Securing demand–response in smart grids against false pricing attacks

Tang, Daogui; Guerrero, Josep M.; Zio, Enrico

Published in:
Energy Reports

DOI (link to publication from Publisher):
10.1016/j.egyr.2024.06.068

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Energy Reports 12 (2024) 892–905

Contents lists available at ScienceDirect

Energy Reports

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

Research paper

Securing demand–response in smart grids against false pricing attacks

Daogui Tang [a,b,c,*], Josep M. Guerrero [d,e,f], Enrico Zio [g,h]

[a] School of Transportation and Logistics Engineering, Wuhan University of Technology, Wuhan, 430063, China
[b] Ningbo Zhoushan Port Group Co., Ltd., Ningbo, 315100, China
[c] Université Paris-Saclay, CentraleSupélec, Laboratoire Génie Industriel, Gif sur Yvette, 91190, France
[d] Center for Research on Microgrids (CROM), Department of Electronic Engineering, Technical University of Catalonia, Barcelona, 08019, Spain
[e] Catalan Institution for Research and Advanced Studies (ICREA), Pg. Lluís Companys 23, Barcelona, 08010, Spain
[f] Center for Research on Microgrids (CROM), AAU Energy, Aalborg University, Aalborg, 9220, Denmark
[g] Centre de Recherche sur les Risques et les Crises (CRC), MINES Paris-PSL University, Sophia Antipolis, France
[h] Energy Department, Politecnico di Milano, Milan, Italy

ARTICLE INFO

Keywords:
Smart grids
Demand–response
False pricing attack
Zero-sum Markov game
Multi-agent reinforcement learning

ABSTRACT

Two-way communication systems in smart grids help to engage consumers in demand–response programs, which can bring many benefits but also make smart grids vulnerable to cyber attacks. In this paper, a cyber attack which aims to disturb the demand–response process by injecting false electricity price signals is considered. A real-time pricing model where the operator has incomplete knowledge of the private demand–response behaviors of the customers is proposed. The vulnerability of the power system to false pricing attacks is analyzed by a Markov decision process, and the dynamic interaction between the attacker and the defender is modeled as a zero-sum Markov game where neither player has full information of the game model. For the solution of the Markov game, a model-free multi-agent reinforcement learning method is proposed to find the Nash Equilibrium policies for both players. The proposed method is applied to the IEEE 34 Node Test Feeder, in which the effect of the defense on mitigating the impact of the attack is demonstrated and different policies of the players given various resources are analyzed. The results shows that the studied cyber attack can cause a maximum of 1.96% unsatisfied load and the proposed defense measure can reduce the unsatisfied load by 50%–70% compared with the cases without defense measures.

1. Introduction

Smart grids (SGs) are the next-generation power infrastructures that integrate cyber networks with physical power grids. Information and communication technology (ICT) enables two-way communications between the two parts, facilitating the control and efficient operation of the SGs. Specifically, bidirectional communication engages customers to improve the efficiency of demand-side management (DSM) in SGs. Customers are able to shape their load profiles through various demand response (DR) programs to improve the reliability and efficiency of the grid (Avordeh et al., 2023). However, despite all the benefits, the increased usage of ICT brings new threats to the power grid. One of the most challenging risks is cyber attacks (Hasan et al., 2023).

The risks associated with the customers' engagement are load altering attacks targeting the DSM and DR programs (Li et al., 2023). In DR programs, customers are guided to reschedule their consumption patterns by responding to the variance of the power utility's incentive or electricity prices (Elsir et al., 2024). With DR programs, various types of loads are vulnerable to cyber attacks (Ayub et al., 2023), e.g., price-responsive loads (Tang et al., 2019b, 2023), frequency-responsive loads (Amini et al., 2016), and loads that respond to direct load control (DLC) command signals (Mohsenian-Rad and Leon-Garcia, 2011). In the present work, the load altering attacks caused by fabricating pricing information, namely false pricing attacks (FPAs) were studied (Tang et al., 2023, 2019a) . The basic idea behind the FPAs is that rational customers following price-based DR programs, e.g., real-time pricing (RTP) and time-of-use (TOU) pricing, will change their consumption scheduling to save electricity expenditures when they receive (potentially false) electricity pricing information, which causes a sudden load increase among some victims and, thus, possibly overload the power system (Tang et al., 2019b).

The advanced metering infrastructure (AMI) plays a key role in the DR programs of the SGs. The AMI is a two-way communication network between customers and utilities that integrates advanced sensors, smart meters, monitoring systems, computer software, and hardware as well

* Corresponding author.
E-mail addresses: tangdaogui@gmail.com, dgtang@whut.edu.cn (D. Tang).

https://doi.org/10.1016/j.egyr.2024.06.068
Received 29 February 2024; Received in revised form 15 June 2024; Accepted 29 June 2024
Available online 8 July 2024

Acronym	
SG	Smart grids
ICT	Information and communication technology
DSM	Demand-side management
DR	Demand-response
DLC	Direct load control
RTP	Real-time pricing
TOU	Time-of-use
FPA	False pricing attack
AMI	Advanced metering infrastructure
WAN	Wide area network
NAN	Neighbor area network
HAN	Home area network
MDP	Markov decision process
MARL	Multi-agent reinforcement learning
NE	Nash equilibrium
PJM	Pennsylvania-New Jersey-Maryland Interconnection
Nomenclature	
$G(V, E)$	Power grid with vertex set V and edge set E
m, m_d, n	Number of nodes, demand nodes and edges
d, d^r, d^n	Total, responsible and nonresponsible demand of a customer
$\lambda_{da}, \lambda_{rt}$	Day-ahead and real-time electricity prices
D_{da}, D_{rt}	Total day-ahead and real-time demand of the power system
pcr	Price change rate
p	Probability of response to electricity price
S	Set of states of the system
$\mathcal{A}^a$, $\mathcal{A}^d$	Action sets of attacker and defender
a^a, a^d	Actions of attacker and defender
$\mathcal{R}^a$, $\mathcal{R}^d$	Reward sets of attackers and defenders
$r^a(s, a^a, a^d)$	Reward of attackers at state s with joint action (a^a, a^d)
$r^d(s, a^a, a^d)$	Reward of defenders at state s with joint action (a^a, a^d)
π^a, π^d	Policies of attackers and defenders
$\mathcal{T}(s'\|s, a^a, a^d)$	System transition probability with the joint actions of attackers and defenders
c_i^a, c_i^d	Attack and defense resources at customer i
C_l^a, C_l^d	Resources limits of attackers and defenders
ls	Load shedding
f_{ij}	Power flow from vertex i to j
x_{ij}	Reactance of distribution lines
θ	Phase angle
$EENS$	Expected energy not supplied

as data management systems (Gungor et al., 2011). The communication networks deployed in the DR process can be divided into three types according to their size and location (Deng et al., 2015b; Kumar et al., 2019): (1) the wide area network (WAN) enables communication between generation and meter data management systems; (2) communication between distribution substations and smart meters is supported by neighborhood area networks (NAN); and (3) home area networks (HAN) are deployed within households to connect smart meters and electrical appliances. Thus, the electricity prices can be sent from the utility to substations via the WAN and then from the substations to smart meters via HAN. In this process, several parts, including the substation, the access point of the HAN and the smart meters, are vulnerable to cyber attacks (Liu et al., 2015). Among these parts, smart meters are the most vulnerable parts because of their simple authentication and encryption procedure, as well as their physical exposure (Tellbach and Li, 2018). Therefore, the present study focuses on FPAs to smart meters.

Recently, several works have addressed the resilience of power system to FPAs. In Acharya et al. (2021), the authors study the vulnerability of DR programs to causative attacks that can cause an increase in monetary cost for a utility company, power curtailment, and over- or under-frequency of power system through modifying the input or training data of an artificial intelligence-based DR program. The authors in Mishra et al. (2016) show that FPAs can cause failures of lines by modifying (reducing) the electricity prices received by some smart meters with a price change rate. Ref. Liu et al. (2015) investigated cyber attacks on guideline day-ahead prices, which can cause economic loss and lead to peak energy loads that increase the stress of the power grid. Furthermore, the attack can cause a large area of power system blackout through cascading effects (Liu et al., 2016). These works are based on the hypothesis that the behavior of customers is deterministic and known to attackers, which is unrealistic. In real life, customers' behavior is complex and normally hard to predict and, thus, is not known to attackers (Wang and Yang, 2018). In addition, attackers and defenders might have limited attack and defense resources (Wang et al., 2019). Therefore, in the present work, it is assumed that the DR behaviors of customers are private and unknown to both the operators and attackers, and both the attacker and defender have limit resources.

In general, there are mainly two strategies aimed to prevent and mitigate the impact of cyber attacks, which can also be used in combination (Wu et al., 2022a). One is to design an online mechanism/device to detect cyber attacks (Yi et al., 2023). For instance, a convolutional neural network-based online detector for RTP attacks is proposed in the authors' previous work (Tang et al., 2023). In Xing et al. (2022), the problem of attack detection under false data injection attacks and multichannel deception attacks is addressed. The other is to find the best policies to defend the power system from attacks by allocating defensive resources (Wu et al., 2022b), which is the aim of the present research.

To model attack and/or protection strategies, many works based on (partially observed) Markov decision process (MDP) approaches, game-theoretic approaches and reinforcement learning approaches have been proposed. For instance, to model the attacker's dynamic policies with the evolution of the environment, the authors in Hao et al. (2016) propose an MDP approach and analyze the attack likelihood based on the obtained policy. From the perspective of defenders, the impact of cyber attacks can be mitigated by timely detection (Drayer and Routtenberg, 2019; Hossain and Rahnamy-Naeini, 2019) and defense (Mo and Sansavini, 2017) of the attacks. Partially observable MDP has been adopted as an effective tool to help defenders detect FPAs in the long term (Liu et al., 2015, 2016). The above MDP approaches are based on the assumption that the complete system model is known to decision makers or can be approximated from observation and experience. For the cases where the full information of the MDP model is unknown, reinforcement learning methods have been introduced (Yan et al., 2016; Chen et al., 2018), where the policies are learned by direct interaction with the environment. The continuous attack sequences of topology attacks are obtained from a Q-learning approach in Yan et al. (2016). In Chen et al. (2018), a Q-learning algorithm with the nearest sequence memory is adopted to learn the optimal attack strategies for attackers who have little information about the power system. However, all these works focus on either attackers or defenders, and neglect the possible interaction between the two.

In other works, game theory has been adopted to model the strategic interaction between attackers and defenders. In Deng et al. (2015a), the interaction between attackers and defenders is formulated as a two-player zero-sum game where the defender tries to minimize the cost of launching an attack by properly allocating the defense resources within the total resource budget. However, the above model is a static game where the interaction between players is a one-shot event and the evolution of the system is ignored. To capture the dynamics of the power grid, dynamic Markov games where defenders interact with attackers with repeated plays to produce probabilistic state transitions

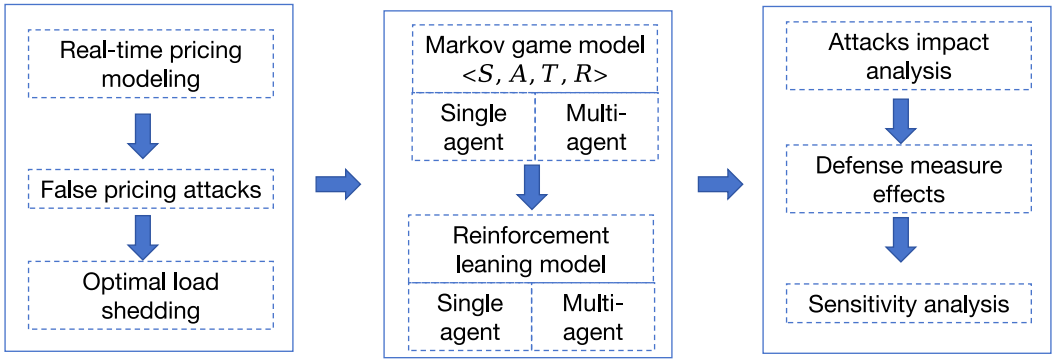

Fig. 1. The research problem-based research flow chart.

have been proposed in a wealth of recent works. For instance, a Markov game in Chen et al. (2016) is proposed to model the intrusion and defense policies for the control of substations. A stochastic game is introduced to determine the optimal resource allocation policies for defenders in Wei et al. (2016). In Zhao et al. (2020), a dynamic game is proposed to model the interactions between jammers and cyber users in a cyber–physical system suffering from jamming attacks. All these game-theoretic approaches assume that the players have full information about the game model, which is normally unrealistic in the SG context. In Ni and Paul (2019), a reinforcement learning approach is proposed to solve a multistage game. In Zhou et al. (2021), to mitigate the impact of sensors and actuator attacks, the authors propose a reinforcement learning approach to solve a two-player, zero-sum differential game. However, the impact of customers' demand response is not considered in these works.

In the present work, the vulnerability of the power system to FPAs is analyzed, assuming that both the operator, who also plays the role of the defender, and the attacker have no full information about the customers' DR model. Then, the interaction between the attacker and the defender is modeled as a Markov game where neither player has full information of the game. A model-free multi-agent reinforcement learning approach was adopted to solve the game and find the best policies for each player. The research problem-based research flow chart is presented in Fig. 1. The flow chart outlines three essential components of the proposed framework: attacks modeling, methodology modeling, and results analysis. In the attacks modeling phase, real-time pricing mechanism is modeled and false pricing attacks are introduced, and an optimal load shedding model is proposed in response to these threats. In the methodology modeling section, the study explores the dynamic interaction between attackers and defenders within a Markov game framework. Additionally, a reinforcement learning framework considering both single and multi-agent scenarios is presented. Finally, the results analysis section offers a comprehensive examination, including the impacts of attacks, the effectiveness of defense mechanisms, and sensitivity analyses. The main contributions of the work can be summarized as follows:

- We introduce a SGs pricing mechanism containing day-ahead and real-time prices. The real-time prices are updated based on the real-time total demand of the system. The customers' DR model is assumed to be private to the SG operator.
- A Markov game framework is established to analyze the dynamic decision process of the false pricing attacker and the system operator (defender), both of whom have incomplete knowledge of the consumer's behavior mechanism.
- Model-free multi-agent reinforcement learning is proposed to find the optimal defense and attack policies for players with different resources.
- The proposed approach is applied to the IEEE 34 node test feeder to demonstrate its relevance, and the vulnerability of the power system and the sensitivity of the attack and defense policies to available resources are analyzed.

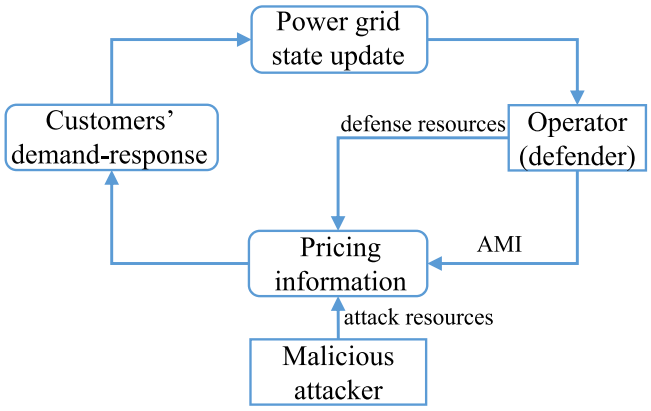

Fig. 2. The work flow of the energy market considered, subject to potential attack and defense actions.

The remainder of the paper is organized as follows. The real-time pricing model is introduced in Section 2. The Markov decision process and the Markov game model are formulated in Section 3. Section 4 introduces the multi-agent reinforcement learning method to solve the Markov game. The proposed framework is applied to an IEEE standard test feeder as a case study, and the results are discussed in Section 5. Finally, the work is concluded and possible future works are discussed in Section 6.

2. Real-time pricing and the FPA problem

This section includes the energy market considered and models the economic behavior of customers in a smart power distribution system. The workflow of the energy market is presented in Fig. 2. First, at each time slot, the SG operator publishes electricity pricing information to customers via AMI. Customers engaging in price-based demand–response programs arrange their real-time consumption by responding to real-time prices. Then, the power flow pattern of the power system is changed due to the rescheduled load. Thus, the state of the power grid is updated, and the operator renews the prices for the next time slot according to the consumption information. Malicious attackers attempt to modify the pricing information of some customers by launching FPAs within their resource limits to change the power system state, whereas the operator (defender) tries to allocate resources to protect the pricing information received by the customers. The detailed models are introduced in the following subsections.

2.1. Power grid model

Consider a smart grid for electric power distribution, represented by a directed graph $G(V, E)$ with nodes set $V = \{v_1, v_2, \ldots, v_m\}$ and edges

set $E = \{e_1, e_2, \ldots, e_n\}$, where m and n represent the number of nodes and edges, respectively. The vertex set V represents the distribution substations, transformers and demand nodes in the power system, and the edge set represents the distribution lines. Each demand node in the power system represents an energy customer, such as households and industrial buildings, who are engaged in price-based DR programs. The power demand d of a demand node consists of two parts, i.e., the price-responsive load d^r, which includes electric vehicles, laundry machines, air conditioners, etc., and the baseline nonresponsive load d^n, which includes the minimum necessary power usage to satisfy the basic life requirement, such as lighting and cooking (Tang et al., 2019b).

2.2. Demand–response model

In the DR program, utility companies release the electricity prices to each smart meter via the AMI. Each customer responds to the price signal by adjusting the energy usage of each electric appliance according to the DR model embedded in the smart meters. The real-time energy usage information is sent to the smart meters and, then, transferred to the utility companies.

In reality, different customers may have different responses to the same price. We use quadratic utility functions (Samadi et al., 2010) to model the different satisfaction levels of the customers to electricity prices:

$$U(d,\omega) = \begin{cases} \omega d - \frac{\beta}{2} d^2, & \text{if } 0 \le d \le \frac{\omega}{\beta} \\ \frac{\omega^2}{2\beta}, & \text{if } d \ge \frac{\omega}{\beta} \end{cases} \tag{1}$$

where d is the power consumption, ω is a characteristic private parameter that may vary among customers, and β is a pre-determined parameter.

The above utility function shows that utility increases when a customer consumes more power at the beginning, as the customer's demand requirement is gradually satisfied. When the power consumption reaches a certain level $\frac{\omega}{\beta}$, the customer needs no more power, and thus, the customers' satisfaction becomes saturated. For each customer, the saturation point is described by the parameter β: the lower β is, the higher the power consumption is needed to reach saturation. Customer types are differentiated by the private parameter ω. With a fixed consumption, the larger ω is, the larger the utility function value is.

The cost of consuming d kW of electricity at the price of λ dollars per kWh is λd dollars per hour. Thus, the welfare of each customer can be represented as (Samadi et al., 2010):

$$W(d,\omega) = U(d,\omega) - \lambda d \tag{2}$$

where $W(d,\omega)$ is the welfare function. λ represents the electricity price, and d is the power consumption. Each customer tries to adjust its power consumption d in response to price λ to maximize its own welfare.

2.3. Pricing mechanism model

One of the difficulties in engaging customers in DR programs is the lack of future pricing information to provide to customers to allow them to properly respond (Mohsenian-Rad and Leon-Garcia, 2010). Thus, in practice, customers are normally provided only with predictive day-ahead guideline prices upon which to schedule their consumption for the next day. In real-time consumption, the customers are given real-time prices. The ideal case is that all customers' real-time consumption matches the day-ahead scheduled consumption, but customers may normally deviate from their scheduling due for various reasons, such as load uncertainty and temporary decisions (Ma et al., 2015; Tarasak, 2011). Thus, we adopt a pricing mechanism containing both day-ahead scheduling and real-time pricing, whereby the day-ahead guideline prices are used by the customers for major energy consumption scheduling of the next day and the real-time pricing is used to address uncertainty in real-time consumption (Ma et al., 2015). This pricing mechanism has been deployed in actual applications in the US energy market (Liu et al., 2015).

2.3.1. Day-ahead scheduling

In day-ahead scheduling, energy suppliers inform customers of the predicted hourly electricity prices for the next day, as estimated from historical information. Each customer determines its hourly consumption to maximize the welfare of the next day:

$$\max_{d^t_{da} \in [d^n, d]} \sum_{t=1}^{24} U(d^t_{da}, \omega_i) - \lambda^t_{da} d^t_{da} \tag{3}$$

where λ^t_{da} and d^t_{da} are the day-ahead guideline price and corresponding consumption at time t, respectively. The day-ahead consumption can be scheduled by solving the optimization problem (3), independently at each time slot.

2.3.2. Real-time pricing

In real-time consumption, customers can determine their consumption either by responding to the real-time prices or using the day-ahead scheduled consumption. Inspired by Ma et al. (2015), we adopt a probability-based decision-making mechanism to help determine whether to respond to real-time prices. The probability of customers responding to real-time prices is calculated by:

$$p_t = k_p \cdot \left| \lambda_{da} - \lambda_{rt} \right| \tag{4}$$

where k_p is a probability parameter. The above formula represents the probability of customers responding to real-time prices, which is proportional to the absolute value of the difference between the real-time price and day-ahead price. This is reasonable since as the difference increases, the customers potentially save more expenditure by being responsive. If customers respond to the real-time price, the optimal consumption $d^{t,*}_{rt}$ can be determined by:

$$\max_{d^t_{rt} \in [d^n, d]} U(d^t_{rt}, \omega_i) - \lambda^t_{rt} d^t_{rt} \tag{5}$$

If a customer does not respond to real-time prices, it may use the day-ahead schedule as its real-time consumption but with uncertainty. We add a noise term δ representing the load uncertainty to the day-ahead schedule, and then, the non-responsive real-time consumption is expressed as:

$$d^t_{rt} = d^t_{da} + \delta^t \tag{6}$$

As in Tarasak (2011), it is supposed that the load uncertainty follows a Gaussian distribution with zero mean and variance σ^2, i.e., $\delta^t \sim N(\mu, \sigma^2)$, where $\mu = 0$.

At the end of each time slot of system operation, energy consumption information is sent to the energy provider, and thus, the real-time price for the next time slot is updated based on the total consumption. If the total consumption is less than the day-ahead scheduled consumption, the energy provider encourages consumption and vice versa. This updated rule is formulated as:

$$\lambda^{t+1}_{rt} = [\lambda^{t+1}_{da} + \eta \cdot (D^t_{rt} - D^t_{da})]^+ \tag{7}$$

where λ^{t+1}_{da} and λ^{t+1}_{rt} are the day-ahead and real-time prices for the time slot $t+1$, respectively, and $\eta > 0$ is a parameter that scales the difference between the day-ahead and real-time total consumption and $[x]^+ = \max(x, 0)$. D^t_{da} and D^t_{rt} represent the total day-ahead and real-time consumption of all customers at time t, respectively. When $D^t_{rt} > D^t_{da}$, consumption is discouraged by setting a larger real-time price than the day-ahead price, i.e., $\lambda^{t+1}_{rt} > \lambda^{t+1}_{da}$; otherwise, consumption is encouraged by setting a smaller real-time price, i.e., $\lambda^{t+1}_{rt} < \lambda^{t+1}_{da}$. When $D^t_{da} = D^t_{rt}$, we have $\lambda^{t+1}_{rt} = \lambda^{t+1}_{da}$.

This updated rule is adopted to reshape customers' energy consumption profile via incentives and encourage customers to use electricity

energy as a day-ahead schedule, to deal with the load uncertainty. The idea to guide customers' consumption behavior through incentives is widely used in electricity market with different formats, e.g., direct load control (Mohsenian-Rad and Leon-Garcia, 2011). Specifically, in the proposed real-time prices updating rule, the real-time prices are updated according to the deviation of the real-time consumption from the day-ahead schedule. This rule is reasonable for the operator since it prevents the energy consumption from deviating from the planned energy supply dramatically, and enables stable operation of the power system. Thus, it is universally adopted in real-time pricing mechanisms with uncertainty (Yi et al., 2017).

2.4. The FPA problem

The previous subsections introduce the pricing and DR mechanism enabled by the two-way communications between utilities and end-users. Communication can be achieved through various communication and protocol technologies, such as ZigBee (IEEE 802.15), WiMAX (IEEE 802.16), Wi-Fi (IEEE 802.11), Mobile-Fi (IEEE 802.20), and Ethernet (IEEE 802.3). The communication among home appliances and smart meters utilizes various communication protocols. One widely used protocol for metering measurement is IEC 61850, which is employed for Intelligent Electronic Devices (IED) in substation communication. This protocol introduces five types of communication services: Abstract Communication Service Interface (ACSI), Generic Object-Oriented Substation Event (GOOSE), Generic Substation Status Event (GSSE), Sampled Measured Value multicast (SMV), and Time Synchronization (T.S.) (Khoei et al., 2022). However, IEC 61850 is vulnerable to several types of cyber attacks in nature, including denial of service, password cracking, and eavesdropping attacks (Youssef et al., 2016).

Another widely used protocol is the IEC 62056 series, which specifies data exchange for electric meter reading, tariffs, and load control (Leszczyna, 2019). Within this series, IEC 62056-5-3 addresses the security of AMI components for data exchange, describing security techniques in the Device Language Message Specification (DLMS) and COmpanion Specification for Energy Metering (COSEM). Although the IEC 62056 standard is robust, it remains vulnerable in several aspects. First, the implementation of these protocols can be flawed. Common issues include poor key management, where weak or poorly managed cryptographic keys can be compromised; software bugs, where implementation flaws can introduce exploitable vulnerabilities; and inconsistent updates, where delays in patching and updating smart meters can leave systems exposed to known exploits. Besides, attackers with physical access to smart meters can bypass software protections. Studies have shown that smart meters are vulnerable to tampering attacks, which violate network integrity by targeting the physical layer (Khoei et al., 2022). Techniques include hardware tampering, which involves direct manipulation of the meter's hardware to alter readings, and side-channel attacks, which exploit electromagnetic emissions or power consumption patterns to extract sensitive information. Furthermore, despite encryption and authentication measures, attackers can still manipulate network traffic through various methods, such as Man-in-the-Middle Attacks and relay attacks. These vulnerabilities highlight the need for continuous improvement in both the technology and practices employed to secure smart metering systems effectively. The main steps to attack a smart meter include network surveying, service and system identification, vulnerability research and verification, and password cracking (Flick and Morehouse, 2010). Network surveying aims to identify the target systems that are network accessible, e.g., through a wireless network or HAN via hardware and software tools, such as Wireshark. Service and system identification can find the services and operating systems running on ports by port scanning. The potentially vulnerable entry points into the smart meters could be scanned by the vulnerability research and verification process. Finally, the attackers can obtain access to the smart meter if they obtain credentials by password cracking. In each step, adequate attack resources, such as hardware and software tools, hackers and economic resources, are needed. An optimal attack policy, within the limited resources acceptable, needs to be found to obtain the maximum reward from the standpoint of an attacker.

To launch a successful attack, the attacker needs to consider the following elements: (1) the target victims and (2) the type of attack. In the present research work, it is assumed that the attacker has limited resources; thus, he needs to determine the attack targets (electricity end-users) according to the dynamic system state. For the second element, there are mainly three types of cyber attacks that target electricity prices in recent research works: scale attacks, delay attacks (Tan et al., 2013) and arbitrary attacks (Giraldo et al., 2016). In the present research, arbitrary attacks are studied, where the attacker can arbitrarily modify electricity prices. In the research works about cyber attacks on real-time prices, there are mainly two kinds of representations of arbitrary attacks. One is expressed as the true prices adding a false signal, i.e., $\widetilde{\lambda} = \lambda_{rt} + \Delta\lambda_{rt}$, as in Ref. Giraldo et al. (2016). In this expression, $\Delta\lambda_{rt}$ is an arbitrary value. Another expression is by a price change rate pcr (Mishra et al., 2016), i.e., $\widetilde{\lambda}_{rt} = (1 - pcr)\lambda_{rt}$, which is adopted in our paper. Then, the home energy management system embedded in the smart meters responds to the false price $\widetilde{\lambda}_{rt}$; thus, the demand patterns of the victims of the FPAs could be possibly changed, which further leads to an increase in the total load demand of the power system. When the total demand is beyond the capacity of the power system, it will suffer from failures due to overload. Since the aim of the attacker is to cause the most severe impact to the power system by modifying the electricity prices and it costs the same amount of resources for the attacker to intrude a smart meter and modify any value of the price change rate (Flick and Morehouse, 2010), it is assumed that the price change rate keeps the maximum for all the intruded smart meters and does not depend on the dynamic state of the system.

Similarly, for the defender, individual prices can be protected by protecting individual smart meters through implementing reinforced security measures at some cost (Amini et al., 2016), e.g., by installing security hardware and software components (Mahmoud et al., 2015). The aim of the defender is to optimally allocate the available budget to prevent and/or mitigate the impact of attacks (Mo and Sansavini, 2017). The interaction between the attacker and the defender can be modeled by a competitive Markov game, in which the attacker and the defender are players. When the defender detects a cyber attack, they can respond to the attack immediately by assigning these resources to some smart meters (located at customers' houses), depending on their feasibility, for instance, asking personnel to update security software remotely or sending them to check the smart meters on the spot. Once the smart meters are protected, the electricity prices received by the customer can be protected. After one action is done by the defender, they can observe the state of the power system and calculate the immediate reward; then, another action can be made according to the system state, by which another group of customers can be protected. After the actions, the defender can observe the state of the system and calculate the immediate reward. This process is repeated until the immediate reward received by the defender no longer changes, i.e., a Nash Equilibrium is achieved. In this way, individual prices can be protected dynamically according to the state of the power system. The formulation of the Markov game model is introduced in detail in the following section.

3. Markov game model formulation

When there are no defense measures, the attack problem can be formulated as MDP. When multiple decision makers (i.e., the attacker and the defender, as in the case of interest) are involved, their interaction and decisions are modeled by a Markov game.

3.1. Basics of MDP and Markov game

An MDP can be described as a tuple $\langle S, \mathcal{A}, \mathcal{T}, \mathcal{R} \rangle$. The finite set S represents the state space of the environment. The finite set $\mathcal{A}$ denotes the action space, and $\mathcal{T}$ stands for the state transition probability. $\mathcal{R} : S \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward function. At each time step, the state of the environment is $s \in S$. The decision-maker (e.g., the attacker) takes an action $a \in \mathcal{A}$, causing the environment to transition to s' with the transition probability $\mathcal{T}(s'|s, a)$, and the decision-maker can receive a reward $r(s, a)$, which belongs to the reward set $\mathcal{R}$. Then, the process moves to the next time step.

The above MDP can be further extended to a Markov game when there are multiple decision-makers: $MG = \langle \mathcal{N}, S, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$, where $\mathcal{N} = \{1, 2, \ldots, n\}$ is a set of players. The finite sets $\mathcal{A}$ and $\mathcal{R}$ contain the set of all actions and reward functions of all players, respectively. $\mathcal{T}(s'|s, a_i, a_{-i})$ represents the next probability of the next state $s' \in S$, given the current state $s \in S$ and the actions of all players. The Markov game is a dynamic game defined as a mathematical model of strategic interactions between independent agents who control a dynamic system (Haurie et al., 2012). The difference between static and dynamic games is that the interaction between the players in a static game is one shot, whereas the interactions in a dynamic game are sequential and repeated.

In the present work, the sequential interactions between the attacker and the defender is modeled as a two-player dynamic game, where they have opposing goals. Thus, the two players have opposite reward functions and form a completely competitive zero-sum game. The detailed elements of the two-player zero-sum Markov game are defined in the following subsections. The proposed dynamic model is more realistic than the static model since the system dynamics are ignored in a static game and the players always choose the same strategies regardless of the system state, which is not wise and cannot achieve an optimal result. However, the players in a dynamic game can choose their actions accordingly as the system state evolves, which enables them to choose the policies that can always benefit them the most.

3.2. Actions of players

The game formulation is associated with the realistic world mainly by the model of players' actions (strategies). For the player, the action is to determine which customer to attack, whereas for the defender, the action is to determine which customer to protect, both considering the limited resources. The attacker wants to optimally allocate the limited resources to each smart meter to cause maximum impact on the power grid. The attacker is able to reduce electricity prices with a price change rate pcr at a cost. For instance, if the real electricity price at time slot t is λ_{rt} and the attacker reduces it with $pcr = 0.3$, the false price received by victims becomes $(1 - 0.3)\lambda_{rt} = 0.7\lambda_{rt}$. Each action of the attacker corresponds to a distribution of pcr, which should satisfy the constraints that the resources are within a limit and that the price change rate cannot be larger than the maximum rate. The action of the attacker can, thus, be formulated as:

$$a^a = [pcr_1, pcr_2, \ldots, pcr_{m_d}] \tag{8}$$

$$\sum_{i=1}^{m_d} c_i^a \le C_l^a \tag{9}$$

$$c_i^a = k_c \cdot pcr_i \tag{10}$$

$$0 \le pcr_i \le pcr_{\max} \tag{11}$$

where m_d denotes the number of smart grid demand nodes, C_l^a represents the resource limit of the attacker and k_c is a cost parameter. This study considers a linear cost function of injecting false prices (Mishra et al., 2016), as expressed by Eq. (10).

The aim of the defender is to protect the smart meters to mitigate the damage incurred by the attack with limited resources. It is assumed that the cost of protecting the prices being modified with a price change rate pcr equals the cost of modifying it. Thus, the action of the defender can be modeled as follows:

$$a^d = [c_1^d, c_2^d, \ldots, c_{m_d}^d] \tag{12}$$

$$\sum_{i=1}^{m_d} c_i^d \le C_l^d \tag{13}$$

where a^d denotes the action of the defender and C_l^d is the resource limit. It is assumed that the defense measures can prevent attackers from accessing smart meters, and thus, the prices of the secured smart meters cannot be modified.

Since both players have limited resources, they can attack/protect limit customers, and the number of customers that they can attack/protect is fixed in one action. Therefore, all possible actions can be obtained for the players during the learning process, which is known as the action space.

3.3. System state and reward

Generally, the expression of states is based on the motivation of the attacker and accessible information for both players in the game; for instance, in Refs. Ni and Paul (2019) and Yan et al. (2016), the state is presented as a combination of the status of all the transmission lines of the power system since the motivation of the attacker is to cause as many overloaded lines as possible through a topological attack. In Ref. Chen et al. (2018), the state of the power system is presented as a vector of power system parameters, such as the voltage angles and amplitudes of all buses since the attacker aims to disturb the stability of the power system. In Ref. Wei et al. (2016), the binary state of the power system is expressed by whether the attack can cause overload to the power system. Based on the above analysis, a *state* should satisfy the following conditions: (1) the state can be quantified by the system parameters accessible to both players; (2) the initial state is a steady state of the system and at least one state can reveal the attacker's motivation; and (3) the state space is finite.

In the present research, a successful attack can cause potential overloads on some distribution lines. Normally, in response to the overload, the SG operator will immediately shed some load to keep the system stable and minimize the impact (Tang et al., 2019b). Therefore, the state of the power system can be defined by whether load shedding occurs, and the immediate reward of the players can be quantified by the loss incurred by performing load shedding.

In the present work, the optimal load shedding is adopted to determine how much load should be shed at each demand node facing potential overload. The DC power flow model is adopted here, as it is commonly used in economic and vulnerability analyses in both power transmission (Mishra et al., 2016; Yuan et al., 2011) and distribution systems (Tang et al., 2019b; Mena et al., 2014) due to its fast solution. The optimal load shedding problem is formulated as follows:

$$\min \quad L = \sum_{i=1}^{m_d} ls_i \cdot w_i \tag{14}$$

$$\text{s.t.} \quad 0 \le ls_i \le d_i \tag{15}$$

$$-f_{ij}^{\max} \le f_{ij} \le f_{ij}^{\max} \tag{16}$$

$$P_{ij}^{\min} \le P_i \le P_{ij}^{\max} \tag{17}$$

$$f_{ij} = (\theta_i - \theta_j) \cdot x_{ij}^{-1} \tag{18}$$

$$\sum_{(i,j)\in e^+} f_{ij} - \sum_{(j,i)\in e^-} f_{ji} = \begin{cases} P_i & i \in P \\ ls_i - d_i & i \in D \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

where ls_i denotes the load shedding at customer i and w_i stands for the customer's loss of shedding per unit load, which quantifies the importance of different types of loads. f_{ij} represents the power flow

through the distribution lines connecting customers i and j. If power flows from i to j, $f_{ij} > 0$; otherwise, $f_{ij} < 0$. P is the power generation, and θ denotes the phase angle. The reactance of a distribution line is represented by x. e^+ and e^- are the lines directed outwards from and inward to node i, respectively.

The detailed physical constraints of the optimization problem can be interpreted as follows. Constraint (15) ensures that the load shedding is no more than the load. Constraint (16) gives the capacity of the distribution lines. Constraint (17) limits the power generation within its minimum and maximum output. Constraints (18) and (19) describe the physical power flow constraints.

After solving the load shedding of each customer, we can also obtain the expected energy not supplied ($EENS$), which is a commonly used metric for measuring the security of the supply of a power system (Wang et al., 2017):

$$EENS = \frac{\sum_{i=1}^{m_d} ls_i}{\sum_{i=1}^{m_d} d_i} \tag{20}$$

The index $EENS$ shows whether there exists an unsatisfied load, and it is adopted here to indicate the state of the system:

$$s = \begin{cases} 1 & EENS = 0 \\ 2 & EENS > 0 \end{cases} \tag{21}$$

State 1 is the normal state where each customer's demand requirement is satisfied, whereas state 2 is the desired state of the attacker. The states are observable to both players since the attacker can compromise the DR management system and obtain the state information as easily as the defender (Barreto and Cárdenas, 2018). In state 1, the attacker chooses an action to inject false prices into some smart meters, and then, the customers respond to the (false) prices with uncertainty. After DR, the system transitions from state 1 to state 2 with the transition probability, and the attacker obtains the immediate reward, which can be calculated as the loss of the unsatisfied load:

$$r^a = L^* = \{\min L = \sum_{i=1}^{m_d} ls_i \cdot w_i : (15)\text{–}(19)\} \tag{22}$$

where r^a stands for the immediate reward and L is the loss of the unsatisfied load due to the attack, which can be calculated by Eqs. (14)–(19).

Since it is a zero-sum game, the reward of the defender equals the loss of the attacker:

$$r^d = -L^* = -\{\min L = \sum_{i=1}^{m_d} ls_i \cdot w_i : (15)\text{–}(19)\} \tag{23}$$

in which r_d represents the immediate reward of the defender.

It should be noted that both the attacker and the defender have incomplete knowledge about the DR behavior of customers, so that the above immediate reward cannot be surely estimated beforehand but can be observed after the attack and defense process.

3.4. Transition probability

In Fig. 2, the state transition process concerning the attack and defense is illustrated. At the beginning of each time slot, the SG operator distributes real-time prices to each end-user customer. The attacker injects false real-time prices with a price change rate into the selected customers, whereas the defender chooses the smart meters to protect. The system state will transition from normal state 1 (i.e., no unsatisfied load) to state 2 if the attack can cause a sufficient load increase that increases the power flow in some distribution lines beyond their capacity. This transition from one state to another is shaped by: (1) the probability of customers responding to real-time prices; (2) attack and defense actions, i.e., attacked and protected customers; and (3) power system parameters, e.g., generation and distribution capacities. It is assumed that both players have no information about the uncertainty of customers' response to the DR model; thus, the state transition probability is unknown to the players.

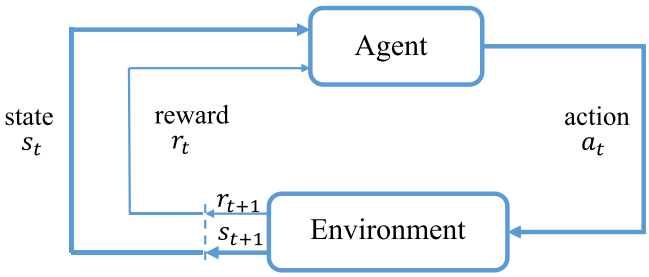

Fig. 3. The general reinforcement learning process through interaction between the agent and the environment (Sutton and Barto, 2011).

4. The proposed reinforcement learning solution

The Markov game model introduced above is incomplete for the players and, thus, cannot be solved by the traditional methods. In the present work, a model-free reinforcement learning approach is proposed, in which the agent learns the best policies by directly interacting with the environment without a model of the environment dynamics.

4.1. The single-agent case

In this paper, first the vulnerability of the power system to FPAs is investigate, so in this subsection, we consider the case without defense measures and solve the optimal policies of the attackers. The single-agent reinforcement learning process is constructed over the MDP. The basic reinforcement learning framework is depicted in Fig. 3, where the learning agent interacts with the environment by selecting an action and then perceives evaluative feedback of the immediate effect of the action, i.e., the immediate reward. The environment evolves to a state that might be the same as or different from the previous one. By taking a sequence of actions from the action space, the agent can assess the value of each action via a trial-and-error procedure and seek a series of actions to produce the maximal cumulative rewards.

4.1.1. Policy

The behavior of an agent is described by its policy, which is either stochastic or deterministic. The policy specifies how the agent chooses its actions, given the current environment state (Buşoniu et al., 2010). The stochastic policy associates different probabilities to each action in a state, whereas the deterministic policy indicates that the agent chooses a specific action to maximize a long-term return. The stochastic policy satisfies the following requirement:

$$\sum_{a \in \mathcal{A}} \pi(s, a) = 1 \qquad \forall s \in S \tag{24}$$

where $\pi(s, a)$ is the probability of choosing action a given state s.

4.1.2. Discounted return

The ultimate goal of the agent is to find a policy that can maximize a long-term return from each time step. Generally, the long-term reward is calculated by a discounted cumulative function of the immediate reward of each learning process:

$$Q = \sum_{t=1}^{n} \gamma^{t-1} r_t(s_t, a_t) \tag{25}$$

where Q stands for the long-term return of the attacker, and $r_t(s_t, a_t)$ represents the immediate reward that the attacker receives at time t with the system state being s_t and the attacker choosing action a_t. $\gamma \in [0, 1)$ is a discount factor that weighs the immediate reward at each time step; $\gamma = 0$ indicates that the agent cares only about maximizing the current immediate reward.

The task of the agent is to maximize the above long-term discounted return while it is only able to receive immediate feedback at each time

step. One way to achieve this is by solving the optimal state–action value function, which is the expected return given that the agent takes action a at state s, and follows policy π:

$$Q^{\pi}(s,a) = E\{\sum_{t=1}^{n} \gamma^{t-1} r_t(s_t,a_t)|s_1 = s, a_1 = a, \pi\} \tag{26}$$

The above function is denoted as the Q-function and the value obtained from the Q-function is called Q-value. Therefore, when a state is observed, the optimal action can be determined by choosing an action with the optimal Q-value:

$$\pi^*(a) = \arg\max_{a\in\mathcal{A}} Q^*(s,a) \tag{27}$$

Traditionally, if the complete model of the MDP is known, the optimal Q-value problem (27) can be solved by dynamic programming (Bellman, 1966). In the case of interest, since we cannot obtain the dynamic model of the environment, a model-free approach, called *Q-learning* (Watkins and Dayan, 1992), is adopted here to learn the Q-function directly.

4.1.3. Q-learning algorithm

Q-learning is an off-policy temporal-difference reinforcement learning algorithm, in which the agent is able to find the optimal Q-values from the learned experience without the complete picture of the environment. In Q-learning, each Q-value of the action-state pair is recorded in a Q-Table, and is initially set to the same arbitrary value and, then, updated iteratively during the learning process. The update rule is as follows:

$$\begin{aligned} Q_{k+1}(s_t,a_t) = Q(s_t,a_t) + \alpha\{r_{t+1} + \gamma \max_a Q(s_{t+1},a) \\ - Q_k(s_t,a_t)\} \end{aligned} \tag{28}$$

where α is the learning rate which controls the speed of the new information overriding the old information. Setting $\alpha = 0$ will stop the learning and make the agent conservative to the initial estimate, whereas a learning rate of one will make the agent focus on the new information and lose the already learned experience. Normally, the learning rate is time-varying and decays over time (Buşoniu et al., 2010).

4.1.4. Exploitation and exploration

In the Q-learning algorithm, the alternative state–action pairs need to be visited infinitely often to converge to the optimal policy rather than to suboptimal policies. Thus, it is necessary for the agent to deal with the trade-off between exploiting the already learned experience, i.e., exploitation policy, and exploring new information, i.e., exploration policy. One simple yet efficient approach to balance such a trade-off is the ϵ−greedy approach, where the agent chooses an action randomly with probability ϵ and takes the learned optimal action with probability $1-\epsilon$.

4.2. The multi-agent case

When there is a defender against the attacker, the interaction between the two players formulates a two-player Markov game, in which the attacker is a leading player and the defender is a passive player. In this case, it is necessary to know which customers are vulnerable to being attacked (i.e., the optimal strategy of the attacker) and the optimal policy for the defender to defend against false pricing attacks. Therefore, both players' optimal policies are considered. To address this, the multi-agent reinforcement learning (MARL), which is generalized from the single-agent case, is adopted here. The main idea of the MARL is similar to the single case in Fig. 3, the main difference being the fact that there are multiple agents acting in the environment and each agent might takes into consideration other agents' effects when making decisions.

A general framework of the application of MARL into a Markov game where two agents, i.e., the attacker and the defender, act on the environment is depicted in Fig. 4. In the game, the attacker and the defender select actions from their action spaces independently and simultaneously from their respective action spaces, according to the ϵ-greedy exploration and exploitation rules: the attacker randomly chooses the actions with probability ϵ and takes the learned optimal actions with probability $1-\epsilon$. Then, each player receives an immediate reward evaluating the value of their actions from the environment, and the environment transitions to the next state with the effect of the joint action. The long-term return of the joint action can be assessed by the Q-function: The long-term return of the joint action can be assessed by the Q-function:

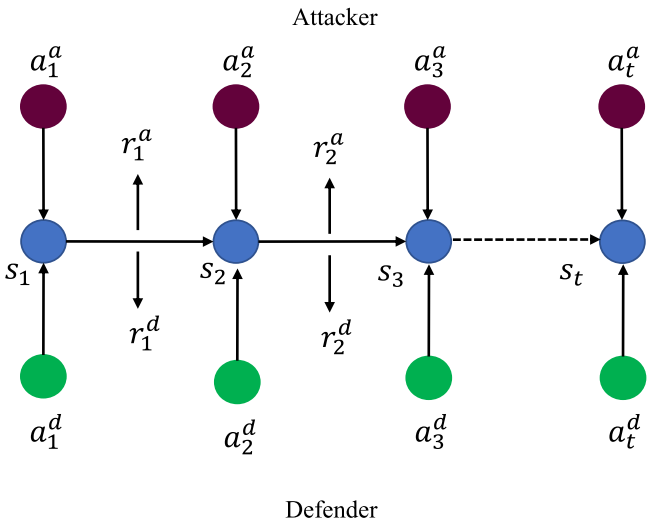

Fig. 4. The reinforcement learning framework for the Markov game.

$$\begin{aligned} Q^{\pi}(s,a^a,a^d) = E\{\sum_{t=1}^{n} \gamma^{t-1} r_t^a(s_t,a_t^a,a_t^d)|s_1 = s, \\ a_1^a = a^a, a_1^d = a^d, \pi^a, \pi^d\} \end{aligned} \tag{29}$$

The expected long-term return of the attacker at each state depends on the joint policy of the players:

$$V_{\pi_a,\pi_d}^a(s) = E\{\sum_{t=1}^{n} \gamma^{t-1} r_t^a(s_t,a_t^a,a_t^d)\} \tag{30}$$

in which $V_{\pi_a,\pi_d}^a(s)$ represents the state value at state s with the attacker and the defender following the strategy π_a and π_d, respectively. $r_t^a(s_t,a_t^a,a_t^d)$ stands for the reward of the attacker at time t with the state s_t and the joint actions of the attacker and defender (a_t^a,a_t^d).

The goal of the players is to find an optimal policy to maximize the above value function given the policies of other players. Traditionally, in a zero-sum game, such an optimal solution can be given by the Nash equilibrium (NE), $\pi^* = (\pi^{a,*},\pi^{d,*})$, which is defined as:

$$V_{\pi^a,\pi^{d,*}}^a \le V_{\pi^{a,*},\pi^{d,*}}^a \le V_{\pi^{a,*},\pi^d}^a \qquad \forall s \in S \tag{31}$$

The NE characterizes a state where no agent has an incentive to deviate from. The existence of NE for a discounted zero-sum game has been proven in many works. Interested readers may refer to Refs. Shapley (1953), Wei et al. (2016) and Daskalakis et al. (2022) for more details. In this paper, the focus is more on the modeling of defense mechanisms against FPAs with game theoretical techniques and quantifying the game parameters. In a fully competitive game, where the complete model cannot be derived, the minimax Q-learning algorithm (Littman, 1994), extended from Q-learning, is proposed to approximate the optimal policies for the players. In particular, the minimax Q-learning algorithm, extends the application of Q-learning from MDP to Markov games and uses a similar temporal-difference rule to update the Q-Table. At state s, the attacker chooses an action a^a according to the ϵ–

greedy policy, which afterward acts on the environment. Then, at state s_{t+1}, the attacker observes the reward of the action and the opponent's action a^d, and updates the Q-Table with the rule:

$$Q_{k+1}(s,a^a,a^d) = Q_k(s,a^a,a^d) + \alpha\{r^a_{k+1} + \gamma V^a(s_{k+1}) - Q_k(s,a^a,a^d)\} \tag{32}$$

where $V^a(s_{k+1})$ is the minimax value of the attacker:

$$V^a(s) = \max_{\pi^a} \min_{a^d\in\mathcal{A}^d} \sum_{a^a\in\mathcal{A}^a} Q(s,a^a,a^d)\pi^a \tag{33}$$

and can be obtained by solving a linear programming problem (Vanderbei et al., 2015). In a zero-sum game, the value function of the defender equals the negative value of that of the attacker, defined in Eq. (30), i.e., $V^d_{\pi_a,\pi_d}(s) = -V^a_{\pi_a,\pi_d}(s)$, and the value of the defender can be calculated by

$$V^d(s) = \min_{\pi^d} \max_{a^a\in\mathcal{A}^a} \sum_{a^d\in\mathcal{A}^d} Q(s,a^a,a^d)\pi^d \tag{34}$$

Thus, the optimal NE policy for the defender can be obtained by solving the above equation. After obtaining the optimal policies for the attacker and defender, they can choose their actions according to the optimal policies. For the attacker, each action is a distribution of the price change rate that is injected into customers' smart meters. Therefore, attackers can choose victims to attack with the optimal policy. On the other hand, the defenders split their defense resources into the customers according to the optimal policy. Furthermore, the expected payoff of the defender at equilibrium can also be used to quantify the vulnerability level of a distribution system against the studied attacks, therefore, uncovering recommendations for system improvement.

The proposed reinforcement learning algorithm can converge to optimal policies, i.e., the NE of the proposed stochastic game. In the present research, we follow the methods in Ni and Paul (2019) to prove the convergence. Eq. (32) can be rewritten in a general form:

$$Q(s,a^a,a^d) = (1-\alpha)Q_k(s,a^a,a^d) + \alpha\{r + \gamma V(s')\} \tag{35}$$

By substituting $\alpha = 1$, the Equation becomes:

$$Q(s,a^a,a^d) = \{r + \gamma V(s')\} \tag{36}$$

The above Equation converges to the Nash Equilibrium if: (1) the action and state spaces are finite; (2) Var(r) is bounded; (3) if $\gamma = 1$, all policies lead to a cost free terminal state with probability 1.

Proof. Subtracting $Q^*(s,a^a,a^d)$ from both sides of Eq. (36), we can obtain:

$$Q(s,a^a,a^d) - Q^*(s,a^a,a^d) = r + \gamma V(s') - Q^*(s,a^a,a^d) \tag{37}$$

Defining $F = Q(s,a^a,a^d) - Q^*(s,a^a,a^d)$ and $F(s,a^a,a^d) = r + \gamma V(s') - Q^*(s,a^a,a^d)$, we have:

$$\begin{aligned}
&\max_{a^a\in A^a} \min_{a^d\in A^d} |F(s,a^a,a^d)| \\
&= \gamma \max_{a^a\in A^a} \min_{a^d\in A^d} |\sum_{s'\in S}[V(s) - V^*(s')]| \\
&\le \gamma \max_{a^a\in A^a} \min_{a^d\in A^d} |\sum_{s'\in S} \max|Q(s,a^a,a^d) - Q^*(s,a^a,a^d)| \\
&= \gamma \max_{a^a\in A^a} \min_{a^d\in A^d} \sum_{s'\in S} V^{\Delta} \\
&= H(V^{\Delta})
\end{aligned} \tag{38}$$

where H is the value iteration operator. It can be seen from the above Equation that F is a contraction, implying that F has a unique solution when $\gamma < 1$, according to the Banach contraction theorem. In addition, because F is a mapping of maximizing $|Q(s,a^a,a^d) - Q^*(s,a^a,a^d)|$, the game converges to the optimal $Q^*(s,a^a,a^d)$ (Ni and Paul, 2019).

The details of the single-agent and multi-agent cases of the reinforcement learning algorithm for solving the proposed Markov game model are given in the pseudo-code in Algorithm 1.

Algorithm 1: Proposed reinforcement learning algorithm

```
The SG operator publishes the day-ahead prices in advance
Customers schedule their consumption for the next day according to (3)
Initialization:
∀s ∈ S, a^a ∈ A^a, a^d ∈ A^d, Q(s,a^a,a^d) = 1, V(s) = 1, π^a(s,a) = 1/|A^a|, let α = 1;
while Trial < MaximumTrial do
    Choose a random action s from A^a with probability ε
    Otherwise, choose action a*
    if defense=true then
        The defender chooses actions a^d with ε−greedy policy
    end
    Customers respond to real-time prices with probability derived by (4)
    Otherwise, determine real-time consumption according to (6)
    The attacker observes the immediate reward, the state s′, and the action of the defender
    if defense=true then
        Update Q-Table according to (32)
        Derive π^a by solving (33)
        Derive π^d by solving (34)
    else
        Update Q-Table according to (28)
        Derive π^a by solving (27)
    end
    Trial = Trial + 1, s = s′, α = α · decay
end
```

5. Case study

We apply the proposed approach to the IEEE 34 Node Test Feeder (Kersting, 1991) and analyze the effects of the FPAs with and without defense measures.

5.1. The test system

The IEEE 34 node test feeder is a radial power distribution system containing a reference bus and 33 demand nodes, which are engaged in the price-based DR program. The day-ahead hourly prices are real data published by the regional transmission organization named PJM (Day-Ahead Hourly, 2019). The customers in the test feeder schedule their consumption for the next day with the day-ahead prices by the proposed welfare maximization method. The private customer utility parameters ω are randomly chosen within [4, 10], with an incremental step of 1.

5.2. Real-time prices

The day-ahead and updated real-time prices, and their corresponding expected demands, are presented in Fig. 5. We suppose that the real-time and the day-ahead prices are the same at the beginning of the day, after which the real-time prices are updated according to the difference between the real-time and the day-ahead scheduled total demands of the power system. From the Figure, we can see that when the price difference is relatively small, e.g., at 1, 2, 3, and 4 h, few customers respond to the real-time prices. The real-time demands at these time slots deviate from the day-ahead scheduled demands due to the load uncertainty. The accumulated deviation, on the other hand, enlarges the difference between λ_{da} and λ_{rt}, which increases the probability of customers responding to real-time prices. If the real-time price is larger than the day-ahead price, e.g., the prices at 5 and 18 h,

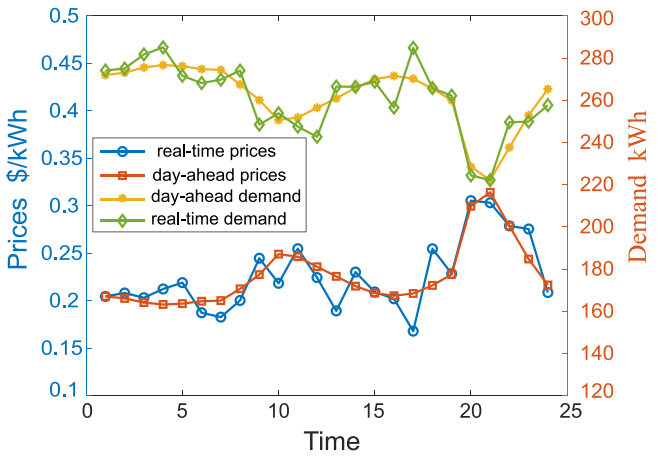

Fig. 5. The prices and corresponding demand.

the demand at the following hours, e.g., 6 and 19 h, is reduced due to the customers' responses to the real-time prices; whereas when $\lambda_{rt} < \lambda_{da}$, e.g., the prices at 7, 13 and 17 h, the consumption at the time slots afterward is increased due to the response. In both cases, the difference between the day-ahead and real-time consumption will decrease. Thus, the proposed real-time pricing mechanism is able to prevent real-time consumption, deviating largely from the planned schedule.

Furthermore, the pricing mechanism sets a natural barrier to possible series FPAs. For instance, an FPA at 17 h may lead to a significant difference between the day-ahead and the real-time demands, which in turn will increase the real-time price at 18 h, and the total demand of the power system will be reduced accordingly, if the attacker does not maintain the previous FPA. Therefore, it takes the attacker more resources to conduct a series of PFAs. In practice, FPAs are continuous and repeated, so the proposed pricing mechanism is helpful to mitigate the impact of FPAs.

In Fig. 5, the total demand of the power system is presented, however, the demand of each customer cannot be demonstrated. In Fig. 6, we take 17 h as an example to show the day-ahead scheduling and the expected real-time demand of each customer. The expected real-time demand is calculated from the results of the simulation. The results show that, in general, real-time consumption is able to attain day-ahead rescheduling for most customers by following the proposed pricing mechanism.

5.3. Results without defense

This study investigates how the attacker carries out FPAs with different attack resources when there are no defense measures. The attacker injects false real-time prices to customers, who then possibly respond to the false prices and increase their loads as expected by the attacker. A large difference between the day-ahead and real-time prices makes customers more likely to respond to the attack. Thus, the most desired opportunity to attack is when the real-time price is much lower than the day-ahead one, which is satisfied at 17 h, as shown in Fig. 5. Therefore, it is supposed that the attacker chooses this time to launch an attack.

Suppose the attacker can inject false prices with a price change rate $pcr \in \{0, 0.5\}$. A straightforward way to determine the action space is to choose demand nodes within the whole power system, where the number of actions is $\binom{m_d}{C_l^a}$, which is computationally prohibitive for large-scale distribution systems. Thus, in the present work, we consider regional attacks where the attacker chooses consecutive demand nodes to attack within the attack resources. Therefore, the number of actions, in this case, is $(m_d - C_l^a + 1)$, which is significantly reduced. Furthermore, FPAs are practically spatially localized (Tan et al., 2013; Zhang et al., 2016), and it is reasonable for the attacker to launch regional attacks, rather than the attacks on the full system (Xiang et al., 2016).

Learning speed is controlled using a relatively large learning rate $\alpha = 0.3$ in the reinforcement learning algorithm and choose actions randomly with the exploration rate $\epsilon = 0.3$ in the learning process (Yan et al., 2016). Meanwhile, discount factor $\gamma = 0.9$ is set to weigh the future rewards (Sutton and Barto, 2011). We first consider the case in which the attacker adopts a pure strategy to obtain a deterministic attack policy and, then, investigate the best attack policies when the attacker is able to inject false prices into different numbers of customers. The selective convergence of the learned Q-values is shown in Fig. 7.

The results in Fig. 7 show that the $Q(s, a^a)$-values start from the same initial value and increase rapidly at the beginning of the learning process, which indicates that the attacker searches the effective attack policy effectively from the beginning of the initial trials. Then, the Q-values are updated at each trial and become saturated at different levels for different attack resources after about 50 trials, showing that the attacker has already learned the optimal policies. The pure strategy attack policies with 2, 4, 6 and 8 demand nodes being attacked are depicted in Fig. 8 with different colors and the detailed attack policies of all cases are listed in Table 1.

In Fig. 8, the attack policies when $C_l^a = 2, 4, 6$ and 8 are covered by red, yellow, blue and green colors, respectively. It can be observed that the attacked demand nodes are mainly in the middle and rear regions of the power system, showing that these parts are the most critical to the cyber attacks. It can be noted that the attack policies have the relationship that $\pi_2^a \subset \pi_6^a \subset \pi_8^a$, showing the consistency of the attack policies in these cases. Combined with Table 1, it can be seen that demand node 832 is attacked in almost all cases. This is because demand node 832 is one of the customers with the largest consumption, which can be seen in Fig. 6. The $EENS$ that results from the different attack policies is reported in Table 1, which shows that the attacks can cause the system to lose 0.31% of the demand when $C_l^a = 2$ and that the $EENS$ increases gradually as more customers are attacked, reaching a maximum impact of 1.96% of unsatisfied load when $C_l^a = 16$.

5.4. Effects of defense measures

In this section, the effects of the defense measures implemented with different resources are discussed. Considering that the attacker and defender use a mixed strategy, the policy is a probability distribution of available actions. First, the case where the defender protects only one demand node is investigated and the results when the attacker can fabricate fake prices for 4, 6, ..., 16 end users are analyzed.

Fig. 9(a) shows the $Q(s, a^a, a^d)$-values of the attacker at each trial of the learning process. From the Figure, in the first 60 trials, the Q-value remains at the initial value when the attacker's resource is 16, which indicates that the first trials cannot find effective policies to cause overloading of the system. After the initial 60 trials, the attacker quickly explores effective actions for causing overloading and finds the optimal policy after approximately 100 trials. For cases with fewer attack resources, it takes the attacker a longer time to explore effective policies and find the optimal strategy. By comparison to the $Q(s, a^a)$-values achieved when no defense measures are in place, in Fig. 7, one can see that it takes the attacker longer time to search for effective actions that can affect the normal operation of the system.

The $EENS$ with and without the implementation of defense measures is presented in Fig. 9(b) and allows for a comparison of the impacts that the attacker can cause to the power system after learning the optimal policies in the two scenarios. By comparison to the case without defense measures, the $EENS$ with $C_l^d = 1$ and $C_l^a = 4$ is approximately 0.17% and the $EENS$ with $C_l^d = 1$ and $C_l^a = 16$ is 1.2%. The defense can reduce the unsatisfied load by 50%–70% compared with the case without defense measures. In each trial, the attacker has to take into consideration the possible action of the defender when he chooses the policy, which makes the attacker unable to choose

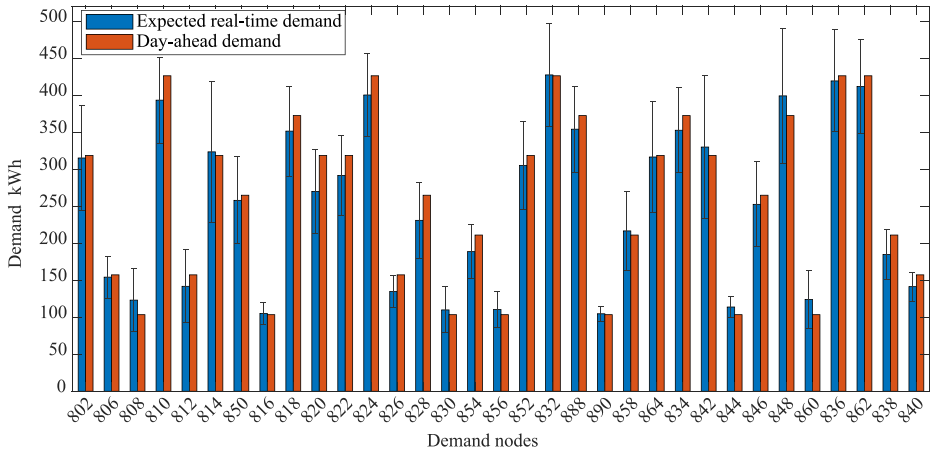

Fig. 6. Day-ahead scheduled and expected real-time demands for each customer at 17 h, with standard deviation error bars.

Table 1
Optimal pure strategies of the attacker with different attack resources.

Attack resources	Attack policy	$EENS$ (%)
2	832, 852	0.31
4	834, 842, 858, 864	0.38
6	832, 852, 854, 858, 888, 890	0.60
8	832, 852, 854, 856, 858, 864, 888, 890	0.90
10	832, 834, 842, 844, 846, 852, 858, 864, 888, 890	1.18
12	832, 834, 836, 838, 842, 844, 846, 848, 858, 860, 862, 864	1.30
14	832, 834, 836, 842, 844, 846, 848, 852, 854, 858, 860, 864, 888, 890	1.55
16	830, 832, 834, 836, 842, 844, 846, 848, 852, 854, 856, 858, 860, 864, 888, 890	1.96

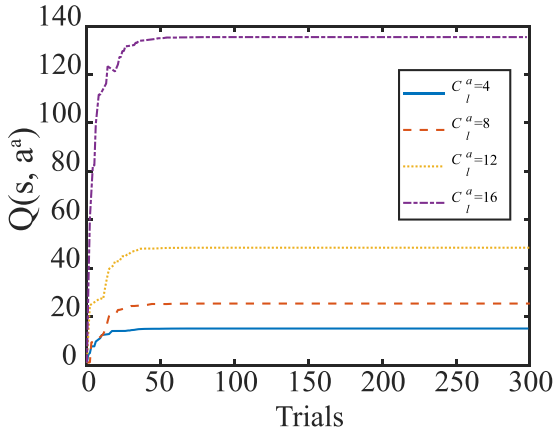

Fig. 7. The convergence of the learned $Q(s, a^a)$-values at each trial with $C_l^a = 4, 8, 12$ and 16.

the optimal actions, unlike in the scenario without defense. From the figure, it can also be seen that when the Markov game reaches the NE, there is still some demand that is not supplied. This is because the attacker has more resources than the defender, thus other smart meters can be the attack target despite of the protected one.

To further analyze the effect of the defense, the resources for attack and defense are varied from 4 to 16 nodes. The results are presented in Fig. 10. It can be observed that the attack can cause the system shed 0.8% of the demand requested when the attacker alters the price of 16 customers and the defender protects 4 customers. With the defense resources increasing or the attack resources decreasing, the unsatisfied demand reduces accordingly. When the defense resources overwhelm the attack resources, the impact of the attack is too foible to cause any load shedding in the power system. From Figs. 9(b) and 10, we can conclude that defending the power system is of significant importance to mitigate the impact of FPAs. The defense can mitigate more than half of the load shedding even with protecting one demand node and the effect can be enhanced when there are more defense resources for the studied system.

5.5. Sensitivity analysis of defense policies

Suppose the defender has resources $C_l^d = 10$: the different optimal policies for allocating the resources to achieve the best mitigation of the attacks are illustrated in Fig. 11(a).

The results show that the defender is more inclined to focus the resources on a specific set of demand nodes, even when the resources are less than those of the attacker, e.g., when $C_l^a = 14$ and 16. The reason is that the defender normally plays a passive role in the game and the aim is to mitigate the damage incurred by the attacker. Thus, even though the defender has fewer resources, the optimal policies focus on the demand nodes that are vulnerable to the attack.

The defender's policies under various defense resources are shown in Fig. 11(b). The results are consistent with the finding in (a) that the defender prefers to focus the resources on a few customers. In addition, combining the results of Fig. 11(a) and (b), the demand nodes secured with a large probability in most cases are from nodes 830 to 846, which are mainly the nodes distributed in the area covered by different colors in Fig. 8, showing the criticalities of these nodes that are protected in the studied system.

5.6. Discussion

In the current study, we introduce a real-time pricing mechanism followed by the formulation of a Markov game to characterize the dynamic interaction between attackers and defenders in false pricing attacks targeting customers' demand–response process. Subsequently, we propose a model-free reinforcement learning method to address the Markov game. Compared to alternative approaches, the proposed model-free reinforcement learning method offers several distinct advantages. First and foremost, it eliminates the need for explicit modeling of the system dynamics, thereby alleviating the burden of accurately

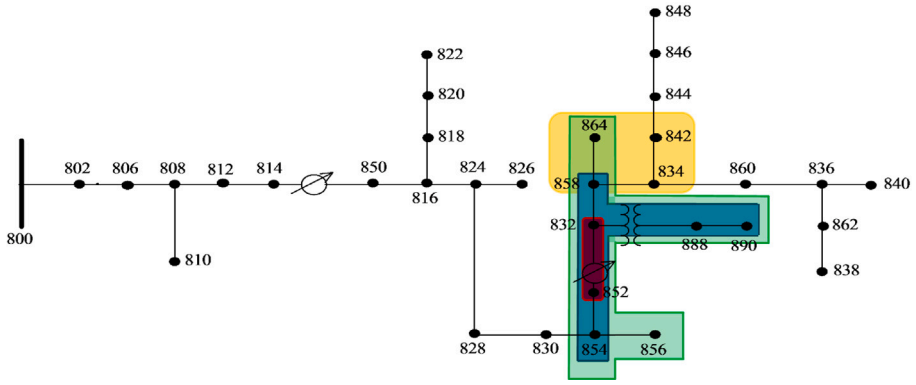

Fig. 8. The selective optimal attack policies of $C_l^a = 2$, 4, 6 and 8. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

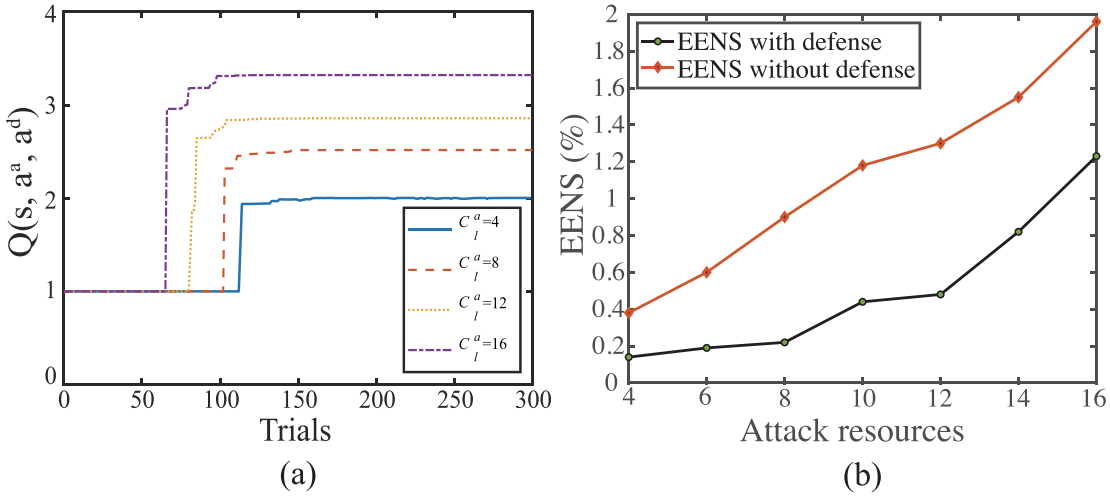

Fig. 9. (a) Convergence of the learned $Q(s, a^a, a^d)$-values at each trial, with attack resources C_l^a =4, 8, 12 16 and defense resources $C_l^d = 1$; (b) The comparison of the $EENS$ with defense resources $C_l^d = 1$ and without defense.

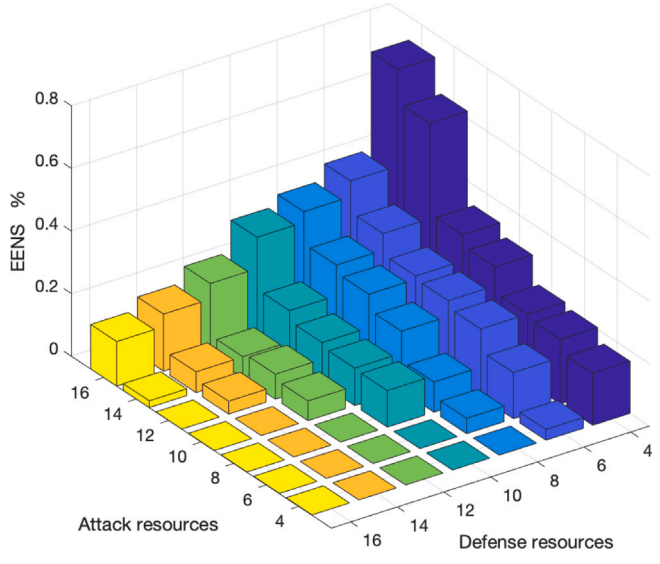

Fig. 10. Impact of the attacks for various attack and defense resources.

specifying the underlying stochastic processes (Liu and Wang, 2021). This feature makes the method particularly well-suited for scenarios where the dynamics of the system are complex or difficult to model accurately.

Furthermore, scaling up to larger systems presents an opportunity to uncover additional challenges and considerations, including the logistical complexities associated with integrating daily pricing updates across a multitude of AMI devices. The model-free nature of reinforcement learning allows for greater flexibility and adaptability in handling diverse and evolving environments. The method can autonomously learn and adapt its behavior based on feedback from the environment, enabling it to effectively cope with uncertainties and non-stationarities inherent in real-world systems. Moreover, by leveraging reinforcement learning techniques, the proposed method has the potential to discover optimal strategies that may not be apparent through traditional analytical or deterministic approaches (Tang et al., 2023; Khalid, 2024). This capability enables the method to achieve superior performance and robustness in challenging and dynamic environments.

The proposed method exhibits versatility across a multitude of scenarios. For example, the envisaged real-time pricing mechanism constitutes a comprehensive framework applicable to diverse neighborhood area networks. Similarly, the Markov game-based defense mechanism introduced holds relevance for users within power distribution systems where malicious cyber attacks have been identified. The dynamic interaction between the malicious attacker and the defender

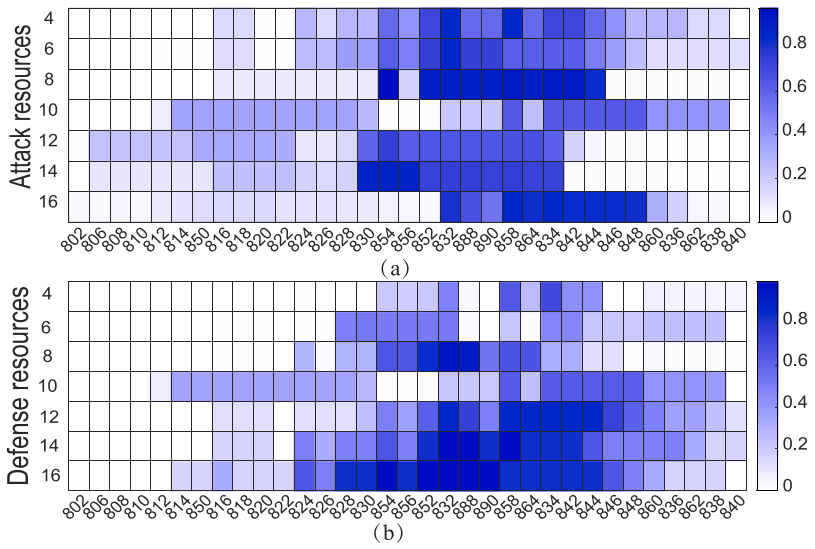

Fig. 11. (a) Optimal defense policies with various attack resources and fixed defense resources $C_l^d = 10$; (b) Optimal defense policies with fixed attack resources $C_l^a = 10$ and various defense resources. The x-axis represents the demand nodes in the considered power system.

unfolds as an iterative process until a NE is attained. Consequently, this iterative approach bolsters the resilience of power distribution systems against cyber threats, thereby fortifying their overall security posture.

6. Conclusion and future work

The integrated cyber network with SG enables two-way communication between customers and the SG operator, which facilitates the engagement of customers in demand response programs. However, in addition to the efficient operation, it gives malicious attackers opportunities to disturb utility services through cyber attacks. In this paper, a particular cyber attack is considered, namely, FPAs, which aim at changing customers' demand patterns by forcing the real-time prices that customers respond to. We first introduce a pricing mechanism containing the guideline day-ahead prices and real-time pricing, which is able to prevent large deviations between the day-ahead and real-time demands. Then, based on the proposed demand–response and pricing mechanism, the vulnerability of the power system to FPAs by an MDP framework is analyzed, where the attacker has incomplete information about the environment and the Q-learning algorithm is proposed to solve it. The defense measures are then introduced to enhance resilience of power systems to FPAs, and the interaction of the attacker and the defender is modeled as a two-player zero-sum Markov game where neither of them has full knowledge of the game model. The Markov game is solved by the proposed multi-agent reinforcement learning algorithm. The results of the application to the IEEE 34 node test feeder show the effect of the defense in mitigating the FPAs and highlight the fact that the policies are sensitive to the players' resources.

Nevertheless, this study is subject to certain limitations. Firstly, the demand–response mechanism is intricate, influenced by numerous factors, including the integration of renewable energy sources. Additionally, it may pose a challenge to directly apply the current proposed method to large-scale systems, given the associated increase in computational burden. Future research endeavors could explore the integration of renewable energy sources into the model, and the application of deep learning methodologies to address scalability challenges. Building upon the insights gleaned from this study, subsequent investigations may also delve into the analysis of various other types of cyber attacks. Besides, it would be interesting to study how to detect false defense agents that act like defense agent at most times and attack the system at its most vulnerable time.

CRediT authorship contribution statement

Daogui Tang: Writing – original draft, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Josep M. Guerrero:** Writing – review & editing, Validation, Supervision. **Enrico Zio:** Writing – review & editing, Validation, Supervision.

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.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by National Key Research and Development Program of China under the Grant No. 2021YFB2601605 and China Scholarship Council under grant number 201700810141.

References

Acharya, S., Dvorkin, Y., Karri, R., 2021. Causative cyberattacks on online learning-based automated demand response systems. IEEE Trans. Smart Grid 12 (4), 3548–3559.

Amini, S., Pasqualetti, F., Mohsenian-Rad, H., 2016. Dynamic load altering attacks against power system stability: Attack models and protection schemes. IEEE Trans. Smart Grid 9 (4), 2862–2872.

Avordeh, T.K., Gyamfi, S., Opoku, A.A., Peprah, F., 2023. Assessing the viability and environmental impact of residential demand response programs: A case study in East Legon, Greater Accra, Ghana. Energy Rep. 10, 4604–4615.

Ayub, M.F., Li, X., Mahmood, K., Shamshad, S., Saleem, M.A., Omar, M., 2023. Secure consumer-centric demand response management in resilient smart grid as industry 5.0 application with blockchain-based authentication. IEEE Trans. Consum. Electron..

Barreto, C., Cárdenas, A.A., 2018. Impact of the market infrastructure on the security of smart grids. IEEE Trans. Ind. Inform. 15 (7), 4342–4351.

Bellman, R., 1966. Dynamic programming. Science 153 (3731), 34–37.

Buşoniu, L., Babuška, R., De Schutter, B., 2010. Multi-agent reinforcement learning: An overview. In: Innovations in Multi-Agent Systems and Applications-1. Springer, pp. 183–221.

Chen, Y., Hong, J., Liu, C.-C., 2016. Modeling of intrusion and defense for assessment of cyber security at power substations. IEEE Trans. Smart Grid 9 (4), 2541–2552.
Chen, Y., Huang, S., Liu, F., Wang, Z., Sun, X., 2018. Evaluation of reinforcement learning-based false data injection attack to automatic voltage control. IEEE Trans. Smart Grid 10 (2), 2158–2169.
Daskalakis, C., Golowich, N., Zhang, K., 2022. The complexity of markov equilibrium in stochastic games. arXiv preprint arXiv:2204.03991.
2019. Day-ahead hourly LMPs. http://dataminer2.pjm.com. (Accessed 1 October 2019).
Deng, R., Xiao, G., Lu, R., 2015a. Defending against false data injection attacks on power system state estimation. IEEE Trans. Ind. Inform. 13 (1), 198–207.
Deng, R., Yang, Z., Chow, M.-Y., Chen, J., 2015b. A survey on demand response in smart grids: Mathematical models and approaches. IEEE Trans. Ind. Inform. 11 (3), 570–582.
Drayer, E., Routtenberg, T., 2019. Detection of false data injection attacks in smart grids based on graph signal processing. IEEE Syst. J..
Elsir, M., Al-Sumaiti, A.S., El Moursi, M.S., 2024. Towards energy transition: A novel day-ahead operation scheduling strategy for demand response and hybrid energy storage systems in smart grid. Energy 130623.
Flick, T., Morehouse, J., 2010. Securing the Smart Grid: Next Generation Power Grid Security. Elsevier.
Giraldo, J., Cárdenas, A., Quijano, N., 2016. Integrity attacks on real-time pricing in smart grids: impact and countermeasures. IEEE Trans. Smart Grid 8 (5), 2249–2257.
Gungor, V.C., Sahin, D., Kocak, T., Ergut, S., Buccella, C., Cecati, C., Hancke, G.P., 2011. Smart grid technologies: Communication technologies and standards. IEEE Trans. Ind. Inform. 7 (4), 529–539.
Hao, Y., Wang, M., Chow, J.H., 2016. Likelihood analysis of cyber data attacks to power systems with markov decision processes. IEEE Trans. Smart Grid 9 (4), 3191–3202.
Hasan, M.K., Habib, A.A., Islam, S., Safie, N., Abdullah, S.N.H.S., Pandey, B., 2023. DDoS: Distributed denial of service attack in communication standard vulnerabilities in smart grid applications and cyber security with recent developments. Energy Rep. 9, 1318–1326.
Haurie, A., Krawczyk, J.B., Zaccour, G., 2012. Games and Dynamic Games, vol. 1, World Scientific Publishing Company.
Hossain, M.J., Rahnamy-Naeini, M., 2019. Line failure detection from PMU data after a joint cyber-physical attack. In: 2019 IEEE Power & Energy Society General Meeting. PESGM, IEEE, pp. 1–5.
Kersting, W.H., 1991. Radial distribution test feeders. IEEE Trans. Power Syst. 6 (3), 975–985.
Khalid, M., 2024. Smart grids and renewable energy systems: Perspectives and grid integration challenges. Energy Strategy Rev. 51, 101299.
Khoei, T.T., Slimane, H.O., Kaabouch, N., 2022. Cyber-security of smart grids: Attacks, detection, countermeasure techniques, and future directions. Commun. Netw. 14 (4), 119–170.
Kumar, P., Lin, Y., Bai, G., Paverd, A., Dong, J.S., Martin, A., 2019. Smart grid metering networks: A survey on security, privacy and open research issues. IEEE Commun. Surv. Tutor. 21 (3), 2886–2927.
Leszczyna, R., 2019. Standards with cybersecurity controls for smart grid-A systematic analysis. Int. J. Commun. Syst. 32 (6), e3910.
Li, J., Li, H., Su, Q., 2023. Dynamic load altering attack detection for cyber physical power systems via sliding mode observer. Int. J. Electr. Power Energy Syst. 153, 109320.
Littman, M.L., 1994. Markov games as a framework for multi-agent reinforcement learning. In: Machine Learning Proceedings 1994. Elsevier, pp. 157–163.
Liu, Y., Hu, S., Ho, T.-Y., 2015. Leveraging strategic detection techniques for smart home pricing cyberattacks. IEEE Trans. Dependable Secure Comput. 13 (2), 220–235.
Liu, Y., Hu, S., Zomaya, A.Y., 2016. The hierarchical smart home cyberattack detection considering power overloading and frequency disturbance. IEEE Trans. Ind. Inform. 12 (5), 1973–1983.
Liu, Z., Wang, L., 2021. FlipIt game model-based defense strategy against cyberattacks on SCADA systems considering insider assistance. IEEE Trans. Inf. Forensics Secur. 16, 2791–2804.
Ma, J., Chen, H.H., Song, L., Li, Y., 2015. Residential load scheduling in smart grid: A cost efficiency perspective. IEEE Trans. Smart Grid 7 (2), 771–784.
Mahmoud, M.M., Mišić, J., Akkaya, K., Shen, X., 2015. Investigating public-key certificate revocation in smart grid. IEEE Internet Things J. 2 (6), 490–503.
Mena, R., Hennebel, M., Li, Y.-F., Ruiz, C., Zio, E., 2014. A risk-based simulation and multi-objective optimization framework for the integration of distributed renewable generation and storage. Renew. Sustain. Energy Rev. 37, 778–793.
Mishra, S., Li, X., Pan, T., Kuhnle, A., Thai, M.T., Seo, J., 2016. Price modification attack and protection scheme in smart grid. IEEE Trans. Smart Grid 8 (4), 1864–1875.
Mo, H., Sansavini, G., 2017. Dynamic defense resource allocation for minimizing unsupplied demand in cyber-physical systems against uncertain attacks. IEEE Trans. Reliab. 66 (4), 1253–1265.
Mohsenian-Rad, A.-H., Leon-Garcia, A., 2010. Optimal residential load control with price prediction in real-time electricity pricing environments. IEEE Trans. Smart Grid 1 (2), 120–133.
Mohsenian-Rad, A.-H., Leon-Garcia, A., 2011. Distributed internet-based load altering attacks against smart power grids. IEEE Trans. Smart Grid 2 (4), 667–674.
Ni, Z., Paul, S., 2019. A multistage game in smart grid security: A reinforcement learning solution. IEEE Trans. Neural Netw. Learn. Syst. 30 (9), 2684–2695.
Samadi, P., Mohsenian-Rad, A.-H., Schober, R., Wong, V.W., Jatskevich, J., 2010. Optimal real-time pricing algorithm based on utility maximization for smart grid. In: 2010 First IEEE International Conference on Smart Grid Communications. IEEE, pp. 415–420.
Shapley, L.S., 1953. Stochastic games. Proc. Natl. Acad. Sci. 39 (10), 1095–1100.
Sutton, R.S., Barto, A.G., 2011. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.
Tan, R., Badrinath Krishna, V., Yau, D.K., Kalbarczyk, Z., 2013. Impact of integrity attacks on real-time pricing in smart grids. In: Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security. pp. 439–450.
Tang, D., Fang, Y., Zio, E., 2019a. A zero-sum Markov defender-attacker game for modeling false pricing in smart grids and its solution by multi-agent reinforcement learning. In: 29th European Safety and Reliability Conference (ESREL2019). pp. 3285–3291.
Tang, D., Fang, Y.-P., Zio, E., 2023. Vulnerability analysis of demand-response with renewable energy integration in smart grids to cyber attacks and online detection methods. Reliab. Eng. Syst. Saf. 235, 109212.
Tang, D., Fang, Y.-P., Zio, E., Ramirez-Marquez, J.E., 2019b. Resilience of smart power grids to false pricing attacks in the social network. IEEE Access 7, 80491–80505.
Tarasak, P., 2011. Optimal real-time pricing under load uncertainty based on utility maximization for smart grid. In: 2011 IEEE International Conference on Smart Grid Communications (SmartGridComm). IEEE, pp. 321–326.
Tellbach, D., Li, Y.-F., 2018. Cyber-attacks on smart meters in household nanogrid: Modeling, simulation and analysis. Energies 11 (2), 316.
Vanderbei, R.J., et al., 2015. Linear Programming. Springer.
Wang, W., Di Maio, F., Zio, E., 2019. Adversarial risk analysis to allocate optimal defense resources for protecting cyber–physical systems from cyber attacks. Risk Anal. 39 (12), 2766–2785.
Wang, J.-S., Yang, G.-H., 2018. Data-driven methods for stealthy attacks on TCP/IP-based networked control systems equipped with attack detectors. IEEE Trans. Cybern. 49 (8), 3020–3031.
Wang, C., Zhang, T., Luo, F., Li, F., Liu, Y., 2017. Impacts of cyber system on microgrid operational reliability. IEEE Trans. Smart Grid 10 (1), 105–115.
Watkins, C.J., Dayan, P., 1992. Q-learning. Mach. Learn. 8 (3–4), 279–292.
Wei, L., Sarwat, A.I., Saad, W., Biswas, S., 2016. Stochastic games for power grid protection against coordinated cyber-physical attacks. IEEE Trans. Smart Grid 9 (2), 684–694.
Wu, Y., Chen, Z., Dang, J., Chen, Y., Zhao, X., Zha, L., 2022b. Allocation of defensive and restorative resources in electric power system against consecutive multi-target attacks. Reliab. Eng. Syst. Saf. 219, 108199.
Wu, S., Jiang, Y., Luo, H., Zhang, J., Yin, S., Kaynak, O., 2022a. An integrated data-driven scheme for the defense of typical cyber–physical attacks. Reliab. Eng. Syst. Saf. 220, 108257.
Xiang, Y., Ding, Z., Zhang, Y., Wang, L., 2016. Power system reliability evaluation considering load redistribution attacks. IEEE Trans. Smart Grid 8 (2), 889–901.
Xing, M., Wang, Y., Pang, Q., Zhuang, G., 2022. Dynamic-memory event-based asynchronous attack detection filtering for a class of nonlinear cyber-physical systems. IEEE Trans. Cybern..
Yan, J., He, H., Zhong, X., Tang, Y., 2016. Q-learning-based vulnerability analysis of smart grid against sequential topology attacks. IEEE Trans. Inf. Forensics Secur. 12 (1), 200–210.
Yi, J., An, H., Xing, Y., Li, J., Zhang, G., Bamisile, O., Yang, K., Xu, Y., 2023. A cyber attack detection strategy for plug-in electric vehicles during charging based on CEEMDAN and Broad Learning System. Energy Rep. 9, 80–88.
Yi, H., Hajiesmaili, M.H., Zhang, Y., Chen, M., Lin, X., 2017. Impact of the uncertainty of distributed renewable generation on deregulated electricity supply chain. IEEE Trans. Smart Grid 9 (6), 6183–6193.
Youssef, T.A., El Hariri, M., Bugay, N., Mohammed, O., 2016. IEC 61850: Technology standards and cyber-threats. In: 2016 IEEE 16th International Conference on Environment and Electrical Engineering. EEEIC, IEEE, pp. 1–6.
Yuan, Y., Li, Z., Ren, K., 2011. Modeling load redistribution attacks in power systems. IEEE Trans. Smart Grid 2 (2), 382–390.
Zhang, X., Yang, X., Lin, J., Xu, G., Yu, W., 2016. On data integrity attacks against real-time pricing in energy-based cyber-physical systems. IEEE Trans. Parallel Distrib. Syst. 28 (1), 170–187.
Zhao, L., Xu, H., Zhang, J., Yang, H., 2020. Resilient control for wireless cyber-physical systems subject to jamming attacks: A cross-layer dynamic game approach. IEEE Trans. Cybern..
Zhou, Y., Vamvoudakis, K.G., Haddad, W.M., Jiang, Z.-P., 2021. A secure control learning framework for cyber-physical systems under sensor and actuator attacks. IEEE Trans. Cybern. 51 (9), 4648–4660.