IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received April 12, 2020, accepted May 15, 2020, date of publication May 22, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2996804

A Countermeasure Against Random Pulse Jamming in Time Domain Based on Reinforcement Learning

QUAN ZHOU[1], YONGGUI LI[2], AND YINGTAO NIU[2]
[1]College of Communications Engineering, Army Engineering University of PLA, Nanjing 210001, China
[2]The Sixty-Third Research Institute, National University of Defense Technology, Nanjing 210001, China

Corresponding author: Yonggui Li (legend64@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant U19B2014, and in part by the Foundation Strengthening Program Area Fund of China under Grant 2019-JCJQ-JJ-212.

ABSTRACT Pulse jamming is one of the common malicious jamming patterns that can significantly reduce the of wireless communication's reliability. This paper investigates the problem of anti-jamming communication in a random pulse jamming environment. In order to obtain the countermeasure in time domain, the Markov decision process (MDP) is employed to model and analyze the above problem, and a time-domain anti-pulse jamming algorithm (TDAA) based on reinforcement learning is proposed. The proposed algorithm learns from the dynamic interaction with the jamming environment to gradually approximate the optimal time-domain strategy. The optimal strategy enables the transmitter to switch between two states, i.e. "active" and "silent", to avoid random pulse jamming. In addition, a state estimation and adjustment method for the random pulse jamming environment is introduced to improve the robustness of the proposed TDAA. Simulation results show that, compared with continuous transmission, the proposed TDAA can effectively reduce the jamming collision ratio and significantly improve the normalized throughput. And compared with transmitting terminal Q-learning algorithm (TTQA), the proposed TDAA has higher time utilization ratio and normalized throughput.

INDEX TERMS Anti-jamming, Markov decision process, Q-learning, reinforcement learning, random pulse jamming.

I. INTRODUCTION

Pulse jamming is a kind of jamming with short duration and large instantaneous power. On the one hand, it can be produced inadvertently by various electrical equipment when they are working. For example, vehicle ignition and aircraft navigation can cause pulse jamming. In addition, nonlinear power devices (e.g., rectifiers, diodes, transformers) in electronic equipment will produce pulse jamming when they work. On the other hand, the jammer can produce malicious pulse jamming by transmitting the jamming signal in a short time and staying shut down in the rest of the time [1]. Both malicious and unintentional pulse jamming can significantly increase the system's bit error rate (BER) or reduce network throughput. For example, the authors in [2] modeled pulse jamming as Bernoulli-Gauss model. Then, by developing a closed form expression for the probability of error for QAM system under pulse jamming, the authors quantitatively analyzed the impact of the pulse jamming on the performance of the QAM system. In [3], the error probability for ASK, PSK, and FSK systems in the presence of a pulse jamming is analyzed by modeling the pulse jamming as a generalized stationary Poisson process. The authors in [4] indicated that the performance in throughput and delay of the ARQ schemes could be affected by the pulse jamming. According to the rule of pulse signal in time domain, pulse jamming can be divided into periodic pulse jamming and random pulse jamming. In [5], the authors proposed a short-period pulse jamming whose duration is far less than the packet length, which can significantly reduce the packet delivery ratio (PDR) of the spread spectrum communication system without significantly changing the detection value of the received signal strength. Authors in [6] demonstrated that periodic pulse jamming can seriously reduce the performance of the IEEE 802.11 media access control (MAC) layer by destroying data frames or affecting the backoff operation of CSMA/CA protocol. With the intelligence level of jammer improves, the well-designed random pulse jamming is more

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

agile and efficient. For example, in [7], a kind of random pulse jamming aimed at the routing operation vulnerability of Ad Hoc network was proposed. The jammer keeps silent when the nodes exchange the handshake information, while at other times, it transmits random pulse jamming to block the channel. Nodes will treat the blocked channel as an available channel, resulting in a severe reduction in network throughput. In [8], the random pulse jamming in satellite communication system occurs only once in each repetition period, while the time of occurrence in the same period is random. Hence, the random pulse jamming not only satisfies a certain duty cycle to achieve the jamming effect, but also makes it difficult to be detected and eliminated.

The existing anti-pulse jamming methods mainly include the "hidden" method based on spread spectrum technology and the "shearing" method based on filtering technology. The authors in [9] proposed a kind of self-encoding spread spectrum technology, which can effectively reduce the symbol error rate (SER) of system under periodic pulse jamming. [10] proposed two self-adaptive filtering algorithms, namely, the incremental and diffusion affine projection sign algorithms, which can reduce the system bit error rate under pulse jamming environment. However, the pulse jamming considered in the above methods has limited jamming bandwidth, low jamming power or periodic pulse signals, while well-designed pulse jamming may have high peak power, the broadband jamming signal which can completely cover communication bandwidth, and random pulses in time domain. More effective anti-jamming methods are needed to deal with the well-designed pulse jamming. In recent years, the development of machine learning provides a new research idea for communication anti-jamming. Reinforcement learning (RL) is one of the methods of machine learning that aims to enable agents to take appropriate actions to get high rewards [11]. RL problems can usually be modeled as Markov decision process (MDP) [12]. Q-learning [13] is a classical model-free reinforcement learning algorithm, which can obtain the optimal strategy without modeling the environment. To solve the communication anti-jamming problem, Q-learning enables legitimate users to learn from the feedback (e.g. throughput or PDR) caused by their own actions (e.g. channel, power or coding mode) in the process of dynamic interaction with the jamming environment, and thus obtain the optimal anti-jamming strategy with the maximum gain or minimum loss. Although Q-learning has been widely used in solving anti-jamming problems [14]–[19], most existing works adopt Q-learning to obtain the frequency domain anti-jamming strategy like optimal channel switching strategy, while the related works on obtaining time-domain anti-jamming strategy by Q-learning is rare.

To solve the limitation of the existing anti-pulse jamming methods, this paper proposes a time-domain anti-pulse jamming algorithm (TDAA) based on reinforcement learning to obtain the optimal time-domain anti-jamming strategy. The proposed algorithm can make the transmitter switches between two states, i.e., "active" or "silent", to avoid random pulse jamming and improve the transmission reliability. The main contributions of this paper are as follows:

- Based on reinforcement learning, a low complexity time-domain anti-pulse jamming algorithm (TDAA) is proposed to to obtain the optimal time-domain anti-jamming strategy.
- A state estimation and adjustment method for pulse jamming environment is designed to improve the robustness of the proposed TDAA.

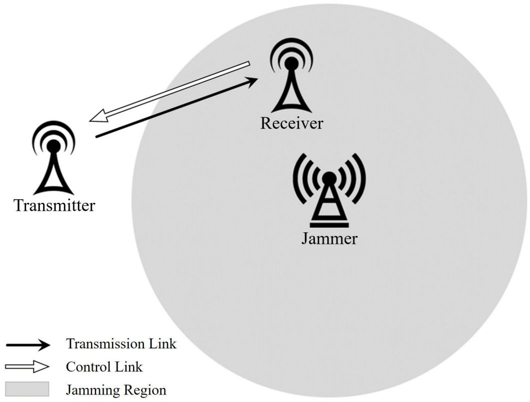


FIGURE 1. System model.

The rest of this paper is organized as follows: Section II presents the system model and problem formulation. In Section III, we introduce the time domain anti-pulse jamming algorithm (TDAA) and the state estimation and adjustment method, and analyze the complexity and convergence of TDAA. The simulation results are discussed in Section IV. And concluding remarks are given in Section V.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

As shown in Fig. 1, the wireless communication system consists of two nodes: a legitimate transmitter and its intended receiver. The system is affected by a malicious jammer, and the jamming region can only cover the receiver. We assume that the transmission time can be divided into discrete timeslots with a length of T_s. Moreover, we consider the timeslot is the minimum time unit for transmission. In other words, if the transmitter starts the transmission, it can last at least one timeslot. In addition, we assume that one sub-frame will be transmitted in one timeslot when the transmitter is active. Each packet is composed of l sub-frames and contains CRC check bits, and the amount of information in each sub-frame is i.

We consider a pulse jammer with prior information about communication frequency and timeslot synchronization frame. The bandwidth of the pulse jamming signal can completely cover the transmission bandwidth, and the duration of a single pulse is equal to the length of the timeslot. To effectively affect the transmission, the pulse jamming should meet a certain duty cycle [20]. In other words, it should

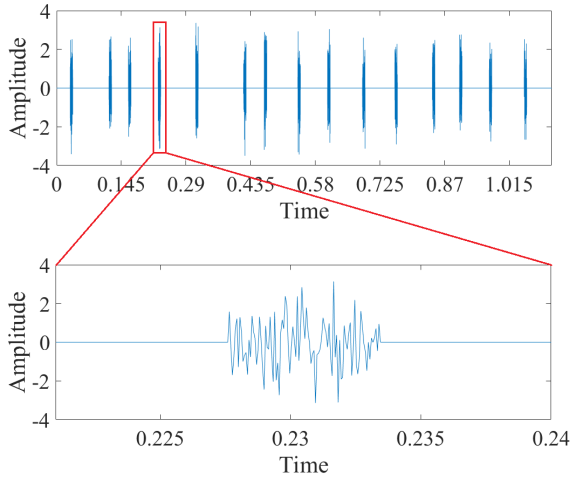


FIGURE 2. Random pulse jamming signal.

occur at least once in a certain number of timeslots. Meanwhile, to reduce the detection probability, each pulse can be launched randomly within a certain time range on the premise of ensuring the duty cycle. Similar to [8], we assume that the jammer defines every N timeslots as one jamming period. Within the same jamming period, a timeslot is selected randomly as the target for pulse jamming according to a specific probability distribution. The above probability distribution is defined as jamming selection distribution (jsd). If the probability density function (pdf) of jsd is $f_J(t)$, then the probability of the jammer to choose the nth timeslot within the kth jamming period, denoted by $p_k(n)$, can be expressed as follows:

$$p_k(n) = \int_{kT_s+(n-1)T_s}^{kT_s+nT_s} f_J(t)dt, \tag{1}$$

where t represents the transmission time. As shown in Fig. 2, each pulse is modeled as shot noise [21] having power spectral density σ^2, and its pulse waveform is equivalent to Gaussian noise in a short time period. Pulse jamming can cause a large number of error bits in the jammed timeslot, and even cause packet loss due to the failure of CRC check. In addition, we assume that the channel noise is not enough to affect the transmission compared with the pulse jamming significantly.

B. PROBLEM FORMULATION

To deal with the uncertainty of jamming signals, the anti-jamming problem can be modeled as a Markov decision process (MDP), which is widely adopted in related works [14]–[19]. The MDP can be defined by a tuple $< \mathcal{S}, \mathcal{A}, P(s'|s,a), r \;\; >$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s'|s,a)$ is the state transition probability and r is the immediate reward of the system.

The state space of the system is defined as follows:

$$\mathcal{S} \triangleq \{(n,j) : n \in \{1, \ldots, N\}; j \in \{0,1\}\}. \tag{2}$$

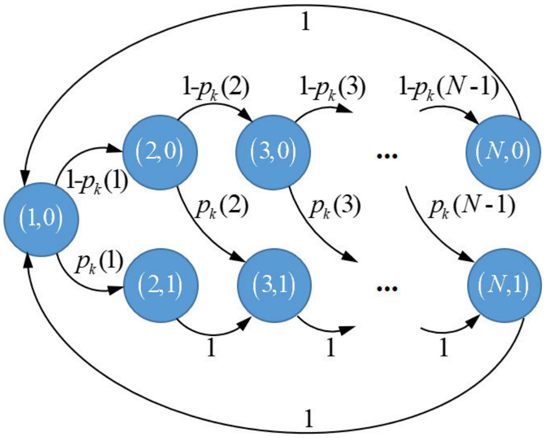


FIGURE 3. State transitions of system.

The system state is defined as a composite variable $s = (n,j) \in \mathcal{S}$, where n represents the sequence number of the timeslot in the jamming period. j is a flag reflecting whether the pulse jamming has occurred in the current jamming period. Specifically, in the first timeslot of each jamming period, j is initialized to 0. From the second timeslot to the last one in the jamming period, $j = 1$ when pulse jamming is sensed in the previous timeslot and $j = 0$ otherwise. Once $j = 1$, the value of j stays 1 for the rest of the jamming period until it is reinitialized to 0 in the next jamming period. According to the above definition, in the same jamming period, $j = 1$ after the pulse jamming is sensed, and $j = 0$ in previous timeslots. In addition, the system can interact with the environment to estimate the size N of the jamming period, which can be used to determine the current system state (estimation method detailed in section III).

The transmitter can perform two actions, i.e., keep active or keep silent. Then, the action space of the transmitter can be defined as follows:

$$\mathcal{A} \triangleq \{a : a \in \{0,1\}\}, \tag{3}$$

where a represents the action of the transmitter, i.e., $a = 0$ when the transmitter keep silent and $a = 1$ otherwise.

Fig. 3 shows the state transitions of the system, where $p_k(n)$ is the timeslot selection probability (tsp) in Eq. (1). If the current state is expressed as $s = (n,j)$ and the next state is expressed as $s' = (n',j')$, then the next sequence number n' can be expressed as follows:

$$n' = \begin{cases} n+1, & \text{if } n < N, \\ 1, & \text{if } n = N. \end{cases} \tag{4}$$

Obviously, the value of n' is only relevant to n. Similarly, j' can be expressed as follows:

$$j' = \begin{cases} 0, & \text{if } n = N, \\ 0, & \text{if } j = 0, g = 0 \,\&\, n < N, \\ 1, & \text{if } j = 0, g = 1 \,\&\, n < N, \\ 1, & \text{if } j = 1 \quad n < N, \end{cases} \tag{5}$$

where g represents the result of jamming sensing, i.e., $g = 1$ when the system senses the pulse jamming and $g = 0$ otherwise. Within the same jamming period, the tsp of each timeslot is independent before the pulse jamming occurs (i.e. $j = 0$), then j' only depends on the tsp of current state; If pulse jamming has occurred before the current timeslot (i.e. $j = 1$), then $j' = 1$ unless n = N, then $j' = 0$. In brief, the value of j' is only relevant to current state s. Hence, the next state s' is only related to the current state s, and independent of the earlier states, i.e., the system state is a Markov chain. As shown in Fig. 3, the probability that the system state transmits from s to s' after taking action a can be expressed as follows:

$$P(s'|s, a) = \begin{cases} 1, & \text{if } n = N, \\ 1, & \text{if } j = 1 \,\&\, n < N, \\ p_k(n), & \text{if } j = 0, j' = 1 \,\&\, n < N, \\ 1 - p_k(n), & \text{if } j = 0, j' = 0 \,\&\, n < N. \end{cases} \tag{6}$$

The immediate reward of the system after the transmitter takes an action a at state s can be defined as follows:

$$r = \begin{cases} E, & \text{if } j = 1 \,\&\, a = 1, \\ E, & \text{if } j = 0, g = 0 \,\&\, a = 1, \\ -L, & \text{if } j = 0, g = 1 \,\&\, a = 1, \\ 0, & \text{if } a = 0. \end{cases} \tag{7}$$

In the above, if $j = 1$ & $a = 1$, i.e., the transmitter takes action of "keep active" after the occurrence of pulse jamming within the jamming period, the system is bound to transmit successfully, then the system gain is E; If $j = 0,\ g = 0$ & $a = 1$, i.e., the transmitter takes the action "keep active" and the pulse jamming has not been sensed within the jamming period, the system is bound to transmit successfully, then the system gain is E; If $j = 0,\ g = 1$ & $a = 1$, i.e., the transmitter takes the action "keep active" when the pulse jamming is sensed in the current timeslot, the transmission will be jammed, then system loss is L; If $a = 0$, i.e., the transmitter takes action of "keep silent", then system gain is obviously 0. As previously assumed, each timeslot can transmit 1 sub-frame, and each packet contains l sub-frames. Thus, we set $E = 1$, which represents the number of sub-frames that transmitted successfully in a single timeslot. In addition, it takes l timeslots for each packet to be transmitted. If the number of pulses in the packet is p, then the number of sub-frames transmitted in the p timeslots with pulse jamming is 0, while the number of sub-frames transmitted in other timeslots without pulse jamming is $l - p$. Hence, we set $L = (l - p)/p$, which represents the average number of lost sub-frames caused by each pulse.

According to the strategy $\pi : \mathcal{S} \mapsto \mathcal{A}$, the system can get the action a that should be taken under any state s. We aim to obtain the optimal strategy π^* that can maximize the average long-term reward. Then, the π^* can be obtained as follows:

$$\max_{\pi} \mathcal{R}(\pi) = \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{\tau=1}^{T} r_\tau\right], \tag{8}$$

where $\mathcal{R}(\pi)$ represents the average long-term reward of the strategy π. τ represents the number of steps, the maximum value of which is T. r_τ represents the immediate reward of step τ. $\mathbb{E}[\cdot]$ is the mathematical expectation operator. According to the definition of immediate reward, the average long-term reward $\mathcal{R}(\pi)$ can represent the average number of sub-frames that transmitted successfully per timeslot. Thus, the average throughput of the system can be expressed as $\mathcal{R}(\pi)\cdot i/T_s$, where T_s is the length of the timeslot and i is the amount of information in each sub-frame. Obviously, the average throughput is also maximized when the system achieves the maximum average long-term reward.

III. TIME-DOMAIN ANTI-PULSE JAMMING ALGORITHM

A. DETAILED DESCRIPTION OF THE ALGORITHM

To get the optimal anti-jamming strategy, a time-domain anti-pulse jamming algorithm (TDAA) based on Q-learning is proposed. Q-learning is the most well-known model-free reinforcement learning (RL) algorithm. And it is a progressive dynamic programming process that can find the optimal strategy step by step [22]. The basic idea of Q-learning is to create and update a Q-table that contains state-action pair values what are called Q-values [23]. In any given system state, the transmitter observes the current state and selects an action based on the current strategy. After performing the selected action, the receiver observes the immediate reward and updates the Q-value. In this way, the system can keep learning from its own action and finally converges to the optimal anti-jamming strategy. Similar to [5], Q-value is updated according to the following Q-function:

$$Q_{\tau+1}(s, a) = Q_\tau(s, a) + \alpha_\tau \left[r + \gamma \max_{a'} Q_\tau(s', a') - Q_\tau(s, a)\right], \tag{9}$$

where r is the immediate reward for taking action a in current state s. s' represents the next state after taking action a. $Q_\tau(s, a)$ is the current Q-value. $r + \gamma \max_{a'} Q_\tau(s', a')$ represent the predicted Q-value. Besides, $0 \le \alpha_\tau < 1$ is the rate factor determining the learning speed, and $0 \le \gamma < 1$ is the attenuation factor reflecting the importance of long-term reward. Eq. (9) can find the temporal difference between the current Q-value and predicted Q-value to calculate the updated Q-value [24].

The system timeslot structure is illustrated in Fig. 4, which consists of a transmission sub-slot, a sensing sub-slot and a learning sub-slot. Let T_0, T_1 and T_2 denote the transmission time, sensing time and learning time respectively. Each iteration is divided into three steps. Firstly, the transmitter takes an action according to the decision of the previous timeslot. Secondly, the receiver senses pulse jamming. Lastly, the receiver adopts the Q-learning algorithm to obtain the action for the next timeslot. the last two steps are performed even when the transmitter remains silent. The pseudo-code of TDAA is shown in Algorithm 1. The system performs the above three steps in each timeslot. Specifically, in the transmission sub-slot, the transmitter takes the action "keep

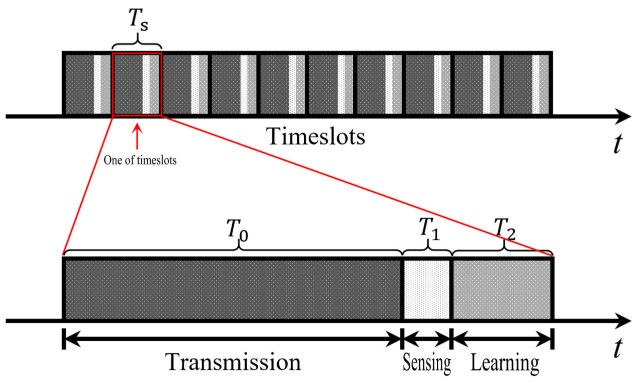


FIGURE 4. Illustration of the system timeslot structure.

active'' or ''keep silent'' according to the decision of the previous learning sub-slot (line 3); In the sensing sub-slot, the receiver senses whether there is pulse jamming in the environment by the broadband spectrum sensing technology [25] (line 4); In the learning sub-slot, The receiver first obtains the immediate reward r and the next state s' according to the first two sub-slots, then updates the Q-value according to Eq. (9) (line 5), and finally obtains the next action and feeds it back to the transmitter (line 6). To select action a', the ε-greedy algorithm [26] is often introduced. In particular, the system choose an action by $a \in \arg\max_a Q(s, a)$ with probability $1 - \varepsilon$, and randomly choose an action with probability ε. The larger ε is, the more obvious the decision-making system's ''exploratory'' intention will be, and the better anti-jamming strategy may be triggered [26]. The learning process will be terminated when all Q-values are convergent or the algorithm reach a certain number of iterations. When all Q-values converge, the system can obtain the optimal strategy indicating an action to be taken at each state such that converged $Q^*(s, a)$ is maximized for all states, which can be expressed as follows:

$$\pi^* = \arg\max{}_a Q^*(s, a). \tag{10}$$

B. JAMMING PERIOD ESTIMATION

In TDAA, the determination of system state is closely related to the size N of jamming period. However, it is hard to know the value of N in advance. Since the jammer only transmits one pulse signal per jamming period, the system can transmit data for a while to estimate the size N of jamming period. The estimated time T_e is defined to represent the transmission time from the beginning of the first pulse to the end of the Kth pulse. The estimated value N_e of N can be calculated as follows:

$$N_e = \text{round}\left[\frac{T_e}{KT_s}\right], \tag{11}$$

where round[·] is rounding operator. K is the number of pulses. T_e/T_s represents the total number of timeslots in the estimated time and satisfies the following equation:

$$\begin{aligned} T_e/T_s &= N \cdot K - (n_1 - 1) - (N - n_k) \\ &= N \cdot (K - 1) + n_k - n_1 + 1, \end{aligned} \tag{12}$$

Algorithm 1 Time-Domain Anti-Pulse Jamming Algorithm (TDAA)

```
1: Initialize: $\alpha, \gamma, \varepsilon \in [0, 1)$, $Q(s, a) \leftarrow 0$.
2: for $\tau = 0, 1, \ldots, T$ do
3:     The transmitter takes the action $a$ at state $s$ based on the previous timeslot.
4:     The receiver senses whether there is pulse jamming in the environment.
5:     The receiver obtains the immediate reward $r$ and the next state $s$' according to the previous two steps, then updates the Q-value according to Eq. (9).
6:     The receiver obtains the next action $a'$ according to the following rules:
         • The receiver chooses the next action $a' \in \arg\max_{a'} Q(s', a')$ with probability $1 - \varepsilon$;
         • The receiver randomly chooses the next action $a'$ with probability $\varepsilon$.
       Then, the receiver sends $a'$ back to the transmitter.
7:     Replace $s \leftarrow s'$ and $a \leftarrow a'$.
8: end for
9: Outputs: $\pi^* = \arg\max{}_a Q^*(s, a)$
```

where $n_1 \in \{1, 2, \ldots, N\}$ represents the sequence number of the first jammed timeslot in the corresponding jamming period, and $n_k \in \{1, 2, \ldots, N\}$ represents the sequence number of the Kth jammed timeslot in the corresponding jamming period. Eq. (12) shows that the timeslots in the estimated time is equal to the different between the timeslots in K complete jamming periods and the timeslots that are not included in the estimated time in the first and last jamming period. When the estimated value N_e converges to the exact value N, i.e. $N_e = N$, the following inequality can be obtained by expanding the rounding operator in Eq. (11):

$$N - 0.5 \le \frac{T_e}{KT_s} < N + 0.5. \tag{13}$$

By substituting Eq. (12) into Eq. (13), we have:

$$\begin{cases} K \ge 2(N - n_k + n_1 - 1), \\ K > 2(-N + n_k - n_1 + 1). \end{cases} \tag{14}$$

Since $4N - 4 \ge 2(N - n_k + n_1 - 1) \ge 2(-N + n_k - n_1 + 1)$, the value of K can be relaxed to $K \ge 4N - 4$, then Eq. (13) will still hold. By substituting $K \ge 4N - 4$ into Eq. (12), we have:

$$\begin{aligned} T_e/T_s &= N \cdot (K - 1) + n_k - n_1 + 1 \\ &\ge 4N^2 - 5N + n_k - n_1 + 1. \end{aligned} \tag{15}$$

Since $4N^2 - 4N \ge 4N^2 - 5N + n_k - n_1 + 1$, the value of T_e/T_s can be relaxed to $T_e/T_s \ge 4N^2 - 4N$, then Eq. (13) will still hold. Since Eq. (13) is a sufficient and necessary condition for $N_e = N$, $T_e/T_s \ge 4N^2 - 4N$ is a sufficient condition for $N_e = N$. In other words, when the total number of timeslots satisfies $T_e > (4N^2 - 4N) \cdot T_s$, the estimated value N_e must have converged to N. Due to the size N of jamming period is a positive integer of finite size, it can be estimated in a finite time.

C. STATE DETERMINATION AND ADJUSTMENT METHOD

Because the system has no prior information about the jamming period, it is difficult to accurately determine the sequence number of the current timeslot in the jamming period (hereinafter referred to as "sequence number"), which affects the determination of the system state. We define state determination error (sde) as $\delta = |n - n_e|$, where n_e is the estimated sequence number and n is the actual sequence number. To accurately determine the state of the system, a state determination and adjustment method is presented in the form of a flowchart in Fig. 5. This method can reduce the sde gradually and finally make the system accurately determine the state of the current timeslot. To be specific, on the premise that the size of jamming period has been accurately estimated, we set the estimated sequence number of initial timeslot to $n_e = 1$. Before $n_e = N$, the system updates the next estimated state $s'_e = (n'_e, j')$ according to Eq. (4) and (5). When $n_e = N$, we judge whether there is only one pulse in the previous N timeslots (including the current timeslot). If it is, the system continues to update the next estimated state $s'_e = (n'_e, j')$ according to Eq. (4) and (5). If it is not, the system adjust the sequence number of next timeslot to $n_e{}' = N$, while the next jamming flag j' still be updated according to Eq. (5). After each adjustment, the sde minus 1. Therefore, after a finite number of adjustments, the estimated sequence number can be equal to the actual sequence number, i.e. $n_e = n$. In other word, the system can accurately determine the state of the current timeslot. If the jammer starts working in the τ_0th timeslot, the actual sequence number $n(\tau_0)$ of the τ_0th timeslot is 1, but the estimated sequence number is $n_e(\tau_0) = (\tau_0 - 1)\mathrm{rem}(N) + 1$, where $\mathrm{rem}(\cdot)$ is the remainder operator. The initial state determination error in τ_0th timeslot can be defined as $\delta_0 = |n(\tau_0) - n_e(\tau_0)| = (\tau_0 - 1)\mathrm{rem}(N)$. It means that it will take $(\tau_0 - 1)\mathrm{rem}(N)$ times adjustments to make $n_e = n$.

D. COMPLEXITY AND CONVERGENCE ANALYSIS

In Algorithm 1, the computational complexity of step 1 and step 9 are obviously $\mathcal{O}(1)$. The main computational complexity of the proposed Algorithm 1 lies in steps 2 to 8, which is about $\mathcal{O}(T)$, where T is the iterations numbers, i.e., the numbers of timeslots. Thus, the total computational complexity can be expressed as $\mathcal{C} = \mathcal{O}(T)$. It means that the proposed algorithm can achieve an optimal solution in polynomial time.

When the learning rate α_t in Eq. (9) is deterministic, nonnegative, and satisfies the following conditions:

$$\alpha_t \in [0, 1), \sum_{t=1}^{\infty} \alpha_t = \infty, \quad \text{and} \sum_{t=1}^{\infty} (\alpha_t)^2 < \infty, \tag{16}$$

the authors in [23] have proved that the Q-learning can fully traverse all system states and converge to the optimal strategy after a finite number of iterations. The proposed time-domain anti-pulse jamming algorithm (TDAA) is based on Q-learning to make decisions in a random pulse jamming environment. Hence, it can converge to the optimal anti-jamming strategy.

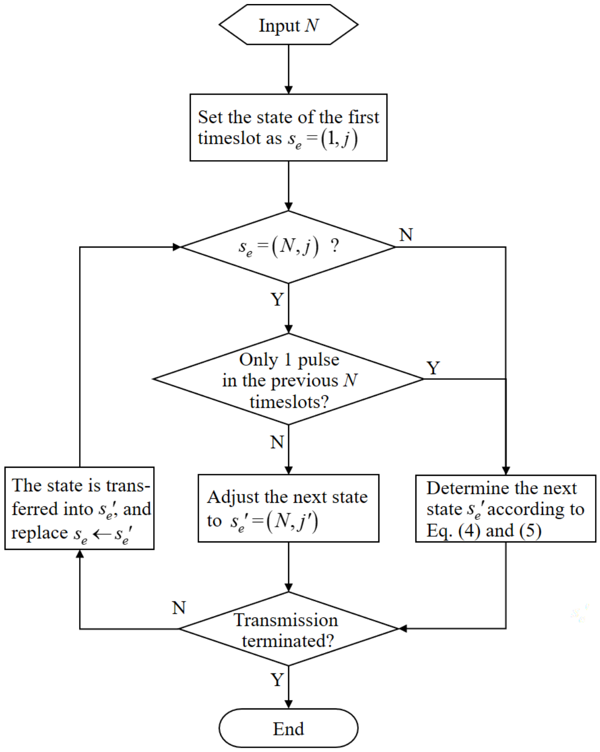

Input N
Set the state of the first timeslot as $s_e = (1, j)$
$s_e = (N, j)$?
N
Y
Only 1 pulse in the previous N timeslots?
Y
N
The state is transferred into s_e', and replace $s_e \leftarrow s_e'$
Adjust the next state to $s_e' = (N, j')$
Determine the next state s_e' according to Eq. (4) and (5)
N
Transmission terminated?
Y
End

FIGURE 5. Flowchart summarizing the state determination and adjustment method.

IV. SIMULATION RESULTS

A. JAMMING SELECTION DISTRIBUTION

To evaluate the performance of the proposed algorithm with different jamming selection distribution (jsd), the normal distribution, uniform distribution and Poisson distribution are simulated in this paper. The simulation parameters of the three distributions are set as follows.

1) NORMAL DISTRIBUTION

As shown in Fig. 6, the jsd is normal distribution, and the probability density function (pdf) in the kth jamming period can be expressed as:

$$f_J(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(t-\mu)^2}{2\sigma^2}], \tag{17}$$

where t represents the transmission time. $\mu = (k - 1/2)N \cdot T_s + \tau_0 \cdot T_s$ represents that the mean of the distribution is the median of the jamming period. τ_0 represents the sequence number of the timeslot when the jammer starts working. σ^2 is the variance. According to the Pauta criterion of normal distribution, the probability that the timeslot selected by the jammer is in the interval $[\mu - 3\sigma, \mu + 3\sigma]$ is approximately 100%. Thus, we set $\sigma = (N \cdot T_s)/10$ to ensure that the kth

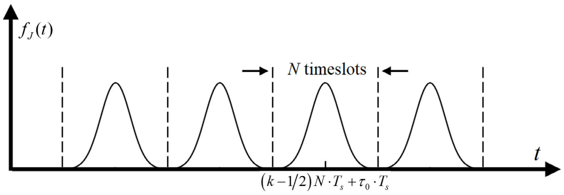


FIGURE 6. Probability density function when the jsd is normal distribution.

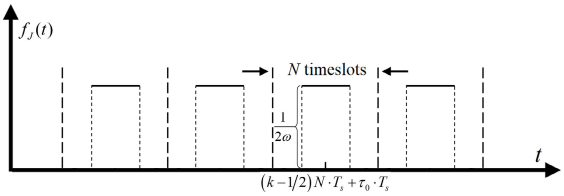


FIGURE 7. Probability density function when the jsd is uniform distribution.

pulse is in the kth jamming period. Then, the probability of pulse jamming in the nth timeslot within the kth jamming period can be expressed as:

$$p_k(n)=\int_{kT_s+(n-1)T_s}^{kT_s+nT_s}\frac{1}{\sqrt{2\pi}\sigma}\exp[-\frac{(t-\mu)^2}{2\sigma^2}]dt, \quad (18)$$

2) UNIFORM DISTRIBUTION

As shown in Fig. 7, the jsd is uniform distribution, and the pdf in the kth jamming period can be expressed as:

$$f_J(t)=\begin{cases}\frac{1}{2\omega T_s}, & \mu-\omega\cdot T_s<t<\mu+\omega\cdot T_s\\ 0, & \text{others},\end{cases} \quad (19)$$

where $\mu=(k-1/2)N\cdot T_s\ +\tau_0\cdot T_s$. We set $\omega=N/5$ to ensure that one timeslot can be selected in the corresponding jamming period. Then, the probability of pulse jamming in the nth timeslot within the kth jamming period can be expressed as:

$$p_k(n)=\begin{cases}\int_{kT_s+(n-1)T_s}^{kT_s+nT_s}\frac{1}{2\omega T_s}dt, & \mu-\omega\cdot T_s<t<\mu\\ & +\omega\cdot T_s\\ 0, & \text{others},\end{cases} \quad (20)$$

3) POISSON DISTRIBUTION

If the selected timeslot obeys Poisson distribution in the time domain, the probability of pulse jamming in the nth timeslot within the kth jamming period can be expressed as:

$$p_k(n)=\frac{\lambda^{n-1}}{(n-1)!}e^{-\lambda}, \quad (21)$$

where $n\in\{1,\ldots,N\}$ is the sequence number of the timeslot in the same jamming period. The parameter of the Poisson distribution is set as $\lambda=N/5$. When the size N of jamming period is 6, 8 and 10 respectively, the probabilities of different timeslots selected by the jammer are shown in Fig. 8.

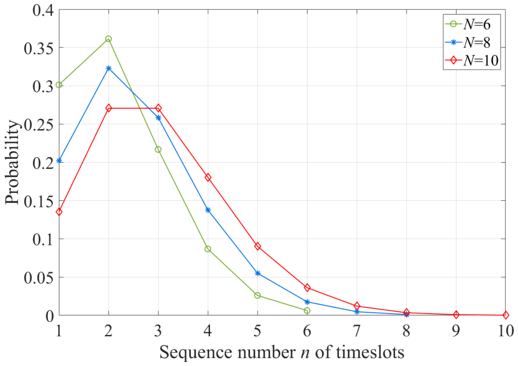


FIGURE 8. Probability of different timeslots being jammed when the jsd is Poisson distribution.

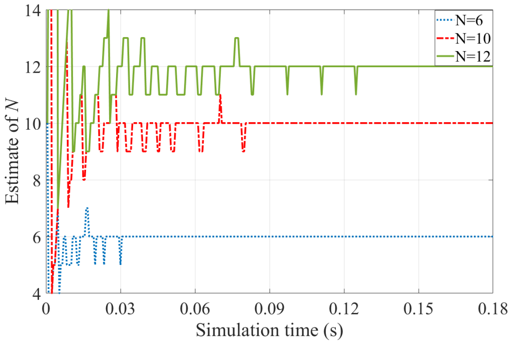


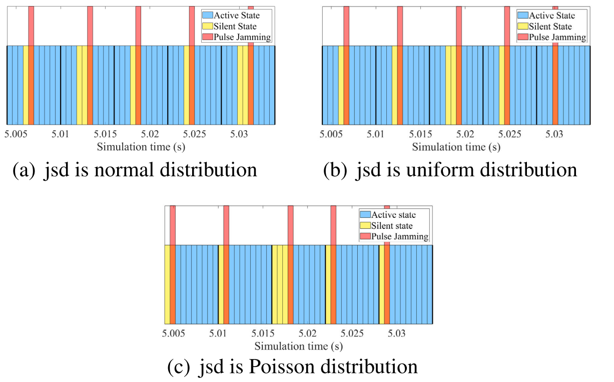
FIGURE 9. Estimation of jamming period *N*.

B. SIMULATION ANALYSIS

In the simulation, we set $T_0=0.5$ms, $T_1=0.04$ms, $T_2=0.06$ms which denote the transmission sub-slot, the detection sub-slot and the learning sub-slot respectively. The number of timeslots for simulations is set as $S=10000$, and the simulation time is $T_{sl}=S\times T_s$. Referring to [16], we set the learning rate $\alpha_t=0.8$, and the discount factor $\gamma=0.6$. Furthermore, the greedy factor is set as $\varepsilon=1/\sqrt{t}$.

It is necessary to accurately estimate the value of N for modeling and solving the anti-pulse jamming problem proposed in this paper. As shown in Fig. 9, when the values of N is 6, 10 and 12, the estimated value N_e converge gradually. According to the analysis in Section III, the estimation of N_e can converge to the exact value N in a finite time, which is confirmed by the simulation results. Thus, in the following simulation, we assume that the system has accurately estimated the value of N in advance. In addition, we temporarily consider that the timeslot when the jammer starts to work is just enough to make the initial state determination error $\tau_0=0$. The influence of different initial state determination errors on the performance of the proposed TDAA will be discussed separately in the following paper.

Fig. 10 shows the diagram of TDAA at convergent state when the jsd is normal distribution, uniform distribution and Poisson distribution. In the figures, the blue, yellow and red areas represent active state of transmitter, silent state of transmitter and pulse jamming respectively. Besides, different

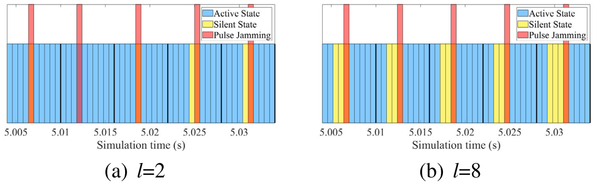


(a) jsd is normal distribution

(b) jsd is uniform distribution

(c) jsd is Poisson distribution

FIGURE 10. Diagram of TDAA at convergent state with different jsds (*N* = 10, *l* = 4).



(a) *l*=2

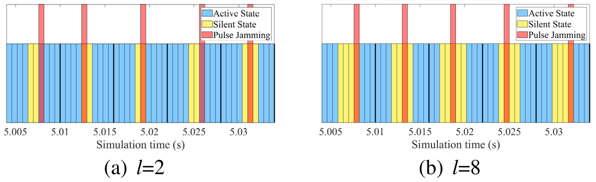
(b) *l*=8

FIGURE 11. Diagram of TDAA at convergent state with different packet lengths (jsd is normal distribution, *N* = 10).

jamming periods are divided by black bold lines. As indicated in the figures, at convergent state, the proposed TDAA can avoid most of the pulse jamming regardless of what the jsd is. In addition, when the size of jamming period is set as $N = 10$, and jsd is normal distribution, Fig. 11 shows the diagram of TDAA at convergent state with different packet lengths. With the increase of packet length of l, the number of sub-frames in each lost packet caused by pulse jamming increases. Thus, as shown in Fig. 11, when l is large, the transmitter keeps silent in more timeslots, ensuring a higher probability of avoiding jamming.

In this section, we compare the proposed TDAA with the following two methods:

- *Continuous transmission*: The transmitter continuously transmits data to the receiver without any anti-jamming measures.
- *Transmitting terminal Q-learning algorithm (TTQA)*: The transmitter independently executes the Q-learning algorithm. Since the pulse jamming signal only covers the receiver, the transmitter cannot sense the pulse jamming. This method uses the ACK mechanism to determine whether the transmission is successful, and thus obtains immediate rewards for different actions. The anti-jamming strategy is based entirely on local learning results of the transmitter. As shown in Fig. 12, when the jsd is normal distribution and $N = 10$, the TTQA can effectively avoid the pulse jamming at convergent state. Moreover, similar to TDAA, with an increase of l, the transmitter keeps silent in more timeslots.



(a) *l*=2

(b) *l*=8

FIGURE 12. Diagram of TTQA at convergent state with different packet lengths (jsd is normal distribution, *N* = 10).

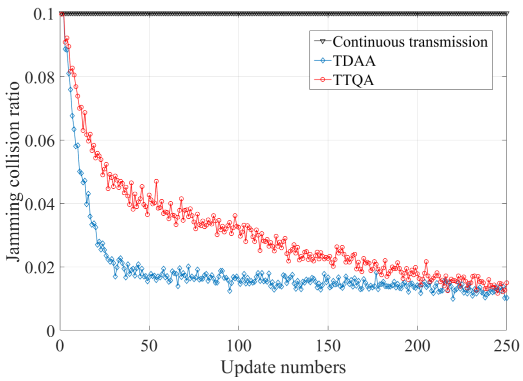


FIGURE 13. Jamming collision ratio of different methods.

To validate the performance of TDAA compared with the above two methods, we introduce the "jamming collision ratio", "time utilization ratio" and "normalized throughput". To describe the ratio of collision with pulse jamming when the transmitter takes a "keep active" action, we define the "jamming collision ratio" as $\rho_j = \tau_{jam}/T_{active}$, where τ_{jam} represents the number of timeslots in which the transmission is jammed, and T_{active} denotes the length of statistics of the timeslot in which the transmitter takes the "keep active" action, which means that the jamming collision ratio ρ_j is calculated every T_{active} active timeslots. To evaluate the time loss caused by the transmitter taking the "keep silent" action, we define the "time utilization ratio" as $\rho_{sl} = \tau_{active}/T_a$, where τ_{active} represents the number of timeslots in which the transmitter takes the "keep active" action. T_a denotes the length of timeslot statistics, which means that the time utilization ratio ρ_{sl} is calculated every T_a timeslots. Due to the average throughput is proportional to the average long-term reward, the "normalized throughput" can be defined as $\rho_{th} = (E \cdot \tau_{active} + (-L) \cdot \tau_{jam})/(E \cdot T_a)$. In our simulation, we set $T_{active} = 20$ and $T_a = 20$. Then, the simulation results are obtained by the mean of 200 independent runs. Besides, we let the jsd is normal distribution and $N = 10$ in the following simulation.

Fig. 13 shows the performance of the jamming collision ratio when the packet length $l = 4$. Compared with the continuous transmission, both TDAA and TTQA can significantly reduce the jamming collision rate to below 0.02. Moreover, the convergence rate of TDAA is higher than TTQA. Besides, Fig. 14 shows the jamming collision ratio of TDAA decreases with growing packet length of l. The reason

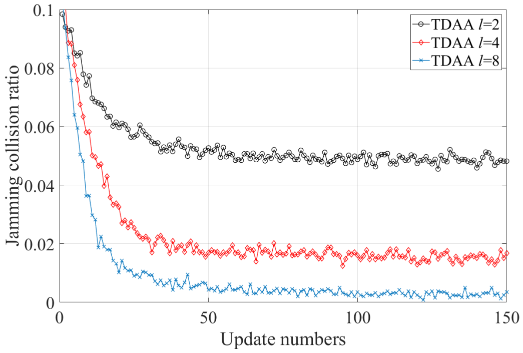


FIGURE 14. Jamming collision ratio of TDAA with different packet lengths.

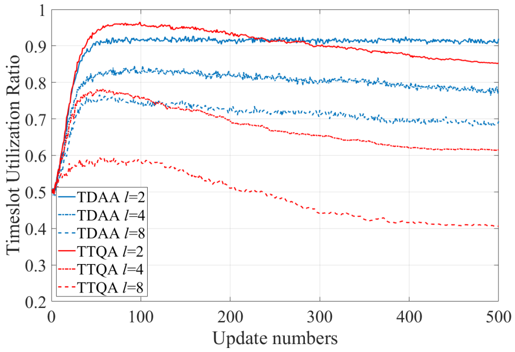
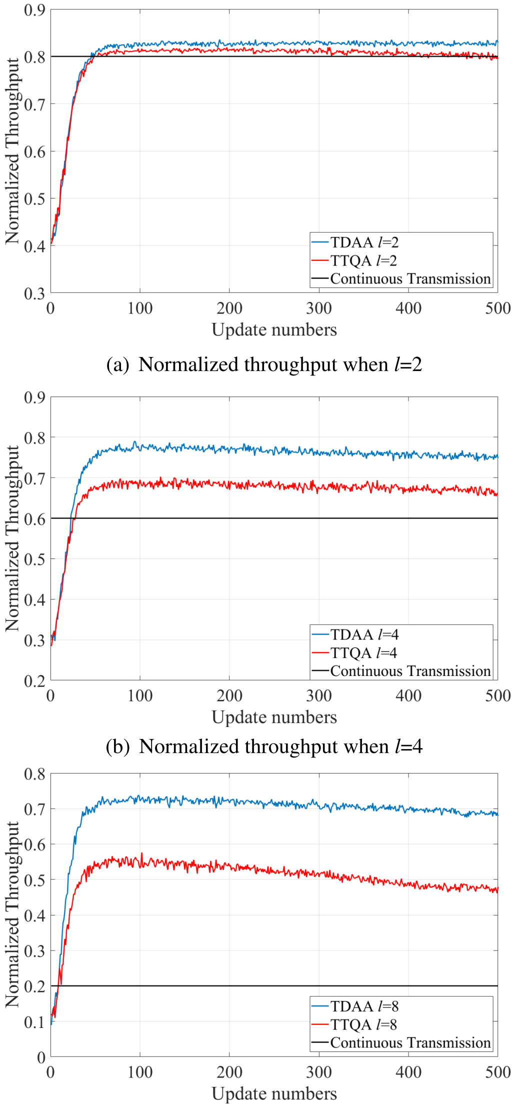


FIGURE 15. Time utilization ratio of TDAA and TTQA.

is that longer packets cause the transmitter to keep silent in more timeslots is shown in Fig. 11.

Fig. 15 compares the time utilization ratio between TDAA and TTQA. When the packet length is the same, the time utilization ratio of TDAA is higher than TTQA. The reason is that the TDAA can keep the transmitter active in the timeslots after the occurrence of pulse jamming in the same jamming period, while the TTQA may still keep the transmitter silent. Moreover, the time utilization ratio of both TDAA and TTQA decreases with growing packet length of l. The reason is that the loss caused by collision with pulse jamming increases with growing packet length of l, and thus both TDAA and TTQA keep the transmitter silent in more timeslots.

Fig. 16 shows the performance of the normalized throughput, it can be seen that the normalized throughput of TDAA at convergent state is higher than both continuous transmission and TTQA when the packet length is 2, 4 or 8. Furthermore, the improvement of the performance increases with growing packet length. The reason is that when the packet is short, the throughput loss caused by packet loss is small, and the throughput gain generated by the ''keep silent'' operation of TDAA for avoiding jamming is 0. In other words, ''zero gain'' does not significantly improve normalized throughput compared with ''small loss''. On the contrary, when the packet is long, the throughput loss caused by packet loss is large, and thus ''zero gain'' can significantly improve normalized throughput compared with ''large loss''. In addition, although the ''silent'' operation of the proposed algorithm results in time loss, the communication efficiency is improved because of the higher normalized throughput compared with the continuous transmission.



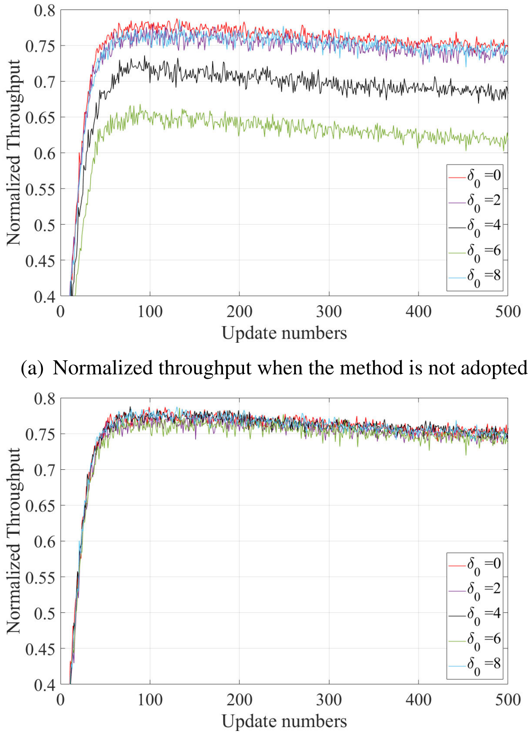
(a) Normalized throughput when l=2

(b) Normalized throughput when l=4

(c) Nnormalized throughput when l=8

FIGURE 16. Normalized throughput comparison when packet lengths are different.

Since the proposed TDAA adopts the state determination and adjustment method mentioned in section 3.3, it is necessary to evaluate the impact of different initial state determination error δ_0 on the performance of the proposed algorithm. According to the simulation Settings, when $N = 10$, there is an initial state determination error $\delta_0 \in [0, 9]$. As shown in Fig. 17(a), when the proposed state determination and adjustment method is not adopted, the convergence values of normalized throughput are reduced to different degrees by different δ_0. The reason is that the system's misjudgment of the state can never be corrected. As shown in Fig. 17(b), when the state determination and adjustment method is adopted, different δ_0 have no obvious impact on the normalized



(a) Normalized throughput when the method is not adopted

(b) Normalized throughput when the method is adopted

FIGURE 17. Normalized throughput of different initial determination errors.

throughput. It can be seen that the proposed state determination and adjustment method can effectively correct the wrong state determination at the initial time, thus improving the robustness of the algorithm.

V. CONCLUSION

In this paper, we investigated the anti-jamming problem under the threat of random pulse jamming. In order to obtain the countermeasure in time domain, Firstly, the anti-jamming problem is modeled as a Markov decision process (MDP). Then, a time-domain anti-pulse jamming algorithm (TDAA) based on reinforcement learning is proposed. The proposed algorithm can continuously learn from the dynamic interaction with the jamming environment, and gradually approach the optimal time-domain anti-jamming strategy that can maximize the system throughput. This strategy enables the transmitter to keep silent in timeslots with high probability of pulse jamming and keep active in other timeslots. In addition, a state estimation and adjustment method for random pulse jamming environment is introduced to improve the robustness of the proposed TDAA. Simulation results show that the proposed TDAA can significantly reduce the jamming collision ratio and improve the normalized throughput compared with the continuous transmission. Compared with TTQA, the proposed TDAA has higher time utilization ratio and normalized throughput.

REFERENCES

[1] R. A. Poisel, *Modern Communications Jamming Principles and Techniques*. Norwood, MA, USA: Artech House, 2011.

[2] M. Ghosh, "Analysis of the effect of impulse noise on multicarrier and single carrier QAM systems," *IEEE Trans. Commun.*, vol. 44, no. 2, pp. 145–147, Feb. 1996.

[3] H. Huynh and M. Lecours, "Impulsive noise in noncoherent M-ary digital systems," *IEEE Trans. Commun.*, vol. 23, no. 2, pp. 246–252, Feb. 1975.

[4] Y. J. Cho and C. K. Un, "Performance analysis of ARQ error controls under Markovian block error pattern," *IEEE Trans. Commun.*, vol. 42, no. 234, pp. 2051–2061, Feb./Apr. 1994.

[5] B. Debruhl and P. Tague, "How to jam without getting caught: Analysis and empirical study of stealthy periodic jamming," in *Proc. 10th Annu. IEEE Commun. Soc. Conf. Sensor, Mesh Ad Hoc Commun. Netw. (SECON)*, Jun. 2013, pp. 496–504.

[6] M. Hall, A. Silvennoinen, and S.-G. Haggman, "Effect of pulse jamming on IEEE 802.11 wireless LAN performance," in *Proc. MILCOM-IEEE Mil. Commun. Conf.*, vol. 4, Oct. 2005, pp. 2301–2306.

[7] J.-J. Lee and J. Lim, "Effective and efficient jamming based on routing in wireless ad hoc networks," *IEEE Commun. Lett.*, vol. 16, no. 11, pp. 1903–1906, Nov. 2012.

[8] N. Noels and M. Moeneclaey, "Performance of advanced telecommand frame synchronizer under pulsed jamming conditions," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.

[9] P. Duraisamy and L. Nguyen, "Self-encoded spread spectrum with iterative detection under pulsed-noise jamming," *J. Commun. Netw.*, vol. 15, no. 3, pp. 276–282, Jun. 2013.

[10] N. I. Jin-Gen and M. A. Lan-Shen, "Distributed affine projection sign algorithms against impulsive interferences," *Acta Electron. Sinica*, vol. 44, no. 7, pp. 1555–1560, 2016.

[11] X. Liu, Y. Xu, L. Jia, Q. Wu, and A. Anpalagan, "Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach," *IEEE Commun. Lett.*, vol. 22, no. 5, pp. 998–1001, May 2018.

[12] Z. Zhou, *Machine Learning*. Beijing, China: Tsinghua Univ. Press (in Chinese), 2016.

[13] H. R. Berenji, "Fuzzy Q-learning for generalization of reinforcement learning," in *Proc. IEEE 5th Int. Fuzzy Syst.*, vol. 3, Sep. 1996, pp. 2208–2214.

[14] B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy, "An anti-jamming stochastic game for cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 877–889, Apr. 2011.

[15] S. Machuzak and S. K. Jayaweera, "Reinforcement learning based anti-jamming with wideband autonomous cognitive radios," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jul. 2016, pp. 1–5.

[16] F. Slimeni, Z. Chtourou, B. Scheers, V. L. Nir, and R. Attia, "Cooperative Q-learning based channel selection for cognitive radio networks," *Wireless Netw.*, vol. 25, no. 4, pp. 4161–4171, 2018.

[17] L. Xiao, X. Lu, D. Xu, Y. Tang, L. Wang, and W. Zhuang, "UAV relay in VANETs against smart jamming with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4087–4097, May 2018.

[18] M. A. Aref, S. K. Jayaweera, and S. Machuzak, "Multi-agent reinforcement learning based cognitive anti-jamming," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[19] M. A. Aref and S. K. Jayaweera, "A cognitive anti-jamming and interference-avoidance stochastic game," in *Proc. IEEE 16th Int. Conf. Cognit. Informat. Cognit. Comput. (ICCI CC)*, Jul. 2017, pp. 520–527.

[20] R. A. Poisel, *Introduction to Communication Electronic Warfare Systems*. Norwood, MA, USA: Artech House, 2002.

[21] A. Hussain, N. A. Saqib, U. Qamar, M. Zia, and H. Mahmood, "Protocol-aware radio frequency jamming in Wi-Fi and commercial wireless networks," *J. Commun. Netw.*, vol. 16, no. 4, pp. 397–406, Aug. 2014.

[22] J.-B. Liang and J.-M. Xu, "A novel contour extraction approach based on Q-Learning," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Aug. 2006, pp. 3807–3810.

[23] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.

[24] N. Van Huynh, D. N. Nguyen, D. T. Hoang, and E. Dutkiewicz, "'Jam me if you can:' Defeating jammer with deep dueling neural network architecture and ambient backscattering augmented communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2603–2620, Nov. 2019.

[25] F. Slimeni, B. Scheers, Z. Chtourou, and V. Le Nir, "Jamming mitigation in cognitive radio networks using a modified Q-learning algorithm," in *Proc. Int. Conf. Mil. Commun. Inf. Syst. (ICMCIS)*, May 2015, pp. 1–7.

[26] M. A. Aref and S. K. Jayaweera, "A novel cognitive anti-jamming stochastic game," in *Proc. Cognit. Commun. Aerosp. Appl. Workshop (CCAA)*, Jun. 2017, pp. 1–4.

QUAN ZHOU was born in Liyang, Jiangsu, China, in 1991. He received the B.S. degree in communication engineering from East China Jiaotong University, Nanchang, in 2014. He is currently pursuing the master's degree in electronics and communication engineering with the Army Engineering University of PLA with a focus in intelligent communication anti-jamming method.

YONGGUI LI was born in Anhui, China, in 1964. He received the M.S. degree in information and communication engineering from the PLA University of Science and Technology, in 2000.

He is currently a Senior Research Fellow with the National University of Defense Technology, China. He has authored or coauthored more than 40 journal and conference papers, and published one book. He is also involved in the research of modern wireless communication and its network intelligence theory and technology, especially wireless communication spectrum sensing and jamming analysis, dynamic spectrum access, adaptive communication algorithm, and so on.

YINGTAO NIU was born in Taian, China, in 1978. He received the M.S. degree from the PLA Commanding Communication Academy, China, in 2005, and the Ph.D. degree from the Institute of Communication Engineering, PLA University of Science and Technology, China, in 2008.

He is currently a Senior Research Fellow with the National University of Defense Technology, China. He has authored more than 40 journal and conference papers. His main research interests are cognitive radio theory and techniques, with particular emphasis on algorithms of signal sensing and communication decision-making algorithm in cognitive radio systems.