Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 53-2 (2020) 2501–2506

Reinforcement Learning Based Anti-Jamming Schedule in Cyber-Physical Systems

Ruimeng Gan * **Yue Xiao** * **Jinliang Shao** ** **Heng Zhang** *** **Wei Xing Zheng** ****

* *National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mail: gruimxq@163.com, xiaoyue@uestc.edu.cn)*

** *School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, 611731, China (e-mail: jinliangshao@126.com)*

*** *School of Science, Jiangsu Ocean University, Lianyungang, Jiangsu, 222005, China (e-mail: Dr.Zhang.Heng@ieee.org)*

**** *School of Computing, Engineering and Mathematics, Western Sydney University, Sydney, NSW 2751, Australia (e-mail: w.zheng@westernsydney.edu.au)*

Abstract: In this paper, the security issue of cyber-physical systems is investigated, where the observation data is transmitted from a sensor to an estimator through wireless channels disturbed by an attacker. The failure of this data transmission occurs, when the sensor accesses the channel that happens to be attacked by the jammer. Since the system performance measured by the estimation error depends on whether the data transmission is a success, the problem of selecting the channel to alleviate the attack effect is studied. Moreover, the state of each channel is time-variant due to various factors, such as path loss and shadowing. Motivated by energy conservation, the problem of selecting the channel with the best state is also considered. With the help of cognitive radio technique, the sensor has the ability of selecting a sequence of channels dynamically. Based on this, the problem of selecting the channel is resolved by means of reinforcement learning to jointly avoid the attack and enjoy the channel with the best state. A corresponding algorithm is presented to obtain the sequence of channels for the sensor, and its effectiveness is proved analytically. Numerical simulations further verify the derived results.

Keywords: Cyber-physical systems (CPSs), reinforcement learning, cognitive radio, softmax method.

1. INTRODUCTION

The interest in cyber-physical systems (CPSs) is becoming prevalent due to its potential applications in smart grid, unmanned aerial vehicles, environmental monitoring, etc. A key characteristic of such CPS applications is the integration of physical processes and wireless communication technologies with the benefit of low cost and energy saving Schenato et al. (2007). But this also leads to the CPSs vulnerable to cyber attacks, such as false data injection attack Liu et al. (2007) and Denial-of-Service (DoS) attack Xu et al. (2019), because of the broadcast nature of wireless networks. It is well known that the CPS performance can be degraded as a result of the presence of such kind of attacks. For instance, in 2008 the switch tracks of trams were attacked, which resulted in four derailments and twelve resultant injuries Iasiello (2013). Therefore, exploring anti-interference approaches in CPSs is necessary to ensure the implementation of reliable data transmission.

The problem of designing strategies to alleviate the attack influence has attracted considerable attention in recent years. For example, the work in Pasqualetti et al. (2013) concentrates on exploring detection methods to counteract the attack effect in real time. Moreover, assuming that the jammer has been detected in CPSs, the problem of ensuring the system stability is investigated in Foroush et al. (2012). Further, taking safety conditions into account, an optimization policy to maximize the system objective is proposed in Amin et al. (2009). An alternative scenario from the viewpoint of the game theory is discussed, where the problem of maximizing the benefit of the sensor and the jammer is considered in Li et al. (2017), respectively. A common hypothesis for the above-mentioned works is that the data transmission between the transmitter and the receiver is over a single wireless channel interfered with an attacker. Besides, although in the setting of multiple channels some policies are proposed to alleviate the attack effect Ding et al. (2017), the channel state in itself is assumed to be independent of time.

10.1016/j.ifacol.2020.12.221

In practice, the channel quality changes with time due to various factors, such as path loss, shadowing, etc. Specially, from the perspective of unlicensed users the channel state varies dynamically with time, since it is uncertain whether the channel is occupied by licensed users Che et al. (2011). In this setting, the cognitive radio (CR) technique is utilized by the sensor to enhance the CPS performance, where the sensor is endowed with the ability of sensing the channel quality and then accessing the corresponding channel Cao et al. (2014). Naturally, an interesting question arises: how to employ the CR technology to enhance the anti-jamming ability from the perspective of the security? Motivated by this, the current paper studies the problem of selecting which channel to transmit the data so as to maximize the CPS performance when the state of each channel varies with time.

Generally, such kind of problem of selecting a sequence of actions among the available choices is typically addressed by using an online learning technique Auer et al. (2002). In this paper, we deal with the problem of selecting the sequence of channels through establishing a new kind of online learning model, referred to as the period online learning (POL) model. Compared with the previous literature, the contributions of this work are summarized as follows.

1 We establish a POL model where the sensor has the ability to sense and access the channel with the help of the CR technique. The objective of the sensor is to jointly avoid the attack and maximize the long-term reward, which is considered in CPSs for the first time as far as we know.
2 We present a novel method based on the reinforcement learning, referred to as the POL algorithm, to explore the sequence of channels for the sensor. Moreover, we prove analytically that the performance of this algorithm is better than that of the softmax (SM) method. Besides, the effectiveness of the POL algorithm is further verified using numerical simulations.

The remainder of the paper is organized as follows. In Section 2, we provide the mathematical model for the problem of selecting the channel for remote state estimation in CPSs in the presence of periodic DoS attacks. The main result about the POL model and algorithm is then presented in Section 3. Simulation results are given in Section 4. Finally, Section 5 concludes the paper.

2. PROBLEM SETUP

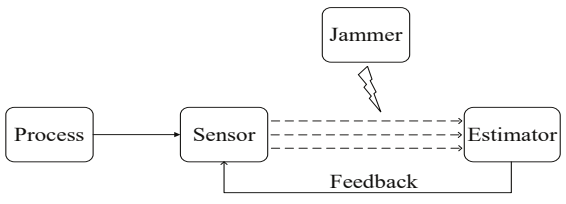

Fig. 1. The schematic of the CPS in the presence of a jammer.

The schematic of a generic single-hop structure for the CPS in the presence of an attacker is depicted in Fig. 1, where the process of the plant is represented by a linear time-invariant (LTI) system. The sensor transmits the observation data to the estimator through wireless channels that are attacked by a jammer, and the state of each channel varies with time. With the help of the CR technique, the sensor can dynamically select the channel available based on the feedback transmitted from the estimator so as to enhance the system performance. In the following, we present the mathematical model of the system illustrated in Fig. 1.

2.1 Local State Estimation

The physical process is described by the following discrete-time dynamic LTI system

$$x_{t+1} = Ax_t + \omega_t, \tag{1}$$

where $x_t \in \mathbb{R}^{n_x}$ denotes the system state, and $\omega_t \in \mathbb{R}^{n_x}$ is the noise that is assumed to be of Gaussian distribution $\mathcal{N}(0, \Sigma_\omega)$. The sensor observes the system state using the following dynamic system

$$y_t = Cx_t + \nu_t, \tag{2}$$

where $y_t \in \mathbb{R}^{n_y}$ is the measurement output at time t, and $\nu_t \in \mathbb{R}^{n_y}$ represents the measurement noise following Gaussian distribution $\mathcal{N}(0, \Sigma_\nu)$. Generally, it is assumed that (A, C) is observable and $(A, \Sigma_\omega^{\frac{1}{2}})$ is stabilizable.

The sensor usually performs the local estimation for x_t before forwarding it to the remote estimator Li et al. (2017). Denote the minimum mean squared error (MMSE) estimate of x_t and its corresponding error covariance, respectively, by $\widehat{x}_t^s$ and P_t^s. Then, we have $\widehat{x}_t^s = \mathbb{E}[x_t|y_1, \cdots, y_t]$ and $P_t^s = \mathbb{E}[(x_t - \widehat{x}_t^s)(x_t - \widehat{x}_t^s)^T|y_1, \cdots, y_t]$, where X^T denotes the transpose of matrix X.

Based on the Kalman filter described in Gu et al. (2006), P_t^s converges to the steady state $\overline{P}$ exponentially, where $\overline{P}$ is the unique positive semi-definite solution of $g \circ h(X) = X$ with $h(X) \triangleq AXA^T + \Sigma_\omega$ and $g(X) \triangleq X - XC^T(CXC^T + \Sigma_\nu)^{-1}CX$. Note that for functions f_1 and f_2 with appropriate domains, $f_1 \circ f_2(x)$ denotes the function composition $f_1(f_2(x))$. For simplicity, like Ding et al. (2017), it is also assumed that the standard Kalman filter is at the steady state, i.e., $P_t^s = \overline{P}$, for each $t \geq 1$.

2.2 Wireless Channel in the Presence of Attacker

Assume that the channel quality varies with time, which makes sense due to the fact that each channel is affected by various factors in different ways, such as path loss, shadowing, etc. For the sake of analysis, as described in Felice et al. (2010), assume that all these factors can be represented by one state, e.g., the throughput. Define the set of M channels available by $\mathcal{C} = \{c_k, k = 1, \cdots, M\}$. And then the state for channel c_k at time t can be expressed as r_{tk}, where $r_{1k}, r_{2k}, \cdots$ are independently and identically distributed (i.i.d), following an unknown distribution with unknown expectation μ_k. Moreover, for throughput across channels the independence is also satisfied, i.e., for any $k_1, k_2 \in [1, M]$ and $t_1, t_2 \geq 1$, $r_{t_1k_1}$ and $r_{t_2k_2}$ are mutually independent, where for two integers a_1 and a_2 with $a_1 \leq a_2$, the notation $[a_1, a_2]$ represents the set of $\{a_1, a_1 + 1, \cdots, a_2\}$.

Similar to Foroush et al. (2012) and Hu et al. (2018), the action of the jammer with DoS attack is assumed to be periodic. Specifically, the action period $\mathcal{T}$ is composed of the active period $\mathcal{T}_{\rm a} = [(n-1)T+1, (n-1)T+T_{\rm a}]$ and the rest period $\mathcal{T}_{\rm r} = [(n-1)T+T_{\rm a}+1, nT]$, where T and $T_{\rm a}$ denote the length of the action period and the active period, respectively, and n represents the period number. Moreover, let β_{tk} be an indication function, where $\beta_{tk} = 1$ denotes that the jammer has the attack on channel c_k at time t, otherwise $\beta_{tk} = 0$. Then, the attack launched at time t is $\beta_t = (\beta_{t1}, \cdots, \beta_{tM})^T$ in a vector form. When the attack is launched at the j-th time in period n, i.e., $t = (n-1)T + j$, the corresponding attack can also be written as $\beta_{n_j} = (\beta_{n_j 1}, \cdots, \beta_{n_j M})^T$. For $n_1, n_2 \geq 1$, based on the characteristic of periodicity, we have $\beta_{n_1 j} = \beta_{n_2 j}, \forall j \in [1, T]$. Taking the energy constraint into account, we assume that $|\beta_{n_j}| \neq M$, where $|\beta_{n_j}| = \sum_{k=1}^{M} \beta_{n_j k}$. Furthermore, when the jammer launches DoS attacks on channel c_k at time t, the channel state is busy and unavailable. This leads to the estimator receiving no data, when the sensor selects to sense this channel. Additionally, it is generally assumed that the amount of time slot is an instant, and the attacker knows the start and end times of each data transmission Roy et al. (2013).

2.3 Dynamical Selection of Channel

With the aid of the CR technique, the sensor can predict the quality of each channel before accessing it. The sensor is assumed to be energy-constrained, and thus can select one channel to sense at each time t. Define θ_{tk} as an indication function, where $\theta_{tk} = 1$ when the sensor selects channel c_k to sense at time t, otherwise $\theta_{tk} = 0$. The corresponding vector form is $\theta_t = (\theta_{t1}, \cdots, \theta_{tM})^T$, and there is only one non-zero element in θ_t at any time t. Moreover, since whether the channel is attacked can be determined based on many metrics, such as packet send ratio and packet delivery ratio (see Xu et al. (2005) for more information), the sensor can sense the state of the channel selected. It is clear that there are only two possibilities for the sensing result: the channel state is either busy or idle, resulting from the jammer launching an attack or no attack, respectively. When $\theta_t^T \beta_t = 1$, the sensor transmits no data and thus the estimator receives no data, otherwise, the sensor communicates with the estimator and the estimator can obtain this data.

After obtaining the data, the estimator transmits r_{tk} as a feedback reward to the sensor. This feedback reward, in turn, helps the sensor make a decision in assessing the quality of the corresponding channel. As a result, at time t the reward fed to the sensor with respect to channel k is given by

$$R_t^{\rm Th} = \begin{cases} 0, & \theta_t^T \beta_t = 1, \\ r_{tk}, & \theta_t^T \beta_t = 0. \end{cases} \tag{3}$$

Note that the estimator sends feedback reward r_{tk} to the sensor through a secure channel. Moreover, for simplicity, it is assumed that the sensor has the knowledge of T.

2.4 Remote State Estimation

Define $\widehat{x}_t$ and P_t as the estimate of $\widehat{x}_t^s$ and its corresponding error covariance, respectively. Then, based on Ren et al. (2013), there holds that

$$\widehat{x}_t = \begin{cases} \widehat{x}_t^s, & \theta_t^T \beta_t = 0, \\ A\widehat{x}_{t-1}, & \theta_t^T \beta_t = 1, \end{cases} \tag{4}$$

and

$$P_t = \begin{cases} \overline{P}, & \theta_t^T \beta_t = 0, \\ h(P_{t-1}), & \theta_t^T \beta_t = 1. \end{cases} \tag{5}$$

2.5 Problem Statement

The objective of the sensor is to avoid the attack and simultaneously to find the high-quality channel, which can be achieved by minimizing

$$J^{\rm Es} = \frac{1}{N} \sum_{t=1}^{N} P_t, \tag{6}$$

and maximizing

$$J^{\rm Th} = \sum_{t=1}^{N} R_t^{\rm Th}. \tag{7}$$

Therefore, the focus is placed on the problem of how to obtain the optimal sequence of channels to minimize $J^{\rm Es}$ and maximize $J^{\rm Th}$ simultaneously. The key challenge lies in that the sensor has no ability to know the best channel and the action of the jammer beforehand. Thus, it is desired to explore the policy to select the channel at each time, such that the sensor can minimize $J^{\rm Es}$ and maximize $J^{\rm Th}$ simultaneously. Even though the SM method can be used to select the channel at each time Kuleshov et al. (2014), the sensor will encounter the jammer. To handle this, we propose the POL algorithm described in next section.

3. PERIOD ONLINE LEARNING ALGORITHM

In this section, the POL model is first established. Then we present the POL algorithm, and prove analytically its effectiveness by comparing with the SM method.

3.1 Period Online Learning Model

The POL model is defined by $\mathcal{M} = \langle \mathcal{P}, \mathcal{C}, \mathcal{R}, r, p \rangle$, where $\mathcal{P} = \{p_i, i = 1, \cdots, T\}$ is the set of players; $\mathcal{R}$ is the set of rewards; r: $\mathcal{P} \times \mathcal{C} \to \mathcal{R}$ is the reward function of players; p: $\mathcal{P} \times \mathcal{R} \to \mathcal{C}$ is the set of probability distributions over the channel set $\mathcal{C}$. The details of $\mathcal{M}$ are presented in the following.

Player: The length of the attack period is mapped to the number of players, and each player $p_i \in \mathcal{P}$ only works at time n_i. For each time t corresponding to the i-th time in period n, selecting which channel to sense and access depends on the information possessed by player i.

Reward: The reward is concerned with the estimation error covariance and the feedback reward. Combining the recursion of P_t in (5) and the characteristic of period of the jammer, for any t, there holds that $P_t \in \mathcal{R}^{\rm Es}$, where $\mathcal{R}^{\rm Es} = \{\overline{P}, h(\overline{P}), \cdots, h^{T_{\rm a}}(\overline{P})\}$. Moreover, from (3), the set of feedback rewards can be expressed as $\mathcal{R}^{\rm Th} = \{r_{tk}, t \geq 1, k \in [1, M]\}$. Therefore, r is a mapping from $\mathcal{P} \times \mathcal{C}$ to $\mathcal{R}^{\rm Es} \times \mathcal{R}^{\rm Th}$. When $\theta_t^T \beta_t = 0$, for any t and $k \in [1, M]$, there holds that $P_t = \overline{P}$ and $R_t^{\rm Th} = r_{tk}$. But when $\theta_t^T \beta_k = 1$, for any t and $k \in [1, M]$, it follows that $P_t = h(P_{t-1})$ and $R_t^{\rm Th} = 0$. The corresponding flow diagram of the POL model can be seen in Fig. 2.

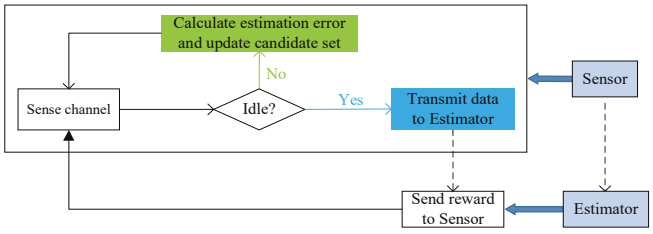

Fig. 2. Flow diagram of POL model at the sides of the sensor and the estimator.

3.2 Period Online Learning Algorithm

For avoiding the attack, there is a need to find the channel attacked first. Define $\mathcal{S}_{ni}^{\mathrm{Ja}}$ as the set of channels attacked, each of which has been detected for player p_i in period n. Once the channel is identified as the one attacked, it will not be considered as an optimal option for the sensor. As a result, with $c_k \in \mathcal{S}_{ni}^{\mathrm{Ja}}$, there holds that

$$c_k \notin \overline{\mathcal{S}}_{ni}^{\mathrm{Ja}}, \tag{8}$$

where $\overline{\mathcal{S}}_{ni}^{\mathrm{Ja}} = \mathcal{C} - \mathcal{S}_{ni}^{\mathrm{Ja}}$ denotes the set of candidate channels for player i. It is noted that $\overline{\mathcal{S}}_{ni}^{\mathrm{Ja}}$ varies over a duration of time until all the sets of channels attacked have been detected, since the sensor just selects one channel to sense at each time. Moreover, for any time j_1 in period n, player j_1 has no knowledge of $\overline{\mathcal{S}}_{nj_2}^{\mathrm{Ja}}$ with $j_1 \neq j_2$. After player i acquires $\overline{\mathcal{S}}_{ni}^{\mathrm{Ja}}$, the main problem encountered by the sensor is how to select a channel belonging to $\overline{\mathcal{S}}_{ni}^{\mathrm{Ja}}$ to maximize the feedback reward.

As mentioned above, the feedback reward r_{tk} can be received by the sensor, when the jammer has no attack on channel k at time t. In fact, for the j-th time in period n, the feedback reward r_{n_jk} might be informed to player j_1 with $j_1 = j$, or to player j_2 for each $j_2 \in [1, T]$. To obtain the optimal channel as early as possible, we focus on the latter case, i.e., the feedback reward r_{n_jk} is announced to all the players.

Based on the feedback reward for each channel, the difference between the SM and the POL algorithms is the probability of selecting the channel. The main characteristics of the SM method are that the possibility of utilizing each channel is proportional to its average reward and that the level of exploration and exploitation can be adjusted by the temperature parameter Iwata (2017). Define the average of the estimate of the throughput for channel c_k at time t as

$$Q_{tk} = \frac{Q_{(t-1)k} N_{tk} + R_t^{\mathrm{Th}}}{N_{tk} + 1}, \tag{9}$$

where N_{tk} denotes the number of channel c_k selected up to time t. Then, at time t the probability of selecting channel c_k can be described as

$$p_{tk} = \frac{e^{\frac{Q_{tk}}{\tau_t}}}{\sum_{k=1}^{M} e^{\frac{Q_{tk}}{\tau_t}}}, \tag{10}$$

where $\tau_t = \gamma \tau_{t-1}$ is the temperature parameter.

However, for the POL algorithm, combining (8), function $\widetilde{p}_t$ assigning the probability over channel set $\mathcal{C} - \mathcal{S}_{ni}^{\mathrm{Ja}}$ can be expressed as

$$\widetilde{p}_{tk} = \frac{p_{tk}}{\sum_{k \notin \mathcal{S}_{ni}^{\mathrm{Ja}}} p_{tk}}. \tag{11}$$

For the sake of convenience, the POL algorithm is summarized as follows.

1) Input A, C, Σ_ω, and Σ_ν; M, T, q, and γ.
2) Initialize $Q_{1k} = 0$, $N_{1k} = 0$, and $\overline{\mathcal{S}}_{1k}^{\mathrm{Ja}} = \emptyset$, for any $k \in [1, M]$; τ_1 and $t = 1$.
3) Select channel c_k with $\widetilde{p}_{tk}$ based on (11), until $n = q$ and $j = T$.
4) For any n and j, when channel c_k is attacked, $P_t^{\mathrm{Es}} = AP_{t-1}A^T + \Sigma_\omega$ and $\overline{\mathcal{S}}_{nj}^{\mathrm{Ja}}$. If channel c_k is not attacked, there holds that $R_t^{\mathrm{Th}} = r_{tk}$, $P_t = \overline{P}$, $Q_{tk} = \frac{Q_{(t-1)k} N_{tk} + R_t^{\mathrm{Th}}}{N_{tk}+1}$ and $N_{tk} = N_{tk} + 1$.
5) Calculate $J^{\mathrm{Th}} = J^{\mathrm{Th}} + R_t^{\mathrm{Th}}$, $J^{\mathrm{Es}} = \frac{1}{t}(J^{\mathrm{Es}} + R_t^{\mathrm{Es}})$, $t = t + 1$, and $\tau_t = \gamma \tau_{t-1}$.

3.3 Analysis on the POL Algorithm

It is well known that there exist the polylogarithmic regret bounds for the SM method Kuleshov et al. (2014). Since in the POL algorithm each player j selects channel c_k with $\widetilde{p}_{tk}$, the POL algorithm also has the polylogarithmic regret bounds. Moreover, the effectiveness of the POL algorithm can be further verified using the following theorem.

Theorem 1. Assume that $\mathcal{S}_{nj}$ has v element, then there holds that

$$\mathbb{E}_{\mathrm{POL}}[P_{(n+1)_1}] \leq \mathbb{E}_{\mathrm{SM}}[P_{(n+1)_1}], \tag{12}$$

and

$$\mathbb{E}_{\mathrm{POL}}[R_{(n+1)_1}^{\mathrm{Th}}] \geq \mathbb{E}_{\mathrm{SM}}[R_{(n+1)_1}^{\mathrm{Th}}]. \tag{13}$$

Proof. Since the difference in both algorithms is the probability of selecting each channel associated with (11), for any $k \in [1, M]$, there holds that $\mathbb{E}_{\mathrm{POL}}[p_{(n+1)_1 k}] = \mathbb{E}_{\mathrm{SM}}[p_{(n+1)_1 k}]$. Without loss of generality, we assign $\mathbb{E}[p_{(n+1)_1 k}]$ as p_k.

For the SM method, the expectation of $P_{(n+1)_1}$ is given by

$$\mathbb{E}_{\mathrm{SM}}[P_{(n+1)_1}] = \frac{|\theta_1|}{M} h(\overline{P}) + \frac{M - |\beta_1|}{M} \overline{P}, \tag{14}$$

where $|\beta_1|$ is the number of channels attacked at the first time in any period n. Similarly, the expectation of the reward is

$$\mathbb{E}_{\mathrm{SM}}[R_{(n+1)_1}^{\mathrm{Th}}] = \frac{M - |\beta_1|}{M} \sum_{\beta_{1k}=0} \frac{p_k}{\sum_{\beta_{1k}=0} p_k} \mu_k. \tag{15}$$

On the other hand, from (11), for any $k \in \overline{\mathcal{S}}_{n1}^{\mathrm{Ja}}$, there holds that $\widetilde{p}_k = (1+\delta) p_k$ with $\delta = \frac{\sum_{k \in \mathcal{S}_{nj}} p_k}{\sum_{k \in \overline{\mathcal{S}}_{nj}} p_k}$. Then for the POL algorithm, we can obtain

$$\begin{aligned}
\mathbb{E}_{\mathrm{POL}}[R_{(n+1)_1}^{\mathrm{Th}}] &= \frac{M - |\beta_1|}{M - v} \sum_{\beta_{1k}=0} \frac{\tilde{p}_k}{\sum_{\beta_{1k}=0} \tilde{p}_k} \mu_k \\
&= \frac{M - |\beta_1|}{M - v} \sum_{\beta_{1k}=0} \frac{(1+\delta) p_k}{\sum_{\beta_{1k}=0} (1+\delta) p_k} \mu_k \\
&= \frac{M}{M - v} \mathbb{E}_{\mathrm{SM}}[R_{(n+1)_1}^{\mathrm{Th}}].
\end{aligned} \tag{16}$$

Moreover, the expectation of $P_{(n+1)_1}$ is given by

$$\mathbb{E}_{\mathrm{POL}}[P_{(n+1)_1}] = \frac{|\beta_1| - v}{M - v} h(\overline{P}) + \frac{M - |\beta_1|}{M - v} \overline{P}. \tag{17}$$

From (13), we have

$$\begin{aligned} &\mathbb{E}_{\mathrm{POL}}[P_{(n+1)_1}] - \mathbb{E}_{\mathrm{SM}}[P_{(n+1)_1}] \\ &= -\frac{(M - |\beta_1|)v}{M(M - v)} (h(\overline{P}) - \overline{P}). \end{aligned} \tag{18}$$

Furthermore, due to $h(\overline{P}) > \overline{P}$ based on Ren et al. (2013), inequality (12) holds.

According to Theorem 1, when measuring the performance of both algorithms within one time slot, the performance of the POL algorithm is better than that of the SM method.

4. NUMERICAL RESULTS

In this section, the performance of the POL algorithm is evaluated and compared through numerical simulations. The parameters for (1) and (2) are designed as $A = 1.5$, $C = 1$, $\Sigma_\omega = 1$, and $\Sigma_\nu = 0.5$ with $\overline{P} = 0.3954$. The number of channels available is $M = 4$. For the POL algorithm, we assign $\tau_1 = 500$ and $\gamma = 0.95$. For facilitating analysis, we assume that r_{tk} follows Gaussian distribution $\mathcal{N}(k, 1)$ corresponding channel k.

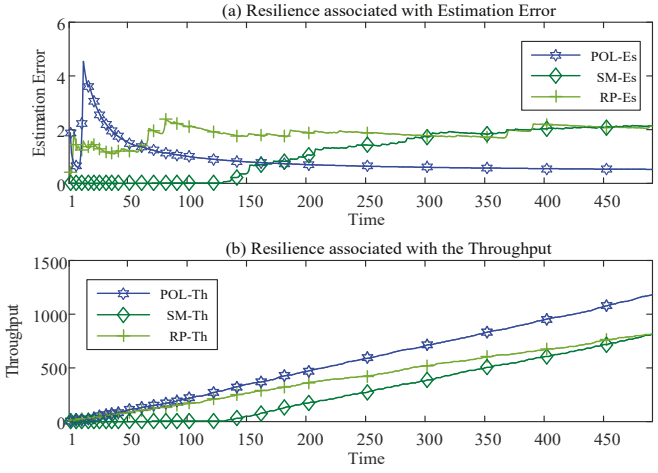

Fig. 3. The performance of the POL algorithm.

In Fig. 3, the performance of the POL algorithm is evaluated, where $\theta_1 = (1, 0, 1, 0)^T$, $\theta_2 = (1, 0, 0, 1)^T$, $\theta_3 = (0, 1, 0, 1)^T$, $\theta_4 = (1, 0, 1, 1)^T$, $\theta_5 = (0, 1, 0, 0)^T$, and $\theta_6 = \theta_7 = (0, 0, 0, 0)^T$ with $T_{\mathrm{a}} = 5$ and $T = 7$. Note that the sensor knows T, but has no knowledge of θ_t and r_{tk}, $\forall t \geq 1$ and $k \in [1, 4]$. It is observed from Fig. 3(a) that the performance of the estimator error under the POL algorithm with $t \geq 180$ is better than the SM algorithm and the random policy (RP), where the RP algorithm implies that the sequence of channels is selected randomly without taking into account the state of each channel and the existence of the jammer. This indicates that the sensor avoids the attack using the POL algorithm, but is caught in the jammer for the other two algorithms. From Fig. 3(b), with $t \geq 50$ it is seen that the throughput under the POL algorithm is better than the SM and the RP algorithms, which indicates that the sensor can find the channel with a better state despite of the presence of the jammer. Moreover, it is observed from Fig. 3 that during the beginning time the performance of the POL algorithm is worse than the other two algorithms. The reason is that during this time period the channels attacked exist in the set of the candidate channels, and each channel is selected with a corresponding probability.

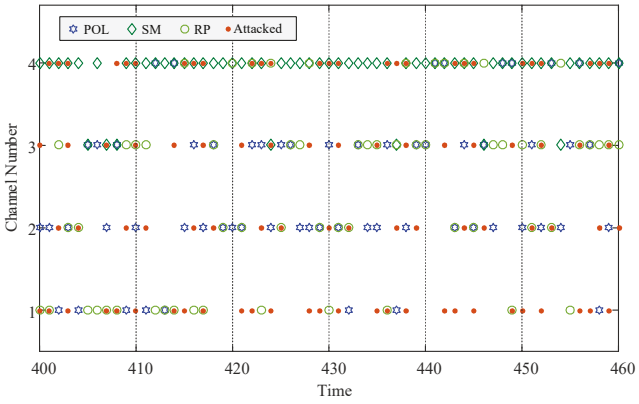

Fig. 4. The sequence of channels selected within [400, 460].

Furthermore, the channel selected at each time corresponding to each algorithm is portrayed in Fig. 4 with $t \in [400, 460]$. It can be seen from Fig. 4 that the SM method has a preference for channel c_4, whereas the POL algorithm prefers channels c_2 and c_3. Although the SM method can find the best channel, the sensor is still attacked. This leads to a lower performance compared with the POL algorithm depicted in Fig. 3. The reason for such result is that the design of the POL algorithm takes jointly the existence of the jammer and the state of the channel into account. For example, when $t = 430$, channel c_4 attacked is selected using the SM method, whereas channel c_3 that is the best channel among the set of channels uninterrupted is chosen using the POL algorithm. Moreover, it is noted that the denominator in (10) will become considerably large with time going on in the practical simulations, which can result in the probability of selecting the channel close to zero. In order to deal with this, a threshold value δ is designed in the POL algorithm and the SM method. Specially, for each $k \in [1, 4]$, probability p_{tk} remains unchanged with $t > t_\delta$, where t_δ denotes that at time t_δ there holds that $p_{t_\delta k} \leq \delta$. Besides, Fig. 6 is given to show the effectiveness of POL algorithm for different kinds of attack modes that are presented in Fig. 5.

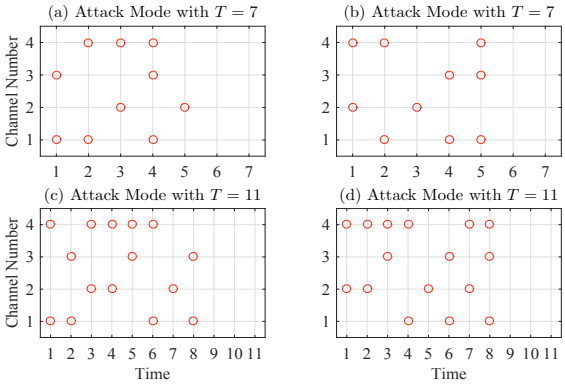

Fig. 5. Different attack modes.

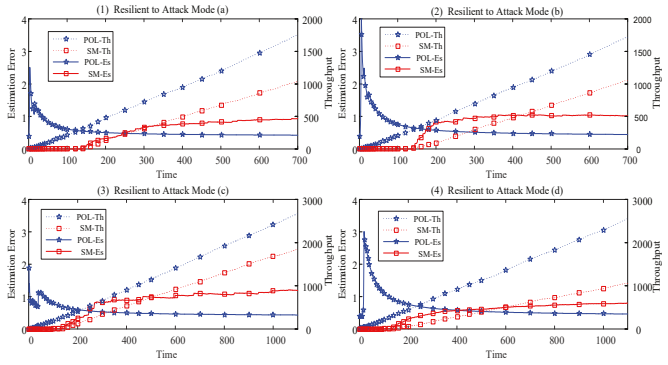

Fig. 6. The performance of the POL algorithm as a function of attack mode.

5. CONCLUSION

In this contribution, we have formulated a novel POL model to help the sensor select the sequence of channels dynamically with the time-variant state of channel. The problem of selecting the sequence of channels to jointly alleviate the attack effect and explore the channel with the best quality has been investigated. To handle this problem, the POL algorithm based on the reinforcement learning has been developed, and the effectiveness of this algorithm has been proved analytically. In addition, the theoretical results have been further verified in numerical simulations, and the effect of the jammer action on the performance of the POL algorithm has been studied.

REFERENCES

L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry. Foundations of control and estimation over lossy networks. *Proc. IEEE*, volume 95, pages 163–187, 2007.

Y. Liu, P. Ning, and M. K. Reiter. False data injection attacks against state estimation in electric power grids. in *Proc. ACM Conf. Computer Commun. Secur.*, Chicago, IL, USA, pages 21–32, 2007.

W. Xu, D. W. C. Ho, J. Zhong, and B. Chen. Event/Self-triggered control for leader-following consensus over unreliable network with DoS attacks. *IEEE Trans. Neural Netw. Learn. Syst.*, DOI: 10.1109/TNNLS.2018.2890119, 2019.

E. Iasiello. Cyber attack: A dull tool to shape foreign policy. in *Proc. Int. Conf. Cyber Confl.*, Tallinn, Estonia, pages 1–18, 2013.

F. Pasqualetti, F. Dörfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Trans. Autom. Control*, volume 58, pages 2715–2729, 2013.

H. Shisheh Foroush and S. Martínez. On event-triggered control of linear systems under periodic denial-of-service jamming attacks. in *Proc. IEEE Conf. Decision Control*, Maui, HI, USA, pages. 2551–2556, 2012.

S. Amin, A. A. Cárdenas, and S. S. Sastry. Safe and secure networked control systems under denial-of-service attacks in *Proc. Lect. Notes Comput. Sci.*, San Francisco, CA, USA, pages 31–45, 2009.

Y. Li, D. E. Quevedo, S. Dey, and L. Shi. SINR-based DoS attack on remote state estimation: A game-theoretic approach. *IEEE Trans. Control Netw. Syst.*, volume 4, pages 632–642, 2017.

K. Ding, Y. Li, D. E. Quevedo, S. Dey, and L. Shi. A multi-channel transmission schedule for remote state estimation under DoS attacks. *Automatica*, volume 78, pages. 194–201, 2017.

Y. L. Che, R. Zhang, and Y. Gong. Opportunistic spectrum access for cognitive radio in the presence of reactive primary users. in *Proc. IEEE Int. Conf. Commun.*, Kyoto, Japan, 2011.

X. Cao, P. Cheng, J. Chen, S. Ge, Y. Cheng, and Y. Sun. Cognitive radio based state estimation in cyber-physical systems. *IEEE J. Sel. Areas Commun.*, volume 32, pages 489–502, 2014.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, volume 47, pages 235–256, 2002.

G. Gu, X. R. Cao, and H. Badr. Generalized LQR control and Kalman filtering with relations to computations of inner-outer and spectral factorizations. *IEEE Trans. Autom. Control*, volume 51, pages 595–605, 2006.

M. D. Felice, K. R. Chowdhury, C. Wu, L. Bononi, and W. Meleis. Learning-based spectrum selection in cognitive radio Ad Hoc networks. *Proc. Int. Conf. Wired/Wireless Internet Commun.*, Lulea, Sweden, pages 133–145, 2010.

S. Hu, D. Yue, X. Xie, X. Chen, and X. Yin. Resilient event-triggered controller synthesis of networked control systems under periodic DoS jamming attacks. *IEEE Trans. Cybern.*, DOI: 10.1109/TCYB.2017.2787740, 2018.

B. Roy and S. Bag. Two channel hopping schemes for jamming resistant wireless communication. in *Proc. Int. Conf. Wirel. Mob. Comput. Netw. Commun.*, Lyon, France, pages 659–666, 2013.

W. Xu, W. Trappe, Y. Zhang, and T. Wood. The feasibility of launching and detecting jamming attacks in wireless networks. In *Proc. Int. Symp. Mobile Ad Hoc Networking Comput.*, Urbana-Champaign, IL, USA, pages 46–57, 2005.

Z. Ren, P. Cheng, J. Chen, L. Shi, and Y. Sun. Optimal periodic sensor schedule for steady-state estimation under average transmission energy constraint. *IEEE Trans. Autom. Control*, volume 58, pages 3265–3271, 2013.

X. Cao, X. Zhou, L. Liu, and Y. Cheng. Energy-efficient spectrum sensing for cognitive radio enabled remote state estimation over wireless channels. *IEEE Trans. Wireless Commun.*, volume 14, pages 2058–2071, 2015.

L. Lyu, C. Chen, C. Hua, and X. Guan. Cognitive radio enabled reliable transmission for optimal remote state estimation in multi-sensor industrial cyber-physical systems. in *Proc. IEEE/CIC Int. Conf. Commun. China*, Shenzhen, China, 2015.

V. Kuleshov and D. Precup. Algorithms for the multi-armed bandit problem. CoRR, volume abs/1402.6028, https://arxiv.org/abs/1402.6028, 2014.

K. Iwata. Extending the peak bandwidth of parameters for softmax selection in reinforcement learning. IEEE Trans. Neural Netw. Learn. Syst., volume 28, pages 1865–1877, 2017.