Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 9973580, 14 pages
https://doi.org/10.1155/2023/9973580

*Research Article*

# Cooperative Game of Energy-Constrained Agents in Wireless Communication Systems through Reinforcement Learning

**Li Guo,[1] Jianhong Wang,[2] Dian Huang,[1] and Shengzhong Feng[1]**

*[1]National Super-Computing Center in Shenzhen, 518000 Shenzhen, Guangdong, China*
*[2]Imperial College London, London SW7 2AZ, UK*

Correspondence should be addressed to Li Guo; guoli@nsccsz.cn

Received 22 December 2022; Revised 15 February 2023; Accepted 18 May 2023; Published 16 June 2023

Academic Editor: Vasudevan Rajamohan

Security issues are always considered in systems with wireless networks. However, few of them investigated covert signals existing on different communication channels to confuse advisories. In this paper, we consider the cooperation between two energy-constrained agents, who could inject covert signals. First, the system performance is measured by Kullback–Leibler divergence (KLD) to avoid much deviation. Then, the cooperative game between two agents is considered, in which two agents share the common goal at confusing advisories. More formally, this cooperative game is formulated as a Markov decision process (MDP) and the most economic strategies are obtained through reinforcement learning (RL) under the imperfect information. Finally, the feasibility of theoretical results is demonstrated on the interconnected New England test system (NETS) as well as its reduced system.

## 1. Introduction

With high flexibility in establishing communications, wireless communication is playing an increasingly vital role in many fields such as aerospace, transportation, mine monitoring, and power systems [1–4]. Meanwhile, the Internet of Things is a concept that has attracted significant attention since the emergence of wireless communication technology. As shown in Figure 1, a wireless communication system involves multiple modules to process the data and interact with the user ends and the communication is necessary for passing on sensor observations to controllers (S-C) and control signals to actuators (C-A) [5]. Therefore, the design of a communication frame is a crucial demand for system functionality. However, the main deficiency of wireless systems is in security because of their strong reliance on the wireless network, through which malicious adversaries could launch attacks to degrade the performance of the systems, or even destroy the systems. There are some common types of attacks including Denial-of-Service (DoS) attacks [6–9], false data injection (FDI) attacks [10–12] and so on, which could result in unacceptable consequences by hampering the critical infrastructure. Hence, security issues have become a crucial factor for wireless systems [13, 14].

To mitigate the defects caused by attacks, various detection methods were proposed. For example, an intelligent system was designed in [15] that could select a proper algorithm in an adaptive way to improve the detection performance. An online cyber-attack detection problem was formulated as a partially observable Markov decision process (MDP) problem, and a solution was proposed by the model-free RL for partially observable MDPs in [16]. An inference algorithm was proposed in [17] for smart grid systems subjected to stealthy attacks.

In [18], the Kullback–Leibler divergence (KLD) was introduced to measure the stealthiness of attacks, which is independent of any specific detection method such as the $\chi^2$ detector. Moreover, some researchers also take constrained energy into consideration out of practicality. In [19], the effect of DoS attacks with the energy budget was investigated and evaluated and the optimal DoS attack scheduling was proposed. A practical stealthily attack model against state estimation in power distribution systems was proposed in [20]. Based on these observations, covert strategies with

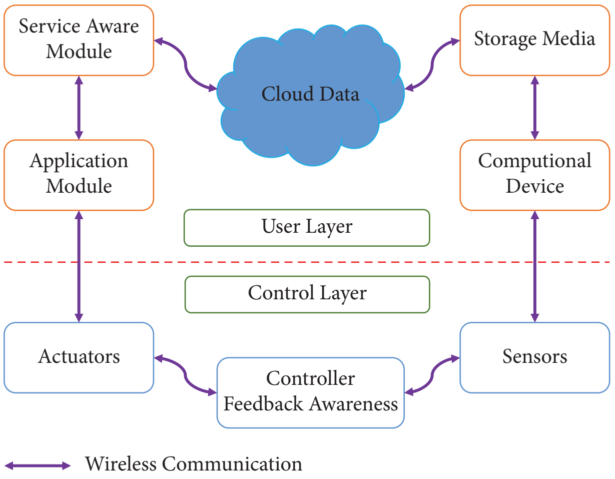


Figure 1: Architecture of a wireless communication system with two layers.

restricted energy to confuse advisories are considered in this paper.

In addition, cooperative games are common in real life, such as the coordination of autonomous vehicles [21] and traffic junctions [22]. In [23], the cooperative game of a discrete-time multiplayer system with control constraints was researched through adaptive dynamic programming techniques, and the graphical Nash equilibrium was studied in [24] by proposing a model-free distributed Q-learning algorithm, achieving attenuation of maximizing the worst-case adversarial inputs. A zero-sum stochastic game with two players was formulated, and a Nash Q-learning method was proposed to solve the Nash equilibrium [25]. Actually, the cooperative game is a special case of nonzero-sum games, where all the players have an identical goal to achieve. Due to the complexity of modern wireless systems such as smart grids and autonomous vehicles, the cooperation of advisories may lead to a huge potential danger, e.g., the manipulated electricity market prices or blackouts. Furthermore, to tackle the computation complexity of wireless systems efficiently, reinforcement learning (RL) is adopted in this paper, which is applicable for obtaining the optimal solution in an uncertain environment [26].

In light of these aforementioned analyses, we consider the Markov cooperative game with independently energy-constrained agents who could emit covert signals via different communication channels of wireless systems. Here, "independent" means that two agents make their own strategies only based on the other's historical information excluding the current one. Moreover, S-C and C-A communication channels of wireless systems as mentioned before are specifically selected as two different communication channels wherein two agents launch the covert signals. Since the environment, i.e., wireless systems, is unknown to the agents, one of the RL methods, named the policy gradient [27], is applied. Different from the previous works that only applied dynamic programming methods (note that the policy gradient method is also categorized as an approximate dynamic programming method) [28–30], our work also involves a multiagent learning process called fictitious play [31, 32]. Moreover, the economic strategies are modeled as Gaussian distribution, which actually is a continuous action space and also motivates us to learn the deterministic policy [33]. Consequently, the most economic strategies for the two agents are investigated when both of them are energy-limited through the policy gradient method. With the previous description, the contributions of this paper are as follows:

(1) A Markov cooperative game between two independent energy-constrained agents is formulated, in which two agents emit covert signals to wireless communication channels and have identical payoff functions

(2) The estimation error covariance of the system state is recalculated, and the system performance with covert signals is analysed and measured by KLD between the normal and modified innovation based on the state estimation error covariance

(3) By guaranteeing the system's performance, the most economic strategies are obtained through the RL method

Furthermore, to guarantee that every agent's strategy converges to the optimal solution, the behaviors of the two agents are modeled as fictitious play, where each agent decides its current action based on the belief of the other's history actions.

The arrangement of the remaining paper is as follows. Section 2 presents the problem formulation including the system model and the covert signal model. The system performance is analysed, and the most economic strategies are obtained through the policy gradient method in Section 3. Simulation is given to illustrate the efficiency of the proposed results in Section 4, followed by the conclusion in Section 5.

*1.1. Notations.* $\mathbb{R}, \mathbb{Z}, \mathbb{N}$ are the sets of real numbers, nonnegative, and positive integers, respectively. $\mathbb{S}_{+}^{n}$ and $\mathbb{S}_{++}^{n}$ denote the set of $n$ by $n$ symmetric positive semidefine and positive define matrices. $k \in \mathbb{Z}$ is the time index. $\mathbb{R}^n$ stands for the $n$ dimensional Euclidean space. $N(\mu, \Sigma)$ denotes the Gaussian distribution with a mean $\mu$ and a covariance $\Sigma$. $\mathbb{E}[\cdot]$ is the expectation of a random variable, and $\mathbb{E}[\bullet|\bullet]$ is the conditional expectation. Tr $[\cdot]$ stands for the trace of a matrix. $(\cdot)^T$ is the transposition. $\|\cdot\|$ and $|\cdot|$ denote the Euclidean norm for a vector and the absolute value, respectively.

## 2. Problem Formulation

*2.1. System Model.* Under the sampled-data control framework, the continuous-time state space and measurement space can be transformed into a discrete-time system with a zero-order holder. The linear discrete system is a general system model considered in the previous works [19, 34–37], which is also the research target of this paper as shown in the following equation:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k, \\ y_k &= Cx_k + v_k, \end{aligned} \tag{1}$$

where $A$, $B$, *$C$ are the constant matrices with appropriate dimensions*; $x_k \in \mathbb{R}^n$ is the state vector with the initial condition $x_0$, obeying the zero-mean Gaussian distribution with covariance $\Omega$, i.e., $x_0 \sim N(0, \Omega)$ and $\Omega \in \mathbb{S}_+^n$; $y_k \in \mathbb{R}^m$ is a vector of sensor measurements; and $u_k \in \mathbb{R}^p$ is the control input. The sequences $\{w_k\}_{k\in\mathbb{N}_{\geq 0}}$ and $\{v_k\}_{k\in\mathbb{N}_{\geq 0}}$ represent the process noises and measurement noises, respectively, both of which are the independent and identically distributed (i.i.d.) Gaussian random vectors with $w_k \sim N(0, Q)$, $v_k \sim N(0, R)$, $Q \in \mathbb{S}_+^n$, and $R \in \mathbb{S}_{++}^m$. Moreover, we assume that $(A, \sqrt{Q})$ is controllable and $(A, C)$ is observable.

The controller employs a Kalman filter, which calculates the minimum mean-square error (MMSE) to estimate and monitor the process state. $\mathbb{I}_k = \{y_0, y_1, y_2, \ldots, y_k\}$ is the information set collected up to time $k$. The controller's local state estimate $\hat{x}_k$ and its corresponding error covariance $P_k$ are calculated, respectively, as $\hat{x}_k \coloneqq \mathbb{E}[x_k \mid \mathbb{I}_k]$ and $P_k \coloneqq \mathbb{E}[(x_k - \hat{x}_k)(x_k - \hat{x}_k)^T \mid \mathbb{I}_k]$.

The optimal state estimate $\hat{x}_k$ of a Kalman filter is generated by the following equations:

$$\begin{aligned} \hat{x}_k^- &= A\hat{x}_{k-1} + Bu_{k-1}, \\ P_k^- &= AP_{k-1}A^T + Q, \\ K_k &= P_k^- C^T \left(CP_k^- C^T + R\right)^{-1}, \\ \hat{x}_k &= \hat{x}_k^- + K_k (y_k - C\hat{x}_k^-), \\ P_k &= (I - K_k C)P_k^-, \end{aligned} \tag{2}$$

where $\hat{x}_k^-$ is a priori MMSE estimate of the state $x_k$ with the initial condition $\hat{x}_0^- = x_0$. $P_k^-$ is the corresponding error covariance with $P_k^- \coloneqq \mathbb{E}[(x_k - \hat{x}_k^-)(x_k - \hat{x}_k^-)^T \mid \mathbb{I}_{k-1}]$. $K_k$ is the Kalman gain, and the innovation $z_k \triangleq y_k - C\hat{x}_k^-$ is a Gaussian process with $z_k \sim N(0, \Sigma_k^z)$ and $\Sigma_k^z = CP_k^- C^T + R$.

It is well known that the Kalman filter converges exponentially fast from any initial condition. Accordingly, it is reasonable to define the steady-state error covariance $P \triangleq \lim_{k \longrightarrow \infty} P_k^-$, where $P$ is the unique positive semidefinite solution of $X = AXA^T + Q - AXC^T (CXC^T + R)^{-1} CXA^T$. Without the loss of generality, the system is assumed to start from $k = -\infty$ which results in a fixed-gain Kalman filter starting from $k = 0$, i.e.,

$$K = PC^T \left(CPC^T + R\right)^{-1}. \tag{3}$$

*2.2. Covert Signal Model.* The security of system information is guaranteed by two agents, who could launch covert signals on two communication channels. We assumed that the two agents (1) know the dimensions of the wireless system, which is public knowledge, (2) can add arbitrarily independent Gaussian noises into C-A and S-C channels, respectively, (3) know the existence of each other, and (4) are mutually independent.

The covert signals are given in the following equation:

$$\begin{aligned} \overline{u}_k &= \tilde{u}_k + \Delta u_k, \\ \tilde{y}_k &= \overline{y}_k + \Delta y_k, \end{aligned} \tag{4}$$

where $\overline{u}_k$ and $\tilde{y}_k$ are the modified control input and measurement output and $\tilde{u}_k = L\hat{\tilde{x}}_k$, $\overline{y}_k = C\tilde{x}_k + v_k$, and $\tilde{x}_k$ are the input, output, and state of the system with covert signals. $\hat{\tilde{x}}_k$ is the posteriori MMSE estimate of $\tilde{x}_k$, and $L$ is a proper matrix such that $A + BL$ is stable. $\Delta u_k \sim N(\mu_k^u, \Sigma_k^u)$ and $\Delta y_k \sim N(\mu_k^y, \Sigma_k^y)$ are the Gaussian signals injected by the two agents on C-A and S-C channels, respectively. We assumed that $w_k$, $v_k$, $\Delta y_k$, and $\Delta u_k$ are mutually independent processes.

*Remark 1.* Since wireless systems in real life are large-scale and vulnerable, it is necessary for agents located at different nodes (positions) to emit external signals to confuse adversaries so as to avoid damage. To fulfil this aim, we consider a cooperative game with an assumption that each agent is able to observe the information of others (e.g. through communications). Moreover, each agent makes its own strategy based on the collected information before each time step of a decision, which is named as an independent decision. In addition, it is practical and reasonable to take the energy budget of a signal injection as a constraint to reach the economic behaviour.

Moreover, the Kalman filter uses the modified measurements of $\tilde{y}_k$ to run. $\tilde{P}_k \coloneqq \mathbb{E}[(\tilde{x}_k - \hat{\tilde{x}}_k)(\tilde{x}_k - \hat{\tilde{x}}_k)^T]$ is defined as the estimate error covariance of the system with covert signals, and $\hat{\tilde{x}}_k^-$ is denoted as the correspondingly prior MMSE estimate of $\tilde{x}_k$. $\hat{\tilde{x}}_k^-$ and $\hat{\tilde{x}}_k$ are obtained from the following recursions:

$$\begin{aligned} \hat{\tilde{x}}_k^- &= A\hat{\tilde{x}}_{k-1} + B\tilde{u}_{k-1}, \\ \hat{\tilde{x}}_k &= \hat{\tilde{x}}_k^- + K\left(\tilde{y}_k - C\hat{\tilde{x}}_k^-\right). \end{aligned} \tag{5}$$

Then, the limited energy budget of agents is considered, which is common in reality. Specifically, the energy distribution as time goes by on the C-A channel is denoted as

$$B^u = \left[B_0^u, B_1^u, B_2^u, \cdots, B_k^u, \cdots\right]^T, \tag{6}$$

where $B_k^u \leq M_u$ and $M_u$ is a constant. Similarly, the energy distribution on the S-C channel is denoted as

$$B^y = \left[B_0^y, B_1^y, B_2^y, \cdots, B_k^y, \cdots\right]^T, \tag{7}$$

where $B_k^y \leq M_y$ and $M_y$ is also a constant. Under the allocation of energy shown previously, the covert signal launched on the C-A channel at instant $k$ is defined as $\Delta u_k (B_k^u)$: $\mathbb{R} \longrightarrow \mathbb{R}^n$. Due to the simplicity of notation, this covert signal is rewritten as $\Delta u_k$. Obviously, $\Delta u_k$ is also restricted to an upper bound because of the existence of an energy upper bound. Then, it is reasonable to give the following definition of covert signals on the C-A channel:

$$\Delta u_k \coloneqq \left[\Delta u_{k,1}, \Delta u_{k,2}, \cdots, \Delta u_{k,n}\right]^T, \tag{8}$$

where $|\Delta u_{k,i}| \leq \widehat{u}$, $1 \leq i \leq n$, and $\widehat{u}$ is a constant. Clearly, these signals obey truncated normal distribution and a similar definition of the covert signal on an S-C channel is given as follows:

$$\Delta y_k \coloneqq \left[\Delta y_{k,1}, \Delta y_{k,2}, \cdots, \Delta y_{k,m}\right]^T, \tag{9}$$

where $|\Delta y_{k,i}| \leq \widehat{y}$, $1 \leq i \leq m$, and $\widehat{y}$ is a constant. Then, the two most expensive covert signals could be given, which are constant vectors and are defined as follows:

$$\begin{aligned} |\widehat{U}| &\coloneqq [\widehat{u}, \widehat{u}, \cdots, \widehat{u}]^T, \\ |\widehat{Y}| &\coloneqq [\widehat{y}, \widehat{y}, \cdots, \widehat{y}]^T. \end{aligned} \tag{10}$$

*Remark 2.* The restrained energy budget is transformed as the restricted injection covert signals, i.e., the equations defined in (8) and (9), although the specific mapping is not discussed.

Clearly, agents aim at confusing the advisories and guaranteeing the system performance simultaneously. The performance is measured by the divergence between the modified innovation $\widetilde{z}_k$ and the normal innovation $z_k$ in this paper, and the KLD, defined as follows, is always adopted to measure the difference between the modified and normal states.

*Definition 3* (see [38]). Let $x_k$ and $y_k$ be two random sequences with joint probability density functions $f_{x_k}$ and $f_{y_k}$, respectively. The KLD between $x_k$ and $y_k$ is defined as follows:

$$D\left(x_k \| y_k\right) = \int_{\left\{\xi_k \,\middle|\, f_{x_k}(\xi_k) > 0\right\}} \log \frac{f_{x_k}(\xi_k)}{f_{y_k}(\xi_k)} f_{x_k}(\xi_k) \mathrm{d}\xi_k. \tag{11}$$

Note that KLD is a non-negative quantity, that is, $(x_k \| y_k) \geq 0$, and it gauges the dissimilarity between the two probability density functions with $D(x_k \| y_k) = 0$ if $x_k = y_k$. When KLD exceeds a certain threshold, an alarm will be triggered, which indicates that the system performance is violated. Agents should guarantee the normal running of the system by keeping KLD between the modified innovation $\widetilde{z}_k$ and the normal one $z_k$ which is small enough. In other words, the agents should ensure that KLD $D(z_k \| z_k)$ is not bigger than a threshold when emitting covert signals.

Obviously, the injection of covert signals will deviate $\widetilde{z}_k$ from $z_k$. This results in widening the disparity between $\widetilde{z}_k$ and $z_k$, i.e., increasing $D(z_k \| z_k)$, which may trigger an alarm. However, there is always an upper bound of $D(z_k \| z_k)$, which means that the agents could ensure that the system performs normally. The related results are provided in the next section.

*2.3. Problem of Interest.* In this paper, there are two agents that launch the covert signals on S-C and C-A channels independently. The economic strategies of the two agents with limited energy budgets are investigated, and this decision process is formulated as a Markov cooperative game with two players. In this Markov cooperative game model, two players have the identical payoff function and the economic strategies are obtained by a RL method, that is, policy gradient.

## 3. Cooperative Game

*3.1. Game Theoretic Framework.* This Markov cooperative game, denoted by $\mathbb{G}$, is characterized by a five-tuple $\langle \mathcal{N}, \mathbb{A}, \mathbb{S}, \mathbb{L}, \mathbb{T}, J \rangle$, where

**Players:** $\mathcal{N} = \{u, y\}$ denotes the set of agents, i.e., two agents on C-A and S-C channels, respectively.

**Action:** $\mathbb{A}^i$ denotes the set of actions of agent $i$ and $i \in \mathcal{N}$. The action of agent $i$ at an instant $k$ is denoted as $a_k^i \in \mathbb{A}^i$ with definitions in (8) and (9) for the two agents.

**State:** $\mathbb{S} = \{s_0, s_1, \ldots, s_k, \ldots\}$ is the set of states, where $s_k$ is the estimated state from the Kalman filter. Note that $s_k = \widehat{x}_k$ before the injection of the first covert signal, otherwise, $s_k = \widehat{\widetilde{x}}_k$.

**Policy (strategy):** $\pi^i\colon \mathbb{S} \longrightarrow \mathbb{A}^i$ is the policy of agent $i$, such that $a_k^i = \pi^i(s_k)$, and $\mathbb{L}^i$ denotes the policy space for agent $i$ with $\pi^i \in \mathbb{L}^i$.

**Transition probability:** $\mathbb{T}(s, a)$ is a transition function that defines transition probabilities between states, i.e., $s_{k+1} \longleftarrow \mathbb{T}(s_k, a_k)$. However, the specific transition probability function is unknown for the agents.

**Payoff:** in a cooperative game, all players have the same payoff function to maximize as follows:

$$J(x, s, \Delta u, \Delta y) = \mathbb{E}^{\pi^u, \pi^y}\left[\sum_{k=0}^{+\infty} \beta^k r_k(x_k, s_k, \Delta u_k, \Delta y_k)\right], \tag{12}$$

where $\Delta u \in \mathbb{A}^u$ and $\Delta y \in \mathbb{A}^y$ denote any possible action of agents, $\beta\colon 0 \leq \beta < 1$ is the discount factor, and $r_k(x_k, s_k, \Delta u_k, \Delta y_k)$ abbreviated as $r_k$ is the immediate reward with the following definition:

$$r_k = (s_k - x_k)^T (s_k - x_k) - D\left(\widetilde{z}_k \| z_k\right) - (\Delta u_k)^T \Delta u_k - (\Delta y_k)^T \Delta y_k. \tag{13}$$

*Remark 4.* Note that $r_k$ not only indicates the constrained covert signals but also reflects the state deviation. Precisely, the item $(s_k - x_k)$ in (13) could reveal the level of deviation about the state trajectory, and the last three items in (13) show that the agents aim at minimizing $D(\tilde{z}_k \| z_k)$ to guarantee the system performance with less energy cost. The discount factor $\beta$ controls how much effect future rewards have on the optimal decisions, with a small value of $\beta$ emphasizing the near-term gain and larger values giving significant weight to future rewards.

*3.2. Fictitious Play.* Players in a game may or may not know some information about other players, especially in a complex system. So, we assumed that the two agents in this paper know the existence of each other but do not know the current action of the other one. Based on the above-mentioned analysis, fictitious play is adopted to describe this kind of an interaction between the two agents. Actually, fictitious play is a model of learning behavior, where players in a game could observe the historical actions made by every other player. Then, each player is able to predict the other players' current action based on these players' previous actions, and then, every individual player could play the best response to other players' historical actions, resulting in the optimal strategy for every player in a game.

*Definition 5.* The best response for each agent in a game $\mathbb{G}$ is defined as

$$\begin{aligned} BR(\Delta \mathrm{u}) &= \left\{\Delta y^{\star} \in \mathbb{A}^{y} | J(x, s, \Delta u, \Delta y) \leq J(x, s, \Delta u, \Delta y^{\star}), \forall \Delta y \in \mathbb{A}^{y}\right\}, \\ BR(\Delta \mathrm{y}) &= \left\{\Delta u^{\star} \in \mathbb{A}^{u} | J(x, s, \Delta u, \Delta y) \leq J(x, s, \Delta u^{\star}, \Delta y), \forall \Delta u \in \mathbb{A}^{u}\right\}. \end{aligned} \tag{14}$$

Furthermore, fictitious play will be elaborated according to the setting of this paper where two agents are considered. Based on the observation of the other agent's actions from initial instant to instant $k-1$, the concept of empirical frequency of the two agents is necessary and is defined as the percentage of stages as follows:

$$\begin{aligned} \alpha_k &= \frac{1}{k} \sum_{j=0}^{k-1} \mathbb{I}\left\{\Delta u_k = \Delta u_j\right\}, \\ \gamma_k &= \frac{1}{k} \sum_{j=0}^{k-1} \mathbb{I}\left\{\Delta y_k = \Delta y_j\right\}, \end{aligned} \tag{15}$$

where $\alpha_k$ is the empirical frequency for the agent $u$, $\gamma_k$ is the one for the agent $y$, and $\mathbb{I}\{\cdot\}$ denotes the indicator function such that $\mathbb{I}\left\{\Delta u_j = \Delta u\right\} = 1$ if $\Delta u_j = \Delta u$, otherwise, $\mathbb{I}\left\{\Delta u_j = \Delta u\right\} = 0$.

According to the empirical frequency, every player could estimate the current action of the other players and then make their own action to maximize the payoff function; that is, every player plays the best response, $\mathrm{BR}(\cdot)$, to the empirical frequency of other players' actions at each instant which is represented by the following equation:

$$\begin{aligned} \mathrm{BR}(\gamma_k) &= \left\{\Delta u_k \left| \max_{\Delta u_k \in \mathbb{A}_u} \sum_{k=0}^{+\infty} \beta^k r_k(x_k, s_k, \Delta u_k, \Delta y) \times \gamma_k \right.\right\}, \\ \mathrm{BR}(\alpha_k) &= \left\{\Delta y_k \left| \max_{\Delta y_k \in \mathbb{A}_y} \sum_{k=0}^{+\infty} \beta^k r_k(x_k, s_k, \Delta u, \Delta y_k) \times \alpha_k \right.\right\}. \end{aligned} \tag{16}$$

*3.3. Deviation of the System State.* To study the deviation of a system state measured by $D(\tilde{z}_k \| z_k)$, the modified innovation $\tilde{z}_k$ needs to be analysed. It is known that the normal innovation $z_k$ obeys the Gaussian distribution, i.e., $z_k \sim N(0, CPC^T + R)$, and the modified innovation $\tilde{z}_k$ still obeys the Gaussian distribution as covert signals are Gaussian with $\tilde{z}_k \sim N(\mu_k^z, \tilde{\Sigma}_k)$. This naturally motivates us to investigate the estimate error covariance at the Kalman filter.

When signals on C-A and S-C channels are modified by two agents, the estimate error covariance is recalculated in the following proposition. First, we define $\overline{P} \coloneqq (I - KC)P$ with $K$ given in (3). According to (2) and (3), one can [34] obtain the following equation:

$$\begin{aligned} \overline{P} &= (I - KC)P(I - KC)^T + KRK^T, \\ P &= A\overline{P}A^T + Q. \end{aligned} \tag{17}$$

Once the covert signals are injected, the system is modified and the injection effect will last continuously until the system stops running. Therefore, it is assumed that the first injection is launched at instant $k$, before which the system was operated under the normal situation with an initial state.

**Proposition 6.** *When covert signals defined in (3) are injected, the estimate error covariance of the Kalman filter can be obtained from the following recursion:*

$$\widetilde{P}_{k+q} = (I-KC)A\widetilde{P}_{k+q-1}A^T(I-KC)^T + KRK^T + (I-KC)Q(I-KC)^T + KE^y_{k+q}K^T + (I-KC)BE^u_{k+q-1}B^T(I-KC)^T, \tag{18}$$

*where*

$$\begin{aligned} E^y_{k+q} &= \mathbb{E}\left[\Delta y_{k+q}\Delta y^T_{k+q}\right] \\ &= \mu^y_{k+q}\left(\mu^y_{k+q}\right)^T + \Sigma^y_{k+q}, \\ E^u_{k+q-1} &= \mathbb{E}\left[\Delta u_{k+q-1}\Delta u^T_{k+q-1}\right] \\ &= \mu^u_{k+q-1}\left(\mu^u_{k+q-1}\right)^T + \Sigma^u_{k+q-1}, \end{aligned} \tag{19}$$

*and $q \geq 0$. Furthermore, (10) could be rewritten as*

$$\widetilde{P}_{k+q} = \overline{P} + E_q + F_q, \tag{20}$$

*where*

$$\begin{aligned} E_q &= \sum_{i=0}^{q}((I-KC)A)^i KE^y_{k+i}K^T\left(A^T(I-KC)^T\right)^i, \\ F_q &= \sum_{i=0}^{q}((I-KC)A)^i(I-KC)BE^u_{k+i}B^T(I-KC)^T\left(A^T(I-KC)^T\right)^i. \end{aligned} \tag{21}$$

For the coherence of logic, the proof of Proposition 6 is seen in Appendix A. Besides, to guarantee that the system performance fluctuated at a certain interval, the following definition is given.

*Definition 7.* The system performance could be guaranteed if $D(\widetilde{z}_k \| z_k) \leq \delta$, where $\delta$ is the threshold.

In this work, since $\widetilde{z}_k$ is an independent Gaussian random variable with $\widetilde{z}_k \sim N(\mu^z_k, \widetilde{\Sigma}_k)$, $D(\widetilde{z}_k \| z_k)$ follows that

$$D\left(\widetilde{z}_k \| z_k\right) = \frac{1}{2}\mathrm{Tr}\left[\Sigma^{-1}\widetilde{\Sigma}_k\right] - \frac{m}{2} + \frac{1}{2}\log\frac{\det(\Sigma)}{\det(\widetilde{\Sigma}_k)} + (\mu^z_k)^T \sum_k^{-1} \mu^z_k. \tag{22}$$

The mean and covariance of $\widetilde{z}_k$ are still definite even though the agents launched the most expensive covert signals defined in (10) at every instant from the first emission. Then, the following proposition is given to show that the system performance could be guaranteed when KLD is utilized as a measurement.

**Proposition 8.** *The state performance with covert signals defined in (4) and (5) could be guaranteed from the sense of KLD by choosing the proper parameters.*

The proof of Proposition 8 is shown in Appendix A.1.

*3.4. Design of the Most Economic Strategies through RL.* In this subsection, we analyse the most economic strategy for two agents. The goal of the agents is not only to make system information covert but also to guarantee the system performance as well as to avoid high energy consumption. Simply speaking, every energy-constrained agent makes their own most economic strategy by maximizing the reward.

Based on the payoff function (12), the following Markov cooperative game is investigated with five constraints.

*Problem 9.* For each agent, $\max_{\Delta u \in \mathbb{A}^u, \Delta y \in \mathbb{A}^y} J(x, s, \Delta u, \Delta y)$

$$s.t.\ s_k = s^-_k + K\left(\widetilde{y}_k - Cs^-_k\right), \tag{23}$$

$$s^-_k = As_{k-1} + B\widetilde{u}_{k-1}, \tag{24}$$

$$\left|\Delta u_{k,i}\right| \leq \hat{u}, 1 \leq i \leq n, \tag{25}$$

$$\left|\Delta y_{k,j}\right| \leq \hat{y}, 1 \leq j \leq m, \tag{26}$$

$$D\left(\widetilde{z}_k \| z_k\right) \leq \delta, \tag{27}$$

where $J(x, s, \Delta u, \Delta y)$ defined in (12) and (23) is the state equation, (24) is the prior MMSE estimate equation, (25) and (26) are the energy budget of the two channels as well as (27) measures the deviation of the system states.

Before deriving the optimal solution to Problem 9, MDP is elaborated first. MDP is a discrete-time stochastic control process, in which an agent decides an action $a_k \in \mathbb{A}$ at each state $s_k \in \mathbb{S}$ emitted from an environment via a probabilistic transition function. The environment gives a reward $r_k$ to measure the performance of an action at each time step. RL is a learning paradigm which aims to find an optimal policy $\pi(a_k|s_k)$ by maximizing the expectation over cumulative long-term rewards. Mathematically, it can be expressed as $\max_{\pi} \mathbb{E}^{\pi}[\sum_{k=0}^{\infty}\beta^k r_k]$.

As mentioned before, the transition probability is unknown for the agents since agents do not know the dynamics of the wireless system. To solve Problem 9, a policy-based RL method, that is, policy gradient, is applied as this methodology is unnecessary to leverage the transition probability distribution within MDP. Moreover, since the covert signal is continuous, we use the deterministic policy gradient (DPG) [39] to learn each agent's policy directly. Actually, the policy-based methods are more useful in continuous space because there are infinite actions and (or) states. In contrast, the value-based approaches such as Q-learning [40] are computationally much more expensive.

To be more specific, different from the value-based methods, the policy gradient directly learns the policy $\pi$, which is modeled as a parameterized function with respect to $\theta$, denoted by $\pi_\theta$. Then, the value of the payoff function is acquired depending on this policy and the maximum payoff could be obtained by optimizing $\theta$, that is, by maximizing the objective function as follows:

$$J(\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{k=0}^{\infty} \beta^k r(s_k, a_k)\right]. \tag{28}$$

The maximum payoff is achieved, and $J(\theta)$ is always denoted as $Q^{\pi_\theta}(s,a)$ given a state-action pair $(s,a)$, i.e., $J(\theta) = Q^{\pi_\theta}(s,a) = \mathbb{E}^{\pi_\theta}[\sum_{k=0}^{\infty}\beta^k r(s_k, a_k)|s_0 = s, a_0 = a]$. Then, since the gradient of $J(\theta)$ with respect to $\theta$ cannot be directly calculated, the policy gradient theorem [41] is used to approximate the gradient such that

$$\nabla_\theta J(\theta) = \mathbb{E}^{\pi_\theta}[Q^{\pi_\theta}(s,a)\nabla_\theta \log \pi_\theta(a\,|\,s)]. \tag{29}$$

Basically, policy gradient algorithms involve two main parts: actor and critic [42]. Precisely, the policy $\pi_\theta(a|s)$ and Q-value $Q^{\pi_\theta}(s,a)$ are called the actor and the critic, respectively, both of which need to be learned.

In the multiagent scenarios [43], the gradient of each agent $i$ can be represented as

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}^{\pi_{\theta_i}}\left[Q_i^{\pi_{\theta_i}}(s, a^i)\nabla_{\theta_i}\log\pi_{\theta_i}(a^i\,|\,s)\right], \tag{30}$$

where $Q_i^{\pi_{\theta_i}}(s,a^i)$ is the estimation of the long-term reward of each agent $i$ for the current policy. Then, the approximate gradient of each agent is reformulated as DPG such that

$$\nabla_{\theta_i} J(\theta_i) = \mathbb{E}^{\pi_{\theta_i}}\left[\nabla_{\theta_i}\pi^i_{\theta_i}(s)\nabla_{a^i} Q_i^{\pi_{\theta_i}}(s,a^i)\Big|a^i = \pi^i_{\theta_i}(s)\right], \tag{31}$$

where $\pi^i_{\theta_i}$ is a deterministic policy of agent $i$, parameterized by $\theta_i$. Consequently, this formulation is applied to learn the policy for each agent. To obtain the most economic strategies$of\ \pi^*_u = \{\Delta u^*_0, \Delta u^*_1, \cdots, \Delta u^*_k, \cdots\}$ and $\pi^*_y = \{\Delta y^*_0,\ \Delta y^*_1, \cdots, \Delta y^*_k, \cdots\}$, DPG is executed through the following steps. First, when the state of the wireless system is $s_k$, agents choose the current actions $\Delta u_k$ and $\Delta y_k$ according to their policies, both of which are added with a unit Gaussian noise $N_k \sim N(0,I)$ to improve exploration, where $I$ is an identity matrix with an appropriate dimension and the policy of each agent is optimized based on its estimated policy gradient (31) and the Q-value $Q_i^{\pi_{\theta_i}}(s,a^i;\omega_i)$ parameterized by $\omega_i$ is updated by the following minimization problem [44]:

$$\min_{\omega_i}\frac{1}{2}\left(r_k + Q_i^{\pi_{\theta_i}}\left(s_{k+1}, a^i_{k+1};\omega_i\right) - Q_i^{\pi_{\theta_i}}\left(s_k, a^i_k;\omega_i\right)\right)^2, \tag{32}$$

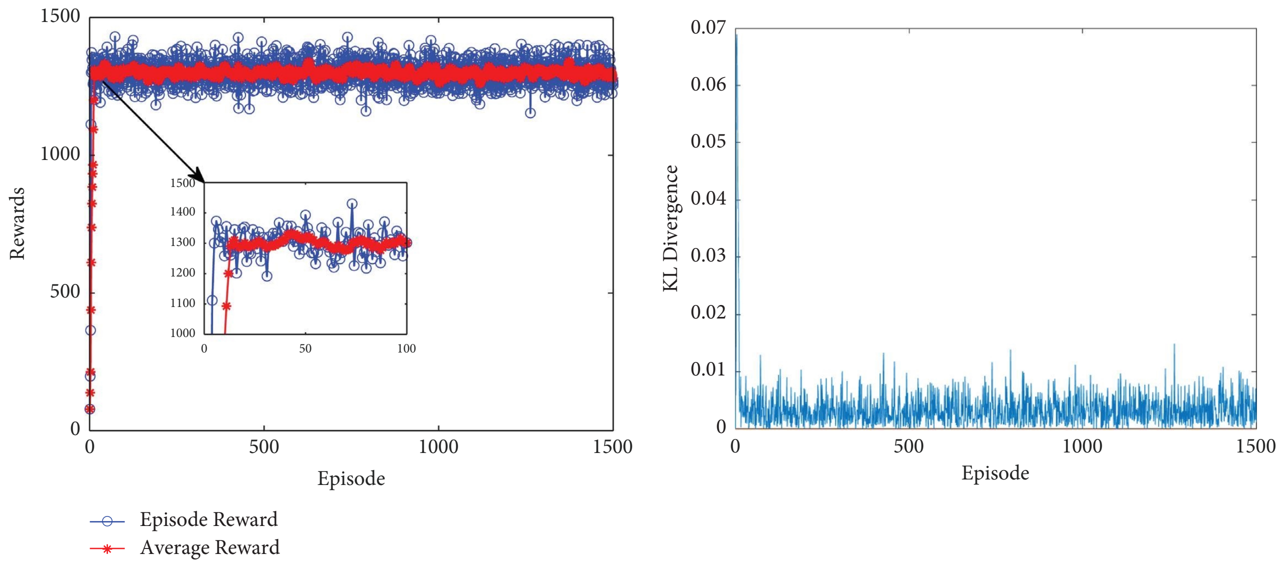
where $a^i_k$ and $a^i_{k+1}$ are the covert signals and sampled from the policy introduced previously. Since the unit Gaussian noise $N_k$ is added, $a^i_k$ and $a^i_{k+1}$ are replaced by $\hat{a}^i_k \coloneqq a^i_k + N_k$ and $\hat{a}^i_{k+1} \coloneqq a^i_{k+1} + N_{k+1}$ in the learning process. In addition, this added unit Gaussian noise during learning is so small that the randomly selected policy will not deviate from the mean too much. On the other hand, the state of the system for each agent implicitly represents the other agent's history actions according to the Markov property, that is, a state is induced from the history actions of agents. For this reason, each agent's policy or Q-value can be seen as a group of parameters, constructing a mapping between its current and historic actions of the other agent. In order to describe such a mapping, the original fictitious play is used and the empirical frequency of each agent up to the current instant is considered as a parameter for DPG. In other words, DPG is an implementation of the fictitious play for the continuous action space.

Based on the abovementioned analysis, the solution to Problem 9 is given in Algorithm 1 and the convergence of this algorithm is shown in Theorem 10 in the sequel. In Algorithm 1, $M$ denotes the number of episodes and the larger $M$ could achieve a better learning and $T$ denotes the length of a learning episode.

Before providing the proof of convergence for Algorithm 1, the following fact needs to be known first. In [45], the authors showed that the actor-critic process is a member of the fictitious play. Therefore, we will mainly discuss the convergence of the Markov cooperative game $\mathbb{G}$ when a fictitious play is introduced. In addition, in [46], a continuous-time embedding of a stochastic zero-sum game with a fictitious play converged to the set of Nash equilibrium strategies was proved. As a result, the convergence of the Markov cooperative game with fictitious play is analysed in Theorem 10 with a similar proof sketch in [46], and some central concepts of [46] should also be noted. Since the agents only can receive the current reward and cannot calculate the future expected discounted payoff, the definition of an auxiliary game for each state was introduced to estimate the long-term payoff (or the future expected discounted payoff). In every auxiliary game, an arbitrary set of the long-term payoff was assumed. That is to say, the payoff for an action in each auxiliary game was given by the immediate payoff plus the expected long-term payoff at the subsequent states. The long-term payoff of each auxiliary game was updated at the rate of $(1/t)$, where $t$ is the calendar time. Obviously, the rate $(1/t)$ becomes slow gradually as time goes on and this is beneficial for players to get close to the equilibrium of the auxiliary game. Conversely, the long-term payoff would converge, which results in that agents' current strategy approaching an equilibrium strategy. Actually, this kind of an idea aligns with the Bellman residual and (32) is a relaxation of the Bellman residual.

```
Randomly initialize the critic $Q^{\pi_{\theta_i}}(s, a^i; \omega_i)$ with parameter $\omega_i$ and the actor $\pi_{\theta_i}(a^i \mid s)$ with parameter $\theta_i$ of each agent $i$, where $a^i$ is
    the covert signal of each agent defined in (8) and (9)
for episode = 1: $M$ do
    Initialize an unit Gaussian process $N_k$ for action exploration
    Acquire the initial state $s_0$
    for $k$ = 1: $T$ do
        Select action $\hat{a}_k^i = a_k^i + N_k$ according to the current policy $\pi_{\theta_i}$ and exploration noise $N_k$
        Execute action $\hat{a}_k^i$ and acquire the immediate reward $r_k$ as well as the new state $s_{k+1}$
        Update the critic by minimizing the loss (32)
        Update the actor policy using the sampled policy gradient (31)
    end for
end for
```

Algorithm 1: DPG algorithm for economic strategies of two agents.



(a)

(b)

Figure 2: The NETS-NYPS 68-bus system. (a) Convergence of Algorithm 1. (b) Evolution of KLD.

**Theorem 10.** *In the Markov cooperative game $\mathbb{G}$, the strategy of each agent, $\pi_s^i$, $i \in \mathcal{N}$, converges to the set of stationary optimal strategies for each state of $s \in \mathbb{S}$.*

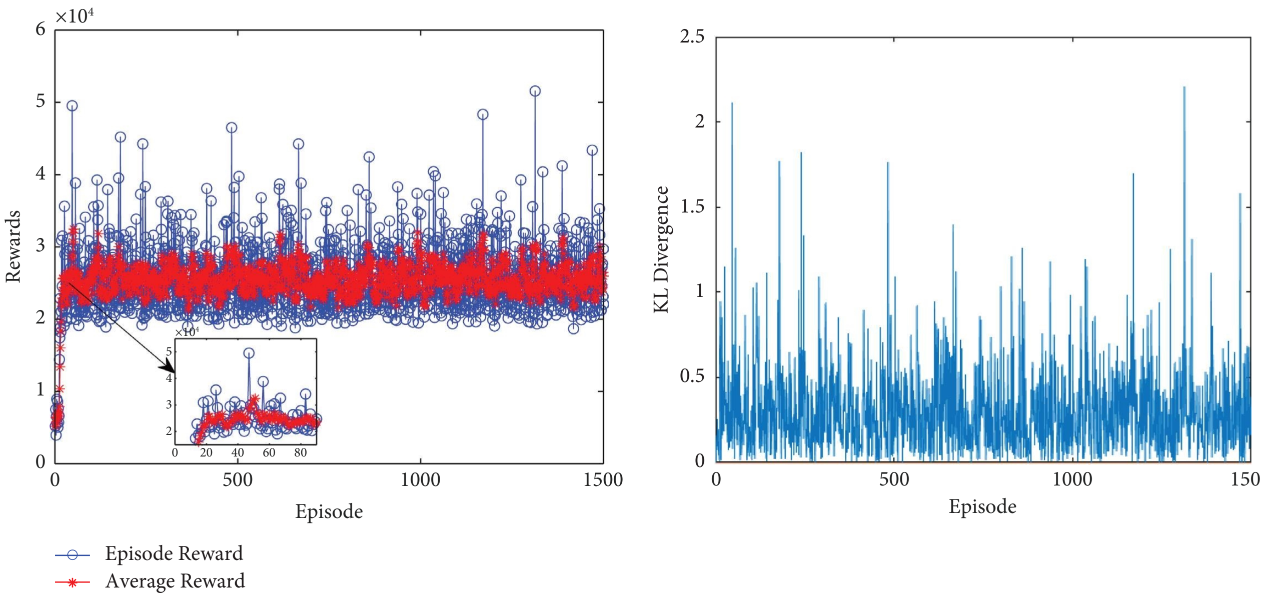
*Proof.* The proof is shown in Appendix B. □

## 4. Simulation

In this section, we assess the performance of the proposed algorithm on the interconnected New England test system (NETS) and the New York power system (NYPS) of the 1970s, where the dimension of $x_k$ is more than 100. In a power system, the loads on buses are the consumption (e.g., electricity) from users, and therefore, our algorithm aims to process these collected user data from the central data hub (cloud data). In order to show the effect of the main results well, Algorithm 1 is first verified over a 68-bus, 16-machine, and a 5-area system, which is an equivalent system of NETS-NYPS with reduced size and is named as the NETS-NYPS 68-bus system [47].

Some parameters in this simulation are given as follows. For agents, the upper bounds of covert signals are chosen to be $\hat{u} = 6$ and $\hat{y} = 6$; that is, the action space for the agent on the C-A channel is $(-6 * \underbrace{[1, 1, \ldots, 1, 1]^T}_{n \text{ dimension}}, 6 * \underbrace{[1, 1, \ldots, 1, 1]^T}_{n \text{ dimension}})$ and the action space for the agent on the S-C channel is $(-6 * \underbrace{[1, 1, \ldots, 1, 1]^T}_{m \text{ dimension}}, 6 * \underbrace{[1,1,\ldots,1,1]^T}_{m \text{ dimension}})$. For the proposed Algorithm 1, the number of episodes is chosen as $M = 1500$ and the length of an episode is $T = 100$, and the learning rate for policy and Q-value are chosen to be 0.01 and 0.001, respectively.

For the NETS-NYPS 68-bus system, the convergence is shown in (a) of Figure 2. In this subfigure, the episode reward in the blue line denotes the average reward of $T = 100$ immediate rewards in one episode and the red line represents the average of all episode rewards up to the current episode. From the beginning instant, both the



(a)

(b)

Figure 3: The interconnected NETS-NYPS. (a) Convergence of Algorithm 1. (b) Evolution of KLD.

episode reward and the average reward increase gradually. After 20 episodes, the average reward floats in a certain interval. This illustrates that Algorithm 1 is convergent and the most economic strategies are obtained. Also, (b) in Figure 2 shows the changing of KLD. Clearly, the value of $D(\tilde{z}_k \| z_k)$ stabilizes at a tiny interval after the initial increase and subsequent decrease. That is to say, the performance of the system could be guaranteed.

For the interconnected NETS-NYPS, Figure 3 shows a similar meaning to Figure 2. However, there are more fluctuations in the former figure than in the latter figure since the interconnected NETS-NYPS is a large system with many states. In (a) of Figure 3, the average reward also varies in a certain range after the initial rise. Similar to (b) of Figure 2, the performance of the system could also be guaranteed even though there are great changes in the value of KLD as shown in (b) of Figure 3. Based on the figures, it is concluded that the simpler the systems, the easier it is to implement it in the guaranteeing of the system performance and in the smoothening of the convergence performance of Algorithm 1.

## 5. Conclusion

This paper addressed the cooperative game between the two energy-constrained agents who injected the covert signals into the two communication channels of the wireless systems, namely, C-A and S-C channels, in order to confuse the advisories. This process is modeled as a Markov game with identical payoffs for both the agents. Since the wireless system is unknown to agents, the policy gradient method is applied to search for the optimal strategy, combined with fictitious play. The system performance with covert signals is measured by KLD when covert signals are launched to confuse the advisories. In addition, the economic strategy is considered to avoid excessive waste of signal injections. The feasibility of theoretical results was validated on both the NETS-NYPS 68-bus system and the interconnected NETS-NYPS. In the future work, this research will be extended to the scenario where attackers will be able to detect the defence strategies (e.g., injecting the covert signals in this paper). In this case, it is necessary to propose an additional predictor to estimate the attackers' behaviours so as to dynamically change the defence strategies. *Also, the problem of co-operation or competition among agents (more than two agents) deserves to be investigated in a similar setting.*

## Appendix

## A. Proof of Proposition 6

When the system is subjected to covert signals in (4), the state equation, output equation, and the state estimate equation should be rewritten as

$$\begin{aligned}\tilde{x}_{k+1} &= Ax_k + B(\tilde{u}_k + \Delta u_k) + w_k,\\ \tilde{y}_k &= Cx_k + v_k + \Delta y_k,\\ \hat{\tilde{x}}_k &= \hat{x}_k^- + K(\tilde{y}_k - C\hat{x}_k^-),\end{aligned} \tag{A.1}$$

where the state equation follows from the assumption that the agent launches signals from instant $k$ and $\hat{x}_k^-$ is also not changed at instant $k$ since $\hat{x}_k^-$ relates to the historic instants, which are the instants before $k$. Then, one has

$$\tilde{x}_k - \hat{\tilde{x}}_k = x_k - \hat{\tilde{x}}_k = (I - KC)(x_k - \hat{x}_k^-) - Kv_k - K\Delta y_k, \tag{A.2}$$

and the estimate error covariance at instant $k$ is

$$\begin{aligned}\widetilde{P}_k &= \mathbb{E}\left[\left(\tilde{x}_k - \hat{\tilde{x}}_k\right)\left(\tilde{x}_k - \hat{\tilde{x}}_k\right)^T\right] \\ &= \overline{P} + K\left[\mu_k^y\left(\mu_k^y\right)^T + \Sigma_k^y\right]K^T,\end{aligned} \tag{A.3}$$

which is derived from $v_k$ and $\Delta y_k$ which are mutually independent Gaussian processes and (17).

At instant $k+1$, the output equation, state estimate equation, and the priori state estimate equation are denoted as

$$\begin{aligned}\tilde{y}_{k+1} &= C\tilde{x}_{k+1} + v_{k+1} + \Delta y_{k+1}, \\ \hat{\tilde{x}}_{k+1} &= \hat{\tilde{x}}_{k+1}^{-} + K\left(\tilde{y}_{k+1} - C\hat{\tilde{x}}_{k+1}^{-}\right), \\ \hat{\tilde{x}}_{k+1}^{-} &= A\hat{\tilde{x}}_k + B\tilde{u}_k.\end{aligned} \tag{A.4}$$

Then, one has

$$\tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1} = A\left(x_k - \hat{\tilde{x}}_k\right) + B\Delta u_k + w_k - Kv_{k+1} - KC\left(\tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1}^{-}\right) - K\Delta y_{k+1}. \tag{A.5}$$

The term $\tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1}^{-}$ could be further evaluated as

$$\tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1}^{-} = A\left(x_k - \hat{\tilde{x}}_k\right) + w_k + B\Delta u_k. \tag{A.6}$$

Hence, (A.5) could be further represented in the form of

$$\tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1} = (I - KC)\left(A\left(x_k - \hat{\tilde{x}}_k\right) + w_k\right) - Kv_{k+1} + (I - KC)B\Delta u_k - K\Delta y_{k+1}. \tag{A.7}$$

and the corresponding estimate error covariance of $\widetilde{P}_{k+1}$ is

$$\begin{aligned}\widetilde{P}_{k+1} &= \mathbb{E}\left[\left(\tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1}\right)\left(\tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1}\right)^T\right] \\ &= (I - KC)A\hat{\widetilde{P}}_k A^T (I - KC)^T + KRK^T + (I - KC)Q(I - KC)^T \\ &\quad + K\mathbb{E}\left[\Delta y_{k+1}\Delta y_{k+1}^T\right]K^T + (I - KC)B\mathbb{E}\left[\Delta u_k \Delta u_k^T\right]B^T (I - KC)^T \\ &= \overline{P} + (I - KC)BE_k^u B^T (I - KC)^T \\ &\quad + \sum_{i=0}^{1} \left(\left((I - KC)A\right)^i KE_{k+i}^y K^T \left(A^T (I - KC)^T\right)^i,\end{aligned} \tag{A.8}$$

where the third equality is obtained by substituting $(A)$ into $\widetilde{P}_{k+1}$. At instant $k+2$, the state equation, output equation, state estimate equation, and the priori state estimate equation are deduced as

$$\begin{aligned}\tilde{x}_{k+2} &= A\tilde{x}_{k+1} + B\tilde{u}_{k+1} + B\Delta u_{k+1} + w_{k+1}, \\ \tilde{y}_{k+2} &= C\tilde{x}_{k+2} + v_{k+2} + \Delta y_{k+2}, \\ \hat{\tilde{x}}_{k+2} &= \hat{\tilde{x}}_{k+2}^{-} + K\left(\tilde{y}_{k+2} - C\hat{\tilde{x}}_{k+2}^{-}\right), \\ \hat{\tilde{x}}_{k+2}^{-} &= A\hat{\tilde{x}}_{k+1} + B\tilde{u}_{k+1}.\end{aligned} \tag{A.9}$$

Then, one has

$$\tilde{x}_{k+2} - \hat{\tilde{x}}_{k+2} = (I - KC)A\left(\tilde{x}_{k+1} - \hat{\tilde{x}}_{k+1}\right) + (I - KC)w_{k+1} + (I - KC)B\Delta u_{k+1} - Kv_{k+2} - K\Delta y_{k+2}, \tag{A.10}$$

and the estimate error covariance $\widetilde{P}_{k+2}$ is obtained as shown in the following equation:

$$\begin{aligned}\widetilde{P}_{k+2} &= \mathbb{E}\left[\left(\widetilde{x}_{k+2} - \widehat{\widetilde{x}}_{k+2}\right)\left(\widetilde{x}_{k+2} - \widehat{\widetilde{x}}_{k+2}\right)^T\right] \\ &= \overline{P} + \sum_{i=0}^{2} ((I-KC)A)^i K E^y_{k+i} K^T \left(A^T (I-KC)^T\right)^i \\ &\quad + \sum_{i=0}^{1} ((I-KC)A)^i (I-KC) B E^u_{k+i} B^T (I-KC)^T \left(A^T (I-KC)^T\right)^i.\end{aligned} \tag{A.11}$$

Similarly, one has

$$\widetilde{P}_{k+q} = \overline{P} + \sum_{i=0}^{q} ((I-KC)A)^i K E^y_{k+i} K^T \left(A^T (I-KC)^T\right)^i, \tag{A.12}$$

where $q \geq 0$.

Therefore, the proof is completed according to the definitions of $E_q$ and $F_q$.

*A.1. Proof of Proposition 8.* Firstly, (A.12) could be rewritten as

$$\begin{aligned}\widetilde{P}_{k+q} - \overline{P} &= \sum_{i=0}^{q} ((I-KC)A)^i K E^y_{k+i} K^T \left(A^T (I-KC)^T\right)^i \\ &\quad + \sum_{i=0}^{q-1} ((I-KC)A)^i (I-KC) B E^u_{k+i} B^T (I-KC)^T \left(A^T (I-KC)^T\right)^i.\end{aligned} \tag{A.13}$$

Moreover, we could have

$$\begin{aligned}\widetilde{P}_{k+q} - \overline{P} &\leq \max_{0\leq i\leq q}\left\|E^y_{k+i}\right\| \sum_{i=0}^{q} ((I-KC)A)^i K K^T \left(A^T (I-KC)^T\right)^i \\ &\quad + \max_{0\leq i\leq q-1}\left\|E^u_{k+i}\right\| \sum_{i=0}^{q-1} ((I-KC)A)^i (I-KC) B B^T (I-KC)^T \left(A^T (I-KC)^T\right)^i \\ &\leq \max_{0\leq i\leq q}\left\|E^y_{k+i}\right\| \left\|K K^T\right\| \sum_{i=0}^{q} ((I-KC)A)^i \left(A^T (I-KC)^T\right)^i \\ &\quad + \max_{0\leq i\leq q-1}\left\|E^u_{k+i}\right\| \left\|(I-KC) B B^T (I-KC)^T\right\| \sum_{i=0}^{q} ((I-KC)A)^i \left(A^T (I-KC)^T\right)^i \\ &\leq \max_{0\leq i\leq q}\left\|E^y_{k+i}\right\| \left\|K K^T\right\| \sum_{i=0}^{q} \widehat{A}\lambda^i + \max_{0\leq i\leq q-1}\left\|E^u_{k+i}\right\| \left\|(I-KC) B B^T (I-KC)^T\right\| \sum_{i=0}^{q-1} \widehat{A}\lambda^i,\end{aligned} \tag{A.14}$$

where the last inequality is based on the fact that there is always a constant matrix $\widehat{A}$ such that $((I-KC)A((I-KC)A)^T)^i \leq \widehat{A}\lambda^i$, where $\lambda = \rho((I-KC)A) < 1$ [34]. Thus, there is always a constant $\delta_1$ such that

$$\begin{aligned}\widetilde{\Sigma}_{k+q}-\Sigma &\le \max_{0\le i\le q}\left\|E^{y}_{k+i}\right\|\left\|KK^{T}\right\|\sum_{i=0}^{q} CA\widehat{A}^{T}A^{T}C^{T}\lambda^{i}\\ &\quad+\max_{0\le i\le q-1}\left\|E^{u}_{k+i}\right\|\left\|(I-KC)BB^{T}(I-KC)^{T}\right\|\sum_{i=0}^{q-1} CA\widehat{A}A^{T}C^{T}\lambda^{i}\\ &\le \left(a\max_{0\le i\le q}\left\|E^{y}_{k+i}\right\|+b\max_{0\le i\le q-1}\left\|E^{u}_{k+i}\right\|\right)CA\widehat{A}A^{T}C^{T}\\ &\le \delta_1 CA\widehat{A}A^{T}C^{T},\end{aligned} \tag{A.15}$$

where $\Sigma = C(A\overline{P}A+Q)C^T+R = CPC^T+R$, $\widetilde{\Sigma}_{k+q} = C(A\widetilde{P}_{k+q-1}A^T+Q)C^T+R$, $a=(\|K\|^2/1-\lambda)$, $b=(\|(I-KC)B\|^2/1-\lambda)$, and $\delta_1 = a\max_{0\le i\le q}\|E^{y}_{k+i}\| + b\max_{0\le i\le q-1}\|E^{u}_{k+i}\|$. Then, according to (A.15), one has

$$\Sigma^{-(1/2)}\left(\widetilde{\Sigma}_{k+q}-\Sigma\right)\Sigma^{(1/2)} \le \delta_1\Sigma^{-(1/2)}CA\widehat{A}A^{T}C^{T}\Sigma^{(1/2)}. \tag{A.16}$$

Furthermore, (A.16) could be deduced as

$$\operatorname{Tr}\left[\Sigma^{-1}\widetilde{\Sigma}_{k+q}\right]-m\le\delta_1\operatorname{Tr}\left[\Sigma^{-1}CA\widehat{A}A^{T}C^{T}\right]. \tag{A.17}$$

In addition, we note that $\widetilde{P}_{k+q}\ge\overline{P}$ always holds according to Proposition 6. Then, one has

$$\log\frac{\det(\Sigma)}{\det\left(\widetilde{\Sigma}_{k+q}\right)}\le 0. \tag{A.18}$$

Also, the upper bound of $(\mu^z_k)^T\sum_k^{-1}\mu^z_k$ denoted as $\delta_2$ could be derived if the agents launch the most expensive covert signals at every instant from the initial injection. Therefore, by (A.17) and (A.18) and $\delta_2$, the system performance is guaranteed according to Definition 7 and (22) subjected to the following condition:

$$\delta_1\le\frac{2(\delta-\delta_2)}{\operatorname{Tr}\left[\Sigma^{-1}CA\widehat{A}A^{T}C^{T}\right]}. \tag{A.19}$$

## B. Proof of Theorem 10

The following three definitions are essential for the proof.

*Definition B.1.* Let $b_1 := \min_{s\in\mathbb{S},a^i\in\mathbb{A}^i} r(s,a^u,a^y)$, $b_2 := \max_{s\in\mathbb{S},a^i\in\mathbb{A}^i} r(s,a^u,a^y)$, $i\in\mathscr{N}$, and $\widehat{B} := [b_1,b_2]$. The best responses of the auxiliary game in state $s$ with the continuation payoff vector $\overrightarrow{u}$ are denoted by

$$\begin{aligned}BR^{u}_{s,\overrightarrow{u}}(\pi^{y}_{s}) &= \operatorname*{argmax}_{\pi^{u,\star}_{s}\in\mathbb{L}^{u}} f_{s,\overrightarrow{u}}(\pi^{u,\star}_{s},\pi^{y}_{s}),\\ BR^{y}_{s,\overrightarrow{u}}(\pi^{u}_{s}) &= \operatorname*{argmax}_{\pi^{y,\star}_{s}\in\mathbb{L}^{y}} f_{s,\overrightarrow{u}}(\pi^{u}_{s},\pi^{y,\star}_{s}),\end{aligned} \tag{B.1}$$

where $\overrightarrow{u}$ is the vector of the long-term payoff $u_s$ and $u_s\in\widehat{B}$ is an arbitrarily long-term payoff in state $s$.

*Remark B.2.* It is noted that some notations in Definition B.1 with subscript $s$, such as $u_s$, $\pi^u_s$, $\pi^y_s$, and $f_{s,\overrightarrow{u}}(\pi^{u,\star}_s,\pi^y_s)$, correspond to the notations defined in the previous sections. More specifically, $u_s=r_k$ (i.e., the immediate reward), $\pi^i_s=\pi^i(s_k)$, and $f_{s,\overrightarrow{u}}=J(x,s,\Delta u,\Delta y)$ (i.e. the payoff), and the same expressions also exist in Definitions B.3 and B.4.

*Definition B.3.* For every state $s\in\mathbb{S}$ at every time $t\ge 1$, the best-response dynamic of agent $i\in\mathscr{N}$ in Markov cooperative game $\mathbb{G}$ evolves according to

$$\begin{aligned}\dot{u}_s(t) &= \frac{f_{s,\overrightarrow{u}(t)}\left(\pi^1_s(t),\pi^2_s(t)\right)-u_s(t)}{t},\\ \dot{\pi}^i_s(t) &\in \mathbb{I}_s(t)\left(BR^{i}_{s,\overrightarrow{u}(\mathrm{t})}\left(\pi^{-i}_s(t)\right)-\pi^i_s(t)\right),\end{aligned} \tag{B.2}$$

where $f_{s,\overrightarrow{u}}(\cdot) := (1-\beta)r(s)+\beta\sum_{s'\in\mathbb{S}}\mathbb{T}_{s,s'}u_{s'}$ is the future expected discounted payoff in state $s$, $s'$ denotes the next state, $-i$ denotes the other agent, and $\mathbb{I}_s(t)$ also denotes the indicator function such that $\mathbb{I}_s(t)=1$ if the agents are in the state $s$ at time $t$, otherwise, $\mathbb{I}_s(t)=0$.

*Definition B.4.* For any stationary optimal strategy $\pi$, the value of state $s\in\mathbb{S}$ denoted as $\mathrm{Val}_s$ satisfies the following equation: $\mathrm{Val}_s=(1-\beta)r(s,\pi)+\beta\sum_{s\in\mathbb{S}}\mathbb{T}^{\pi}_{s,s'}\mathrm{Val}_{s'}$.

Based on these definitions, the sketch proof of Theorem 10 is given as follows. According to Definitions B.1 and B.3, the term $|f_{s,\overrightarrow{u}(t)}(\pi^u_s(t),\pi^y_s(t))-u_s(t)|$ in (B.2) is bounded, which could infer that $|\dot{u}_s(t)|\longrightarrow 0$ as $t\longrightarrow\infty$. In other words, the long-term payoff $\overrightarrow{u}$ moves slowly. This could conversely verify that $|f_{s,\overrightarrow{u}(t)}(\pi^u_s(t),\pi^y_s(t))-v(g_{s,\overrightarrow{u}})|\longrightarrow 0$, where $v(g_{s,\overrightarrow{u}})$ is the value of the cooperative game and $g_{s,\overrightarrow{u}}$ denotes the auxiliary game with payoff $f_{s,\overrightarrow{u}}$. Then, the optimal strategies could be derived according to [46] and the references therein.

In fact, the dynamical system (B.2) could be considered as a feedback system in which $f_{s,\overrightarrow{u}(t)}(\pi^u_s(t),\pi^y_s(t))$ transforms the economic strategy to payoff and the best response to the current belief about the other agent transforms the payoff back to the strategy.

## Data Availability

The model and data used to support the findings of this study can be downloaded from the following website: https://electricgrids.engr.tamu.edu/electric-grid-test-cases/new-england-68-bus-test-system/.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

The author would like to thank China Postdoctoral Science Foundation for their support (Grant no.: 2022M710921).

## References

[1] M. Cheng, M. Cardei, M. Cardei, J. Sun, and D. Du, "Topology control of ad hoc wireless networks for energy efficiency," *IEEE Transactions on Computers*, vol. 53, no. 12, pp. 1629–1635, 2004.

[2] X. Gao, X. Zhu, J. Li et al., "A novel approximation for multi-hop connected clustering problem in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 2223–2234, 2017.

[3] Y. Liu, J. Peng, J. Kang, A. M. Iliyasu, D. Niyato, and A. A. A. El-Latif, "A secure federated learning framework for 5G networks," *IEEE Wireless Communications*, vol. 27, no. 4, pp. 24–31, 2020.

[4] S. Li, Y. Yılmaz, and X. Wang, "Quickest detection of false data injection attack in wide-Area smart grids," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2725–2735, 2015.

[5] J. Cao, X. Zhu, and S. Sun, "Age of loop oriented wireless networked control system: communication and control co-design in the FBL regime," in *Proceedings of the IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, New York, NY, USA, May 2022.

[6] L. Guo, T. Cui, H. Yu, and F. Hao, "Stability of networked control system subject to denial-of-service," *Science China Information Sciences*, vol. 64, no. 2, 3 pages, Article ID 129203, 2021.

[7] L. Guo, H. Yu, and F. Hao, "Event-triggered control for stochastic networked control systems against Denial-of-Service attacks," *Information Sciences*, vol. 527, pp. 51–69, 2020.

[8] Y. Li, R. Huang, and L. Ma, "Hierarchical-attention-based defense method for load frequency control system against DoS attack," *IEEE Internet of Things Journal*, vol. 8, no. 20, 15530 pages, Article ID 15522, 2021.

[9] N. Zhao, P. Shi, W. Xing, and J. Chambers, "Observer-based event-triggered approach for stochastic networked control systems under denial of service attacks," *IEEE Transaction Control Network System*, vol. 8, no. 1, pp. 158–167, 2021.

[10] L. Hu, Z. Wang, Q. Han, and X. Liu, "State estimation under false data injection attacks: security analysis and system protection," *Automatica*, vol. 87, pp. 176–183, 2018.

[11] W. Qi, Y. Hou, G. Zong, and C. Ahn, "Finite-time event-triggered control for semi-markovian switching cyber-physical systems with FDI attacks and applications," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 6, pp. 2665–2674, 2021.

[12] M. Ghaderi, K. Gheitasi, and W. Lucia, "A blended active detection strategy for false data injection attacks in cyber-physical systems," *IEEE Trans Control Netw Syst*, vol. 8, no. 1, pp. 168–176, 2021.

[13] M. Masdari, S. Bazarchi, and M. Bidaki, "Analysis of secure LEACH-based clustering protocols in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 36, no. 4, pp. 1243–1260, 2013.

[14] C. Liang, K. Qiu, Z. Zhang, J. Yang, Y. Li, and J. Hu, "Towards robust and stealthy communication for wireless intelligent terminals," *International Journal of Intelligent Systems*, vol. 37, no. 12, 11814 pages, Article ID 11791, 2022.

[15] C. Wu and W. Li, "Enhancing intrusion detection with feature selection and neural network," *International Journal of Intelligent Systems*, vol. 36, no. 7, pp. 3087–3105, 2021.

[16] M. Kurt, O. Ogundijo, C. Li, and X. Wang, "Online cyber-attack detection in smart grid: a reinforcement learning approach," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 5174–5185, 2019.

[17] M. Esmalifalak, H. Nguyen, and R. Zheng, "Stealth false data injection using independent component analysis in smart grid," *Proc.IEEE Int. Conf. Smart Grid Commun.*, pp. 244–248, 2011.

[18] R. Zhang and P. Venkitasubramaniam, "False data injection and detection in LQG systems: a game theoretic approach," *IEEE Trans Control Netw Syst*, vol. 7, no. 1, pp. 338–348, 2020.

[19] H. Zhang, P. Cheng, L. Shi, and J. Chen, "Optimal dos attack scheduling in wireless networked control system," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 3, pp. 843–852, 2016.

[20] R. Deng, P. Zhuang, and H. Liang, "False data injection attacks against state estimation in power distribution systems," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2871–2881, 2019.

[21] T. Keviczky, F. Borrelli, K. Fregene, D. Godbole, and G. J. Balas, "Decentralized receding horizon control and coordination of autonomous vehicle formations," *IEEE Transactions on Control Systems Technology*, vol. 16, no. 1, pp. 19–33, 2008.

[22] J. Wang, Y. Zhang, and T. Kim, "Shapley q-value: a local reward approach to solve global reward games," in *Proceedings of the 34th AAAI conf Artificial Intelligence*, Hillton, NY, USA, February 2019.

[23] H. Jiang, H. Zhang, X. Xie, and J. Han, "Neural-network-based learning algorithms for cooperative games of discrete-time multi-player systems with control constraints via adaptive dynamic programming," *Neurocomputing*, vol. 344, pp. 13–19, 2019.

[24] K. Vamvoudakis and J. Hespanha, "Cooperative Q-learning for rejection of persistent adversarial inputs in networked linear quadratic systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 4, pp. 1018–1031, 2018.

[25] K. Ding, Y. Li, D. Quevedo, S. Dey, and L. Shi, "A multi-channel transmission schedule for remote state estimation under DoS attacks," *Automatica*, vol. 78, pp. 194–201, 2017.

[26] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MA, USA, 2nd edition, 2017.

[27] R. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[28] Y. Yang, H. Modares, K. Vamvoudakis, W. He, C. Z. Xu, and D. C. Wunsch, "Hamiltonian-driven adaptive dynamic programming with approximation errors," *IEEE Transactions on Cybernetics*, vol. 52, no. 12, 13773 pages, Article ID 13762, 2022.

[29] Y. Yang, Y. Pan, C. Xu, and D. C. Wunsch, "Hamiltonian-driven adaptive dynamic programming with efficient experience replay," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 10, pp. 1–13, 2022.

[30] Y. Yang, B. Kiumarsi, H. Modares, and C. Xu, "Model-free $\lambda$-policy iteration for discrete-time linear quadratic regulation $\lambda$-policy iteration for discrete-time linear quadratic regulation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 2, pp. 635–649, 2023.

[31] G. W. Brown, "Iterative solution of games by fictitious play," *Act. Anal. Prod Allocation.*vol. 13, no. 1, p. 374, 1951.

[32] U. Berger, “Brown’s original fictitious play,” *Journal of Economic Theory*, vol. 135, no. 1, pp. 572–578, 2007.

[33] D. Silver, G. Lever, and N. Heess, “Deterministic policy gradient algorithms,” in *Proceedings of the 31 st International Conference on Machine Learning*, vol. 32, Beijing, China, June 2014.

[34] L. Guo, H. Yu, and F. Hao, “Optimal allocation of false data injection attacks for networked control systems with two communication channels,” *IEEE Trans Control Netw Syst*, vol. 8, no. 1, pp. 2–14, 2021.

[35] H. Yu, J. Shang, T. Chen, and T. Chen, “Stochastic Stochastic event-based LQG control: An analysis on strict consistencyvent-based lqg control: an analysis on strict consistency,” *Automatica*, vol. 138, Article ID 110157, 2022.

[36] Y. Li, Y. Yang, T. Chai, and T. Chen, “Stochastic detection against deception attacks in CPS: performance evaluation and game-theoretic analysis,” *Automatica*, vol. 144, Article ID 110461, 2022.

[37] J. Lu and D. Quevedo, “A jointly optimal design of control and scheduling in networked systems under denial-of-service attacks,” *Automatica*, vol. 148, Article ID 110774, 2023.

[38] R. M. G. Ferrari and A. M. H. Teixeira, “Safety, security and privacy for cyber-physical systems,” *Lecture Notes in Control and Information Sciences*, John Wiley, Hoboken, NJ, USA, 2021.

[39] Y. Liu, W. Zhang, F. Chen, and J. Li, “Path planning based on improved deep deterministic policy gradient algorithm,” in *Proceedings of the 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference*, Chengdu, China, March 2019.

[40] C. Watkins and P. Dayan, “Q-learning,” *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[41] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, MS, USA, 2018.

[42] V. Konda and J. Tsitsiklis, “Actor-critic algorithms,” *Advances in Neural Information Processing Systems*, vol. 12, pp. 1008–1014, 2000.

[43] R. Lowe, Y. Wu, and A. Tamar, “Multi-agent actor-critic for mixed cooperative-competitive environments,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 6379–6390, 2017.

[44] R. Sutton, D. McAllester, S. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” *Neural Info Processing Syst*, vol. 12, pp. 1057–1063, 1999.

[45] D. Leslie and E. Collins, “Generalised weakened fictitious play,” *Games and Economic Behavior*, vol. 56, no. 2, pp. 285–298, 2006.

[46] D. Leslie, S. Perkins, and Z. Xu, “Best-response dynamics in zero-sum stochastic games,” *Journal of Economic Theory*, vol. 189, Article ID 105095, 2020.

[47] A. Singh and B. Pal, *Dynamic Estimation and Control of Power System*, Academic Press, Cambridge, MS, USA, 2018.