Cascading Failure Attacks in the Power System: A Stochastic Game Perspective

Weixian Liao, *Graduate Student Member, IEEE*, Sergio Salinas, *Member, IEEE*, Ming Li, *Associate Member, IEEE*, Pan Li, *Member, IEEE*, and Kenneth A. Loparo, *Life Fellow, IEEE*

***Abstract*—Electric power systems are critical infrastructure and are vulnerable to contingencies including natural disasters, system errors, malicious attacks, etc. These contingencies can affect the world's economy and cause great inconvenience to our daily lives. Therefore, security of power systems has received enormous attention for decades. Recently, the development of the Internet of Things (IoT) enables power systems to support various network functions throughout the generation, transmission, distribution, and consumption of energy with IoT devices (such as sensors, smart meters, etc.). On the other hand, it also incurs many more security threats. Cascading failures, one of the most serious problems in power systems, can result in catastrophic impacts such as massive blackouts. More importantly, it can be taken advantage by malicious attackers to launch physical or cyber attacks on the power system. In this paper, we propose and investigate cascading failure attacks (CFAs) from a stochastic game perspective. In particular, we formulate a zero-sum stochastic attack/defense game for CFAs while considering the attack/defense costs, budget constraints, diverse load shedding costs, and dynamic states in the system. Then, we develop a Q-CFA learning algorithm that works efficiently in power systems without any *a priori* information. We also formally prove that the convergence of the proposed algorithm achieves a Nash equilibrium. Simulation results validate the efficacy and efficiency of the proposed scheme by comparisons with other state-of-the-art approaches.**

***Index Terms*—Cascading failure attacks (CFAs), Nash equilibrium, Q-CFA learning algorithm, stochastic games.**

I. Introduction

ELECTRIC power systems are critical infrastructure and the failure of these systems can lead to severe economic, social, and security consequences. Thus, the security of these systems is crucial. The recent development of the Internet of Things (IoT) technologies helps traditional electric power systems to be transformed into smart grids, and offers tremendous promise of future smart grids [1]. In particular, IoT technologies enable power systems to support various network functions throughout the generation, transmission, distribution, and consumption of energy by incorporating IoT devices (such as smart sensors, actuators and smart meters), as well as by providing the connectivity, automation, etc. [2]. However, the use of such IoT devices also brings new security challenges. The security of power systems has now been further aggravated by various malicious cyber attacks that can be launched on the IoT devices such as denial-of-service (DoS) attacks [3], false data injection attacks [4], unobservable cyber attacks through topology errors [5], etc. Due to the expansive geographical coverage and complex interdependencies among system components, protecting the power system is data and computing intensive and hence extremely challenging [6].

Cascading failures are a very concerning security problem in the power system. They are system failures where the failure of a system component can trigger the successive components and a series of unpredictable chain events in the system that can possibly result in a large-scale collapse of the system. Taking the cascading failure in transmission networks [7] as an example, when a transmission line fails, it will shift its load it has been supplying to the other lines that share the same bus with it. Those connected lines may be pushed beyond their line capacities, become overloaded, and further shift their loads to other lines. Such sudden load spikes could induce overloaded lines into loss of service due to the operation of the protection system or failure, which quickly spreads to other lines before the system operator can conduct any countermeasures, hence finally taking down the entire system in a very short period of time [8]. This is exactly what happened in the 2003 Northeastern blackout, where the failure of a critical transmission line triggered a cascade of failures, resulting in shutting down a portion of the power system that affected more than 55 million people in the Eastern U.S. and Canada [9]. Cascading failures have attracted intensive attention because of their criticality in the power system operations. Chen *et al.* [7] proposed a hidden failure model to assess the cascading dynamics in power systems. Rahnamay-Naeini *et al.* [10] constructed a probabilistic model for cascading failures while retaining key physical attributes and operating characteristics of bulk power systems.

Manuscript received July 8, 2017; revised September 7, 2017; accepted September 21, 2017. Date of publication October 10, 2017; date of current version December 11, 2017. The work of P. Li was supported by the U.S. National Science Foundation under Grant CNS-1602172 and Grant CNS-1566479. The work of M. Li was supported by the U.S. National Science Foundation under Grant CNS-1566634 and Grant ECCS-1711991. *(Corresponding author: Pan Li.)*

W. Liao, P. Li, and K. A. Loparo are with the Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: weixian.liao@case.edu; lipan@case.edu; kal4@case.edu).

S. Salinas is with the Department of Electrical Engineering and Computer Science, Wichita State University, Wichita, KS 67260 USA (e-mail: salinas@cs.wichita.edu).

M. Li is with the Department of Computer Science and Engineering, University of Nevada, Reno, NV 89557 USA (e-mail: mingli@unr.edu).

Digital Object Identifier 10.1109/JIOT.2017.2761353

As cascading failures can lead to catastrophic damages in the power system and possibly take down the entire system, there is strong motivation for attackers to launch deliberate attacks by taking advantage of it, which we call "cascading failure attacks (CFAs)." For example, a malicious attacker can launch CFAs to trip the critical transmission lines and in turn induce a massive cascading failure [11]. Motter and Lai [12] studied cascading-based attacks on complex networks. Zhu *et al.* [13] assessed the line vulnerability and attack strategies from an attacker's perspective in the smart grid. Yan *et al.* [14] also investigated the topology and cascading attacks in the smart grid. These previous works mainly focus on the impact of cascading attacks but do not consider the defense strategies in such systems. In fact, analyzing CFAs in the power system is a very challenging problem because of the unpredictable cascading effect, the complex interactions between the attacker and the system operator, the extremely high problem dimensionality in a large-scale system, and so on [15].

In this paper, we explore CFAs in the power system, from a game theoretic perspective. Specifically, defending critical infrastructures against malicious attacks requires system operators to make optimal decisions about where to deploy limited resources to improve system resilience against adversaries. Game theory can naturally be used to provide the system operators with guidance on strategies for infrastructure protection [16]–[19]. For instance, Salmeron *et al.* [17] formulated the competition between a defender and an attacker as a leader–follower game. Chen *et al.* [18] proposed a static game framework for defending the power system against deliberate attacks. Rao *et al.* [19] studied a Stackelberg game while taking both infrastructure survival probability and costs into account. These works consider the competition between the attacker and defender as a one-time event. However, power system protection can be a continuous process where an attacker and a defender interact with each other many times across different dynamic states [20]. For example, the nationwide power system in Yemen suffered from repeated attacks on transmission lines in 2014, which very soon left Yemen in total darkness [21]. Therefore, an attack-defense interaction model that explicitly considers the temporal aspects of the dynamic system states and the long-term effects is indispensable.

To this end, we formulate a zero-sum stochastic game to characterize the long-term interactions between an attacker and a system operator in CFAs. Specifically, we consider that an attacker deploys limited resources to disrupt the components in the power system, such as transmission lines and substations, through either physical attacks or cyber attacks on the IoT devices. Maximizing the amount of load shedding due to disruption is usually adopted as the objective of the attacker in previous studies. However, loads on different transmission lines are of different importance to the system, and each transmission line contributes differently to the overall system reliability and security [7]. Therefore, we consider that the attacker's objective is to maximize the total cost of the load shedding that is defined as a nondecreasing function of the total amount of shedding load, making the problem more challenging. On the other hand, a system operator deploys limited resources to minimize the total cost of load shedding by taking actions such as reinforcing a vulnerable transmission line or repairing a damaged line. Because the objectives of the attacker and the system operator are conflicting, we model the interactions in dynamic environments between two players as a zero-sum stochastic game.

Stochastic games are difficult to solve due to the possible large problem dimensionality and their stochastic nature. Value iteration and policy iteration [22], such as iteratively improving the value functions or policies, respectively, have been developed to solve this problem. Unfortunately, such dynamic programming-based algorithms need to enumerate all the system states, the number of which is obviously too large in a large-scale power system for the solution to be tractable. Thus, these algorithms suffer from the well known "curse of dimensionality" problem [23]. Furthermore, although such approaches are proven to converge to the optimum, they are under the assumption that all the dynamic system parameters, for example, reward functions and transition probabilities, are always available to the players, which may not always hold in practice, especially to the attacker in the power system. Some previous works on stochastic game analysis also assume complete *a priori* system information. Instead of having such strong assumptions, we develop a Q-CFA learning algorithm to solve our stochastic game which can address the dimensionality problem and does not need any *a priori* system information. The intuition behind the learning process is that learning through past experience facilitates more intelligent decision making and performance optimization.

The main contributions of this paper are briefly summarized as follows.

1) We formulate a zero-sum stochastic game for an attacker and a system operator while considering the attack/defense costs, limited resources, and diverse load shedding costs in the system.
2) We propose a Q-CFA learning algorithm that works efficiently without having *a priori* knowledge of all system information.
3) The proposed scheme is formally proved to converge fast and achieve the Nash equilibrium.
4) Simulation results demonstrate that the proposed scheme achieves convergence and has much better performance than the benchmark algorithms.

The rest of this paper is organized as follows. Section II introduces our system models in detail, including DC power network model, cascading hidden-failure model, as well as the threat and defense models. We formulate the zero-sum stochastic game in the dynamic environment in Section III, which is solved by the proposed Q-CFA learning algorithm in Section IV. In Section V, we conduct extensive simulations to validate the convergence and efficiency of our scheme, followed by the conclusion drawn in Section VI.

II. System Models

In this section, we introduce DC power network model, cascading hidden failure model, as well as the threat and defense models used in this paper, respectively.

A. DC Power Network Model

We consider a power network consisting of $\mathcal{N} = \mathcal{G} \cup \mathcal{D}$ buses and $\mathcal{L} = \{1, \ldots, l, \ldots, L\}$ transmission lines. We assume that each bus is either a generation bus, denoted by $g \in \mathcal{G}$, or a load bus, denoted by $d \in \mathcal{D}$. Bus n_1 is identified as the reference bus. Similar to [24] and [25], we use DC power flow approximation of the AC system. Denote by $\Theta = [\theta_1, \ldots, \theta_n, \ldots, \theta_N]^T$, $\mathbf{P^G} = [p_1^G, \ldots, p_g^G, \ldots, p_G^G]^T$, and $\mathbf{D} = [d_1, \ldots, d_d, \ldots, d_D]$ as the bus voltage angle vector, the real power generation vector and the load demand vector, respectively (note that $N = |\mathcal{N}|$, $G = |\mathcal{G}|$, and $D = |\mathcal{D}|$). Then, the DC power flow equations are formulated as

$$\mathbf{P}^{\text{inj}} = \mathbf{K_g}\mathbf{P^G} - \mathbf{K_d}\mathbf{D} \tag{1}$$

$$\Theta = \mathbf{B}\mathbf{P}^{\text{inj}} \tag{2}$$

$$f(l) = b_{ij}\big(\theta_i - \theta_j\big) \tag{3}$$

where $\mathbf{P}^{\text{inj}} = [p_2^{\text{inj}}, \ldots, p_n^{\text{inj}}, \ldots, p_N^{\text{inj}}]^T$ is a vector of nodal injection power for buses $2, \ldots, N$, $\mathbf{K_g}$ is the bus-generation incidence matrix, and $\mathbf{K_d}$ is the bus-load incidence matrix. θ_i and θ_j are the phase angles of bus i and bus j, respectively, that are connected by transmission line l. $f(l)$ is the real power flow on line l. $\mathbf{B}$ is the $N \times N$ system susceptance matrix, in which $b_{ii} = -\sum_{j \in S_i, j \neq i}(1/x_{ij})$ and $b_{ij} = (1/x_{ij})$, where x_{ij} is the reactance between bus i and bus j. Notice that in this DC power network model, (1) is the power balance constraint. Equation (2) calculates the phase angles for all the buses, which is used for the power flow calculation on each line in the network as shown in (3).

B. Cascading Hidden Failure Model

Cascading failures are system failures where the failure of a system component triggers the successive components and possibly spreads among the entire system. Hidden failures are among the top reasons for cascading failures in the power system [7], [10]. Particularly, hidden failure remains undetected until it is triggered by another system failure [26]. In this paper, we study the line protection hidden failure by considering how protective relays work. Protective relays are designed to trip the circuit breakers on the transmission lines when any fault is detected. They may incorrectly trip a transmission line with a load-dependent probability [7], which may in turn lead to more and more lines tripped due to the increased load, i.e., the cascading failure. Hidden failures are undetectable during normal operation but will be exposed as a direct consequence of other system disturbances such as a sudden attack or natural disasters. For example, a malicious attacker can launch false data injection attacks on selected IoT devices, or physically sever critical transmission lines, and in turn induce a massive cascading hidden failure [11]. Such sudden disturbances may cause the protective relay systems to inappropriately and incorrectly disconnect circuit elements. In particular, when transmission line l trips, hidden failures on all the lines connected with it will be exposed such that those lines are then exposed to incorrect tripping probabilistically because of the redistribution of the loads from the tripped line [27]. Furthermore, if an exposed line trips, then the lines that are connected to this tripped line will be further exposed and subject to tripping probabilistically as well, which could eventually cause a cascade of failures and in the worst case, may spread across the entire power system and result in a blackout.

In order to quantify the effects by the cascading failure, we follow a general cascading hidden failure model in [7]. Specifically, we consider line protection hidden failures in the power system. Assuming that an attacker launches a successful attack and takes down a transmission line l, it will trigger the cascading effects in the power system. That is, lines that are connected to this tripped line will be exposed because of the load redistribution from the tripped line, which may result in the total flow through the remaining lines to be larger than the nominal capacities. Based on the observations in NERC events [28], the probability for an exposed line to be tripped incorrectly is very low and considered as a constant p, when the load on this line is below its rated capacity, denoted by $F^{\max}(l)$, and increases linearly to 1 as the load approaches $1.4F^{\max}(l)$. Furthermore, when the load on the line is or above $1.4F^{\max}(l)$, this line will be tripped immediately for security purposes. Thus, the probability of an exposed line tripping incorrectly, also known as the load-dependent probability, defined as $P_t(l)$, is

$$P_t(l) = \begin{cases} p, \text{ if } 0 \leq f(l) \leq F^{\max}(l) \\ \frac{5(1-p)f(l)+7pF^{\max}(l)-5F^{\max}(l)}{2F^{\max}(l)}, \\ \text{if } F^{\max}(l) \leq f(l) \leq 1.4F^{\max}(l) \\ 1, \text{ if } 1.4F^{\max}(l) \leq f(l). \end{cases} \tag{4}$$

Based on (4), we are able to determine if exposed lines will be further tripped after the initial line tripping as a chain of cascading effects. If the exposed line trips, then the lines that are connected to this new tripped line will be further exposed and tripped based on (4). Therefore, we can model the potential spread of cascading hidden failure in the power system by the protective relays in all the transmission lines. Notice that our framework accounts for the possibility that lines that are not connected to failed lines may also fail. For example, let us assume that line l_1 is connected to l_2, l_2 is connected to l_3, but l_1 is not connected to l_3. When line l_1 is tripped, l_2 may be tripped, and l_3 could further be tripped based on its load-dependent probability. Therefore, in each round the number of tripped lines is a variable and can be greater than 1. This procedure will go on until there are no further line trippings in the system, then the system will conduct the optimal power flow for the current system configuration, which will be clear in Section III after we formulate the zero-sum stochastic game.

C. Threat Model

In the power system, an attacker aims to disrupt the system by either physical attacks such as severing transmission lines, damaging critical infrastructure like transmission towers, or cyber attacks on IoT devices, e.g., false data injection attacks and DoS attacks on sensors [4]. We also assume that the attacker has the knowledge of the system topology and that is able to launch a combination of cyber and physical attacks

that can affect many more components across geographical locations. The attacks can be launched on any components of the power system. Without loss of generality, in this paper we use attacks on transmission lines as an example, which are one of the most common and far-ranging targets in the power system [29].

We first define two binary variables as follows:

$$\epsilon(l) = \begin{cases} 1, & \text{if line } l \text{ is attacked} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

$$\delta(l) = \begin{cases} 1, & \text{if line } l \text{ is exposed} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\epsilon(l)$ is equal to 1 if the transmission line l is attacked by the attacker, and $\delta(l)$ is equal to 1 if line l is exposed according to the cascading hidden failure model in Section II-B.

For practicality, we assume that the malicious attacker has limited resources to launch attacks. Specifically, it can only attack a limited number of transmission lines in one action. Therefore, the attacker's action is constrained by

$$\sum_{l \in \mathcal{L}} \mathbf{1}_{\epsilon(l)=1} = b_a \tag{7}$$

where b_a denotes the attacker's limited resources, i.e., the maximum number of transmission lines that can be attacked in one action, and $\mathbf{1}_A$ is an indicator function that is equal to 1 when the event A is true and zero otherwise.

Subject to resource constraints, the objective of the attacker is to cause the most damage to the power system. In the past, damage is simply measured as the total amount of loads that have to be shed due to line failures [8]. However, because different loads may have different adverse impacts on the power system, it is more appropriate if we use the *costs of load shedding* as the objective of the attacker instead of the *amount of load that is shed*. To this end, we denote the cost function for the transmission line l as $u_l(\cdot)$, which is a nondecreasing function with regard to the shed load on the transmission line l, i.e., $\hat{d}(l)$. Consequently, the objective of the attacker is to maximize the total cost of the loads that are shed in the power system, i.e., to maximize $\mathcal{U} = \sum_{l \in \mathcal{L}} u_l(\hat{d}(l))$.

D. Defense Model

Similarly, a system operator, who could be the power system operator or a third-party system protector, aims to protect the power system from the attack. For illustrative purposes, we define the available actions by the system operator as repairing a damaged line or compromised IoT devices, or reinforcing an important line such that

$$\beta(l) = \begin{cases} 1, & \text{if line } l \text{ is repaired or reinforced} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $\beta(l)$ indicates if the system operator chooses to repair the transmission line l or reinforce it. We note that by reinforcement, we mean that the system operator can reinforce the protection on a specific line by adding physical barriers or deploying additional security personnel. A system operator can also adopt new malicious data analysis schemes that can detect compromised PMUs and perform healing actions such as firmware updates to defend cyber attacks. Since some particular lines are more likely to start a cascading failure, system operators are willing to allocate more resources to these lines to enhance the security of their system. We also assume that the system operator has limited resources to protect the power system, that is,

$$\sum_{l \in \mathcal{L}} \mathbf{1}_{\beta(l)=1} = b_o \tag{9}$$

where b_o denotes the system operator's limited resources, i.e., the maximum number of transmission lines that it can repair or reinforce in one action. Besides, the objective of the system operator of the power system is to find the best strategy that minimizes the total cost of load shedding in the power system, i.e., to minimize $\mathcal{U} = \sum_{l \in \mathcal{L}} u_l(\hat{d}(l))$.

Therefore, as the objectives of the system operator and the attacker are conflicting and two players compete with each other through dynamic system states, we formulate a zero-sum stochastic game that will be introduced in the next section.

III. Zero-Sum Stochastic Game for CFAs

As presented above, the objectives of the attacker and that of the system operator in CFAs are opposite to each other. Therefore, in this section, we formulate a zero-sum stochastic game for the attacker and the system operator in the power system.

Before delving into details of the formulation for the zero-sum stochastic game, we first briefly introduce stochastic games. In game theory, a stochastic game is a dynamic game with probabilistic transitions played by several players [30], which can be considered as an extension of Markov decision processes [31]. The game is played in a sequence of stages. Specifically, at the beginning of each stage, the game is in a given state and players select actions independently and simultaneously based on their own resources and constraints at the current state, and each player will then receive an *immediate reward* that results from the chosen actions and the current state. Thereafter, the game moves to a new random stage, the transition probability of which is determined by both actions from the players and the previous state. This procedure repeats continuously for a number of stages and each player endeavors to maximize their *long-term reward*, that is defined as the discounted sum of the *immediate rewards* at all stages.

A. States, Actions, and State Transitions

By considering the interactive competition between the attacker and the system operator, we now formulate the CFAs as a stochastic game $\boldsymbol{G}$. In this game $\boldsymbol{G}$, there are a set of system states, denoted by $\mathcal{S}$, in which each state $s \in \mathcal{S}$ is a vector that denotes the current status of all the transmission lines. we use time-slot-based system as the temporal resolution in our model [20]. Without loss of generality, we define the status of each transmission line as "up," denoted by u, or "down," denoted by w, when the line is functioning well or malfunctioning after being attacked, respectively. The stochastic game proceeds in a time-slotted fashion. Specifically, in each time slot, each player will choose an action based on

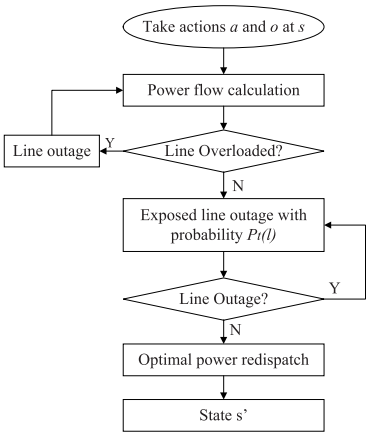

Fig. 1. Flow chart for the power system after being attack.

the current system state so as to optimize its own objective. We denote by $\mathcal{M}_A(s)$ and $\mathcal{M}_O(s)$ the set of all the possible actions that the attacker and the system operator can take at state s, respectively. As discussed in Sections II-C and II-D, for the attacker, each $a \in \mathcal{M}_A(s)$ indicates the set of transmission lines to be attacked. On the other hand, for the system operator, each $o \in \mathcal{M}_O(s)$ refers to a set of transmission lines to be repaired (if not working) or reinforced (if still working but vulnerable to attacks). Each action $a \in \mathcal{M}_A(s)$ and $o \in \mathcal{M}_O(s)$ will be selected by the attacker and the system operator, respectively, in each state s, with a certain probability denoted by $\pi_a(s)$ and $\pi_o(s)$.

Recall that each player selects their actions independently and simultaneously in each stage. We denote p_{uwr} and p_{uw} as the probabilities that a functioning transmission line fails upon attack with and without reinforcement by the system operator in the same time slot, respectively. Similarly, we denote p_{wua} and p_{wu} as the probabilities that a nonfunctioning line recovers after repair with and without being attacked in the same time slot, respectively. The following constraints must be satisfied, $0 \le p_{uwr} < p_{uw} \le 1$ and $0 \le p_{wua} < p_{wu} \le 1$ and in practice, these probabilities can be obtained by either conducting simulations or observing historical records. We can see that these probabilities determine the transition probability $T(a, o, s, s')$ from state s to state s' under the actions a and o by the attacker and the system operator, respectively. For example, suppose at the beginning all lines in the system are up and there are no actions from the attacker or the system operator. Then, when the attacker and the system operator choose the same line to attack and reinforce, respectively, the probability for the power system to remain in the same state is $1-p_{uwr}$. Similarly, when the attacker attacks a line l and the system operator chooses to reinforce another line l', the probability for the system to move to another state where only line l is down is p_{uw}.

B. Immediate Rewards

As mentioned before, the objectives of the attacker and the system operator are opposite; maximizing/minimizing the total cost of the load shedding in the power system. At each stage of the game, both attacker and system operator will receive an *immediate reward* defined by the actions taken by them (the attacker a, the system operator o) at state s. For example, the immediate reward for the attacker, denoted by $U_A(a, o, s)$, is the total cost for load shedding.

We show in Fig. 1 what happens sequentially in one stage of the game where the attacker and the system operator take actions a and o, respectively, at state s. Particularly, after both players take actions, some transmission lines might be tripped, and hence the system immediately adjusts according to the power equations (1)–(3) [32]. Then the system checks whether there are any lines overloaded. If so, the protective relays trip the overloaded lines and the system readjusts accordingly until there are no overloaded lines. Otherwise, the exposed lines, which share the same bus with the tripped lines, are tripped with probability $P_t(l)$, based on the cascading model in Section II-B. The cascading effect continues until there are no line outages. Finally, the power system performs security constrained optimal power flow, which is formulated as an optimization problem to minimize the total cost of load shedding, i.e., $\mathcal{U}(a, o, s)$, in the current configuration of power system

$$\text{minimize} \quad \mathcal{U}(a, o, s) = \sum_{l \in \mathcal{L}} u_l\left(\hat{d}_l\right)$$

$$\text{s.t.} \quad \sum_{g \in \mathcal{G}} P_g + \sum_{l \in \mathcal{L}} \hat{d}_l - \sum_{l \in \mathcal{L}} d_l = 0 \tag{10}$$

$$P_g^{\min} \le P_g \le P_g^{\max} \quad \forall g \in \mathcal{G} \tag{11}$$

$$-F^{\min}(l) \le f(l) \le F^{\max}(l) \quad \forall l \in \mathcal{L} \tag{12}$$

$$0 \le \hat{d}_l \le d_l \quad \forall l \in \mathcal{L} \tag{13}$$

where (10) is the power balance constraint, (11) is the generation capacity constraint for each generation unit, (12) limits the maximum power flow on each transmission line, and (13) indicates that the shed load cannot exceed the original load on the load bus. After solving the above minimization problem, we can shed loads when necessary. In practice, utility companies have several tools to reduce users' load demands. For example, industrial users, which account for 60% of total energy consumption [33], often have contracts with utility companies where they commit to reduce their load after a request from the utility companies in exchange for reduced energy prices; under real-time energy pricing, all users can be incentivized to reduce their energy consumption by significantly increasing prices; and system operators can disconnect complete sections of the power system by opening switches.

Therefore, we have the immediate rewards for the attacker and the system operator, known as the payoff of the game at state s given by $\mathcal{U}(a, o, s)$ for all $a \in \mathcal{M}_A(s)$ and $o \in \mathcal{M}_O(s)$. Notice that this framework can also account for the case where the system is disconnected into nonconnected islands. Specifically, when the system is disconnected to several islands, both players still take actions in the whole system subject to the limited resources b_a and b_o. After there are no more line trippings in the system, we conduct the OPF for every island and then the system state transits to the next state.

Because the objective function is convex and all the constraints are linear, this problem can be easily solved and we can obtain the *immediate rewards* for each player at any system

status. Actions a and o executed at state s will bring the system state to the next state, resulting in further immediate rewards, i.e., $\mathcal{U}(a', o', s')$, at the next state s'. Thus, actions taken at dynamic states will finally accrue a long-term reward as the game continues. The objective of both players is to obtain the optimal expected long-term reward, which will be discussed next.

IV. Optimal Strategies of the Stochastic Game

In this section, we first present the definition of optimal strategies. Then, we develop a Q-CFA learning algorithm to find the optimal strategies for the zero-sum stochastic game.

A. Optimal Strategies

We refer to the *optimal strategies* as the mixed strategies of all actions chosen by the players that maximize their expected long-term rewards [34]. In this paper, we consider the case of stationary policies where the action selection probabilities, i.e., $\pi_A(s)$ and $\pi_O(s)$, do not change over time. In other words, we are interested in finding the stationary policies for each player at each state s.

From the attacker's point of view, we let $V_A(s)$ denote the attacker's expected long-term reward under the optimal strategies when the game starts at state s, and $Q_A(a, o, s)$ as the expected long-term reward for taking action a while the system operator selects the action o when the game starts at state s. Specifically, we have

$$V_A(s) = \max_{\pi_A(s)} \min_{\pi_O(s)} \sum_{a \in \mathcal{M}_A(s)} \sum_{o \in \mathcal{M}_O(s)} \pi_a(s) Q_A(a, o, s) \pi_o(s) \tag{14}$$

where $\pi_A(s) = \{\pi_a(s) | a \in \mathcal{M}_A(s)\}$, $\pi_O(s) = \{\pi_o(s) | o \in \mathcal{M}_O(s)\}$, and

$$Q_A(a, o, s) = \mathcal{U}(a, o, s) + \gamma \cdot \sum_{s' \in \mathcal{S}} V_A(s') \cdot T(a, o, s, s'). \tag{15}$$

$V_A(s)$ and $Q_A(a, o, s)$ are also called the *value* of the state $s \in \mathcal{S}$ and the *quality* of the state s given actions a and o, respectively, for the attacker. $T(a, o, s, s')$ is the state transition probability from state s to state s' after taking actions a and o. Here the *maxmin* function can be interpreted as follows. Because our game is a fully competitive stochastic game where each player selects an action independently and simultaneously at each system state, we need opponent-independent algorithms to solve this problem [35]. The *maxmin* function makes (14) opponent-independent in which the attacker attempts to maximize its own expected long-term reward under the worst case assumption that the system operator will always endeavor to minimize the payoff. Besides, note that (15) states that $Q_A(a, o, s)$ is equal to the immediate reward plus the discounted expected optimal value attainable from the next state s'. In (15), $\gamma \in [0, 1)$ is a discount factor that represents how much impact the current decisions can have on the long-term reward. Particularly, when γ equals 0, the game becomes a one-time-event game [17]–[19]. When γ is larger than 0, a smaller value of γ emphasizes more the immediate rewards and a larger γ gives higher weight to the future rewards.

Similarly, the system operator's expected long-term reward under the optimal strategies when the game starts at state s, denoted by $V_O(s)$, is

$$V_O(s) = \min_{\pi_O(s)} \max_{\pi_A(s)} \sum_{a \in \mathcal{M}_A(s)} \sum_{o \in \mathcal{M}_O(s)} \pi_a(s) Q_O(a, o, s) \pi_o(s) \tag{16}$$

where $Q_O(a, o, s)$ is the expected long-term reward for taking action o while the attacker selects the action a, known as the *quality* of the state s for the system operator, and is formulated as

$$Q_O(a, o, s) = \mathcal{U}(a, o, s) + \gamma \cdot \sum_{s' \in \mathcal{S}} V_O(s') \cdot T(a, o, s, s'). \tag{17}$$

We note that generally $V_A(s) \leq V_O(s)$ due to weak duality, where $V_A(s)$ and $V_O(s)$ correspond to the primal problem and the dual problem, respectively. However, in a zero-sum stochastic game, strong duality holds and we have $V_A(s) = V_O(s) = V(s)$ [36, Sec. 5.4.5]. Consequently, the optimal solutions computed individually by the two players, i.e., $\pi^*_{\mathbf{A}}(s)$ and $\pi^*_{\mathbf{O}}(s)$, are the best responses to each other. We denote by $\pi^*(s) = \{\pi^*_{\mathbf{A}}(s), \pi^*_{\mathbf{O}}(s)\}$ the *optimal strategy pair* [37], which is known as the *Nash equilibrium point* in a stochastic game and defined as follows.

Definition 1 (*Nash Equilibrium):* In a zero-sum stochastic game $\mathbf{G}$, the Nash equilibrium for any state $s \in \mathcal{S}$ is an optimal strategy pair $\pi^*(s) = \{\pi^*_{\mathbf{A}}(s), \pi^*_{\mathbf{O}}(s)\}$ satisfying

$$V^{\pi^*(s)}(s) \geq V^{\{\pi_{\mathbf{A}}(s), \pi^*_{\mathbf{O}}(s)\}}(s)$$
$$V^{\pi^*(s)}(s) \leq V^{\{\pi^*_{\mathbf{A}}(s), \pi_{\mathbf{O}}(s)\}}(s).$$

Therefore, by finding the Nash equilibrium for each state s, we can obtain the attacker's and the system operator's optimal strategies, specifically, the probability mass distributions on their action sets $\mathcal{M}_A(s)$ and $\mathcal{M}_O(s)$, which result in the optimal expected long-term reward for the attacker and the system operator, respectively.

From the attacker's perspective, the optimal strategies $\pi^*_A(s)$ ($s \in \mathcal{S}$) can be obtained by solving (14) using algorithms like "value iteration" [22]. Particularly, at the kth iteration, for each $s \in \mathcal{S}$, the attacker needs to solve the following problem:

$$\begin{aligned}
V^k_A(s) &= \max_{\{\pi_a(s)\}} \min_{o \in \mathcal{M}_O(s)} \sum_{a \in \mathcal{M}_A(s)} Q^k_A(a, o, s) \\
&\quad \times \pi_a(s) \\
\text{s.t.} \quad Q^k_A(a, o, s) &= \mathcal{U}(a, o, s) + \gamma \cdot \sum_{s' \in \mathcal{S}} V^{k-1}_A(s') \\
&\quad \times T(a, o, s, s') \\
\sum_{a \in \mathcal{M}_A(s)} Q^k_A(a, o, s) &\geq V^{k-1}_A(s) \\
\sum_{a \in \mathcal{M}_A(s)} \pi_a(s) &= 1 \\
\pi_a(s) &\geq 0 \quad \forall a \in \mathcal{M}_A(s)
\end{aligned}$$

where $V^k_A(s)$ is the value of the state s in the kth iteration. The basic idea of value iteration is that it iteratively estimates the value of $Q_A(a, o, s)$ and $V_A(s)$ using (14) and (15) for each $s \in \mathcal{S}$ in each iteration until convergence. The optimal

strategies can then be obtained after scanning all the available states and action spaces. The system operator can find its optimal strategies $\pi^*_{\mathbf{O}}(s)$ ($s \in \mathcal{S}$) by following a similar approach, which is omitted here due to space limit.

Value iteration has been proved to converge to the optimal results in stochastic games [38]. However, it assumes that the system information, such as the state transition probabilities $T(a, o, s, s')$'s, is *a priori* knowledge for both players, which may not be the case in most practical applications. Moreover, this algorithm needs to enumerate all the system states and available actions in each iteration in order to obtain the optimal strategies. Nevertheless, the number of states and actions grows exponentially with the number of transmission lines, which obviously makes such algorithms infeasible for large-scale system applications.

B. Q-CFA Learning Algorithm

In order to account for the drawbacks of previous algorithms, we develop a machine learning based method named Q-CFA learning algorithm that is based on the minimax-Q learning framework [34]. The proposed algorithm can gradually learn the optimal strategies without having any *a priori* knowledge of system information such as the state transition probabilities, i.e., $T(a, o, s, s')$'s. Besides, unlike value iteration and other previous algorithms, it does not need to scan all the states and actions in each iteration, and hence is very efficient for power system applications.

The main idea of the proposed algorithm is as follows. Different from that in (15), we rewrite the quality of state s for the attacker under actions a and o by the attacker and the system operator, respectively, i.e., $Q_A(a, o, s)$, at the kth iteration into

$$Q^k_A(a, o, s) = (1 - \alpha(k)) \cdot Q^{k-1}_A(a, o, s) + \alpha(k) \cdot \left[\mathcal{U}(a, o, s) + \gamma V^{k-1}_A(s')\right] \tag{18}$$

where $\alpha(k) = (1/k + 1)$ is the learning rate that decays over time, and s' is the next state after actions are executed in the current state s. In other words, $Q^k_A(a, o, s)$ is updated by mixing the previous Q-value with a correction from the new estimate at a learning rate $\alpha(k)$. Then, the value of state s at the kth iteration, i.e., $V^k_a(s)$, can be updated accordingly by (14). Note that the quality and the value of state s for the system operator can be updated in the same fashion.

Specifically, because of their limited resources, both attacker and system operator only have a limited number of actions at each stage of the game, which could be very diverse at different states. At the beginning of each state s_k, the algorithm first checks whether the current state has been observed in previous stages. If so, then both players use the previous profiles at state s_k to initialize parameters such as the action sets, along with Q and V values. Otherwise, the algorithm initializes all the variables, and then adds the current state s_k into the *observation history set* denoted by H_s that contains profiles at all the past states. Subsequently, each player chooses an action. In particular, with a probability of p_{exp}, the attacker and the system operator choose to explore their available action spaces, i.e., $\mathcal{M}_A(s)$ and $\mathcal{M}_O(s)$, respectively, and uniformly and randomly select actions. This process is called *exploration*. On the other hand, with a probability of $1 - p_{\text{exp}}$, they choose to take the same actions selected in the previous initialization step, that is called *exploitation*. The intuition here is that the players in Q-learning can either randomly try out one of the available action profiles to possibly achieve higher reward in the long run, namely exploration, or attempt to maximize the reward by choosing the best known action, namely exploitation [39]. Looking into (18), the Q-CFA learning algorithm only uses the previous predicted state value, i.e., $V^{k-1}_a(s)$, which avoids enumerating all the possible future states for current state s. After both players take actions, they obtain their *immediate rewards*, update their Q and V function values, policies $\pi^*_A(s_k)$ and $\pi^*_O(s_k)$, and learning rate $\alpha(k)$, respectively, and then update the profiles for state s_k in the observation history set H_s. Thereafter, the game transits to the next state s_{k+1}. This procedure goes on until the policies in all states have converged. The details of the proposed Q-CFA learning algorithm are described in Algorithm 1.

Notice that in order to update the profiles for each state, i.e., $(\pi^*_A(s_k), \pi^*_O(s_k))$, $V_A(s_k)$, and

Algorithm 1 Q-CFA Learning Algorithm

1: **At State** s_k**,** $k = 0, 1, ...$
If state s_t has been observed in any previous iteration, i.e., $s_t \in H_s$
initialize π_a, π_o, Q, V with the recorded profiles in H_s
Otherwise,
generate action sets $\mathcal{M}_A(s_k)$ and $\mathcal{M}_O(s_k)$,
initialize $Q(a, o, s_k) \leftarrow 1$, for all $a \in \mathcal{M}_A(s_k)$ and $o \in \mathcal{M}_O(s_k)$,
initialize $\pi_A(s_k) \leftarrow \frac{1}{|\mathcal{M}_A(s_k)|}$ and $\pi_O(s) \leftarrow \frac{1}{|\mathcal{M}_O(s_k)|}$,
2: **Choose an action pair** $\{\pi_{\mathbf{a}}, \pi_{\mathbf{o}}\}$ **at state** s_k**:**
With probability p_{exp}, uniformly and randomly select an action in the action sets;
Otherwise, return the action pair $\{\pi_a, \pi_o\}$ obtained in the initialization;
3: **Learn and Update:**
Update $Q^k_A(a, o, s_k)$ according to (18), and $Q^k_O(a, o, s_k)$ similarly
Update the optimal strategies $\pi^*_A(s_k)$ and $\pi^*_O(s_k)$ by

$$\pi^*_A(s_k) \leftarrow \arg\max_{\pi_A(s)} \min_{\pi_O(s)} \sum_{a \in \mathcal{M}_A(s_k)} \sum_{o \in \mathcal{M}_O(s_k)} \pi_a(s_k) Q^k_A(a, o, s_k) \pi_o(s_k),$$

$$\pi^*_O(s_k) \leftarrow \arg\min_{\pi_O(s_k)} \max_{\pi_A(s_k)} \sum_{a \in \mathcal{M}_A(s_k)} \sum_{o \in \mathcal{M}_O(s_k)} \pi_a(s_k) Q^k_O(a, o, s_k) \pi_o(s_k)$$

Update $V_A(s_k)$ and $V_O(s_k)$ according to (14) and (16),
Update $\alpha(k+1) \leftarrow \frac{1}{k+1}$;
4: **The system transits to the next state** s_{k+1}**;**
5: **If all states' policies have converged, stop; otherwise, go to step 1.**

$V_O(s_k)$, we need to solve the subproblem of $\max_{\pi_A(s_k)} \min_{\pi_O(s_k)} \sum_{a\in\mathcal{M}_A(s_k)} \sum_{o\in\mathcal{M}_O(s_k)} \pi_a(s) Q_A^k(a,o,s_k)\pi_o(s_k)$ in the learning process, which turns out to be a matrix game where the strategies of the attacker and system operator form the rows and columns of the matrix, respectively, with payoffs $Q_A^k(a,o,s)$ and $Q_O^k(a,o,s)$ and we have that $Q_A^k(a,o,s) = Q_O^k(a,o,s_k) = Q^k(a,o,s_k)$. Therefore, we formulate the matrix game as

$$\max_{\pi_A(s_k)} \min_{\pi_O(s_k)} \sum_{a\in\mathcal{M}_A(s_k)} \sum_{o\in\mathcal{M}_O(s_k)} \pi_a(s_k) Q^k(a,o,s_k)\pi_o(s_k). \quad (19)$$

However, the above optimization problem cannot be solved directly. In order to achieve the optimal strategies, i.e., $(\pi_A^*(s_k), \pi_O^*(s_k))$, we begin by assuming that the attacker's strategies are fixed. Then the problem is reduced to

$$\min_{\pi_O(s_k)} \sum_{a\in\mathcal{M}_A(s_k)} \pi_a(s_k) Q^k(a,o,s_k) \sum_{o\in\mathcal{M}_O(s_k)} \pi_o(s_k). \quad (20)$$

As $\sum_{a\in\mathcal{M}_A(s_k)} \pi_a(s_k) Q^k(a,o,s_k)$ is a vector, the solution to problem (20) is equivalent to searching for the smallest element in the vector, i.e., $\min_i [\sum_{a\in\mathcal{M}_A(s_k)} \pi_a(s_k) Q^k(a,o,s_k)]_i$. Thereafter, the matrix game (19) can be reformulated as

$$\max_{\pi_A(s_k)} \min_i \left[\sum_{a\in\mathcal{M}_A(s_k)} \pi_a(s_k) Q^k(a,o,s_k) \right]_i. \quad (21)$$

Next, we define $x = \min_i [\sum_{a\in\mathcal{M}_A(s_k)} \pi_a(s_k) Q^k(a,o,s_k)]_i$ and we have that $[\pi_a(s_k) Q^k(a,o,s_k)]_i \geq x$. Therefore, problem (19) can be further rewritten as

$$\begin{aligned} \max_{\pi_A(s_k)} \ & x \\ \text{s.t.} \ & \left[\sum_{a\in\mathcal{M}_A(s_k)} \pi_a(s_k) Q^k(a,o,s_k) \right]_i \geq x & (22) \\ & \sum_{a\in\mathcal{M}_A(s_k)} \pi_a(s_k) = 1 & (23) \\ & \pi_a(s_k) \geq 0 \quad \forall a \in \mathcal{M}_A(s_k). & (24) \end{aligned}$$

Finally, we can transform this to a linear programming (LP) problem by viewing x as another variable

$$\begin{aligned} \max_{\pi'} \ & \mathbf{0}_{\text{aug}}^T \pi' \\ \text{s.t.} \ & Q'\pi' \leq \mathbf{0} & (25) \\ & \sum_{a\in\mathcal{M}_A(s_k)} \pi_a(s_k) = 1 & (26) \\ & \pi_a(s_k) \geq 0 \quad \forall a \in \mathcal{M}_A(s_k) & (27) \end{aligned}$$

where $\pi' = [\pi_{\mathbf{a}}(\mathbf{s_k}), x]^T$, $Q' = ([\mathbf{01}] - [\mathbf{Q^k}(\mathbf{a},\mathbf{o},\mathbf{s_k})\mathbf{0}])$. $\mathbf{0}_{\text{aug}}^T = [\mathbf{0}^T 1]$ is used to augment the original variable vector $\pi_{\mathbf{a}}(\mathbf{s_k})$ by viewing x as another variable so that we can transform the problem into the standard form of an LP. Because (25) is an LP, we can find the optimal solution of the matrix game. Furthermore, as we optimally solve the subproblem, our algorithm converges to the Nash equilibrium of the game, which is proved in the next section.

C. Proof of the Nash Equilibrium

In what follows, we prove that our proposed algorithm converges to the Nash equilibrium in the formulated zero-sum stochastic game. The general idea is that, we first prove the convergence of our algorithm, then prove that the obtained result is the Nash equilibrium of the game as defined in Section IV-A.

Before we prove the convergence of the proposed algorithm, we have the following assumptions and lemma [40].

Assumption 1: Every state and action have been visited infinitely often.

Assumption 2: The learning rate, $\alpha(k)$, satisfies the following conditions.

1) $1 < \alpha(k) < 1$.
2) $\sum_{k=0}^{\infty} (\alpha(k))^2 < \infty$.

Lemma 1 (Conditional Averaging Lemma): Under Assumptions 1 and 2, the process $V(k+1) = (1-\alpha(k))V(k) + \alpha(k)\omega(k)$ converge to $\mathbb{E}(\omega|h(k),\alpha(k))$, where $h(k)$ is the history at time stamp k.

Then, we arrive at a theorem for the convergence of our algorithm.

Theorem 1: In the proposed Algorithm 1, for any state $s \in \mathcal{S}$, the attacker's and the system operator's policies, i.e., $\pi_A(s)$ and $\pi_O(s)$, converge to the Nash equilibrium point.

Proof: In Algorithm 1, we have that the decaying learning rate $\alpha(k)$ is equal to $(1/k+1)$. Therefore, we can see that $0 < \alpha(k) < 1$, and $\sum_{k=1}^{\infty} (\alpha(k))^2 = \sum_{k=1}^{\infty} (1/k+1)^2 < \sum_{k=1}^{\infty} (1/k+1)1/k)) = \sum_{k=1}^{\infty} (1/k - 1/k + 1) < \infty$.

For the attacker, by substituting (18) into (14), we get that for any $s \in \mathcal{S}$

$$\begin{aligned} V_A^k(s) &= \max_{\pi_A(s)} \min_{\pi_O(s)} \sum_{a\in\mathcal{M}_A(s)} \sum_{o\in\mathcal{M}_O(s)} \pi_a(s) \\ &\quad \cdot \Big[(1-\alpha(k)) \times Q_A^{k-1}(a,o,s) + \alpha(k) \\ &\quad \cdot \Big(\mathcal{U}(a,o,s) + \gamma V_A^{k-1}(s') \Big) \Big] \times \pi_o(s) \\ &= (1-\alpha(k)) V_A^{k-1}(s) + \alpha(k) \max_{\pi_A(s)} \min_{\pi_O(s)} \sum_{a\in\mathcal{M}_A(s)} \\ &\quad \times \sum_{o\in\mathcal{M}_O(s)} \pi_a(s) \Big(\mathcal{U}(a,o,s) + \gamma V_A^{k-1}(s') \Big) \pi_o(s). \end{aligned}$$

Define a mapping function T^k as

$$\begin{aligned} T^k V_A^k(s) = \mathbb{E}_{s'} \Bigg[& \max_{\pi_A(s)} \min_{\pi_O(s)} \sum_{a\in\mathcal{M}_A(s)} \sum_{o\in\mathcal{M}_O(s)} \pi_a(s) \\ & \times \Big(\mathcal{U}(a,o,s) + \gamma V_A^{k-1}(s') \Big) \pi_o(s) \Bigg]. \end{aligned}$$

According to the conditional averaging lemma, we can know that as the iterations in Algorithm 1 continue, $V_A^k(s)$ converges to $T^k V_A^k(s)$.

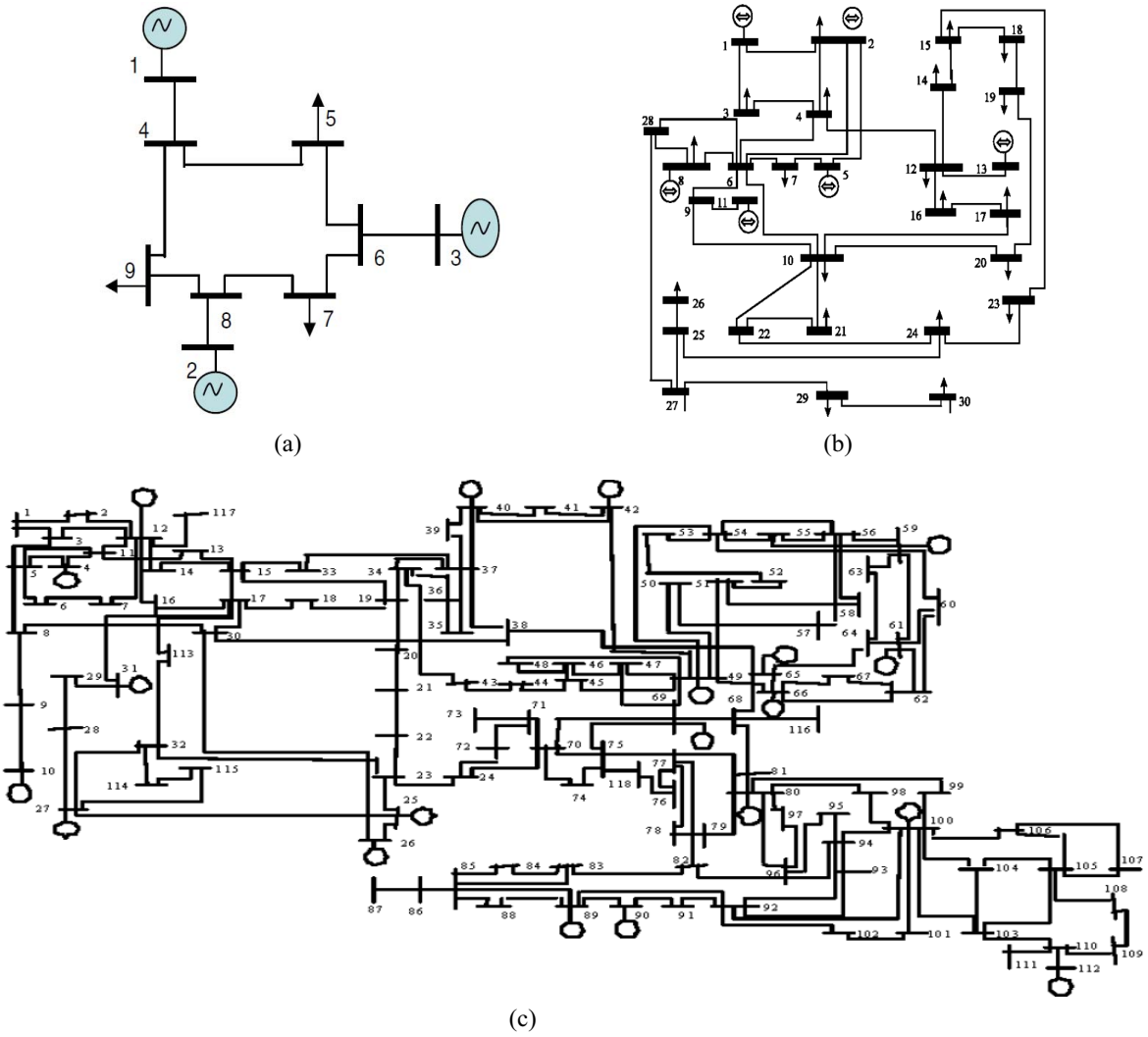

Fig. 2. IEEE standard bus systems. IEEE (a) 9-bus system, (b) 30-bus system, and (c) 118-bus system.

Next, we show that $T^k V_A^k(s)$ converges to the optimal value. Specifically, we can rewrite $T^k V_A^k(s)$ into

$$\begin{aligned} T^k V_A^k(s) &= \max_{\pi_A(s)} \min_{\pi_O(s)} \sum_{a\in\mathcal{M}_A(s)} \sum_{o\in\mathcal{M}_O(s)} \pi_a(s) \\ &\quad \times \sum_{s'\in\mathcal{S}} T\big(a,o,s,s'\big)\Big(\mathcal{U}(a,o,s)+\gamma V_A^{k-1}\big(s'\big)\Big) \\ &\quad \times \pi_o(s) \\ &= \max_{\pi_A(s)} \min_{\pi_O(s)} \sum_{a\in\mathcal{M}_A(s)} \sum_{o\in\mathcal{M}_O(s)} \pi_a(s) \\ &\quad \times \left(\mathcal{U}(a,o,s)+\gamma \sum_{s'\in\mathcal{S}} V_A^{k-1}\big(s'\big) T\big(a,o,s,s'\big)\right) \\ &\quad \times \pi_o(s). \end{aligned}$$

We define another mapping function Z^{k-1} as

$$\begin{aligned} Z^{k-1}V_A^{k-1}(s) = \pi_a(s)\Bigg(&\mathcal{U}(a,o,s)+\gamma \sum_{s'\in\mathcal{S}} V_A^{k-1}\big(s'\big) \\ &\times\ T\big(a,o,s,s'\big)\Bigg)\pi_o(s). \end{aligned}$$

Z^{k-1} has been proved to be a contraction mapping in [41]. Therefore, $T^k V_A^k(s)$ is a contraction mapping as well.

Thus, we have

$$\begin{aligned} T^k\Big(V_A^k\Big)^*(s) &= \sum_{a\in\mathcal{M}_A(s)} \sum_{o\in\mathcal{M}_O(s)} \pi_a^*(s) \\ &\quad \cdot \left(\mathcal{U}(a,o,s)+\gamma \sum_{s'\in\mathcal{S}} V_A^{k-1}\big(s'\big) T\big(a,o,s,s'\big)\right) \\ &\quad \times \pi_o^*(s) \\ &= \Big(V_A^k\Big)^*(s) \end{aligned}$$

which means that $(V_A^k)^*(s)$ is the fixed point of T^k. According to [40, Th. 1], $V_A^k(s)$ converges to $(V_A^k)^*(s)$, i.e., $V^*(s)$, with Probability 1.

Similarly, we can prove that $V_O^k(s)$ converges to $V^*(s)$ with Probability 1 as well. Thus, this theorem directly follows. ■

V. Simulation Results

In this section, we conduct extensive simulations to demonstrate the efficacy and efficiency of the proposed scheme. We

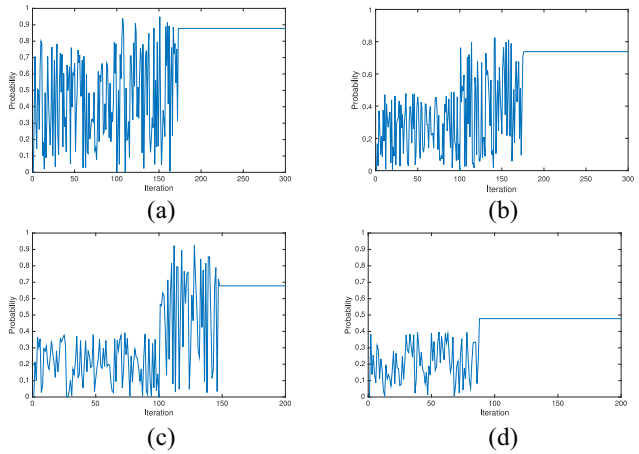

Fig. 3. Learning curves of the attacker and the system operator in the IEEE 9-bus system. (a) Attacker's strategy on line 7 at state 0. (b) System operator's strategy on line 7 at state 0. (c) Attacker's strategy on line 3 at state 7. (d) System operator's strategy on line 7 at state 7.

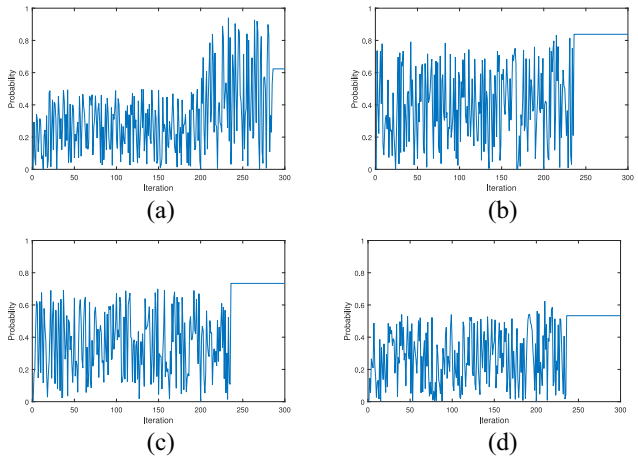

Fig. 4. Learning curves of the attacker and the system operator in the IEEE 30-bus system. (a) Attacker's strategy on line 27 at state 0. (b) System operator's strategy on line 29 at state 0. (c) Attacker's strategy on line 16 at state 27. (d) System operator's strategy on line 27 at state 27.

first demonstrate the convergence of our proposed Q-CFA algorithm in different systems. Then, we analyze the system operator's optimal strategies in different scenarios. Finally, we compare the system operator's expected long-term cost in our scheme with that in other existing schemes.

A. Convergence of Q-CFA

We first study the convergence of the proposed Q-CFA algorithm using the IEEE standard 9-bus, 30-bus, and 118-bus systems, respectively, and the MATPOWER toolbox [42]. As IEEE 118-bus test system does not include flow limits, we employ the flow limits in [43, Table 3] (the transmission line data). In Fig. 2 we show the configuration of standard IEEE bus systems used in our experiments. To initialize the simulation, we set the transition probabilities $p_{uw} = 0.5$, $p_{uwr} = 0.3$, $p_{wu} = 0.5$, and $p_{uwa} = 0.3$, the discounting factor $\gamma = 0.3$ and the exploration probability $p_{\text{exp}} = 0.6$. For illustrative purposes, we consider that the resources of each player are normalized to one, particularly, each player can affect one transmission line in one time slot. Because each transmission line is of different importance to the entire system, we set different load shedding cost for each line. Specifically, we define the load shedding cost as a linear function of the amount of shed loads on line l and is given by

$$u_l(\hat{d}_l) = c_l \hat{d}_l \tag{28}$$

where c_l is a given positive constant for line l. We conduct experiments on a desktop with a 3.41 GHz i7-6700 CPU, 16-GB RAM and a 1-TB hard disk drive. To demonstrate the convergence of our proposed Q-CFA, we show in Figs. 3–5 the learning curves of the system operator's and the attacker's strategies at certain states in the IEEE 9-bus, 30-bus, and 118-bus systems, respectively. For instance, lines 3 and 7 are the most important lines in the IEEE 9-bus system, which become the main targets in the players' optimal strategies as shown in Fig. 3. In particular, the attacker and the system operator tend to attack and defend, respectively, the transmission line 7 when the game starts. It indicates that when all

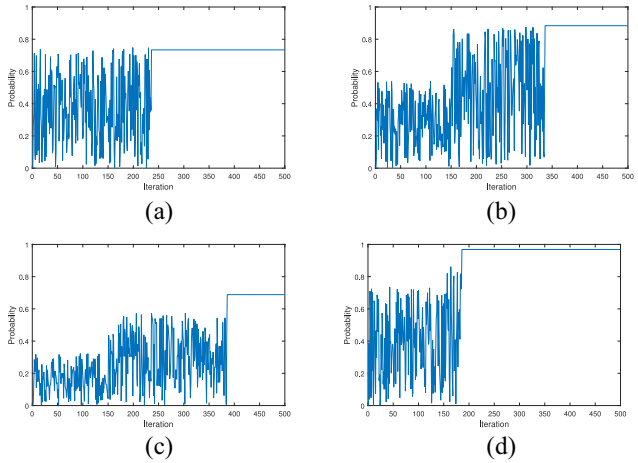

Fig. 5. Learning curves of the attacker and the system operator in the IEEE 118-bus system. (a) Attacker's strategy on line 9 at state 0. (b) System operator's strategy on line 7 at state 0. (c) Attacker's strategy on line 8 at state 9. (d) System operator's strategy on line 9 at state 9.

the transmission lines are well functioning, the most critical line in the IEEE 9-bus system is the line 7. As the iteration goes by, both attacker and defender's strategies converge and the obtained strategies are stationary, which means the mixed strategies do not change over time. When the state of the game transits to state 7 where line 7 is malfunctioning, as shown in Fig. 3(c) and 3(d), we can see that the system operator is more likely to repair line 7 but the attacker more likely turns to attack line 3. We can also observe similar results in the IEEE 30-bus and 118-bus systems. Noticeably, from Figs. 3–5 we can find that both players' strategies converge within 200, 250, and 400 iterations in the IEEE 9-bus, 30-bus, and 118-bus systems, respectively. Since we have proved that the converged strategies are the Nash equilibrium points, the results in the simulation are optimal under dynamic environments. Moreover, from a game-theoretic perspective, the strategies obtained by our proposed algorithm will serve as guidance for the system operator to deploy either reinforcement or repair on system components in different system configuration under the

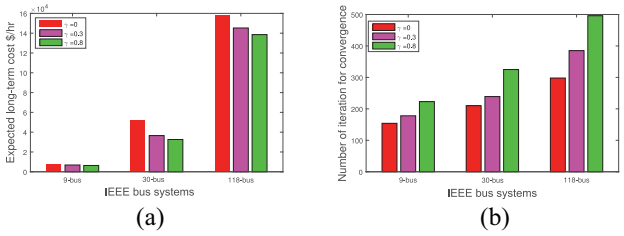

Fig. 6. Strategy analysis for stochastic game. (a) Performance analysis with regard to different γ. (b) Convergence analysis with regard to different γ.

condition that the attacker targets the most critical system components. By doing so, the system operator can reduce the risk of having cascading failures, and hence the expected long-term costs.

Besides, the computing time of our proposed algorithm is dominated by the solution of an LP, i.e., (25)–(27), at every iteration. For example, one iteration of our algorithm for the 118-bus system takes 2.72 s of which 2.57 s are due to the solution of the LP. Therefore, the computational complexity of our algorithm depends on the size of the LP and the number of iterations. Specifically, according to (25)–(27), the number of variables and constraints in the LP, depends on the number of actions, which grows linearly with the number of lines in the system. By employing the simplex algorithm, the complexity of solving one LP is a polynomial function of the number of lines in the system. Moreover, according to our simulation results, the number of iterations of our algorithm also grows linearly with the number of lines in the system. Therefore, the overall computational complexity of our algorithm is a polynomial function of the number of lines in the system. Moreover, in Figs. 3–5, the number of iterations needed for convergence does not linearly increase as the number of system buses increases, which makes our algorithm scalable even for large systems.

B. Strategy Analysis

Next, we analyze the system operator's optimal strategies in the stochastic game when the discount factor γ varies, with γ being equal to 0, 0.3, and 0.8. Recall that $\gamma \in [0, 1)$ represents the impact that current decisions can have on the long-term reward. Particularly, when γ equals 0, the game becomes a static game. When γ is larger than 0, a smaller value of γ emphasizes more on the immediate rewards and a larger γ gives a higher weight to the future rewards. In Fig. 6(a), compared with the results in the static game where $\gamma = 0$, the performance in the stochastic games where $\gamma > 0$ is much better. This is because in the stochastic games, players not only care about current rewards, but also take the future rewards into consideration. By considering both current and future rewards, players are able to obtain optimal expected long-term rewards. In addition, we can see that the higher γ is, the lower expected long-term load shedding cost the system operator can achieve. This is because when γ increases, the system operator places more emphasis on the future states and can better react to the dynamic environments, which results in more savings in the long-term cost. On the other hand, Fig. 6(a) and 6(b) together demonstrate the tradeoff between performance and computational cost. As shown in Fig. 6(b), the number of iterations needed for convergence increases as γ increases. This is because when we emphasize more on the future rewards, it takes more iterations to search for the optimal solution.

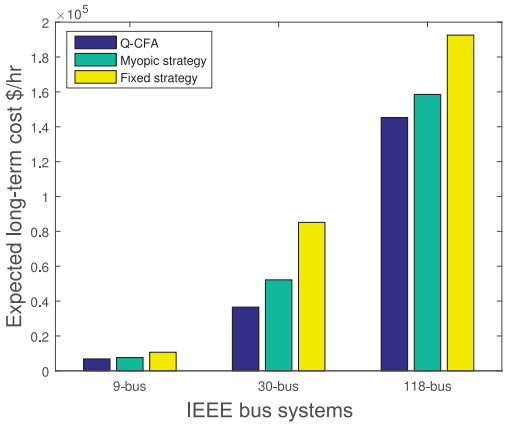

Fig. 7. Performance comparison among three strategies.

C. Performance Comparison

Finally, from the system operator's perspective, we compare the performance of the optimal strategies obtained by our Q-CFA algorithm with that of two other strategies, i.e., the fixed strategy and the myopic learning strategy. In particular, in the fixed strategy, the system operator will draw an action o uniformly from the available action space, i.e., $\mathcal{M}_O(s)$, for each state s. In the myopic learning strategy where the game is a static game ($\gamma = 0$), the system operator only considers immediate rewards and ignores the impact of the current action on future rewards. Note that it is of paramount importance to select initiating events in each algorithm because it allows the attacker to determine if the initial event can cause a cascading failure. In the three benchmark algorithms, the selections of "important line" are different. In particular, our proposed scheme optimizes the expected long-term rewards, so the selection of initiating events takes the opponent's strategy and the dynamic environments into consideration. However, as the myopic strategy is a static-game strategy, selection of initiating event only considers the opponent's strategy in current state and the strategy can be explained as trying to launch a one-time attack to cause cascading failure and achieve the maximum immediate reward. On the other hand, the fixed strategy is a uniform strategy for comparison. So the selection of initial event is uniformly distributed. We compare the optimal expected long-term cost in these three strategies in Fig. 7.

We can find that the optimal costs obtained by our proposed Q-CFA and the myopic learning strategy are much lower than that obtained by the fixed strategy. This is because both of our proposed Q-CFA and the myopic learning strategy try to minimize the attacker's maximal reward, while the fixed strategy only uniformly chooses actions from the available action set without taking the opponent's possible strategies into consideration. In addition, because our Q-CFA algorithm optimizes the

expected long-term reward while the myopic learning strategy only focuses on optimizing the strategies at the current state, our scheme outperforms the myopic learning strategy in the long run. Therefore, as a power system operator, adopting our proposed Q-CFA algorithm to defend the power system can both adapt to the dynamic state changes and attacker's intelligent strategies, which results in the best performance in the long run.

VI. Conclusion

The IoT technologies have brought both new features and significant security challenges to power systems. In this paper, we have investigated CFAs in power systems. Specifically, we have formulated a zero-sum stochastic game to analyze the interactions between an attacker and a system operator in dynamic environments for power systems. This problem is very complex and computationally intensive. Different from the previous work where complete enumeration of the system states is required, making the algorithms computationally intractable for large-scale power system applications, we propose an efficient Q-CFA learning algorithm that only searches certain related possible actions for each player in the game, making the scheme scalable with fast convergence. We have also theoretically proven that the proposed algorithm achieves the Nash equilibrium. Moreover, considering that real-time statistics and sensitive data like system transition probabilities may not be accessible in practice, which unfortunately is an indispensable assumption in previous algorithms, our scheme works efficiently without requiring *a priori* knowledge of the system transition states. Simulation results show that by considering the system dynamics and the opponent's possible strategies, the optimal policy obtained by our proposed Q-CFA algorithm can achieve much better performance compared to several benchmark schemes.

References

[1] (Jan. 2017). *Internet of Things and the Myth of the Killer App*. [Online]. Available: https://www.metering.com/magazine_articles/smart-grid-and-iot/

[2] Y. Saleem, N. Crespi, M. H. Rehmani, and R. Copeland, "Internet of Things-aided smart grid: Technologies, architectures, applications, prototypes, and future directions," *arXiv preprint arXiv:1704.08977*, Apr. 2017.

[3] S. Liu, X. P. Liu, and A. El Saddik, "Denial-of-Service (DoS) attacks on load frequency control in smart grids," in *Proc. IEEE PES Innov. Smart Grid Technol. (ISGT)*, Feb. 2013, pp. 1–6.

[4] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. Inf. Syst. Security*, vol. 14, no. 1, pp. 1–13, May 2011.

[5] A. Ashok and M. Govindarasu, "Cyber attacks on power system state estimation through topology errors," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, Jul. 2012, pp. 1–8.

[6] "Managing big data for smart grids and smart meters," IBM Corporat., Armonk, NY, USA, White Paper, May 2012.

[7] J. Chen, J. S. Thorp, and I. Dobson, "Cascading dynamics and mitigation assessment in power system disturbances via a hidden failure model," *Int. J. Elect. Power Energy Syst.*, vol. 27, no. 4, pp. 318–326, May 2005.

[8] I. Dobson, B. A. Carreras, V. E. Lynch, and D. E. Newman, "Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization," *Chaos Interdiscipl. J. Nonlin. Sci.*, vol. 17, no. 2, Jun. 2007, Art. no. 026103.

[9] B. Liscouski and W. Elliot, "Final report on the Aug. 14, 2003 blackout in the United States and Canada: Causes and recommendations," U.S. Dept. Energy, Washington, DC, USA, Tech. Rep. 40(4), Dec. 2004.

[10] M. Rahnamay-Naeini, Z. Wang, N. Ghani, A. Mammoli, and M. M. Hayat, "Stochastic analysis of cascading-failure dynamics in power grids," *IEEE Trans. Power Syst.*, vol. 29, no. 4, pp. 1767–1779, Jul. 2014.

[11] L. Liu, M. Esmalifalak, Q. Ding, V. A. Emesih, and Z. Han, "Detecting false data injection attacks on power grid by sparse optimization," *IEEE Trans. Smart Grid*, vol. 5, no. 2, pp. 612–621, Mar. 2014.

[12] A. E. Motter and Y.-C. Lai, "Cascade-based attacks on complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 66, no. 6, pp. 1–4, Dec. 2002.

[13] Y. Zhu, J. Yan, Y. Tang, Y. L. Sun, and H. He, "Joint substation-transmission line vulnerability assessment against the smart grid," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 1010–1024, May 2015.

[14] J. Yan, H. He, X. Zhong, and Y. Tang, "Q-learning-based vulnerability analysis of smart grid against sequential topology attacks," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 200–210, Jan. 2017.

[15] Z. J. Bao, Y. J. Cao, G. Z. Wang, and L. J. Ding, "Analysis of cascading failure in electric grid based on power flow entropy," *Phys. Lett. A*, vol. 373, no. 34, pp. 3032–3040, Aug. 2009.

[16] A. J. Holmgren, E. Jenelius, and J. Westin, "Evaluating strategies for defending electric power networks against antagonistic attacks," *IEEE Trans. Power Syst.*, vol. 22, no. 1, pp. 76–84, Feb. 2007.

[17] J. Salmeron, K. Wood, and R. Baldick, "Analysis of electric grid security under terrorist threat," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 905–912, May 2004.

[18] G. Chen, Z. Y. Dong, D. J. Hill, and Y. S. Xue, "Exploring reliable strategies for defending power systems against targeted attacks," *IEEE Trans. Power Syst.*, vol. 26, no. 3, pp. 1000–1009, Aug. 2011.

[19] N. S. V. Rao *et al.*, "Cyber and physical information fusion for infrastructure protection: A game-theoretic approach," in *Proc. Int. Conf. Inf. Fusion*, Istanbul, Turkey, Jul. 2013.

[20] C. Y. T. Ma, D. K. Y. Yau, X. Lou, and N. S. V. Rao, "Markov game analysis for attack-defense of power networks under possible misinformation," *IEEE Trans. Power Syst.*, vol. 28, no. 2, pp. 1676–1686, May 2013.

[21] O. Adaki. (Jun. 2014). *Attack on Power Lines Leaves Yemen in Total Darkness*. [Online]. Available: http://www.longwarjournal.org/threat-matrix/archives/2014/06/aqap_attack_on_power_lines_lea.php

[22] T. E. S. Raghavan and J. A. Filar, "Algorithms for stochastic games, a survey," *Zeitschrift für Oper. Res.*, vol. 35, no. 6, pp. 437–472, Nov. 1991.

[23] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vols. 1–2, 2nd ed. Belmont, MA, USA: Athena Sci., Jan. 2007.

[24] J. Salmeron, K. Wood, and R. Baldick, "Worst-case interdiction analysis of large-scale electric power grids," *IEEE Trans. Power Syst.*, vol. 24, no. 1, pp. 96–104, Feb. 2009.

[25] Y. Wang and R. Baldick, "Interdiction analysis of electric grids combining cascading outage and medium-term impacts," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2160–2168, Sep. 2014.

[26] A. G. P. David, C. Elizondo, J. de la Ree, and S. Horowitz, "Hidden failures in protection systems and its impact on power system wide-area disturbances," in *Proc. IEEE Power Eng. Soc. Win. Meeting*, Jan. 2001, pp. 710–714.

[27] J. S. Thorp, A. G. Phadke, S. H. Horowitz, and S. Tamronglak, "Anatomy of power system disturbances: Importance sampling," *Int. J. Elect. Power Energy Syst.*, vol. 20, no. 2, pp. 147–152, Feb. 1998.

[28] J. Chen, J. S. Thorp, and M. Parashar, "Analysis of electric power system disturbance data," in *Proc. 34th Annu. Hawaii Int. Conf. Syst. Sci.*, Washington, DC, USA, Jan. 2001, pp. 738–744.

[29] X. Liu, K. Ren, Y. Yuan, Z. Li, and Q. Wang, "Optimal budget deployment strategy against power grid interdiction," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 1160–1168.

[30] A. Neyman and S. Sorin, *Stochastic Games and Applications*, vol. 570. Dordrecht, The Netherlands: Springer, 2003.

[31] J. Filar and K. Vrieze, *Competitive Markov Decision Processes*. New York, NY, USA: Springer, 2012.

[32] S. Tamronglak, S. H. Horowitz, A. G. Phadke, and J. S. Thorp, "Anatomy of power system blackouts: Preventive relaying strategies," *IEEE Trans. Power Del.*, vol. 11, no. 2, pp. 708–715, Apr. 1996.

[33] *International Energy Outlook*, U.S. Energy Inf. Admin., Washington, DC, USA, May 2016. [Online]. Available: https://www.eia.gov/outlooks/ieo/

[34] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. 11th Int. Conf. Mach. Learn.*, vol. 157. New Brunswick, NJ, USA, Jul. 1994, pp. 157–163.
[35] L. Buşoniu, R. Babuška, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.
[36] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, Mar. 2004.
[37] J. C. Harsanyi and R. Selten, *A General Theory of Equilibrium Selection in Games*, vol. 1. Cambridge, MA, USA: MIT Press, Jun. 1988.
[38] G. Owen, *Game Theory*. New York, NY, USA: Academic, Jul. 1982.
[39] Y. Gwon, S. Dastangoo, C. Fossa, and H. T. Kung, "Competing mobile network game: Embracing antijamming and jamming strategies with reinforcement learning," in *Proc. IEEE Conf. Commun. Netw. Security (CNS)*, Oct. 2013, pp. 28–36.
[40] C. Szepesvári and M. L. Littman, "A unified analysis of value-function-based reinforcement-learning algorithms," *Neural Comput.*, vol. 11, no. 8, pp. 2017–2060, Nov. 1999.
[41] J. Hu and M. P. Wellman, "Multiagent reinforcement learning: Theoretical framework and an algorithm," in *Proc. ACM Int. Conf. Mach. Learn.*, Jul. 1998, pp. 242–250.
[42] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas, "MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Trans. Power Syst.*, vol. 26, no. 1, pp. 12–19, Feb. 2011.
[43] *IEEE 118 Bus Case Flow Limits*, Illinois Inst. Technol., Chicago, IL, USA, 2004. [Online]. Available: http://motor.ece.iit.edu/data/JEAS_IEEE118.doc

Weixian Liao (GS'13), photograph and biography not available at the time of publication.

Sergio Salinas (S'07–M'10), photograph and biography not available at the time of publication.

Ming Li (A'14), photograph and biography not available at the time of publication.

Pan Li (GS'06–M'09), photograph and biography not available at the time of publication.

Kenneth A. Loparo (S'75–M'77–SM'89–F'99–LF'16) photograph and biography not available at the time of publication.