) to improve the limitations of slow convergence of RL [Mnih *et al.*, 2015], has shown to be well suitable for power system emergency control, which is a dynamic, sequential decision-making problem under-uncertainty. For example, the Deep Q Network (DQN) algorithm [Huang *et al.*, 2019] and the Deep Deterministic Policy Gradient (DDPG) algorithm [Chen *et al.*, 2020] have been exploited to enhance the adaptiveness and timeliness of generator dynamic braking under-voltage load shedding, and emergency frequency control. Moreover, multi-agent DDPG algorithms are used to conduct load frequency control of the multi-area power system in the continuous action domain, and the experimental results demonstrate the superior performance of the multi-agent framework [Yan and Xu, 2020]. In addition, a meta-learning method combined with DRL algorithms is applied to solving emergency control problems in the context of extremely limited data availability [Nikoloska and Simeone, 2021].

Although the DRL-based emergency control schemes have

*Corresponding Author

shown dominant performance compared to conventional model-based approaches, with the integration of information and communication technologies, modern power systems are facing potential threats (e.g., cyber-attacks), and thus, it puts forward higher requirements for the security and reliability of the DRL model itself. For the machine learning community, the vulnerability of DRL algorithms has been preliminarily studied via developing various attack methods. Huang et al. [Huang *et al.*, 2017] altered the observation by the Fast Gradient Sign Method (FGSM), which resulted in a significant performance decline of the DRL algorithms. Moreover, Lin et al. [Lin *et al.*, 2017] implemented Carlini & Wagner (CW) attack to generate adversarial samples, which disturbed observations only at 25% of time steps but produced the same result as FGSM. On the other hand, machine learning algorithms for the regression and classification tasks in power systems have been recognized. Chen et al. [Chen *et al.*, 2018] first showed the impacts of adversarial attacks against ML-based power system tasks, including classification of power quality disturbances and forecast of building loads. In addition, the adversarial attack methods proposed in [Li *et al.*, 2021] successfully decrease the accuracy of Deep Neural Networks-based false data detection approaches and thus may cause significant security issues.

However, there is no research about the vulnerabilities of DRL-based emergency control systems in the literature, which is of great importance to be investigated before using the system in practice. Therefore, designing adversarial attacks against the DRL-based emergency control system and analyzing their system impacts is essential to assess its vulnerability. After that, the corresponding defense mechanisms against adversarial attacks for DRL-based emergency control schemes are also critical problems to ensure the stability of cost-effective operation in the power system.

To this end, this paper aims to fill the knowledge gap and investigate the fundamental limitations of existing approaches through the following novel contributions:

–To the best of the authors' knowledge, this is the first work that comprehensively investigates the vulnerability of the DRL-based power system emergency control schemes. In particular, this paper focuses on the potential threats in the test phase. Five recovery-targeted (RT) adversarial attack methods are designed for gradient-based approaches that aim to prevent the system from restoring to a stable state effectively.

–Based on the explored vulnerabilities, two distance-based robust defense mechanisms are proposed to enhance the robustness of the DRL-based emergency control model under various adversarial attacks. Furthermore, four evaluation indicators are designed to assess the effectiveness, stealthiness, instantaneity, and transferability of the tested attacks as well as the performance of the defense methods.

Experiments are conducted based on the world's first open-source platform, Reinforcement Learning for Grid Control (RLGC), which was designed to develop and benchmark DRL algorithms for power system control. The proposed attacks and defense mechanisms against SAC-, PPO- and DQN-based emergency control models are tested with both white-box and black-box settings.

The remainder of this paper is organized as follows: Section 2 introduces power system emergency control and formulates the problem as an MDP model; Section 3 explains the proposed RT adversarial attacks and robust defense (RD) mechanisms in detail; Section 4 demonstrates the performance of the proposed approaches based on the platform of RLGC; Concluding remarks are provided in Section 5.

2 Background and Problem Definition

2.1 Power System Emergency Control

Power system security control, including the two main stages of preventive and emergency control before and after contingencies occur, plays a key role in maintaining the safety and stability of power system operation. The target of emergency control is to restore the system state from an unstable to a stable one. For large-scale power systems, the emergency control problem is a highly non-linear and non-convex optimal decision-making problem [Huang *et al.*, 2019], which can be expressed as:

$$\begin{array}{ll} \min\limits_{a_t} & \int_{T_0}^{T_c} F\left(x_t, a_t\right) dt \\ \text{s.t.} & g\left(x_t, a_t\right) = 0, \\ & h\left(x_t, d_t, a_t\right) \leq 0, \\ & a_t^{\min} \leq a_t \leq a_t^{\max}, \\ & t \in [T_0, T_c] \end{array} \tag{1}$$

where $F(\cdot)$ is the cost function of the power grid emergency control; $g(\cdot)$ and $h(\cdot)$ represent the constraint relationships;x_t represents state vector of the power system, consisting of voltage magnitudes and angles; a_t represents the vector of controls available to the operator, such generator speed; d_t represents disturbance vector that could occur in the grid. In this problem, the main objective function is to minimize cost during the time horizon $T_c - T_0$ by a_t.

For this nonlinear time-varying system, traditional emergency control methods can not achieve real-time, and are difficult to obtain the optimal decision, especially in the context of a large-scale power system with high renewable energy penetration. To this end, the DRL algorithms are considered one of the most effective solutions in the literature to address these issues.

2.2 Emergency Control as an MDP

In this work, we focus on one of the emergency control technologies, UVLS, which has been implemented in the open-source platform RLGC. For the ULVS, load shedding is the ultimate countermeasure to prevent widespread outages and recover the power system from an insecure state to a normal one. The UVLS scheme reduces the load percentage of the high load area on the bus by a set of distributed controllers. In this study, the UVLS problem can be formulated as a Markov Decision Process (MDP) consisting of four parts $< \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P} >$, where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$, $\mathcal{P}$ represent the state space, the action space, the reward function, and the state transition probability, respectively. In an MDP, the agent takes observations of environment states and executes actions generated by its policy π to the environment; in turn, the environment provides feedback rewards to the agent. During the interaction

with the environment, the agent constantly adjusts its policy to achieve the best decision according to rewards. Given that the state transition probability is unknown, the state, reward, and actions can be defined as:

Observation and State. The voltage amplitude on the observation bus, represented by $V(t)$, and the remaining load percentage on the control bus, represented by $L(t)$, are selected as the observation o_t; and the last T simulation steps' observations are regarded as the states, that is, the input of DRL agent's network:

$$\begin{aligned} o_t &= \{V(t), L(t)\}, \\ s_t &= \{o_{t-T+1}, o_{t-T+2}, ..., o_t\} \end{aligned} \tag{2}$$

Action. The action is selected as the percentage of load shedding on the control bus, which quantity decides the dimension of a_t. The action can be discrete or continuous:

$$a_t = \left\{a_t^1, a_t^2, ..., a_t^i, ..., a_t^N\right\}, \quad i \in 1, 2, ..., N \tag{3}$$

where i represents the ith control bus, N is the number of control buses and a_t^i is the control action of the ith bus. In the discrete action space, a_t^i is set as a value at a fixed interval, and in the continuous space, a_t^i is a arbitrary value of a reasonable range.

Reward. Given the state s_t and action a_t, the reward function is a feedback from the environment that can be represented as:

$$r_t = \begin{cases} -C, & \text{if} \quad V_i(t) < 0.95 \\ & \text{and} \quad t > T_{pf} + 4 \\ r(V_i(t), \Delta P_j, u_{ivld}), & \text{otherwise} \end{cases} \tag{4}$$

where C is a large penalty given to the reward in this step if the limit of Transient Voltage Recovery Criteria (TVRC) (i.e., the voltage recovery to 0.95 within 4s) can not be satisfied; T_{pf} refers to the instant of fault clearing; $V_i(t)$ is the bus voltage amplitude of bus i in the power grid; $\Delta P_j(t)$ refers to the total load shedding of load bus j at time t; u_{ivld} represents a penalty when the load on a specific bus has been reduced to zero at the previous moment, but the DRL agent still provides load shedding action. Note that the reward is always negative and will get a great penalty when the action violates constraints. The specific definitions are illustrated in [Huang *et al.*, 2019].

Overall, the DRL-based UVLS problem can be formulated as an optimization problem to find the optimal control policy π, which gets maximum cumulative expected rewards during time horizon $T_c - T_0$, as shown below:

$$\max_{\theta} \quad E_{\mathcal{P},\pi}\left[\sum_{t=T_0}^{T_c} r\left(s_t, a_t\right)\right] \tag{5}$$

2.3 Adversarial Attack Against the UVLS

Adversarial attacks against the DRL-based UVLS model aim to influence the effectiveness of the control action via manipulating the states in the test phase. Let π_θ denote the policy. The target of an adversarial attack at every time step is to find an adversarial example s_t^{adv} in the ϵ neighborhood of s_t (use l_p norm to define the distance between them) that can minimize the reward, which can be formulated as an optimization problem as follows:

$$\min_{s_t^{adv}} \quad r\left(s_t, a_t^{adv}\right) \tag{6}$$

To solve the above optimization problem (6), we use the gradients descent method to minimize the loss function. In a white-box attack, the loss function J of the adversarial state s_t^{adv} is designed to realize the optimization objective:

$$\min_{\left\|s_t^{adv} - s_t\right\|_p < \epsilon} J\left(s_t^{adv}; s_t, a_t, \theta\right) \tag{7}$$

Similarly, the transfer-based black-box attack is also developed to generate the adversarial sample that minimizes J. Nevertheless, it uses the network parameters of the known DRL model θ to generate s_t^{adv} against the unknown DRL model with parameters θ'.

3 Proposed Attacks and Defense Mechanisms

For the application of DRL-based UVLS, there are significant differences in adversarial examples generation mechanisms between pixel values investigated in the ML community and power system observations for the following reasons. First, the objective function of adversarial attack against the DRL-based UVLS attempts to influence the decisions of the percentage of load shedding on the control buses during test periods so as to induce more severe stability issues such as cascading failure or even widespread blackout. However, most adversaries against classification models in the ML community aim to misclassify images. Second, observations in the DRL-based UVLS model are not as unified as pixel information but rather the voltage and load percentage information with more complex physical meanings. For example, bus voltages are mutually restrained, making it more challenging to manipulate observations considering stealthiness.

On this basis, integrated with the characteristics of power systems, we propose the concept of recovery-targeted adversarial attacks for the DRL-based emergency control schemes as well as the corresponding distance-based robust defense mechanisms to improve the robustness of the model. An overview of the RT-attacks and DB-defense mechanisms for the DRL-based emergency control scheme proposed in this paper is outlined in Fig.1.

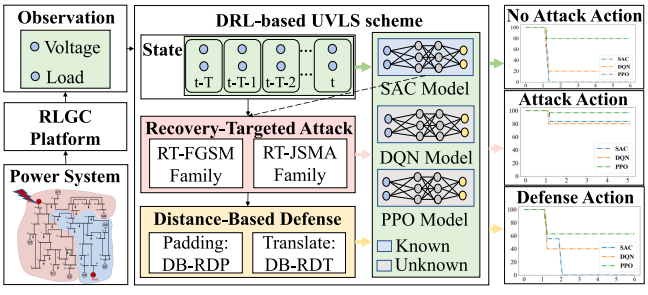

Figure 1: An overview of the proposed RT-attacks and DB-defenses for the DRL-based emergency control schemes.

3.1 Recovery-Targeted Adversarial Attacks

In the DRL-based UVLS scheme, the output of DRL models is the action, which refers to the percent of load shedding on control buses. To this end, the targeted adversarial action a_{adv} that the attacker would like the system operator to execute should realize the following optimization objective:

$$a_{adv} = \text{argmin}_{a \in \mathcal{A}} r(s, a) \tag{8}$$

Note that in this work, we simplify the procedure to calculate the optimal targeted adversarial action via solving the above problem and directly consider the actions of no load shedding for UVLS, denoted by a_{nls}, to ensure that the system will not recover to a stable state after an emergency.

On this basis, we propose adversarial attacks against the UVLS scheme, including the recovery-targeted FGSM (RT-FGSM) family and recovery-targeted Jacobian-based Saliency Map Attack (RT-JSMA) family.

RT-FGSM Family

FGSM [Goodfellow *et al.*, 2014] is one of the simplest and fastest gradient-based attack methods that crafts adversarial examples by calculating gradients of model outputs to inputs [Wang and He, 2021]. For the UVLS, the objective functions J_1 of the proposed RT-FGSM can be defined as:

$$J_1 = \frac{1}{N} \sum_{i=1}^{N} \left(a_t^i - a_{nls}\right)^2 \tag{9}$$

Then, the FGSM adversarial example can be designed by minimizing J_1 with a one-step update:

$$s_t^{adv} = f\left(s_t - \omega \cdot \text{sign}\left(\nabla_{s_t} J_1\right)\right), \tag{10}$$

where $f(\cdot)$ is a function to keep state satisfy physical constraints of power systems, such as load constraints and voltage constraints; ω denotes a small positive value. On this basis, the proposed RT loss function J_1 can be employed in all FGSM families, such as Iterative Fast Gradient Sign Method (I-FGSM) [Kurakin *et al.*, 2016], Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [Dong *et al.*, 2018].

RT-JSMA Family

Different from the FGSM, the JSMA algorithm [Papernot *et al.*, 2016] is a targeted attack. Meanwhile, JSMA pursues to reduce the number of modified input dimensions as much as possible. Thus, we attempt to make agents' actions on a particular control bus more prone to decision errors than other buses through JSMA attacks. The JSMA algorithm [Papernot *et al.*, 2016] is another type of gradient-based attack, which uses l_0 norm as the distance metric.

Aiming at the DRL-based UVLS scheme, the implementation process of the proposed RT-JSMA algorithm can be divided into the following three steps. First, we define the loss function as the l_1 norm between the action at time t and no load shedding action a_{nls}, then calculating the forward derivative as the Jacobian matrix:

$$\frac{\partial J_2^i}{\partial s_t^j} = \left[\frac{\partial ||a_t^i - a_{nls}||_1}{\partial s_t^j}\right] i \in 1, ..., N; j \in 1, ..., M \tag{11}$$

where s_t^j is the jth value of the state s_t, M is the dimension of the state. Then, the second step is to construct adversarial saliency graph $G_{s_t}^b$:

$$G_{s_t}^b[j] = \begin{cases} 0, \text{ if } \frac{\partial J_2^b}{\partial s_t^j} < 0 \text{ or } \sum_{i \neq b} \frac{\partial J_2^i}{\partial s_t^j} > 0 \text{ or } j \notin \Gamma \\ \frac{\partial J_2^b}{\partial s_t^j} \left| \sum_{i \neq b} \frac{\partial J_2^i}{\partial s_t^j} \right|, \text{ otherwise} \end{cases} \tag{12}$$

where b is the targeted attack bus; Γ is a searching domain, $\Gamma = \{1, ..., T\}$. Then, we find two states dimensions, $s_t^{m_1}$, $s_t^{m_2}$, which have the greatest impact on the objective function J_2^b for the current s_t. Note that searching domain Γ will remove m_1, m_2, if $s_t^{m_1}$, $s_t^{m_2}$ are beyond limits $f(\cdot)$.

$$\begin{aligned} m_1 &= \text{argmax}\, G_{s_t}^b, \\ m_2 &= \text{argmax}_{m_2 \neq m_1} G_{s_t}^b \end{aligned} \tag{13}$$

The adversarial state is iteratively updated until it reaches the maximum distortion γ. During the iteration, $s_t^{m_1}$, $s_t^{m_2}$ are modified with an appropriate disturbance β.

$$\begin{aligned} s_t^{adv}(0) &= s_t, \\ s_t^{adv}(k+1) &= f\left(s_t^{adv}(k), s_t^{adv,j}(k) - \beta\right), \\ j &\in \{m_1(k), m_2(k)\}, \\ k &< \gamma M \end{aligned} \tag{14}$$

It is imperative to highlight that, in this study, we propose a new adversarial attack called recovery-targeted Momentum-JSMA (RT-MJSMA) that combines RT-JSMA with the idea of momentum. Different from RT-JSMA, RT-MJSMA uses the previous gradient information to modify the state at each iteration step k and calculates the adversarial saliency graph $G_{s_t}^b$ defined by Eq.16 , where g_j^i is jth row, ith column element of g defined by Eq.15. The whole algorithm of the proposed RT-MJSMA is given in Appendix B.2.

$$\begin{aligned} g(0) &= 0, \\ g(k+1) &= \mu \cdot g(k) - \frac{\nabla_{s_t^{adv}(k)} J_2}{\left\| \nabla_{s_t^{adv}(k)} J_2 \right\|_1} \end{aligned} \tag{15}$$

$$G_{s_t}^b[j] = \begin{cases} 0, \text{ if } g_j^b < 0 \text{ or } \sum_{i \neq b} g_j^i > 0 \text{ or } j \notin \Gamma \\ g_j^b \left| \sum_{i \neq b} g_j^i \right|, \text{ otherwise} \end{cases} \tag{16}$$

Note that all the above-proposed approaches can be directly used as transfer-based attacks for the black-box settings.

3.2 Distance-based Robust Defense Methods

To mitigate the threats arising from the attacks mentioned above, it is of great importance to investigate the corresponding defense mechanisms to enhance the robustness of the DRL-based UVLS model so that it can still produce reliable control actions under various types of attacks. To this end, we propose two distance-based robust defense methods, named the distance-based robust defense padding method (DB-RDP) and the distance-based robust defense translate method (DB-RDT), which utilize the characteristics of historical states to modify inputs s_t. The detail can be seen in Appendix B.3.

DB-RDP

The key idea of the proposed DB-RDP method is to pad the last observation o_t with historical observation. If s_t is attacked, the attacker is more likely to modify the recent observation o_t, especially in iterative attacks and local attacks. This observation has the greatest influence on action decisions, and its value is also an important indicator of the reward function. Meanwhile, the historical observation is similar with o_t if the state is not attacked due to the short simulation sample time. Thus, the furthest observation from o_t^{adv} is more likely to be the nearest true value of o_t and l_2 norm is selected as the distance metric in order to find the furthest observation o_{t-T+d1}, as shown below:

$$d1 = \text{argmax} \left\| o_t - o_{t-T+j} \right\|_2, j \in 1, ..., T \tag{17}$$

Then we pad o_t with o_{t-T+d1} and the defense observation can be represented as:

$$s_t^{def} = \{o_{t-T+1}, o_{t-T+2}, ..., o_{t-1}, o_{t-T+d1}\} \tag{18}$$

Moreover, if no attack is applied in the DRL model, s_t^{def} is similar to s_t, which makes the DRL-based UVLS scheme still perform well in no attack setting.

DB-RDT

The proposed DB-RDT method differs from the distance-based padding method in constructing the defense observation. More specifically, it exchanges o_t with o_{t-T+d1}:

$$s_t^{def} = \{o_{t-T+1}, ..., o_t, ..., o_{t-1}, o_{t-T+d1}\} \tag{19}$$

To study the feasibility of proposed defensive methods, we use parameter p to control the percentage of our defense methods in the test episode.

$$D\left(s_t^{adv}; p\right) = \begin{cases} D\left(s_t^{adv}\right) & \text{with probability } p \\ s_t^{adv} & \text{with probability } 1-p \end{cases} \tag{20}$$

where $D(\cdot)$ is the function corresponding to our defense methods.

Note that this work proposes two distance-based robust defense methods, which utilize the characteristics of historical states to modify inputs states. If the attacker is aware of the proposed attack mitigation scheme and factors this in the attack planning, one possibility is that the attacker will choose to reduce the attack preference and make more uniform changes to all state variables, resulting in a significant decline in attack performance; another possibility is that the attacker will make more preferred changes to a historical observation, but the modified observations may not influence the performance of our methods. Moreover, the design of defense probability p also ensures the dynamic characteristics and makes it difficult for the attackers to bypass it.

4 Experiments and Analysis

This section discusses results related to the proposed adversarial attack and defense methods on three DRL-based emergency control schemes (i.e., SAC-, DQN- and PPO-based) with four designed evaluation indicators. All the experiments have been conducted based on the RLGC open-source platform [Huang *et al.*, 2019] with the IEEE 39-bus system [Athay *et al.*, 1979]. The detailed information is shown in Appendix A and C.

4.1 Experiment Setup

Baselines

We use three DRL baselines exposed by OpenAI Lab, including DQN for the discrete action space, PPO and SAC for the continuous action space. All networks are publicly available[1]. In our setting, the state includes voltage magnitudes at buses 4, 7, 8, 18, and the percentage of remaining load served by buses 4, 7, and 18. The action is load shedding percentage at buses 4, 7, and 18. The model output is a binary vector in the discrete action space, where 0 represents no load shedding actions, and 1 represents 20% load shedding actions at corresponding buses. In the continuous action space, the action value is within -0.5 to 0, meaning the specific load shedding percentage value. It should be noted that simulations will break early when the voltage in observation buses is below the standard recovery for more than 4 seconds.

Evaluation Metrics

Considering the perspectives of effectiveness, stealthiness, and instantaneity of our adversarial attack methods, four evaluated metrics are given as follows: **(1) Average Episodes Reward (AER).** It represents the average cumulative rewards under the attack within test episodes; **(2) Bad Recovery Ratio (BRR).** It compares the voltage recovery with the standard recovery [Huang *et al.*, 2019]. We record the duration if the voltage recovery line is lower than the standard recovery line; **(3) Average Modification of Observation (AMO).** It evaluates the stealthiness of adversarial attacks. We calculate the l_1 norm between the adversarial state s_t^{adv} and the original state s_t; **(4) Average Time Consumption (ATC).** It evaluates the instantaneity of adversarial attacks. Because DRL decision-making in emergency control is real-time, adversarial example generations can not be time-consuming. Therefore, we introduce this indicator to evaluate the average time to make adversarial samples in each test step.

4.2 Adversarial Attacks Performance

This subsection aims at analyzing: 1) the influence of white-box attack on SAC model; 2) the influence of transfer-based black-box attack on PPO and DQN models.

White-box Attacks. In the white box cases, the severe impacts of the proposed RT-adversarial attacks and their comparison with existing un-targeted attacks are given in Table 1. As can be seen, compared with the case of no attack, injecting the adversarial examples generated via the proposed RT-methods can significantly decrease the effectiveness of the executed control actions, indicated by approximately nine times lower and 35.78% higher metrics values in terms of the average AER and average BRR across all five RT-attacks. In particular, the RT-IFGSM and RT-MJSMA methods have reduced their AER metric values by more than ten times. Meanwhile, conventional attacks only result in the reduction of AER by between 1.5 and 3 times, and the BRR almost no change. Nevertheless, the metric values of average perturbation AMO and generation time ATC for the proposed RT-attacks and conventional attacks can be retained at the same level. To sum up, it can be demonstrated that our RT-attacks

[1] https://github.com/hill-a/stable-baselines

Attack Methods	Attack Performance				DB-RDP Defense Performance				DB-RDT Defense Performance			
	AER	BRR	AMO	ATC	AER	BRR	AMO	ATC	AER	BRR	AMO	ATC
no attack	-2301.33	0.007	0.000	0.000	-2414.66	0.007	0.000	0.000	-2414.66	0.007	0.000	0.000
FGSM	-5778.01	0.009	0.060	**0.001**	-5230.06	**0.007**	0.059	**0.001**	-3974.78	**0.007**	0.067	**0.001**
IFGSM	-5718.77	0.009	0.060	**0.001**	-5171.41	**0.007**	0.062	0.003	-4515.40	**0.007**	0.065	**0.001**
MIFGSM	-7113.66	0.009	0.060	0.003	-5867.57	**0.007**	0.061	0.003	-4623.80	**0.007**	0.062	0.003
JSMA	-9538.21	0.009	0.060	0.075	-7102.08	**0.007**	0.060	0.075	-6638.73	0.008	0.060	0.077
MJSMA	-9538.57	0.009	0.060	0.077	-7651.68	**0.007**	0.060	0.077	-7023.86	0.008	0.060	0.078
RT-FGSM	-17201.57	0.269	0.060	**0.001**	-5409.06	**0.007**	0.057	**0.001**	-11904.38	0.143	0.059	**0.001**
RT-IFGSM	**-27157.80**	**0.394**	0.060	0.003	-8709.32	0.159	**0.052**	0.003	**-2778.83**	0.032	**0.044**	0.004
RT-MIFGSM	-22988.47	0.369	0.060	0.003	-6626.75	0.132	0.056	0.008	-10273.40	0.145	0.056	0.003
RT-JSMA	-23470.70	0.362	0.060	0.077	**-2594.81**	0.082	0.060	0.073	-10297.06	0.176	0.060	0.078
RT-MJSMA	-23461.20	0.361	0.060	0.076	-2648.40	0.083	0.060	0.077	-11741.90	0.177	0.060	0.082

Table 1: Performances of white-box attacks and defense methods in the SAC-based UVLS model.

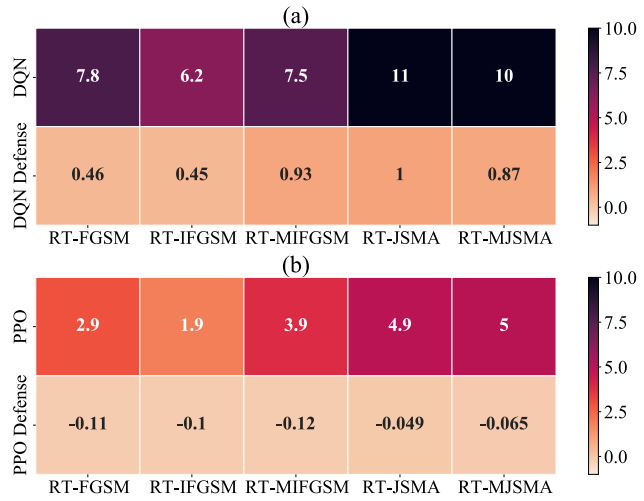

Figure 2: Attack and defense results in the black-box setting for the DQN- and PPO-based UVLS models. The values on each square represent the decline multiple of AER compared with no attack.

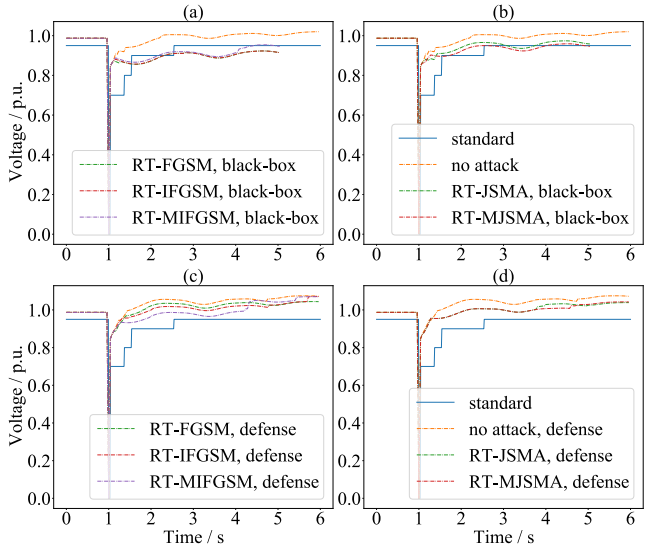

Figure 3: Recovery lines under black-box attacks (a,b) and defense (DB-RDP) (c,d) in the DQN-based UVLS model.

have more practical significance in exploring the vulnerabilities of the DRL-based emergency control model.

Black-box Attacks. To study transferability, we use RT-attack examples generated based on the SAC model (continuous action space) into the DQN model (discrete action space) and the PPO model (continuous action space). As shown in Fig.2, black-box RT-attacks still perform pronounced attack effects in DQN and PPO with an average of 8.46 times and 3.71 times decreasing rewards of AER. In addition, the BRR of them are increased by average 27.25% and 14.98%, respectively, when compared with the cases of no attack. One of the key conclusions stemming from the results is: black-box attacks against DQN are more severe than PPO, indicating that the transferability of continuous transfer into discrete action space is higher than the opposite. Furthermore, the proposed RT-JSMA family exhibits stronger transferability with average 2.692 higher rewards decreasing than the RT-FGSM family. The performances of RT-attacks in the DQN model under one test episode are presented in Figure 3 (a) and (b). From the recovery line, the black-box attack effect is not up to the white-box attack, which is consistent with common sense.

4.3 Defense Methods Performance

This part aims to evaluate the performance of the proposed defense mechanisms under both the white-box and black-box RT-attacks. As indicated in Table 1, the metric values of AER have significantly improved after applying the DB-RDP and DB-RDT methods with about 3.75 and 3.08 times average increases across different attacks, respectively. Moreover, Fig. 2 shows that the proposed DB-RDP successfully mitigates the impacts of all RT-attacks in the black-box setting, as indicated by the approximately 7.76 and 3.81 times decreased AER values for the DQN- and PPO-based UVLS models, respectively. In addition, we observe that all recovery lines are all over the blue standard recovery line from Fig. 3 (c) and (d). Overall, the above results illustrate that the proposed defense mechanisms can enhance the robustness of the DRL-based emergency control model to obtain an effective control action to

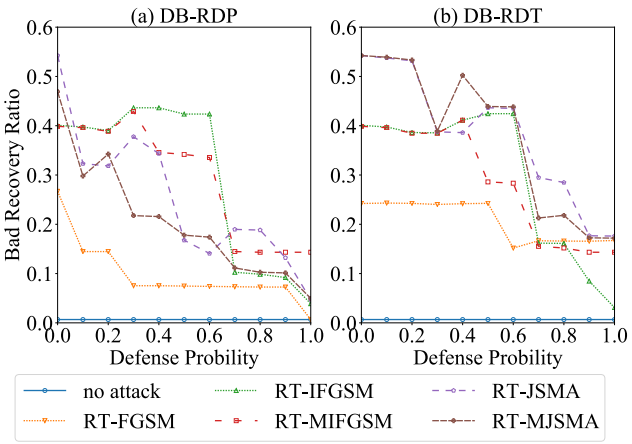

Figure 4: BRR results with different defense application probabilities p under different white-box attacks.

recover the power system back to a stable state.

Ablation Study. Furthermore, we conduct a series of ablation experiments to study the impact of defense application proportion p on white-box adversarial attacks. The defense results of DB-RDP and DB-RDT on the SAC model are presented in Fig. 4 (a) and (b), respectively. As the proportion of defense application p increases, attacks' effectiveness gradually weakens or is even invalid. For the DB-RDP method, the BRR of the DRL-based UVLS scheme drops to less than 5% when it is fully applied (i.e., $p = 1$).

5 Conclusion and Future Work

This paper explores the vulnerability of DRL-based emergency control for low carbon power systems, which aims to fill the knowledge gap between the ML and power system communities. In particular, we design five recovery-targeted adversarial attacks against the DRL-based UVLS scheme, named RT-FGSM family and RT-JSMA family. Considering characteristics of observations in the power system, we propose two distance-based robust defense methods, namely DB-RDP and DB-RDT. Based on the open-source platform, RLGC, case studies are conducted to evaluate the performance of the proposed adversarial attacks on the SAC model with the white-box setting and, the DQN model, the PPO model with the black-box setting. The key insights observed from the experiment results include: (1) DRL-based emergency control scheme indeed has security risk, and the proposed RT-attack algorithms can lead to more severe impacts regarding their effectiveness, stealthiness, instantaneity, and transferability; (2) the proposed active defense mechanisms can alleviate the vulnerability brought by the attacker, and at the same time, these defense mechanisms do not weaken the model performance in the attack-free DRL model.

In the future, we will explore the vulnerability considering power system physical constraints and extend to a wider range of power system applications. Furthermore, the proposed approaches can also be generalized to other complex domains (e.g., smart manufacturing systems, intelligent transportation systems) by considering their physical dynamics or characteristics.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 52161135201, U20A20159, and 62103371.

References

[Athay *et al.*, 1979] T Athay, Robin Podmore, and Sudhir Virmani. A practical method for the direct analysis of transient stability. *IEEE Transactions on Power Apparatus and Systems*, (2):573–584, 1979.

[Banijamali and Amraee, 2018] Seyed Sohrab Banijamali and Turaj Amraee. Semi-adaptive setting of under frequency load shedding relays considering credible generation outage scenarios. *IEEE Transactions on Power Delivery*, 34(3):1098–1108, 2018.

[Chen *et al.*, 2018] Yize Chen, Yushi Tan, and Deepjyoti Deka. Is machine learning in power systems vulnerable? In *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–6, 2018.

[Chen *et al.*, 2020] Chunyu Chen, Mingjian Cui, Fangxing Li, Shengfei Yin, and Xinan Wang. Model-free emergency frequency control based on reinforcement learning. *IEEE Transactions on Industrial Informatics*, 17(4):2336–2346, 2020.

[Dong *et al.*, 2018] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Gottlob *et al.*, 2002] Georg Gottlob, Nicola Leone, and Francesco Scarcello. Hypertree decompositions and tractable queries. *Journal of Computer and System Sciences*, 64(3):579–627, May 2002.

[Gottlob, 1992] Georg Gottlob. Complexity results for nonmonotonic logics. *Journal of Logic and Computation*, 2(3):397–425, June 1992.

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

[Huang *et al.*, 2019] Qiuhua Huang, Renke Huang, Weituo Hao, Jie Tan, Rui Fan, and Zhenyu Huang. Adaptive power system emergency control using deep reinforcement learning. *IEEE Transactions on Smart Grid*, 11(2):1171–1182, 2019.

[Kundur *et al.*, 2000] P Kundur, GK Morison, and L Wang. Techniques for on-line transient stability assessment and

control. In *2000 IEEE Power Engineering Society Winter Meeting. Conference Proceedings (Cat. No. 00CH37077)*, volume 1, pages 46–51. IEEE, 2000.

[Kurakin *et al.*, 2016] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016.

[Levesque, 1984a] Hector J. Levesque. Foundations of a functional approach to knowledge representation. *Artificial Intelligence*, 23(2):155–212, July 1984.

[Levesque, 1984b] Hector J. Levesque. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 198–202, Austin, Texas, August 1984. American Association for Artificial Intelligence.

[Li *et al.*, 2021] Jiangnan Li, Yingyuan Yang, Jinyuan Stella Sun, Kevin Tomsovic, and Hairong Qi. Conaml: Constrained adversarial machine learning for cyber-physical systems. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, pages 52–66, 2021.

[Lin *et al.*, 2017] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3756–3762, 2017.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[Nebel, 2000] Bernhard Nebel. On the compilability and expressive power of propositional planning formalisms. *Journal of Artificial Intelligence Research*, 12:271–315, 2000.

[Nikoloska and Simeone, 2021] Ivana Nikoloska and Osvaldo Simeone. Black-box and modular meta-learning for power control via random edge graph neural networks. *arXiv preprint arXiv:2108.13178*, 2021.

[Papernot *et al.*, 2016] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[Wang and He, 2021] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021.

[Yan and Xu, 2020] Ziming Yan and Yan Xu. A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system. *IEEE Transactions on Power Systems*, 35(6):4599–4608, 2020.

[Zhang *et al.*, 2018] Jingyi Zhang, Chao Lu, Chen Fang, Xiang Ling, and Yong Zhang. Load shedding scheme with deep reinforcement learning to improve short-term voltage stability. In *2018 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia)*, pages 13–18. IEEE, 2018.