Electrical Power and Energy Systems 131 (2021) 107113

Contents lists available at ScienceDirect

International Journal of Electrical Power and Energy Systems

journal homepage: www.elsevier.com/locate/ijepes

Reinforcement-learning-based dynamic defense strategy of multistage game against dynamic load altering attack

Youqi Guo, Lingfeng Wang [*], Zhaoxi Liu, Yitong Shen

Department of Electrical Engineering, University of Wisconsin-Milwaukee, Milwaukee 53211, USA

ARTICLE INFO

Keywords:
Power grid security
Dynamic load altering attacks
Reinforcement learning
Game theory
Cyber risk modeling
Optimal load shedding
Cyber-physical resilience

ABSTRACT

As the current power grid is highly interconnected and more information and communication technologies (ICTs) are being deployed recently, it could be the target of malicious cyber-physical attacks. Dynamic load altering attacks (D-LAAs), as a special case of load altering attacks, could be performed to interfere the demand response and ultimately force certain generators off-line. Cascading failures due to transmission line overloads may also be triggered. In this paper, we propose a new dynamic defense strategy against D-LAAs through a multistage game between the attacker and the defender which is solved by minimax-q learning. Different from the static game, the multistage game considers the attacker and defender's action sequences and the optimal strategies at each state are learned. After each time step, the cascading failure is measured, and the load shedding is used as the feedback for the attacker to generate the next action strategy. The performance of the proposed model is evaluated on the IEEE 39-bus system. Comparisons between the dynamic defense strategy and the passive defense strategy are conducted, and the results verify the advantage of the proposed dynamic defense strategy. To improve the power system resilience, this defense strategy can be deployed in advance when such cyber-physical attacks are anticipated.

1. Introduction

Ensuring cybersecurity of modern power grids has become a national priority with the smart grid initiative. The use of information and communication technologies (ICTs) has not only enhanced the efficiency and reliability of the power grids but also created new vulnerabilities if they are not accompanied by advisable security reinforcements [1,2]. Various vulnerabilities may leave some sectors of the power system to a wide range of cyber-physical attacks [3]. As a real example, attackers remotely switched off the breakers in a series of substations by pre-installed malware, resulting in a widespread outage in the Ukrainian power system in late 2015. This blackout is the first publicly acknowledged incident caused by cyber-attack which is even more destructive than natural disasters [4]. Furthermore, identifying and mitigating such risks are instrumental in improving the resiliency of power grids [5]. Thus, considering the increasing cyber-physical threats to the modern power system, it is imperative that we understand the risks resulting from cyber-physical attacks and thus implement effective security strategies against them.

Load Altering Attack (LAA) is a representative cyber-physical attack with the aim to maliciously control and alter a group of remotely accessible yet unsecured controllable loads. A successful LAA can disturb the balance between the power demand and supply, causing frequency and angle instability and consequently system blackout through circuit overflow or generator tripping. The potential vulnerable loads to LAAs can be frequency-responsive loads [6,7], data center's computational load [8], loads with direct load control (DLC) which is one of the most common demand side management programs [9,10], etc.

LAAs can be categorized into static load altering attack (S-LAA) (which is mainly focused on the amount of vulnerable loads) and dynamic load altering attack (D-LAA) (which is additionally concerned with the trajectory of the changes that are made in the vulnerable loads). Reference [11] introduces and models S-LAA in smart grids, and the studies in [12–14] address the prevention or detection of LAAs. Unlike these investigations, reference [15] introduces, characterizes and classifies D-LAAs as a new class of cyber-physical attacks against the power grids. In [16], the authors present a protection scheme using energy storage systems to improve the power grid's reaction to D-LAAs.

Game theory is oftentimes used to help people understand the

[*] Corresponding author.
E-mail address: l.f.wang@ieee.org (L. Wang).

https://doi.org/10.1016/j.ijepes.2021.107113
Received 7 December 2020; Received in revised form 24 February 2021; Accepted 12 April 2021
Available online 10 May 2021

situations in which decision-makers interact, e.g., between attackers and defenders. There is a wide range of situations to which game theory can be applied: political candidates competing, companies competing in business, bidders bidding in an auction, and so on [17]. Various games are formulated to illuminate different economic, political, engineering phenomena, such as general sum, zero-sum and potential games. Recently, researchers recognized the critical role of game-theoretic approaches in power grid security. The security games introduce an analytical framework with a rich mathematical basis for modelling the interactions between intentional attackers whose aim is to disrupt the power grid and operators defending it [18,19]. The games in power grid security are classified into two categories: *static* and *dynamic* games. The static game can be considered as a one-shot process, which means players only take one action. A wealth of research [20–25] has emerged on the static defense schemes against malicious attacks in the smart grid. In [20], the authors present a comprehensive and quantitative static game framework for the power system security problem. Under this framework, a new criterion is derived to seek reliable defense strategies. In [21], a zero-sum static game model is proposed to provide security policies in the cyber layer with corresponding resilient control in the physical layer. In [22], Farraj et al. analyze the cyber switching attacks and corresponding mitigation method by the zero-determinant strategy in an iterative game. The strategy allows the electric power utility (EPU) to stabilize the power grid in the face of cyber switching attacks. A game equilibrium is obtained by a zero-sum static game between intentional attackers and defenders to provide a reliable fusion-based defense scheme for the communication network of power systems in [23]. In [24], the effect of the compromised active power measurements on the electricity price is quantified. This situation is modeled as a zero-sum game between the defender and the attacker who performs the bad data injection attack on the measurements. For defending against denial-of-service (DoS) attacks, Li et al. [25] investigate the interaction between the sensor nodes and adversaries.

On the other hand, dynamic games have been a largely underexplored domain in the power grid security area. Most existing work mentioned previously are focused on static games or static defenses without considering dynamic processes. In dynamic or multistage games, attackers can compromise multiple components in a time sequence [26]. For some practical cases, to obtain the maximum profit or achieve the attack objective, attackers have to take offensive actions one by one based on the defender's protection policy and the next steady state of system. Note that full knowledge and observation of the target system are required for the players. In [27], the authors propose a stochastic game to protect the power system against coordinated cyber-physical attacks. Although two states are considered, the game proposed in [27] is more like a one-shot game because the attacker can only target one element at a time. There is no more dynamic evolution in this game. Ma et al. consider a multi-act dynamic game in the electricity market for defending against jamming attacks in [28]. Dynamic programming is adopted to solve the game. To carry out the recursions, knowing the model of environment is necessary. In [29], a q-learning method is devised to solve a multistage game. The attacker's actions are considered while the defender's actions are pre-defined rather than evolving by interacting with the attacker's action and system state.

Furthermore, machine learning methods are being applied to address cyber-physical security issues in power systems for attack detection, analysis of defense strategy, and fault diagnosis. In [30], a deep-learning-based algorithm to detect power theft and false data injection (FDI) attack on real-time measurements is proposed. The authors in [31] use Q-learning to analyze vulnerabilities of the power system in sequential topological attacks. Wang et al. [32] develop a deep learning method for fault diagnosis of power plants. A hierarchical deep domain adaptation (HDDA) approach is proposed to apply a classifier with labeled data under one loading condition to detect faults with unlabeled data under another loading condition.

Thus far, the focus in power grid security against malicious attacks has been mainly on static defense schemes. In contrast, in this paper, we address a new dynamic defense policy against multistage D-LAA, which is concerned with dynamic interactions between the attacker and defender. The attack-defense interaction is modeled by a two-player zero-sum multistage game and the solution is obtained based on minimax-q learning. Unlike dynamic programming solution given in [28] that requires exact knowledge of the model of environment, the proposed minimax-q learning based solution in this paper goes from experience to policies by learning a model rather than needing a model. The main contributions of this paper are summarized as follows:

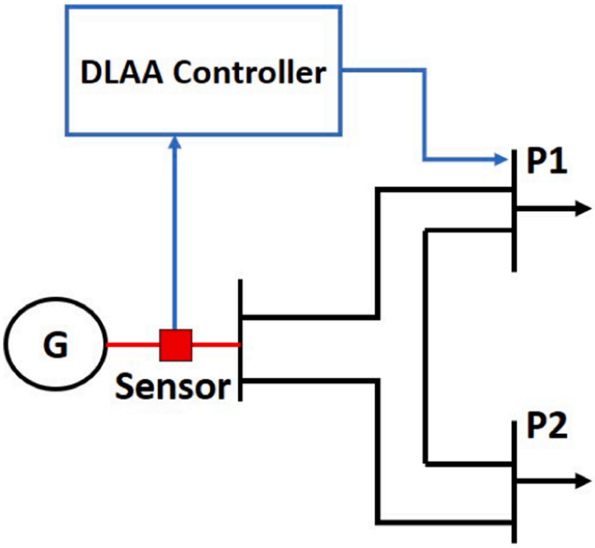

Fig. 1. Single-point closed-loop D-LAA.

- The one-shot dynamic load altering attacks (D-LAA) in [15] is extended to a sequence attack. The corresponding cascading failures caused by D-LAA are studied holistically. It allows the attacker takes offensive actions one by one based on the states of system and adversary's protection policy to achieve much higher attack objective.
- A two-player zero-sum multistage game considering both dynamic of the attacker and the defender is proposed.Different from the one-shot games that lack dynamic evolvement of the attack-defense sequence and passive defense strategies where the evolvement of defender's actions is neglected in the existing literature, a minimax-q learning scheme is adopted in this paper to effectively find out the optimal defense sequence against chronological D-LAA considering dynamic interactions between the attacker and the defender. This is also the main difference between the proposed dynamic defense and the existing research.
- This dynamic defense strategy is compared with the static (passive) defense policy. The simulation results show that the power grid with the proposed defense strategy does have lower load loss due to D-LAAs.

The rest of this paper is organized as follows: Section II presents the related preliminaries and the game model is formulated in Section III. Analysis of the minimax-q learning solution is presented in Section IV. In Section V, simulation setup, results and analysis are presented. Concluding remarks and potential future directions are given in Section VI.

2. Related Preliminaries

In this section, some related preliminaries are presented including the mathematical model of D-LAAs, optimal load shedding problem and cascading failures.

2.1. Dynamic Load Altering Attack

2.1.1. D-LAA Implementation Principle

The basic threat model is adopted from reference [15]. As mentioned, D-LAA is concerned with the volume as well as the trajectory of the changes in the vulnerable load. In a closed-loop D-LAA, referring to Fig. 1, the attacker tries to manipulate the vulnerable load (P1) with constant monitoring at the sensor bus for the grid conditions. Although there are various approaches to measure the grid conditions and alter the load, in this paper we limit our scope to the power system frequency obtained from the installed frequency sensors and frequency-responsive loads. A successful D-LAA can be conducted only if there are sufficient potential vulnerable loads to be compromised. The attack objective is to deviate the frequency from the system's nominal value and eventually push one generator off-line. To implement a D-LAA, there are three main steps that the attacker must undertake:

- Install the frequency monitor at the sensor bus and constantly send frequency acquisitions to the D-LAA controller. In general, it is not difficult to monitor the frequency of power system using an inexpensive commercial sensor.
- Based on the mechanism of the attack controller and the feedback signal, calculate the amount of vulnerable load which needs to be compromised at the victim bus.
- Remotely control and alter the victim load at the amount that is calculated in the last step. The feasibility of remotely altering the load is discussed in [33].

2.1.2. Attack Model

In power systems, theoretically, the power flow between buses i and j is a nonlinear function of bus voltages and the impedance of transmission lines. The active power flow can be given as follows:

$$P_{ij} = V_i V_j \left[G_{ij} cos\left(\phi_i - \phi_j\right) + B_{ij} sin\left(\phi_i - \phi_j\right) \right] \tag{1}$$

where V is the voltage magnitude, ϕ is the phase angle in the corresponding bus, and G and B are the real and imaginary parts of the impedance, respectively. Note that G_{ij} can be considered as zero because the resistance of the transmission lines is significantly less than the reactance in practice. Furthermore, the difference of voltage phase angle between two buses is small and the voltage magnitude in each bus is very close to unity in the per-unit system. Thus, further approximation for the power flow equation can be written as:

$$P_{ij} = B_{ij}\left(\phi_i - \phi_j\right). \tag{2}$$

Specifically, consider a power system with $\mathscr{G}$ generator buses and $\mathscr{L}$ load buses. Let then $\mathscr{N} = \mathscr{G} \cup \mathscr{L}$ represents the set of all buses in this grid. For a bus $i \in \mathscr{N}$, the total amount of power flow can be separated into the power injection of the generator P_i^G at bus $i \in \mathscr{G}$ and power absorbed by load P_i^L at bus $i \in \mathscr{L}$. Defining δ_i as the voltage phase angle of the i-th generator bus, θ_i as the voltage phase angle at i-th load bus and B_{ij} as the admittance value between buses i and j, the linearized power flow equations based on Eq. (2) can be written as:

$$P_i^G = \sum_{j\in\mathscr{G}} B_{ij}\left(\delta_i - \delta_j\right) + \sum_{j\in\mathscr{L}} B_{ij}\left(\delta_i - \theta_j\right), \tag{3}$$

$$-P_i^L = \sum_{j\in\mathscr{G}} B_{ij}\left(\theta_i - \delta_j\right) + \sum_{j\in\mathscr{L}} B_{ij}\left(\theta_i - \theta_j\right). \tag{4}$$

To model the dynamic behavior of each generator, the swing equations are used for the generator bus:

$$\dot{\delta}_i = \omega_i, \tag{5}$$

$$M_i\dot{\omega}_i = P_i^M - P_i^G - D_i^G\omega_i, \tag{6}$$

where ω_i is the rotor angular frequency deviation of generator bus i, M_i is the rotor inertia of each generator, P_i^M is the mechanical power input and D_i^G represents the damping coefficient. Note that P_i^M and D_i^G must be positive.

Specifically, the turbine-governor controller and the load–frequency controller can be integrated together as a proportional-integral (PI) controller, aimed at maintaining the rotor angular frequency at its nominal level to affect the mechanical power input [34]. The PI controller is represented as:

$$P_i^M = -\left(K_i^P\omega_i + K_i^I\int_0^t \omega_i\right), K_i^P, K_i^I > 0, \tag{7}$$

where K_i^P and K_i^I are the proportional and integral controller coefficients, respectively. As a result, the rotor frequency dynamics in Eq. (6) can be rewritten by expressing the mechanical power for each generator in terms of frequency deviation ω_i, as defined in Eq. (7). It becomes:

$$M_i\dot{\omega}_i = -\left(K_i^P\omega_i + K_i^I\int_0^t \omega_i\right) - P_i^G - D_i^G\omega_i. \tag{8}$$

According to Eq. (3), we obtain:

$$M_i\dot{\omega}_i = -\left(K_i^P + D_i^G\right)\omega_i - K_i^I\delta_i - \sum_{j\in\mathscr{G}} B_{ij}\left(\delta_i - \delta_j\right) - \sum_{j\in\mathscr{L}} B_{ij}\left(\delta_i - \theta_j\right). \tag{9}$$

In this way, expressions (5), (4), (9) formulate the complete dynamical model and can be written as the following linear state-space descriptor system:

$$\begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M \end{bmatrix}\begin{bmatrix} \dot{\delta} \\ \dot{\theta} \\ \dot{\omega} \end{bmatrix} = \begin{bmatrix} 0 & I & 0 \\ B^{LG} & B^{LL} & 0 \\ -\left(K^I + B^{GG}\right) & -B^{GL} & -\left(K^P + D^G\right) \end{bmatrix}\begin{bmatrix} \delta \\ \theta \\ \omega \end{bmatrix} + \begin{bmatrix} 0 \\ P^L \\ 0 \end{bmatrix}, \tag{10}$$

where B is the imaginary part of the admittance matrix:

$$B_{bus} = \begin{bmatrix} B^{GG} & B^{GL} \\ B^{LG} & B^{LL} \end{bmatrix}. \tag{11}$$

Now, we consider a single-point closed-loop D-LAA that is performed at victim load bus v and the frequency sensor is installed at a generator bus s aiming to push this particular generator off-line. Suppose a proportional-integral controller is used by the attacker, creating a large deviation while less load is needed. Let K_p^L and K_I^L denote the attack controller's proportional and integral gains at the generator bus (sensor bus) s, respectively. We can write the compromised power consumption level $\overline{P}_v^L$ at victim bus v:

$$\overline{P}_v^L = P_v^L - K_p^L\omega_s - K_I^L\int_0^t \omega_s. \tag{12}$$

As a result, the system dynamics subjects to the above D-LAA becomes

$$\begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & M \end{bmatrix}\begin{bmatrix} \dot{\delta} \\ \dot{\theta} \\ \dot{\omega} \end{bmatrix} = \begin{bmatrix} 0 & I & 0 \\ B^{LG} - K_I^L & B^{LL} & -K_P^L \\ -\left(K^I + B^{GG}\right) & -B^{GL} & -\left(K^P + D^G\right) \end{bmatrix}\begin{bmatrix} \delta \\ \theta \\ \omega \end{bmatrix} + \begin{bmatrix} 0 \\ P^L \\ 0 \end{bmatrix}. \tag{13}$$

From Eq. (13), when the system is under attack, the attacker can affect the system dynamics and compromise the system stability by adjusting the attack controller matrices K_p^L and K_I^L. In particular, the system

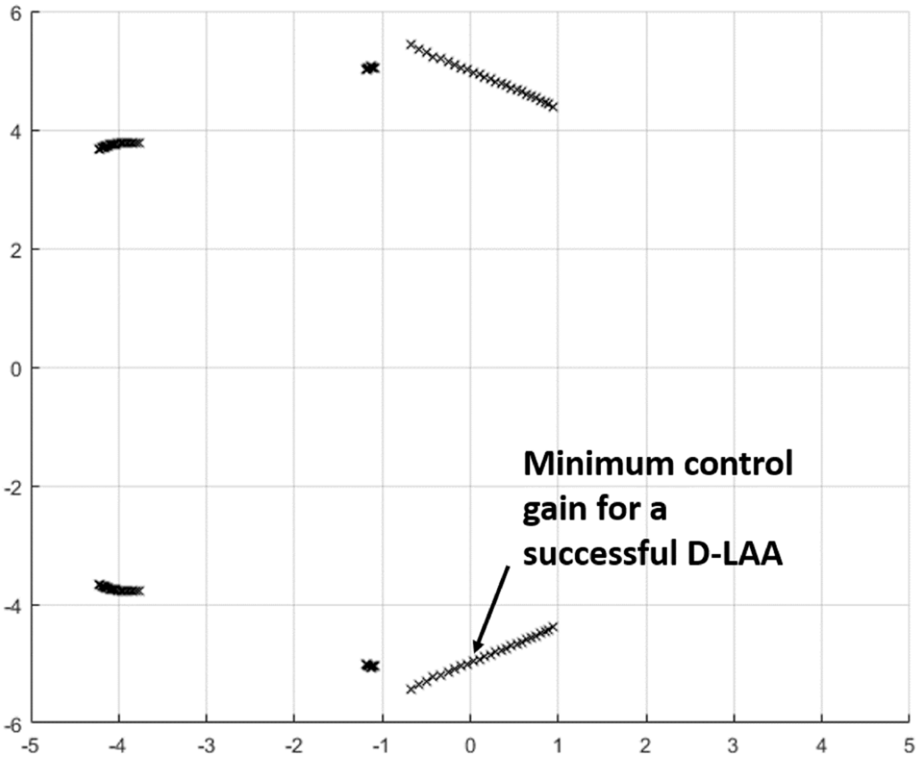

Fig. 2. Root locus plot of power system under a D-LAA attack.

becomes unstable if the attacker is capable of moving the system poles to the right-half complex plane assuming the vulnerable load is large enough. Considering generators are generally equipped with various relays in the modern power system [35], a single-point closed-loop D-LAA may push the generator off-line and disconnect it from the grid.

2.1.3. *Control Scheme of Closed-loop D-LAA*

It is worth mentioning that PI control in (7) is not the only option to successfully implement D-LAA. In general, the closed-loop D-LAA can be viewed as a frequency controller which makes the compromised loads react to frequency deviants in the opposite direction of the normal demand response for frequency regulation. It has been discussed in [15] that "*the attacker may use a bang-bang, P, PI, or PID controller, or any other more complex feedback control system mechanism*" for the closed-loop D-LAA, and a P control model is used in [15] to formulate the D-LAA. In this work, the P control based D-LAA attack model in [15] is extended to a more general PI control based model. The models provide effective methods for the modeling of malicious D-LAA actions against the frequency control of the grid. The proposed model keeps the direction of D-LAA actions opposite to the normal frequency control of the grid. Furthermore, the focus of this work is not to design attack controller but rather to develop a dynamic defense framework plan by minimax-q learning against such multistage attacks. This method and idea are not affected by the controller selection and can potentially be extended to other multistage attacks.

2.1.4. *D-LAA Implementation*

The D-LLA is launched by altering the remotely controllable demands instead of the outputs of the generation units in the grid. Referring to Fig. 1 the adversary only needs to hack into the remote load control systems to adjust the power consumption trajectory by constantly monitoring the frequency signals to implement D-LAA. Such remote load control systems extensively exist in demand response programs. Specifically, an attacker may aim to compromise command signals in Direct Load Control (DLC) programs that often involve two-way communications between the power system operator and loads or aggregators [36]. The adversary may utilize the vulnerability in any of these communications infrastructures to gain direct and remote access and control over the load through the load control mechanism. These loads that are potentially vulnerable to D-LAA attack include air conditioners [37], building lighting system [38], water heaters [39] and electric vehicles [40]. For example, considering the heating, ventilation and air conditioning (HVAC) demand in buildings, after intruding into the communication between the building and grid operator, the attacker could generate desired aggregated load profile by orchestrated periodic on/off signals to each component, e.g., air conditioner and fan.

To set the parameters in the controller-based model of D-LAA, the attacker needs to know or estimate the system dynamic model including the system frequency control settings and grid topology, which do not change frequently and are considered constant during the attack in this paper. The only signal that needs to be updated in real time by the attacker when implementing the closed-loop D-LAA is the frequency signal. Thus, following the work in [15], it is assumed in this paper that the attacker can constantly monitor the frequency signals via the attacker's installed sensors or by hacking into an existing monitoring infrastructure of the grid.

2.2. *Optimal Load Shedding*

As mentioned, when a system is attacked and the topological structure of the system is changed, such as a generator being off-line, a transmission line being tripped by relay/man or a system partition being caused, load shedding must be performed to regain stability. Considering a power system with n buses the optimal load shedding problem can be formulated as a constrained optimization problem with the physical constraints of the power flows [27,41,42]:

$$\min_{z_g, z_l} = \sum_{i=1}^{n} w_{li} z_{li}, \tag{14}$$

subject to,

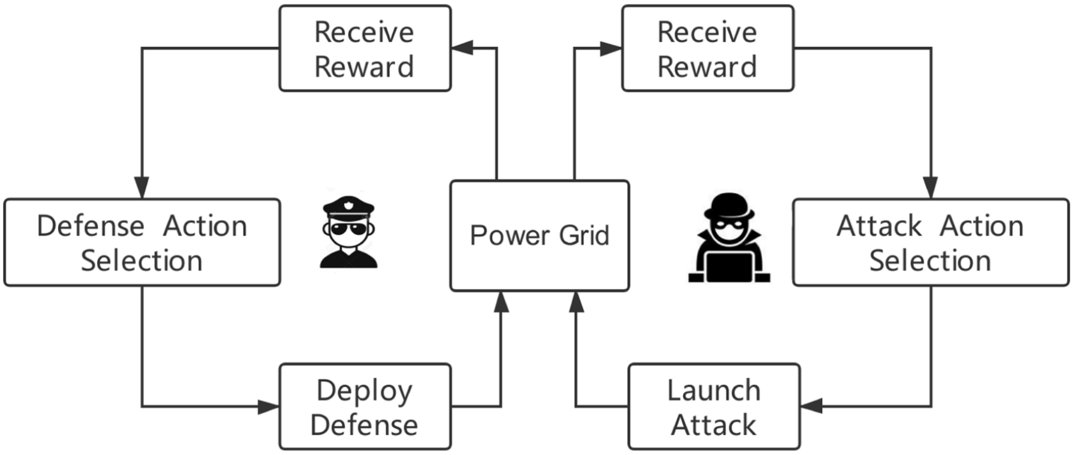

Fig. 3. The interaction between players and the system.

$$\Lambda' B \Lambda \phi - (p+z) = 0, \quad (15)$$

$$p_{gmin} \leqslant p_g + z_g \leqslant p_{gmax}, \quad (16)$$

$$z_{gmin} \leqslant z_g \leqslant z_{gmax}, \quad (17)$$

$$p_{lmin} \leqslant p_l - z_l \leqslant p_l, \quad (18)$$

$$\phi_{min} \leqslant \Lambda \phi \leqslant \phi_{max}, \quad (19)$$

where $w_l = [w_{l1}, w_{l2}, \ldots, w_{ln}]^T$ is the weight vector representing the relative importance of different load buses; vector $z = [z_g; z_l]$, in which vector z_g refers to the re-dispatched power at each generation bus; vector z_l is the load to be shed at each load bus; vector $p = [p_g; p_l]$, in which vector p_g represents the original active power output at each generation bus; vector p_l is the demand at each load bus; vector ϕ represents the phase angle at each bus; Λ is the incidence matrix for the topology of the grid; and B is the diagonal matrix of the transmission-line admittances. Constraint (15) represents the power balance at each bus; constraint (16) is the power output limit of the generation; constraints (17) and (18) are the constraints of the generation redispatch and load shedding respectively; Constraint (19) limits the phase angle difference of the connected buses of each transmission line in the grid.

2.3. Cascading Failures and D-LAA in Sequence

A successful single D-LAA with the aim to disconnect a generator may cause cascading failures during the post-attack stage. Due to the excessive load demand after attack, load shedding is an inevitable option for the system operator [11]. The optimal load shedding technology has been discussed and presented in Section 2.2. After the load shedding is carried out, a DC power flow analysis is performed to check for overloads on the transmission lines. If a transmission line is overloaded by over 50%, it will be tripped by the operator. Then the balance between the generation and demand is checked again and these steps are repeated until entering into the next steady state.

On the other hand, for causing a more severe damage to the power system such as more load shedding or generation losses, the attacker may perform the D-LAAs in sequence (one-by-one). For example, the attacker may perform a D-LAA to force a generator to be disconnected from the grid and trigger cascading failures. Then, based on the current state and system topology of the post-attack stage, the new proper victim and sensor buses are selected and another D-LAA can be performed. The attacker may repeat this process until the attack goal is achieved.

There is a main concern for the D-LAA sequence: *how to choose the best attack controller gain for each step?* For the ease of analysis, it is assumed the frequency sensor is placed at the generator bus, that is, the sensor bus is always one of the generator buses, and all portions of loads are controllable at each vulnerable load bus. As mentioned previously, the attacker may destabilize the system by changing the controller gain matrix K_P^L. From the control perspective, the locations of system poles change with the increase of K_P^L and once the pole(s) are moved to right-half plane the system becomes unstable. Fig. 2 shows how the root locus analysis helps the attacker find the minimum attack gain.

The minimum vulnerable load that must be compromised can be calculated by Eq. (12). If the minimum amount of load is not larger than the total load at this load bus, the selections of victim load bus and sensor bus are feasible. The attacker tries to make the least effort to achieve the attack goal, so for the same sensor bus when there are two feasible victim buses the attacker tends to choose the one with less minimum compromised load. Once a D-LAA is successfully performed and the cascading failures are triggered, the system enters into the next steady state and the attacker may choose the new feasible victim and sensor buses to conduct the next attack. In simulations, we can change the entries of matrix B in Eq. (13) based on the current system topology because $B_{ij} = 0$ if the transmission line between buses i and j is tripped.

3. Game-theoretic Analysis of Attack-defense Interactions

In this section, the behaviors of the attacker and defender are modeled using a two-player zero-sum multistage game. As introduced in Section 1, game theory helps people understand the interactions between the decision-makers. For the analysis of power system security, the attacker and defender are considered as two decision-makers or players. The attacker can be hackers, organized terrorists or other criminals. The defender is the system administrator who monitors the power system network and implements security measures. The attacker intends to cause the maximum damage to power grid while the defender strives to minimize the impact. Thus, the defender's gain is regarded as the opposite of the attacker's gain. In the attack, the adversary may compromise components in sequence instead of at the same time, in order to cause more damage and decrease the risk of being detected. Similarly, the defender has to change defense actions with a dynamic attacker. Thus, both the attacker and defender have to adjust their actions based on the observation of their past actions and current states. In this way, the attack-defense game falls exactly into the category of two-player zero-sum games.

This game can be considered as a 5-tuple $(\mathscr{S}^{\mathscr{S}}, \mathscr{A}^A, \mathscr{A}^D, \mathbf{R}^A, \mathbf{R}^D)$ Markov game, where.

- $\mathscr{S}^S \overset{def}{=} \{s_1, \ldots, s_{Ns}\}$ denotes the system's state space;
- $\mathscr{A}^A \overset{def}{=} \{a_1, \ldots, a_{N_A}\}$ denotes the attacker's action space;

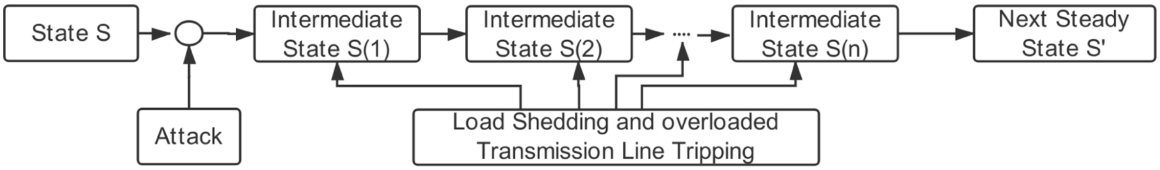

Fig. 4. Transition from one state to the next steady state.

- $\mathscr{A}^D \overset{def}{=} \{d_1, \dots, d_{N_D}\}$ denotes the defender's action space;
- $\mathbf{R}^A = [R^A_{a,d}(s)]_{N_A \times N_D}$ denotes the attacker's expected reward associated with attack action $a \in \mathscr{A}^A$ against defense action $d \in \mathscr{A}^D$ in state $s \in \mathscr{S}^S$; and
- $\mathbf{R}^D = [R^D_{a,d}(s)]_{N_A \times N_D}$ denotes the defender's expected reward associated with defense action $d \in \mathscr{A}^D$ against attack action $a \in \mathscr{A}^A$ in state $s \in \mathscr{S}^S$.

Fig. 3 illustrates a typical player-system interaction for the two-player game. The attacker obtains system state s and takes the attack action a, and will receive reward R_A. Meanwhile, the defender will conduct the same process and receive reward R_D.

3.1. Action Spaces

Attacker's target is to implement D-LAA aiming to disconnect one generator and cause cascading failures. Attacker's action $a \in \mathscr{A}^A$ means trying to force one generator disconnecting from the main grid at a time step. The defender's action is related to protect a generator bus. However, the defender can restrain this attack on generators by protecting the load on corresponding victim buses. As discussed in Section II-B, a successful D-LAA needs a victim bus and a sensor bus. The defender may follow the same method in Section II-B to obtain all vulnerable loads that can be potentially controlled by the attacker. Thus, protecting the victim load is an effective method against D-LAA. That is, the physical meaning of protection action is to protect the potential victim load rather than protecting these generators. Currently, the load can be protected by implementing reinforced security measures, e.g., adding hardware and software based security components, at both the communication level [43] and device level [44,45]. For example, reference[43] proposes a method in which the administrator can temporarily revoke the certificate of some nodes. In this way, these nodes are excluded from the grid's communication network, that is, the attacker is not able to remotely alter these loads.

In this paper, it is assumed that when a defense action is performed on a load bus, the load at the protected bus P^L_v cannot be manipulated by the attacker. Note that not all loads can be targeted by attacker to implement D-LAA because some loads are traditional types which cannot be remotely manipulated.

3.2. System States

This game is played over a finite state space denoted by $\mathscr{S}^S$. States are formulated as a combination of the statuses of all transmission lines of the power grid. For each state s, the status of the transmission lines is represented by a binary number '1' or '0'.

$$S_t\left(l\right) = \begin{cases} 1 & \text{ifline}l\text{worksproperly} \\ 0 & \text{ifline}l\text{isout}-\text{of}-\text{service}, \end{cases} \tag{20}$$

Fig. 4 shows the transition process from one state to the next steady state through some intermediate states. At state s, attack (D-LAA in this paper) is launched and one generator is pushed to be disconnected from the grid due to instability. The load is shed based on the model in Section 2.2 and the demand is balanced. Then, a DC power flow analysis is applied to decide if any overloads occur on the transmission lines. Generally, the transmission line is tripped if the overload exceeds 50%. The system repeats this process until the generation and demand is balanced and there is no overloaded transmission line. Please see Section 2.3 for more details.

3.3. Attacker and Defender's Policies and Rewards

There are two players in the game: the attacker and the defender. At state s, the players choose their respective actions $a \in \mathscr{A}^A$ and $d \in \mathscr{A}^D$ independently, and immediately receive rewards $R^A_{a,d}(s)$ and $R^D_{a,d}(s)$, respectively. In this zero-sum game, the defender's expected reward is opposite to the attacker's expected reward, denoted by $R^A_{a,d}(s) = -R^D_{a,d}(s)$. The rewards for players are assigned following the conditions given as:

$$R^A_{a,d}\left(s\right) = \begin{cases} 0 & \text{ifloadshedding} z(t) = z(t-1) \\ 1 & \text{if} z(t-1) < z(t) < N \\ 10 & \text{if} z(t) \geqslant N, \end{cases} \tag{21}$$

and

$$R^D_{a,d}\left(s\right) = \begin{cases} 0 & \text{ifloadshedding} z(t) = z(t-1) \\ -1 & \text{if} z(t-1) < z(t) < N \\ -10 & \text{if} z(t) \geqslant N, \end{cases} \tag{22}$$

where z represents the total load shedding caused by the D-LAA attack and N is the attack objective.

Now we have specified the immediate rewards of the attacker and the defender at each state, but have not indicated how these rewards are aggregated into an overall payoff. The most commonly used aggregation method is the discounted-sum reward. For an attack action a and a defense action d, the discounted-sum rewards of the attacker and defender considering deterministic state transition are represented as:

$$Q_A\left(s, a, d\right) = \sum_{t=0}^{\infty} \gamma^t R^A_{a,d}\left(s\left(t\right)\right), \tag{23}$$

$$Q_D\left(s, a, d\right) = \sum_{t=0}^{\infty} \gamma^t R^D_{a,d}\left(s\left(t\right)\right), \tag{24}$$

where Q_A and Q_D represent game values for the attacker and the defender, respectively; and $\gamma \in (0, 1)$ is the discount factor. A smaller value of γ implies the agent emphasizes the immediate reward while a larger value indicates more concerns about future rewards. For a given state s, the attacker's strategy is defined as probability distributions over action space $\mathscr{A}^A$, i.e.,

$$\pi^A\left(s\right) = [Pr(a(s) = a_1), Pr(a(s) = a_2), \dots, Pr(a(s) = a_{N_A})]^T, \tag{25}$$

which satisfies $\sum_{i=1}^{N_A} Pr(a(s) = a_i) = 1 \Big| a_i \in \mathscr{A}^A$. Similarly, the defender's strategy is given as:

$$\pi^D\left(s\right) = [Pr(d(s) = d_1), Pr(d(s) = d_2), \dots, Pr(d(s) = d_{N_D})]^T \tag{26}$$

and $\sum_{i=1}^{N_D} Pr(d(s) = d_i) = 1 \Big| d_i \in \mathscr{A}^D$.

When only one entry of the strategies described above is nonzero (and equal to 1), π^A and π^D are called pure strategy and players always adopt this action at state $s(t)$. Otherwise, they are mixed strategies which are adopted in this paper. Note that in this multi-stage game, the attacker and defender choose their different targets in time sequence until the attack objective is achieved.

3.4. Nash Equilibrium

In this game, the defender tries to minimize the discounted sum of expected reward Q_D while the attacker aims to maximize it. *Nash equilibrium* is a common solution to solve the players' optimal strategies for such a Markov game [17,46]. Nash equilibrium is a state that no player has a unilateral incentive to change actions as that would reduce their rewards, that is, each agent plays best response to their opponents. For the proposed game model, a Nash equilibrium can be mathematically defined as follows:

Definition 1. *In the proposed zero-sum two-player Markov game, a Nash equilibrium is a pair of mixed optimal strategies* (π_A^*, π_D^*) *for all mixed strategies* π_A *and* π_D *for all states* $s \in S$

$$Q_A(s, \pi_A^*, \pi_D^*) \geqslant Q_A(s, \pi_A, \pi_D^*), \tag{27}$$

$$Q_D(s, \pi_A^*, \pi_D^*) \geqslant Q_D(s, \pi_A^*, \pi_D). \tag{28}$$

For such a two-player game, it is proved that unique Nash Equilibrium exists in stationary strategies (for all t) by Shapley [47]. That is, the mixed optimal attack/defense strategies can be solved for each state instead of each time t. In general, the stationary optimal strategy can be solved recursively by dynamic programming if environment of the model is known such as in [27,28].

4. Proposed Solution Approach

4.1. Minimax-q Learning

In this section, we propose a new dynamic defense solution for the two-player zero-sum Markov game based on the minimax-q learning approach. Our objective is to characterize the attacker's and the defender's Nash equilibrium strategies for each state $s \in \mathscr{S}^S$ and their attack/defense actions in time sequence, where all players are rational and tend to maximize their own benefits. The attacker and the defender are completely competitive and do not cooperate with each other. Minimax-q learning [48] is used in conjunction with Markov games. As a modification of q-learning which just considers the opponent as part of the environment, this algorithm treats the Q function not just from the state/action pairs to values, but from the state/action/action to values, i.e., $Q(s, a, d)$. Thus, both players' actions and their interactions are modeled more explicitly. The minimax-q learning algorithm is adopted to approach real unknown *state value function* $V(s)$ by interacting with the environment and then players obtain the optimal Nash strategies by the learned state value function. Furthermore, a *state-action value function* (Q function) is to quantify the performance for a player to apply a particular action following a policy π in a state. From the defender's perspective, the Q function can be defined as:

$$Q_D\left(s,a,d\right) = \left(1-\alpha_t\right)Q_D\left(s,a,d\right) + \alpha_t\left(R^D + \gamma\sum_{s'\in S} V_D\left(s'\right)\right), \tag{29}$$

and the state value function $V(s)$ is defined as:

$$V_D\left(s'\right) = \min_{\pi_D}\max_{a}\sum_{a} Q\left(s',a,d\right)\pi_D\left(s'\right), \tag{30}$$

where α_t denotes the learning rate for adjusting the step size. To improve the convergence rate of this algorithm, a polynomial learning rate is adopted as $1/t^\beta$ where $\beta \in (1/2, 1)$.

Note that in Eq. (30), minmax is adopted to find the best response instead of playing actions with the highest Q in [29]. Eq. (30) can be converted to a linear constraint optimization problem to obtain the optimal strategy at state s:

$$\begin{gathered}\min_{\pi_D} V_D\left(s\right), \\ s.t. V_D\left(s\right) \geqslant \sum_{d} Q\left(s,a,d\right)\pi_D\left(s\right), \forall a \in \mathscr{A}^A.\end{gathered} \tag{31}$$

Similarly, the attacker's state value function and Q function can be dually derived:

$$Q_A\left(s,a,d\right) = \left(1-\alpha_t\right)Q_A\left(s,a,d\right) + \alpha_t\left(R^A + \gamma\sum_{s'\in S} V_A\left(s'\right)\right), \tag{32}$$

$$V_A\left(s'\right) = \min_{\pi_A}\max_{d}\sum_{d} Q\left(s',a,d\right)\pi_A\left(s'\right). \tag{33}$$

The optimal strategy of the attacker can also be obtained by linear programming in (33):

$$\begin{gathered}\max_{\pi_A} V_A\left(s\right), \\ s.t. V_A\left(s\right) \geqslant \sum_{a} Q\left(s,a,d\right)\pi_A\left(s\right), \forall d \in \mathscr{A}^D.\end{gathered} \tag{34}$$

The procedure to compute the Nash equilibrium at each state and the attack/defense sequence are detailed in Algorithm 1.

Algorithm 1. Minimax-q Learning Algorithm

```
Initialize $Q_0(s,a,d)$, $V(s)$, $\pi_A$, and $\pi_D$
Obtain feasible D-LAA target for initial state discussed in Section 2.3 as action space for attack/defense
Define exploration probability $\epsilon$ and learning rate $\alpha$
for number of episodes do
  while Attack objective is not reached do
    Select current state $s$
    if Generated random number $< \epsilon$ then
      Take random attack and defense action
    else
      Take attack and defense action based on Q-table
    end if
    Execute actions
    Calculate load shedding by (14), overloads, and cascades
    Determine next $s'$
    Assign reward by Eqs. (21) and (22)
    Update state-action value function $Q$ by Eqs. (29) and (32)
    Solve state value functions (30) and (33) by linear programming and update $V(s)$ and $\pi_A(s)\pi_D(s)$
    Update feasible D-LAA target for attack/defense's action space
    Update $s = s'$
  end while
end for
Find optimal strategies and sequences of actions for attacker and defender
```

In the proposed algorithm, the game starts with the initialization of Q function, state value function V and attacker/defender's policy. Then the system is evaluated to obtain feasible D-LAA attack discussed in Section 2.3. Note that for simulations, instead of observing the root-locus plot, we can analytically obtain the minimum compromised load by gradually increasing controller K_P^L until the system is unstable. In the beginning, the initial state is assumed at the normal operation condition, that is, all transmission lines and generators are active and work properly. A ϵ-*greedy* strategy is also adopted to balance the exploration and exploitation [49]. With ϵ-*greedy*, the agent plays a random action with a probability $0 < \epsilon < 1$, instead of making the best decision given in the Q-function. With the execution of the actions, a certain generator is disconnected from the power grid due to the D-LAA if the corresponding

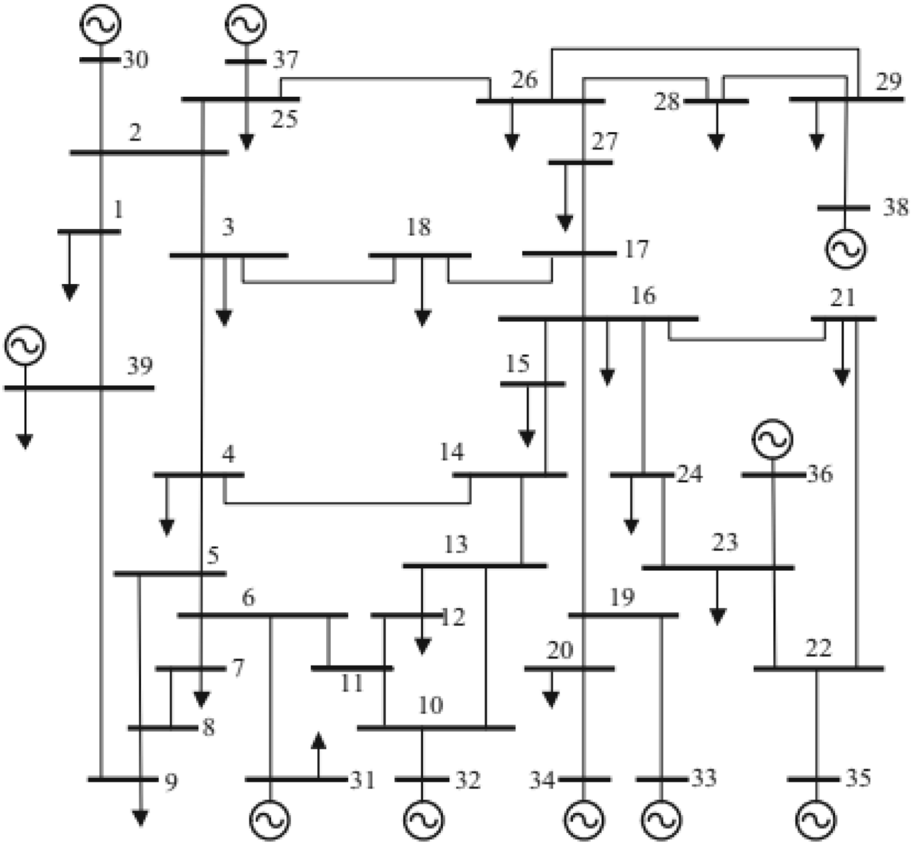

Fig. 5. IEEE 39-bus system.

load is not protected by the defender. Load shedding of the current state is calculated and cascading overloads on the transmission lines may be triggered until the system enters into the next steady state s'. Instant rewards are assigned to the attacker and the defender by Eqs. (21) and (22) and the value of Q-function is updated. Strategies $\pi_A(s)$ and $\pi_D(s)$ and state value V are solved by linear programming. Then, based on the new topology of system and state, feasible targets for the attack/defense's action spaces are decided. The game is repeated until the attack objective is reached. Ideally, if the process above (from step 4 to step 20) repeats for enough times, i.e., Q matrix is updated at each state by enough times, the players will learn the real complex relationships between the actions and outcomes. Thus, such relationships are reinforced in this process and eventually the players find their optimal Nash equilibrium strategies. Note that the optimal attack/defense sequence is not unique and in this study we evaluate the performance by computing the average impact to the system.

Furthermore, this defense strategy is not a real-time one but is more like a pre-stipulated plan against low-probability high-impact attacks, such us D-LAA in this paper, to minimize the damage. According to the features of D-LAA attacks, the defender can find an optimal policy for each state against potential vicious attacker if both play rationally. The defense strategy can be deployed in advance when such attacks are anticipated to improve the resilience of the power system.

In this paper, the proposed work focuses only on the D-LAA scenario. Nevertheless, because the proposed defense method is based on minimax-q learning which works with Markov game, this work can be extended to other multistage attacks, e.g., Load Redistribution (LR) attacks and line switching attacks, as long as the action spaces and cascading failures are redesigned following specific attack mechanisms. The power grid operator may make multiple such stored plans based on different types of attacks and defense actions.

4.2. Discussion on Computational Complexity

The computational complexity is $\mathcal{O}(S^2 M_A M_D)$ per iteration in Algorithm 1, where M_A and M_D are the numbers of strategies for the attacker and the defender, respectively. Because single-point D-LAA is considered in this paper, i.e., the attacker and the defender select one bus to compromise and protect at one time, $M_A = \binom{1}{A}$ and $M_D = \binom{1}{D}$, where A and D represent the numbers of total possible attack and defense actions, respectively. It can be seen that the computational complexity increases linearly with more attack/defense options. As for S, more possible states will cause relatively quicker increase of the computational complexity.

5. Simulation Results and Analysis

Now we evaluate the performance of the proposed minimax-q learning for this two-player zero-sum Markov game on the IEEE 39-bus system that consists of 46 transmission lines and 10 generators. The results of dynamic defense strategy may provide useful insight for grid operators to improve the resiliency of power systems. Comparisons with the existing passive and dynamic defense strategies are conducted to illustrate the importance of deploying the proposed dynamic strategy against D-LAA.

5.1. System Parameters

Fig. 5 shows the IEEE 39-bus system based on a 10-machine New-England power network. There are 10 generators, 46 transmission lines and 19 loads. There are two loops in the simulation: episodes and runs. The episodes loop is the main loop in which the attacker and the defender interact to learn the optimal policy. The attacker and the

Table 1
Simulation Parameters for IEEE 39-bus system

No.	Parameter	Value
1	Number of Generators, $\mathscr{G}$	10
2	Total Transmission lines	46
3	Discount Factor, γ	0.8
4	Learning Rate Coefficient, β	0.7
5	Initial Exploration Probability, ϵ	0.9
6	Number of Episodes	1000
7	Number of Runs	50
8	Maximum Iteration per Episode	100
9	Total Capacity	6245 MW
10	Attack Objective	at least 50% load shedding

Table 2
Minimum portion of vulnerable load that must be compromised at initial state.

	Sensor Bus					
Victim Bus	**30**	**32**	**33**	**34**	**36**	**39**
4	62	92.5	79.1	69	125	46.2
6	4.9	0.91	1.2	3.7	3.6	128
7	0.72	12.4	0.6	64.5	5.1	5.1
12	73.9	23.5	48.6	77.2	89.5	89
18	146	8.5	117	222	189	46.5
19	48.1	0.77	7.4	1.5	0.62	66.8
23	280	1.9	15.6	2.8	1.9	72
29	4.6	58.5	12.7	4.7	4	0.54

defender complete a bunch of actions in sequence. As the number of episodes increases, the attacker tends to approach the optimal policy. At the end of the episodes, the attacker and the defender reach the Nash equilibrium point. A number of runs are conducted to deduce different Nash equilibria. Therefore, the whole game simulation is conducted for many runs. Each run includes a number of episodes. The number of episodes is the required number of trials for the agent in the learning process. The initial exploration rate ϵ is 0.9 and decreases 10% every 20 episodes to ensure the convergence. Other simulation parameters are given in Table 1.

5.2. Selection of Vulnerable Bus and Attacker/defender's Action Space

As discussed in Section II, not all loads can be considered vulnerable to D-LAAs. Some loads are traditional ones and may not even have smart meters or any demand response equipment, which the attacker cannot remotely manipulate. In this case, we assume that only eight load buses have vulnerable loads. They can potentially become victim buses, i.e., $\mathscr{V} = \{4, 6, 7, 12, 18, 19, 23, 29\}$. On the other hand, according to [50], generators $\{31, 35, 37, 38\}$ represent nuclear stations which are fully protected. Thus, frequency sensors are assumed to be placed only at $\mathscr{S} = \{30, 32, 33, 34, 36, 39\}$ that are considered as fossil and hydro stations. Thus, the attacker's action space is $\mathscr{A}^A = \{30, 32, 33, 34, 36, 39\}$. Table 2 shows the minimum portion of the vulnerable load that must be compromised to guarantee a successful D-LAA at the initial state. We assume K_I^L a pre-tuned parameter and there is no need to change it for simplicity of calculation. The highlighted cells indicate the attacker could launch D-LAA on the corresponding sensor bus and victim bus. For the initial state, the attacker is not able to compromise generator 34 because there are not enough loads to be manipulated for the given vulnerable buses. Therefore, for the initial state, the attacker can perform D-LAA to disconnect generators $\{30, 32, 33, 36, 39\}$. Furthermore, at each visited unique state, Table 2 is updated for the next selection of the attack target. Based on the same table, the defender also decides the protection action that should be taken. The defender's action space is denoted as $\mathscr{A}^D = \{30, 32, 33, 34, 36, 39\}$. As mentioned previously, the physical meaning of the protection action is not to protect these generators but to protect the corresponding potential victim load. For example, at the initial state, when the defender selects action "30", it means the load on the corresponding victim bus 7 is protected.

5.3. Game-theoretic Attack/Defense

Figs. 6 and 7 show the convergence curves of the optimal number of attacker/defender's actions. After adequate learning and exploration, we can see that both players reach the optimal number of actions. Among 50 independent runs, the attacker needs three actions in sequence to achieve the objective and the defender also needs the same number of actions to minimize the load shedding caused by D-LAAs. The average computing time per run is about 564s. We need to emphasize again that the defense strategy is not real-time but off-line trained prestipulated plan against the low-probability high-impact D-LAA. The defense plan can be deployed in advance when such attacks are anticipated. Therefore, the proposed strategy can adequately meet the time requirement of practical applications.

Fig. 8 depicts the convergence of the total load shedding. The average load shedding converges to around 3400 MW after 50 runs. From the three figures, we can portray how the algorithm works especially in the intermediate process. In the early stage of learning process,

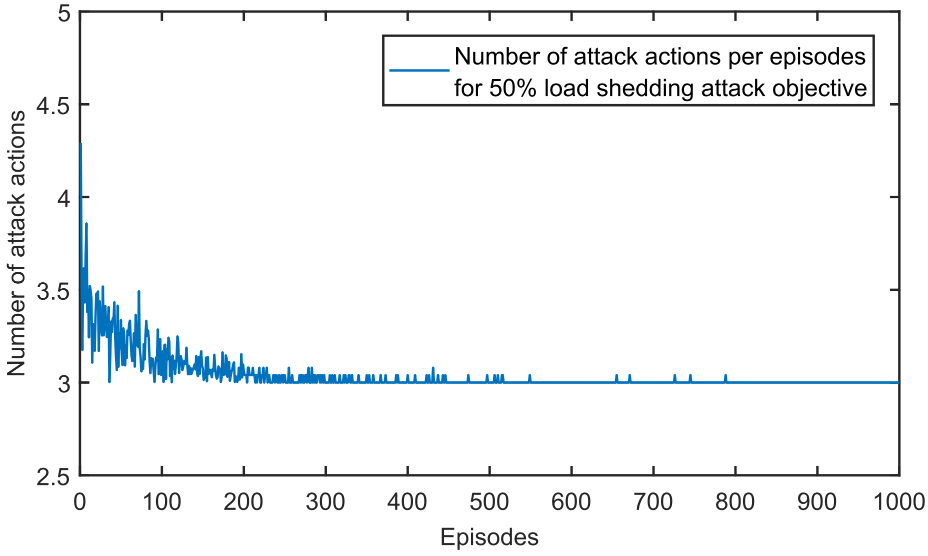

Fig. 6. Convergence of the defender's number of actions.

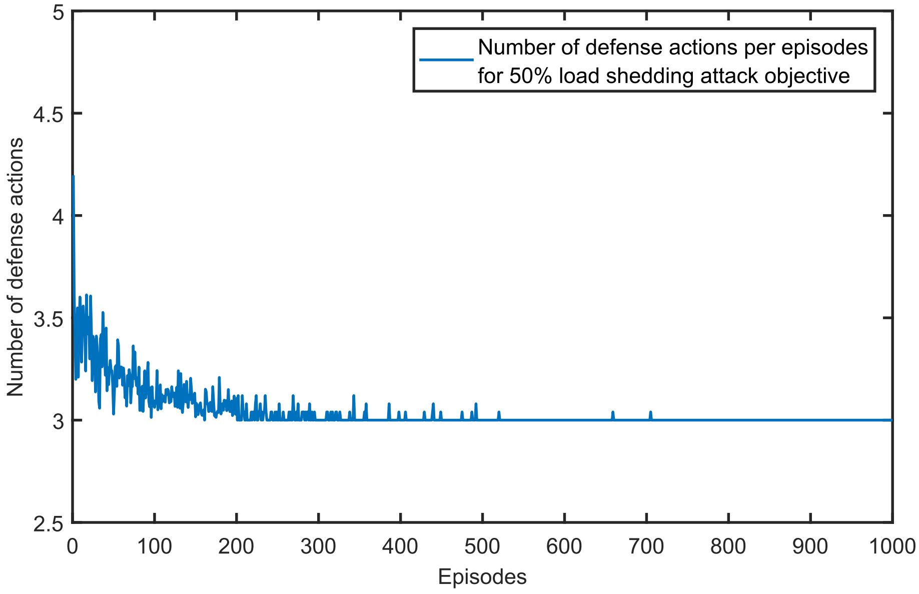

Fig. 7. Convergence of the attacker's number of actions.

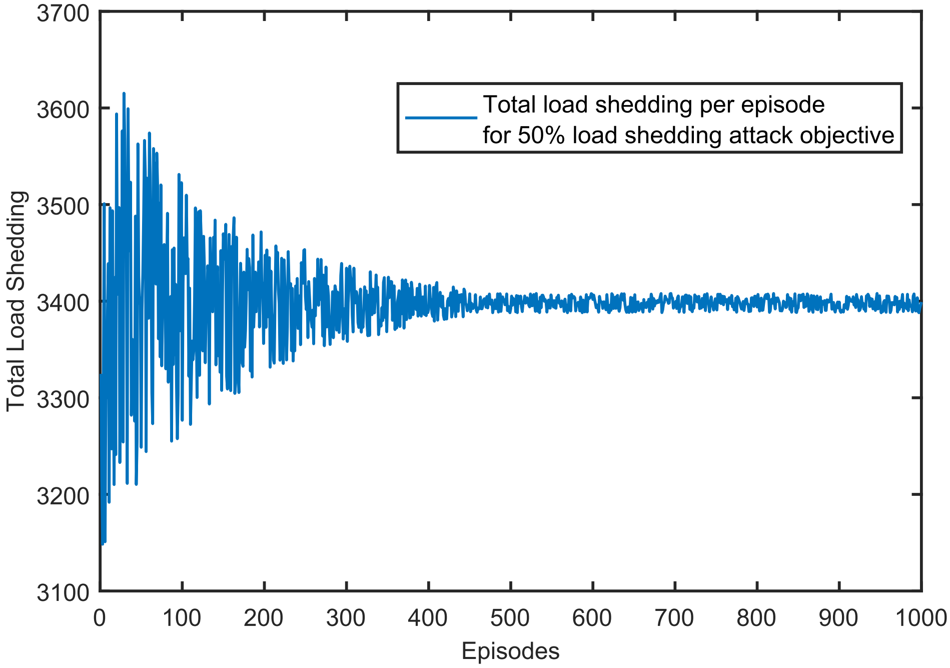

Fig. 8. Convergence of the total load shedding.

because of the large exploration rate ∊ and inaccurate Q matrix, both the attacker and defender take actions randomly or wrongly. Thus, they may take more additional steps and the total amounts of load shedding is not stable. With the decrease of ∊ and the update of Q matrix, the curves gradually converge. At the end of the process, the Q matrix is updated for enough times and the players learn the real relationships between the actions and outcomes. The policies at each state converge to the optimal Nash Equilibrium strategies.

Figs. 9 and 10 show the attacker's and defender's optimal policies when Nash Equilibrium is reached at each unique state. For this situation, both the attacker and defender have no unilateral incentive to alter their actions, because they have maximized their profits. The physical meaning of the Nash Equilibrium status for this case is that the defender can minimize the damage (load shedding) if they both play rationally their optimal strategies. The system operators are advised to adopt these strategies for each possible state against the D-LAA. Specifically, there is no need to place any defensive strategies for some states because there are not enough vulnerable loads to alter for disconnecting generators from the power grid. Note that mixed strategy at some states. Regarding the actual implementation in practice, the operator may change the defense plan according to the probabilities of the optimal policy. For instance, at state 9, the probabilities of defender's action on generators (30, 32, 33, 34, 36, 39) are (0, 0.165, 0, 0.835, 0, 0), respectively. Thus, the system administrator may plan to take protective actions for bus 34 with a probability of 0.835 and protect bus 32 with a probability of 0.165 at each interval of the actions. The results provide useful information for power system operators to thwart dynamic load altering attacks.

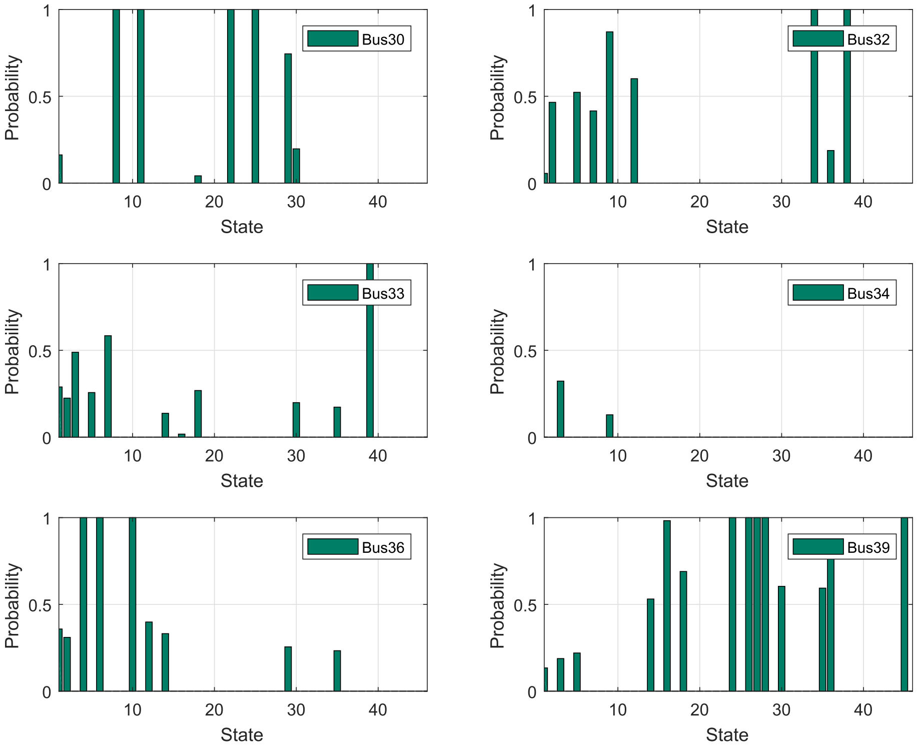

Fig. 9. Probability of attacker's action at each state.

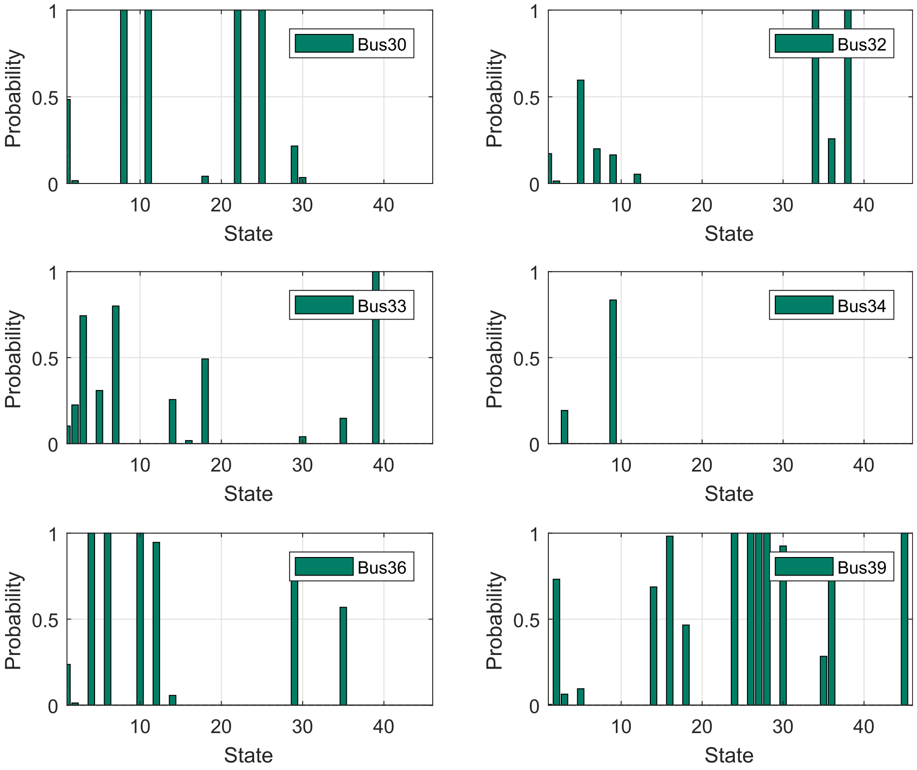

Fig. 10. Probability of defender's action at each state.

The defensive action sequences are shown in Table 3. For this game, total 11 unique defense sequences are found and the average load shedding can be calculated as 3398.2 MW. As mentioned in Section 5.2, the defense action sequence indicates the protected generators identified by the defender while the actual actions of the defense strategy are to protect the corresponding victim load buses. For example, for the first run, this action sequence indicates that the defender tries to protect the loads on victim buses (7, 19, 23).

Table 3
Defender’s action sequences for dynamic defense strategy.

Runs	Defense sequence	Physical meaning (protected buses)	Total load shedding (MW)
1	(33, 39, 30)	(7, 19, 23)	3421.1
2	(30, 39, 32)	(7, 29, 4)	3709.6
3	(39, 32, 34)	(29, 12, 18)	3315.2
.. . …	. .. …	. .. …	. .. …
25	(39, 30, 32)	(29, 19, 12)	4005
26	(33, 32, 30)	(7, 29, 6)	3321.9
.. . …	. .. …	. .. …	. .. …
49	(36, 34, 33)	(19, 18, 6)	3200.2
50	(39, 32, 34)	(29, 12, 18)	3315.6

Table 4
Attacker’s action sequences for passive defense strategy I.

Runs	Attack action sequence	Total load shedding (MW)
1	(32, 39, 36)	3856.4
2	(32, 39, 36)	3856.4
3	(32, 36, 30)	3725
.. . …	. .. …	. .. …
25	(32, 39, 36)	3856.4
26	(32, 39, 36)	3856.4
.. . …	. .. …	. .. …
49	(32, 36, 30)	3725
50	(32, 39, 36)	3856.4

Table 5
Attacker’s action sequences for passive defense strategy II.

Runs	Attack action sequence	Total load shedding (MW)
1	(32, 36, 34)	3564.7
2	(32, 34, 36)	3649.2
3	(32, 36, 34)	3564.7
.. . …	. .. …	. .. …
25	(32, 30, 34)	3425.7
26	(32, 34, 36)	3649.2
.. . …	. .. …	. .. …
49	(32, 36, 34)	3564.7
50	(32, 34, 36)	3649.2

Table 6
Attacker’s action sequences for passive defense strategy III.

Runs	Attack action sequence	Total load shedding (MW)
1	(36, 39, 30)	3992
2	(39, 32, 36)	3710.3
3	(36, 39, 30)	3992
.. . …	. .. …	. .. …
25	(36, 39, 30)	3992
26	(36, 39, 30)	3992
.. . …	. .. …	. .. …
49	(39, 32, 36)	3710.3
50	(36, 39, 30)	3992

5.4. Comparison with Passive Defense Strategy

To illustrate the importance of dynamic defense strategy, we compare our results with the passive defense strategy in this section. For a passive strategy, the defensive actions are predefined and the attacker is trained to find the optimal attack strategy in the presence of the passive defender. Considering the limited resources the operator has, we assume only two loads can be protected at a time. In this case study, three different predefined protected load sets are considered: (7, 29), (4, 29) and (6, 7). They are denoted as passive defense case I, II and III respectively. We adopt a similar algorithm by calculating the largest value instead of solving minimax in Eq. (33). The attack objective is to cause at least 50% load shedding.

Table 7
Total load shedding of different defense strategies.

Proposed dynamic defense	3398.2 MW
Passive defense I	3827.5 MW
Passive defense II	3601.9 MW
Passive defense III	3894.5 MW
Dynamic defense in [51]	3709.6 MW

Because the defense strategy is passive and unchangeable, we analyze the performance from the attacker’s perspective. Table 4–6 show the attack sequences of different runs and the total load shedding. It is found in Table 7 that the attacker’s action converges to a sequence of three actions, and the total amounts of load shedding for the passive defense I, II and III are 12.6%, 6.0% and 14.6% more than that obtained by the dynamic defense strategy, respectively. The comparison shows the proposed dynamic defense method is more effective against the single point D-LAA.

5.5. Comparison with Dynamic Defense Strategy

In this section, the proposed dynamic defense strategy in this paper is compared with the dynamic strategy in [51]. In [51], the dynamic defense strategy is obtained by the pre-calculated worst-case dynamic attack, which ignores the adversarial game between the rational attacker and defender, and their future expected gains. This is the main difference between the proposed models in this paper and [51]. To compare by same standards, the attack objective is still at least 50% load shedding. The last row of Table 7 shows that the total amount of load shedding by applying the defense plan in [51] is 9.2% higher than that obtained by the proposed strategy in this paper. One reason of the result is that the outcome of two players’ game, i.e., the attacker and defender, is not always the best for one of them but inclines a Nash equilibrium mentioned in Section 3.4. Thus, the defense strategy derived by the worst-case dynamic attack, i.e., unilateral optimal attack, results in the worse outcome because the interaction between two rational players in each state of the Markov game for D-LAA is not considered. In general, the proposed model formulates a more complex and realistic game considering two rational players’ game, which leads to better performance for the defense against D-LAA.

6. Concluding remarks

In this paper, we propose a novel reinforcement-learning-based dynamic defense solution against the single point D-LAA in power grid, where considering the attacker/defender’s action sequence. We have derived the D-LAA in time sequence considering cascading failures at each state. A two-player zero-sum Markov game is formulated to analyze the complex interactions between the attacker and the defender, in which all players are rational and tend to maximize their own benefits. The proposed minimax-q algorithm is applied to derive the attacker/defender’s Nash equilibrium strategies. The IEEE 39-bus system is used to test the proposed algorithm and evaluate the dynamic defense strategy against D-LAA. Simulation results are compared with the existing passive and dynamic defense strategies, which indicates the proposed dynamic strategy exhibits a better performance. The system operator is informed to enforce the optimal dynamic defense strategy at each state in advance to improve the power system resiliency. In future work, distributed algorithms will be developed to further enhance the effectiveness of the defense strategy, such as the learning automata including linear reward-inaction and linear reward-penalty.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the U.S. National Science Foundation under Award ECCS1711617.

References

[1] Metke AR, Ekl RL. Security technology for smart grid networks. IEEE Transactions on Smart Grid 2010;1(1):99–107. https://doi.org/10.1109/TSG.2010.2046347.

[2] Moslehi K, Kumar R. A reliability perspective of the smart grid. IEEE Transactions on Smart Grid 2010;1(1):57–64. https://doi.org/10.1109/TSG.2010.2046346.

[3] Y. Guo, L. Wang, Cybersecurity analysis and improvement of bilinear systems against false data injection attacks, in: Proc. IEEE Power Energy Society Innovative Smart Grid Technologies Conf. (ISGT), 2020, pp. 1–5. doi:10.1109/ISGT45199.2020.9087740.

[4] Liang G, Weller SR, Zhao J, Luo F, Dong ZY. The 2015 Ukraine blackout: Implications for false data injection attacks. IEEE Trans. Power Syst. 2017;32(4): 3317–8. https://doi.org/10.1109/TPWRS.2016.2631891.

[5] NERC, Risc report on resilience, Tech. rep., NERC (2018).

[6] G.S. Ledva, S. Peterson, J.L. Mathieu, Benchmarking of aggregate residential load models used for demand response, in: Proc. IEEE Power Energy Society General Meeting (PESGM), 2018, pp. 1–5. doi:10.1109/PESGM.2018.8585847.

[7] Molina-Garcia A, Bouffard F, Kirschen DS. Decentralized demand-side contribution to primary frequency control. IEEE Trans. Power Syst. 2011;26(1):411–9. https://doi.org/10.1109/TPWRS.2010.2048223.

[8] Zeng W, Zhang Y, Chow M. Resilient distributed energy management subject to unexpected misbehaving generation units. IEEE Trans. Industr. Inf. 2017;13(1): 208–16. https://doi.org/10.1109/TII.2015.2496228.

[9] Mortaji H, Ow SH, Moghavvemi M, Almurib HAF. Load shedding and smart-direct load control using internet of things in smart grid demand response management. IEEE Trans. Ind. Appl. 2017;53(6):5155–63. https://doi.org/10.1109/TIA.2017.2740832.

[10] Haring TW, Mathieu JL, Andersson G. Comparing centralized and decentralized contract design enabling direct load control for reserves. IEEE Trans. Power Syst. 2016;31(3):2044–54. https://doi.org/10.1109/TPWRS.2015.2458302.

[11] Mohsenian-Rad A, Leon-Garcia A. Distributed internet-based load altering attacks against smart power grids. IEEE Transactions on Smart Grid 2011;2(4):667–74. https://doi.org/10.1109/TSG.2011.2160297.

[12] Marnerides AK, Smith P, Schaeffer-Filho A, Mauthe A. Power consumption profiling using energy time-frequency distributions in smart grids. IEEE Commun. Lett. 2015;19(1):46–9. https://doi.org/10.1109/LCOMM.2014.2371035.

[13] Mellucci C, Menon PP, Edwards C, Ferrara A. Load alteration fault detection and reconstruction in power networks modelled in semi-explicit differential algebraic equation form. In: Proc. American Control Conf. (ACC); 2015. p. 5836–41. https://doi.org/10.1109/ACC.2015.7172254.

[14] Pan T, Mishra S, Nguyen LN, Lee G, Kang J, Seo J, Thai MT. Threat from being social: Vulnerability analysis of social network coupled smart grid. IEEE Access 2017;5:16774–83. https://doi.org/10.1109/ACCESS.2017.2738565.

[15] Amini S, Pasqualetti F, Mohsenian-Rad H. Dynamic load altering attacks against power system stability: Attack models and protection schemes. IEEE Transactions on Smart Grid 2018;9(4):2862–72. https://doi.org/10.1109/TSG.2016.2622686.

[16] Di Giorgio A, Giuseppi A, Liberati F, Ornatelli A, Rabezzano A, Celsi LR. On the optimization of energy storage system placement for protecting power transmission grids against dynamic load altering attacks. In: Proc. 25th Mediterranean Conf. Control and Automation (MED); 2017. p. 986–92. https://doi.org/10.1109/MED.2017.7984247.

[17] Osborne MJ. An Introduction to Game Theory. Oxford University Press; 2004.

[18] T. Alpcan, T. Basar, A game theoretic approach to decision and analysis in network intrusion detection, in: Proc. 42nd IEEE Int. Conf. Decision and Control (IEEE Cat. No.03CH37475), Vol. 3, 2003, pp. 2595–2600 Vol. 3. doi:10.1109/CDC.2003.1273013.

[19] Law YW, Alpcan T, Palaniswami M. Security games for risk minimization in automatic generation control. IEEE Trans. Power Syst. 2015;30(1):223–32. https://doi.org/10.1109/TPWRS.2014.2326403.

[20] Chen G, Dong ZY, Hill DJ, Xue YS. Exploring reliable strategies for defending power systems against targeted attacks. IEEE Trans. Power Syst. 2011;26(3): 1000–9. https://doi.org/10.1109/TPWRS.2010.2078524.

[21] Zhu Q, Basar T. Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: Games-in-games principle for optimal cross-layer resilient control systems. IEEE Control Syst. Mag. 2015;35(1):46–65. https://doi.org/10.1109/MCS.2014.2364710.

[22] Farraj A, Hammad E, Daoud AA, Kundur D. A game-theoretic analysis of cyber switching attacks and mitigation in smart grid systems. IEEE Transactions on Smart Grid 2016;7(4):1846–55. https://doi.org/10.1109/TSG.2015.2440095.

[23] Chen P, Cheng S, Chen K. Smart attacks in smart grid communication networks. IEEE Commun. Mag. 2012;50(8):24–9. https://doi.org/10.1109/MCOM.2012.6257523.

[24] Esmalifalak M, Shi G, Han Z, Song L. Bad data injection attack and defense in electricity market using game theory study. IEEE Transactions on Smart Grid 2013; 4(1):160–9. https://doi.org/10.1109/TSG.2012.2224391.

[25] Li Y, Shi L, Cheng P, Chen J, Quevedo DE. Jamming attacks on remote state estimation in cyber-physical systems: A game-theoretic approach. IEEE Trans. Autom. Control 2015;60(10):2831–6. https://doi.org/10.1109/TAC.2015.2461851.

[26] Z. Ni, S. Paul, X. Zhong, Q. Wei, A reinforcement learning approach for sequential decision-making process of attacks in smart grid, in: Proc. IEEE Symp. Series Computational Intelligence (SSCI), 2017, pp. 1–8. doi:10.1109/SSCI.2017.8285291.

[27] Wei L, Sarwat AI, Saad W, Biswas S. Stochastic games for power grid protection against coordinated cyber-physical attacks. IEEE Transactions on Smart Grid 2018; 9(2):684–94. https://doi.org/10.1109/TSG.2016.2561266.

[28] Ma J, Liu Y, Song L, Han Z. Multiact dynamic game strategy for jamming attack in electricity market. IEEE Transactions on Smart Grid 2015;6(5):2273–82. https://doi.org/10.1109/TSG.2015.2400215.

[29] Ni Z, Paul S. A multistage game in smart grid security: A reinforcement learning solution. IEEE Transactions on Neural Networks and Learning Systems 2019:1–12. https://doi.org/10.1109/TNNLS.2018.2885530.

[30] He Y, Mendis GJ, Wei J. Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism. IEEE Transactions on Smart Grid 2017;8(5):2505–16. https://doi.org/10.1109/TSG.2017.2703842.

[31] Yan J, He H, Zhong X, Tang Y. Q-learning-based vulnerability analysis of smart grid against sequential topology attacks. IEEE Trans. Inf. Forensics Secur. 2017;12(1): 200–10. https://doi.org/10.1109/TIFS.2016.2607701.

[32] Wang X, He H, Li L. A hierarchical deep domain adaptation approach for fault diagnosis of power plant thermal system. IEEE Trans. Industr. Inf. 2019;15(9): 5139–48. https://doi.org/10.1109/TII.2019.2899118.

[33] Ciavarella S, Joo J, Silvestri S. Managing contingencies in smart grids via the internet of things. IEEE Transactions on Smart Grid 2016;7(4):2134–41. https://doi.org/10.1109/TSG.2016.2529579.

[34] Glover JD, Sarma MS, Overbye TJ. Power System Analysis and Design. Cengage Learning; 2009.

[35] Vieira JCM, Freitas W, Xu Wilsun, Morelato A. Performance of frequency relays for distributed generation protection. IEEE Trans. Power Delivery 2006;21(3):1120–7. https://doi.org/10.1109/TPWRD.2005.858751.

[36] Kiliccote S, Lanzisera S, Liao A, Schetrit O, Piette M. Fast dr: Controlling small loads over the internet. Proc. ACEEE Sum. Study Energy Efficien. Build. 2014: 196–208.

[37] Yao L, Lu H. A two-way direct control of central air-conditioning load via the internet. IEEE Trans. Power Delivery 2009;24(1):240–8. https://doi.org/10.1109/TPWRD.2008.923813.

[38] S.A. Raziei, H. Mohscnian-Had, Optimal demand response capacity of automatic lighting control, in: Proc. IEEE PES Innovative Smart Grid Technologies Conf. (ISGT), 2013, pp. 1–6. doi:10.1109/ISGT.2013.6497854.

[39] Vanthournout K, D`hulst R, Geysen D, Jacobs G. A smart domestic hot water buffer. IEEE Transactions on Smart Grid 2012;3(4):2121–7. https://doi.org/10.1109/TSG.2012.2205591.

[40] Masuta T, Yokoyama A. Supplementary load frequency control by use of a number of both electric vehicles and heat pump water heaters. IEEE Transactions on Smart Grid 2012;3(3):1253–62. https://doi.org/10.1109/TSG.2012.2194746.

[41] Otomega B, Van Cutsem T. Undervoltage load shedding using distributed controllers. IEEE Trans. Power Syst. 2007;22(4):1898–907. https://doi.org/10.1109/TPWRS.2007.907354.

[42] Q. Wang, X. Cai, W. Tai, Y. Tang, A multi-stage game model for the false data injection attack against power systems, in: Proc. and Intelligent Systems (CYBER) 2018 IEEE 8th Annual Int. Conf. CYBER Technology in Automation, Control, 2018, pp. 1450–1455. doi:10.1109/CYBER.2018.8688306.

[43] Ma R, Chen H, Huang Y, Meng W. Smart grid communication: Its challenges and opportunities. IEEE Transactions on Smart Grid 2013;4(1):36–46. https://doi.org/10.1109/TSG.2012.2225851.

[44] Mahmoud MMEA, Mišic J, Akkaya K, Shen X. Investigating public-key certificate revocation in smart grid. IEEE Internet of Things Journal 2015;2(6):490–503. https://doi.org/10.1109/JIOT.2015.2408597.

[45] R. Hassan, M. Abdallah, G. Radman, F. Marco, S. Hammer, J. Wigington, J. Givens, D. Hislop, J. Short, S. Carroll, Under-frequency load shedding: Towards a smarter smart house with a consumer level controller, in: Proc. IEEE Southeastcon 2011, 2011, pp. 73–78.

[46] Alpaydin E. Introduction to Machine Learning. Cambridg, MA: MIT Press; 2012.

[47] Shapley LS. Stochastic games. Proceedings of the national academy of sciences 1953;39(10):1095–100.

[48] Littman ML. Markov games as a framework for multi-agent reinforcement learning. In: Machine learning proceedings 1994. Elsevier; 1994. p. 157–63.

[49] M. Tokic, Adaptive ε-greedy exploration in reinforcement learning based on value differences, in: Annual Conference on Artificial Intelligence, Springer, 2010, pp. 203–210.

[50] Pai A. Energy Function Analysis for Power System Stability. Springer; 1989.

[51] Hasan S, Dubey A, Karsai G, Koutsoukos X. A game-theoretic approach for power systems defense against dynamic cyber-attacks. International Journal of Electrical Power & Energy Systems 2020;115:105432. https://doi.org/10.1016/j.ijepes.2019.105432.