applied sciences

MDPI

Article

An Intelligent Access Channel Algorithm Based on Distributed Double Q Learning

Guoliang Zhang, Yingtao Niu *, Yonggui Li and Liping Zhao

The Sixty-Third Research Institute, National University of Defense Technology, Nanjing 210007, China
* Correspondence: niuyingtao78@hotmail.com

Abstract: Aiming at the problem of mutual interference between nodes and external malicious jamming in wireless communication networks, this paper proposes an intelligent communication anti-jamming algorithm based on distributed double Q-learning. First of all, the proposed algorithm anti-jamming problem is modeled as a time-frequency two-dimensional optimization problem. The decoupled relationship between subframe and channel is used. Each node uses two Q-learning. According to the automatically received feedback signal, users choose their subframe and channel to avoid malicious external jamming and mutual interference between users. It maximizes the sum of all user throughput. Simulation results show that the proposed algorithm can effectively shorten the convergence time and improve the performance of the system.

Keywords: communication anti-jamming; double Q-learning; distributed learning; channel access

check for updates

Citation: Zhang, G.; Niu, Y.; Li, Y.; Zhao, L. An Intelligent Access Channel Algorithm Based on Distributed Double Q Learning. *Appl. Sci.* **2022**, *12*, 10815. https://doi.org/10.3390/app122110815

Academic Editor: Juan-Carlos Cano

Received: 17 September 2022
Accepted: 21 October 2022
Published: 25 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the explosive growth of wireless communication devices, wireless communication presents networked development [1,2]. Channel access is one of the key problems to be solved in wireless communication networks in order to prevent interference caused by multiple users using the same channel at the same time [3]. In addition, due to the openness of electromagnetic wave propagation space [4], wireless communication networks are susceptible to external malicious jamming [5,6]; as an important threat to the security of wireless communication, it mainly affects and destroys the detection of useful signals by electromagnetic radiation signals, which seriously degrades the system performance. Existing communication anti-jamming measures are mainly extended spectrum communication technology [7], which can effectively deal with conventional jamming such as multi-tone and blocking. With the application of artificial intelligence [8] to the field of communication jamming, the jamming modes are characterized by diversity, dynamics and intelligence [9,10]. The conventional extended spectrum communication can no longer ensure the reliability of communication. In the face of a more complex and harsher electromagnetic environment, how to use more "clever" algorithms to avoid mutual interference and external malicious jamming, to ensure reliable communication between users, is a wireless communication network problem to be solved urgently.

In order to solve the problem of mutual interference between different users, the traditional approach is to formulate access criteria for users through access protocol [11]. However, the current protocols do not consider the existence of external jamming. Considering future more complex wireless network scenarios, an artificial intelligence algorithm is a worthwhile idea. Machine learning is an important kind of artificial intelligence algorithm, which can be divided into supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning according to learning methods [12,13]. The paper mainly studies how to use reinforcement learning [14] to solve the communication anti-jamming problem. The main idea of reinforcement learning is that in the face of an unknown environment, the agent can take actions, conduct trial and error experiments, interact with the environment, constantly

learn from the environment feedback and update and improve its action strategy, so as to learn the optimal strategy and make the return value of the agent reach the optimal value. Multi-agent reinforcement learning algorithms can be divided into centralized Q-learning [15] and distributed Q-learning [16] according to whether they depend on the central controller. Centralized Q-learning is aimed at cooperative reinforcement learning scenarios with a central controller, which is the learning subject. Distributed Q-learning is aimed at the non-centralized cooperative reinforcement learning scenario, where all communication nodes are learning subjects. Reference [17] proposes a multi-user cooperative anti-jamming algorithm, but the convergence effect is slow, which is difficult to cope with the real-time demand for decision making under the battlefield. Reference [18] designs a Markov game framework and proposes a multi-user joint anti-jamming algorithm. The algorithm can shorten convergence time, but only for a few users. Many users will lead to a dimensional disaster problem, resulting in difficult convergence. Reference [19] considers the influence of distance on mutual disturbance and proposes a multi-agent Q-learning algorithm based on distance metrics. References [17–19] are all centralized Q-learning. Centralized Q-learning requires centralized learning and distributed execution. It requires the central controller to perform all the operations. When there are many users, the response speed will slow down. Reference [20] proposes a multi-user reinforcement learning algorithm, but each user adopts the single-user reinforcement learning method [21] without considering the cooperation and interference between users. The convergence speed is slow. In reference [22], a novel collaborative distributed Q-learning mechanism is designed for resource-constrained communication devices, so that they can find idle subframes for their transmission and take the occupancy order of subframes as the global competition cost. This algorithm only considers the time domain subframes, without considering the frequency domain channel. Reference [23] presents a distributed Q-learning based on the decoupling (DQL) scheme. Each device separately uses two Q-learning mechanisms to find a spare subframe and code book (a collection of all the available frequency hopping codes). The scheme has strong scalability and can expand the number of users to hundreds, but it does not consider the external disturbance problem. References [20–23] are all distributed Q-learning algorithms. As each user conducts distributed computing and distributed executing, and with moderate hardware performance requirements, it displays strong extendibility and processing capacity.

Distributed Q-learning was applied to multi-user scenarios anti-interference decision making. The paper puts forward a distributed double Q-learning algorithm. Each communication node uses two Q-learning. According to the automatically received feedback signal, users choose the optimal subframe and channel at the same time to avoid malicious external jamming and mutual interference between the users and improve the network overall throughput and reduce the dimension of action set. To improve the convergence of the algorithm. The main contributions of this paper are as follows:

1. A distributed double Q-learning communication anti-jamming algorithm (DDQL) is proposed, which can effectively avoid the interference between users and external malicious jamming. The algorithm reduces the dimension of action set and shortens the convergence time of the algorithm.
2. It is proved that the channel subframe binary optimization problem can be equivalently transformed into two unary optimization problems.

The rest of the article is organized as follows. The system model is presented in Section 2, while the problem is formulated. Section 3 introduces the anti-jamming algorithm of intelligent communication based on distributed double Q-learning, while the simulation results and analyses are given in Section 4. The final part is the summary.

2. System Model and Problem Formulation

2.1. System Model

In order to facilitate the research, this paper makes the following assumptions:

- As shown in Figure 1, we consider a wireless communication network. There are N users, one base station, and one jammer. The number of channels available is M. The set of channels available is $\mathbf{C} = [C_1, \ldots, C_m, \ldots C_M]$. Each user distributed computing

and distributed execution, and hardware performance requirements are not high with good scalability and strong processing power.

- A MAC frame consists of K subframes, and each subframe consists of B time slots. Therefore, the time slot of a MAC frame is $N_{slot} = KB$. Each subframe is followed by a feedback signal from the base station. The content of the feedback signal is the current subframe and the current channel occupancy information, so that all users can make full use of the global resource occupancy information.
- The jamming mode of the jammer is multi-channel intelligent blocking interference [23]. The jammer can quickly sense the channel occupied by the user and select $Z(Z < M)$ channels that have been occupied for the longest time in the previous multiple subframes for jamming. The jamming channel set can be defined as $\mathbf{J} = \{J_1, \cdots, J_z, \cdots, J_Z\}$.
- It is assumed that the subframe and time slot of the system are synchronized, and the subframe is the basic transmission unit in this paper. Each user has L packets to be transmitted. Each time slot can successfully transmit one packet.

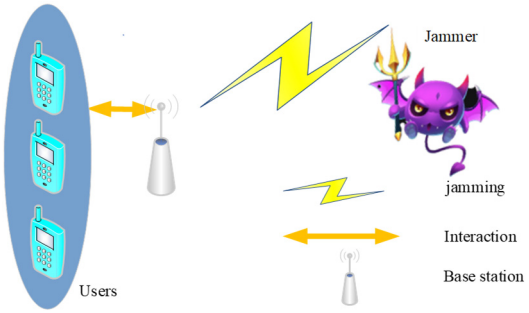


Figure 1. System model.

2.2. Problem Formulation

In this section, we first formulate a throughput optimization problem to maximize the sum of throughput of all users by selecting the optimal subframes and signal malicious jamming information on the interactive jammer userbase station channels for each user. We propose a collaborative decoupling distributed double Q-learning algorithm to solve this problem.

Due to the interference between users and the malicious jamming of the external jammer, the communication between users will fail. We introduce a binary variable $\xi_{n,m,k} \in \{0,1\}$. The binary variable represents the result of subframe-channel selection for the nth user. If nth user selects the mth channel in the kth subframe, then $\xi_{n,m,k} = 1$, and otherwise $\xi_{n,m,k} = 0$. The number of users occupying the mth channel in the kth subframe is defined as $l_{m,k} = \sum_n \xi_{n,m,k}$. The users' set of successful transmission is $S(n') \triangleq \{n' | \sum_v 1(l_{m,k} = 1)\xi_{n,m,k} = 1\}$, where n' is users that choose the idle channel; $1(\cdot)$ indicates function. When the events of $(\cdot)$ is true, the $1(\cdot) = 1$. If user n occupies the mth idle channel in the kth subframe, then user n can successfully transmit the packet to the base station. Assume that the base station can only successfully receive users in user set $S(n')$. Then, a binary variable $\beta_{n,m,k}$ is introduced to represent whether user n can make the base station successfully receive the packet by selecting the mth channel in the kth subframe. If $n \in S(n')$, then $\beta_{n,m,k} = 1$, otherwise, $\beta_{n,m,k} = 0$. The goal of the paper is to maximize the sum of throughput of all users. In this paper, it is assumed that the throughput is the ratio of the number of successful user communications to the number of subframes. The optimization problem can be formulated as follows:

$$\max_{\{\beta_{n,m,k}\}} O = \sum_{n=1}^{N} \beta_{n,m,k}/K \tag{1}$$

$$\begin{gathered}\text{s.t. } \beta_{n,m,k} \in \{0,1\} \\ \forall n \in [1,2,\ldots,N] \\ \forall m \in [1,2,\ldots,M] \\ \forall k \in [1,2,\ldots,K]\end{gathered}$$

According to the literature [23], Equation (1) is a non-convex function, so it is usually difficult to directly obtain the optimal solution. In order to maximize the system throughput, the channel conflicts between users should be minimized and the channel without malicious jamming should be selected. For each user, an idle channel should be selected in a subframe. Ideally, each user selects a unique subframe and channel at a time to send its packets. The selected channel is not jammed by jammers. However, this ideal is almost impossible to achieve with completely random access. However, access control methods based on machine learning (ML) algorithms (such as DQL algorithm) are expected to solve the problem (1). In this paper, the solution of Equation (1) can be found automatically by the distributed DQL algorithm deployed on all users, so that each user can find the optimal selection result of available subframes and channels, and avoid the interference between users and the malicious jamming of external jammers as much as possible.

The process of joint channel and subframe selection is modeled as a two-element Markov problem, where each user is the learning subject. Its state space is the currently available channel subframe combination:

$$\mathbf{S_{k,m}} = \{(k,m)\}, k \in (1,2,\ldots,K), m \in (1,2,\ldots,M) \tag{2}$$

The action space of channel and subframe joint selection Markov is defined as the joint action of channel and subframe that can be selected:

$$\mathbf{A}_{k,m} = \{(k,m)\}, k \in (1,2,\ldots,K), m \in (1,2,\ldots,M) \tag{3}$$

The current reward function of user n:

$$R_n(k_n,m_n) = \begin{cases} 1 & (k_n,m_n) \neq (k_j,m_j) \,\&\, (k_n,m_n) \neq (k_i,m_i)(i \in N/n) \\ -1 & \text{otherwise} \end{cases} \tag{4}$$

where, (k_n,m_n) denotes the subframe and channel selected by user n. (k_j,m_j) denotes the subframe and channel where the jamming is located. (k_i,m_i) denotes the subframe and channel selected by other users except user n.

Furthermore, we model the channel and subframe selection process as two Markov decision processes (MDP), and equivalently prove that the two-element joint selection optimization problem of the subframe channel is equivalent to two unitary optimization problems of the subframe channel, and each access subframe and channel are, respectively, treated as the states of different MDP. In the Markov decision process, each agent constantly interacts with the environment, queries the Q table according to the current state of the environment, selects the optimal action, so as to obtain the return brought by the action, and then enters the next state.

The current states of the two Markov processes defining access subframe and channel are, respectively, each access subframe and channel, so the two state spaces are:

$$\mathbf{S_k} = \{1,2,\ldots,K\} \tag{5}$$

$$\mathbf{S_m} = \{1,2,\ldots,M\} \tag{6}$$

The current action spaces of the two Markov processes defining access subframe and channel are, respectively, each access subframe and channel, so the two action spaces are:

$$\mathbf{A_k} = \{1,2,\ldots,K\} \tag{7}$$

$$\mathbf{A_m} = \{1,2,\ldots,M\} \tag{8}$$

In the existing literature, the reward function $R_1(n,k)$ of user n in the kth time slot is called binary classical Q-learning, and its expression is as follows:

$$R_1(n,k) = \begin{cases} +1 & \text{if transmission succeeds} \\ -1 & \text{otherwise} \end{cases} \tag{9}$$

We define the occupancy intensity of a subframe as the number of users that a user intends to transmit packets in the same subframe; that is $\mathbf{ZL} = [ZL_1, ZL_2, \ldots, ZL_K]$, where $\mathbf{ZL}$ is a one-dimensional vector of $1 \times K$. On the basis of occupancy intensity, we define the Markov process reward function for accessing subframes as $R_1(n,k)$:

$$R_1(n,k) = \begin{cases} +1 & \text{when } \mathbf{ZL}(k) \leq (M - \mathrm{Z}) \\ -\mathbf{ZL}(k)/K & \text{when } \mathbf{ZL}(k) > (M - Z) \end{cases} \tag{10}$$

where, $R_1(n,k)$ is the reward function for the user n to select a subframe, $n \in \{1, \ldots, N\}$, $k \in \{1, \ldots, K\}$, $m \in \{1, \ldots, M\}$. The number of channels jammed by the jammer is Z. $\mathbf{ZL}(k)$ represents the user occupancy intensity of the kth subframe. That is, the number of users who intend to transmit packets in the same subframe. If the user occupancy intensity in the kth subframe is less than or equal to the number of all channels $M - Z$, then $R_1(n,k)$ equals 1, indicating that the transmission is successful; Otherwise, it equals $-\mathbf{ZL}(k)/K$, indicating that the transmission fails.

The Markov process reward function of access channel is defined as $R_2(n,m)$:

$$R_2(n,m) = \begin{cases} +1 \text{ when } n \in S(n') \\ -1 \text{ when } n \notin S(n') \end{cases} \tag{11}$$

where, $R_2(n,m)$ is the reward function for the user n to select a channel, $k \in \{1, \ldots, K\}$,$m \in \{1, \ldots, M\}$. The set of users who occupy idle channel collection is $S(n')$. $n \in S(n')$ represents user n select the mth channel that is idle channel in the kth subframe. There is no other user interference and jamming, otherwise the user n in the kth subframe select the mth channel where other user or jammer occupying. It can lead to channel conflict or disturbance of communication signals, lead to the failure signals. If the mth channel selected by user n in the kth subframe is an idle channel, then $R_2(n,m)$ is equal to 1, indicating that the transmission is successful; Otherwise, it is equal to -1, indicating that the transmission fails.

3. Anti-Jamming Algorithm of Intelligent Communication Based on Distributed Double Q-Learning

To solve the problem shown in Equation (1), we propose an anti-jamming algorithm based on distributed double Q-learning. Each user uses two similar Q-learning algorithms to select channel and sub-frame, respectively, according to the occupancy strength of channel and subframe. In the Markov decision process, each agent constantly interacts with the environment, queries the Q table according to the current state of the environment, selects the optimal action, so as to obtain the return brought by the action, and then enters the next state. As mentioned earlier, the goal of the proposed algorithm is to maximize the sum of the throughput of all users in the wireless communication network.

First, we prove that two unary Q learning algorithms and a binary Q learning algorithm (subframe is row, channel is column) are equivalent; that is, we prove whether the binary optimization problem of the joint selection of subframe and channel is equivalent to two unary optimization problems of subframe selection and channel selection. According to the optimization theory [24], the necessary and sufficient condition for equivalence is that the two independent variables are separable.

Prove that $Q_2(n,m) = Q_n(k_n, m_n)$, $Q_1(n,k)$ is the Q value of the selected subframe, $Q_2(n,m)$ is the Q value of selected channel. $Q_n(k_n, m_n)$ is the Q value of selected subframe and channel together.

In Equation (4), $(k_n, m_n) \neq (k_j, m_j)$& $(k_n, m_n) \neq (k_i, m_i)(i \in N/n)$ indicates that the user n does not suffer from inter-user interference or external malicious jamming. It fits the characteristics of the element in $S(n')$. The mth channel selected by user n in the kth subframe is the idle channel.

Therefore, Equation (9) can be expressed as:

$$R_n(k_n, m_n) = \begin{cases} 1 & n \in S(n') \\ -1 & \text{otherwise} \end{cases} \tag{12}$$

$R_1(n,k)R_2(n,m)$ is:

$$R_1(n,k)R_2(n,m) = \begin{cases} +1 & \text{when } n \in S(n') \\ -1 & otherwise \end{cases} \tag{13}$$

According to Equations (12) and (13):

$$R_1(n,k)R_2(n,m) = R_n(k_n, m_n) \tag{14}$$

According to the literature [14], the relationship between Q value and return function is as follows:

$$Q(s,a) = r(s'|s,a) + \gamma \sum_{s' \in S} P^a_{ss'} \max_{a \in A} Q(s', a') \tag{15}$$

where, s' is the state of the next subframe, and $P^a_{ss'}$ is the state transition probability that is transferred from state s to state s' after action a is taken by state s. $r(s'|s,a)$ represents the immediate reward transferred to state s' after action a is taken by state s. The action taken in this paper only considers the current reward and does not consider the future reward. That is, the discount factor γ of the future reward is 0, so Equation (15) can be expressed as:

$$Q(s,a) = r(s'|s,a) \tag{16}$$

According to Equations (14) and (16):

$$Q_1(n,k)Q_2(n,m) = Q_n(k_n, m_n) \tag{17}$$

As can be seen from Equation (17), k and m are separable, so the algorithm proposed in this paper is equivalent to the subframe and channel joint algorithm. Therefore, we transform a two-element optimization problem of joint selection of subframe channels into two unary optimization problems of the separate selection of the subframe and channel.

Each user needs to maintain two one-dimensional Q tables, namely user-subframe Q table and user-channel Q table. We denote the $t-th$ iteration Q value of user n transmitting a packet in the kth subframe as $Q_{1,t}(n,k)$, and the $t-th$ iteration Q value of user n transmitting a packet in the mth channel as $Q_{2,t}(n,m)$.

According to the literature [25], Q value can be updated as follows:

$$Q_{1,t+1}(n,k) = (1-\alpha)Q_{1,t}(n,k) + \alpha * R_1(n,k) \tag{18}$$

$$Q_{2,t+1}(n,m) = (1-\alpha)Q_{2,,t}(n,m) + \alpha * R_2(n,m) \tag{19}$$

where, α is the learning rate, and our proposed algorithm takes actions that only consider the current reward and do not consider the future reward. The discount factor of the future reward γ is 0, because once the user successfully communicates, their needs are satisfied

The base station can obtain the channel collision status of users in each time slot through preamble detection [26]. The subframe occupancy intensity and the channel collision state can be broadcast to all users through the base station, reducing the interaction between users, signaling overhead and user edge computation complexity. The simplest way to implement the Q-learning algorithm is to use the ε-greedy method of trade-off ex-

ploration; that is, each user chooses the action with the highest Q value with the probability of $1-\varepsilon$, and randomly chooses an action in the action space with the probability of ε. The optimal policy of this user is to execute his actions according to the ε-greedy policy. The greedy policy selects the actions with the highest Q value in subframe-Q and channel-Q tables with a probability of $1-\varepsilon$, namely:

$$\pi_n^*(k,m) = \begin{cases} \underset{k,m}{\text{argmax}}\{Q_1(n,k), Q_2(n,m)\} & \text{Probability of } 1-\varepsilon \\ \forall_k, \forall_m & \text{Probability of } \varepsilon \end{cases} \tag{20}$$

For the user n, the specific process of the proposed algorithm is shown in Algorithm 1. In the first iteration, the values of the two Q tables are all initialized to 0, and each user randomly selects any subframe and channel. After receiving the feedback signal from the base station, Equations (10) and (11) are used to calculate the reward $R_1(n,k)$ and $R_2(n,m)$ brought by the action selected by the user, and then Equations (18) and (19) are used to update the corresponding Q values in the time slot and channel table. After each iteration of the Q table, the user selects the channel corresponding to the maximum Q value in the channel-q table, and then selects the subframe corresponding to the maximum Q value in the subframe -Q table to send the packet. When the number of iterations does not reach the maximum number of iterations or the convergence result is not reached, the DDQL algorithm will continue to update iteratively.

The processing flow of user n DDQL algorithm is shown in Algorithm 1. We first initialize the parameters, the learning factor α, the maximum number of iterations T, and then set all Q values in the subframe-Q table and channel-Q table of each user to 0. In the first iteration ($t=1$), each user selects a channel in a random manner and sends a packet on a subframe selected in a random manner. Later, when each user feedback signal received from the base station, through the Equations (18) and (19), update the subframe Q table, respectively, and channel—Q value in the table, where each iteration after the update, the user choosing the Q value of the highest channel—Q value and the corresponding channel, and select in the table in a subframe—Q value of the highest in the corresponding subframe to transmit packets. Until the number of iterations T reaches the maximum number of iterations L or the system converges, the DDQL algorithm will continue to be executed until each user can obtain the best subframes and channels, thus obtaining the optimal solution to problem (1).

Algorithm 1: The user n cooperative distributed double Q learning algorithm

1: Initialization: $Q_{1,,t}(n,k)=0$, $Q_{2,,t}(n,m)=0$, $t=1$, $\forall k \in [1,\ldots,K]$, $\forall m \in [1,\ldots,M]$.
Set the learning factor α and the maximum number of iterations T.
2: **repeat**
3: **for** $\forall k, m$ **do**
4: According to the base station feedback signal, users calculate their own reward $R_1(n,k)$ and $R_2(n,m)$ according to the Equations (10) and (11).
5: Update $Q_{1,t+1}(n,k)$ and $Q_{2,,t+1}(n,m)$ according to the Equations (18) and (19).
6: **End for**
7: According to Equation (20), the optimal strategy is found.
8: **until** the number of iterations t has reached T or the convergence has been reached.

4. Simulation Results and Discussion

4.1. Parameter Settings

We set the parameters related to simulation in Table 1 as follows:

Based on the simulation setup, we set up the five channels and 50 subframes. The jammer occupies the three channels every time. Each subframe has 100 slots. The above assumes each time slot transmission of a packet and each user has a 100-packet transmission. Therefore, the user in a MAC frame only need to finish the communication to successfully

transmit a subframe. The channel number is far less than the number of users. Users can also communicate.

In this paper, simulation experiments are carried out to verify the performance of the proposed algorithm (DDQL) under traditional interference linear frequency sweep interference and multi-channel intelligent blocking jamming. As shown in Figure 2a, linear frequency sweep jamming is a dynamic jamming that periodically scans the target channel according to the MAC frame by narrow-band strong power jamming signals. It is simple and efficient, and widely used in the actual battlefield environment. The number of simultaneous blocked channels of the parameter jammer channels Z is 1. As shown in Figure 2b, the multi-channel jamming blocking channel is an upgrade of linear frequency sweep interference. The jammers can sense the channel used by the communication users in the last transmission subframe, so as to focus on interference in the next subframe, and take the MAC frame as periodic interference, and increase the number of jamming channels Z to 3.

Table 1. Settings of Model related parameter.

Parameters	Value
Communication users N	10–150
Available channels M	5
Subframes K	50
Slots B	100
Packet to be transmitted L	100
Number of blocked channels Z	3
Learning α	0.9
Discount factor γ	0
Greedy index ε	$1/\sqrt{t}$
Number of iterations T	300

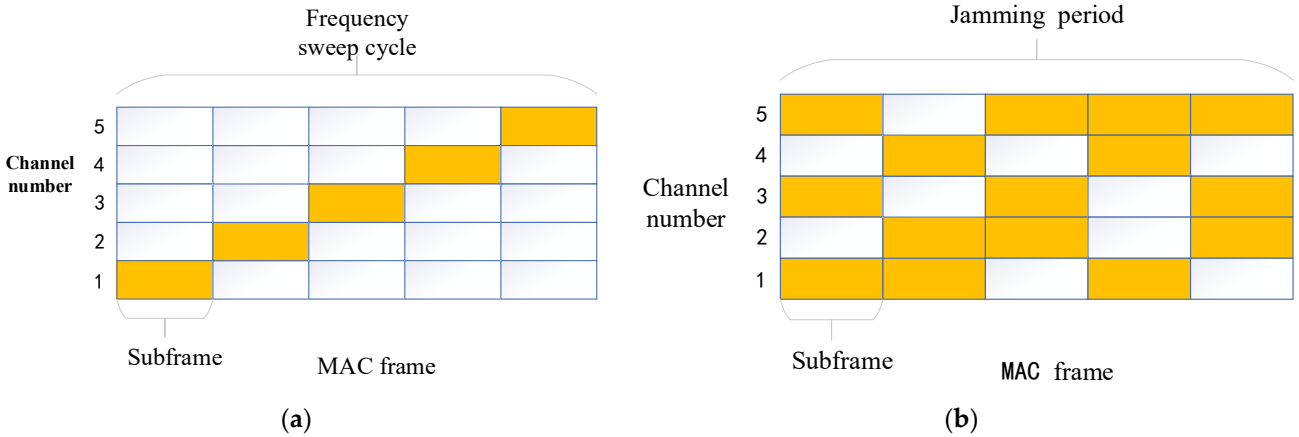


Figure 2. (**a**) Linear frequency sweep jamming. (**b**) Multi-channel intelligent blocking jamming.

4.2. Analysis of Simulation

In this paper, simulation verification is carried out on the MATLAB platform. In order to highlight the superiority of the proposed algorithm, we compare it with the other two algorithms.

- Channel selection algorithm: based on Sense and random selection. Before selecting a channel, the user can sense whether there is interference in the channel. If no interference is detected in the current transmission channel, the user still transmits in the same channel; otherwise, the user randomly switches to other channels in the next action slot. Here, no information is exchanged between users.

- Binary independent Q learning algorithm (ILQ): each user executes the Q-learning algorithm independently only through the local learning results to make transmission decisions and never considers the decision of other users.

For multi-channel intelligent blocking jamming, we consider the impact of the number of users in the network on the normalized throughput of the network. We calculate the maximum access load $\rho = (M - Z) * K = 100$ that can be borne by a frame. The maximum number of users can be allowed to access in a frame. We control the number of users within the range of 10–150, as shown in Figure 3. It can be seen that within the maximum access load of the system, the normalized throughput increases with the increase in the number of users. When the maximum access load exceeds; that is, the number of users exceeds 100, the normalized throughput decreases significantly, because the number of users increases. The interference between users due to the frequency conflict leads to the decrease in the overall network throughput.

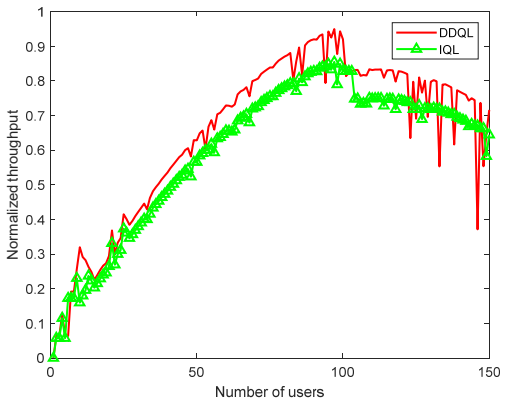


Figure 3. Normalized throughput versus number of users.

To investigate the effect of learning factors a on the convergence effect, we set the learning factors of the proposed algorithm DDQL and the independent Q-learning algorithm IQL to $\alpha = 0.2$ and $\alpha = 0.1$.As shown in Figure 4, in both algorithm DDQL and algorithm IQL, the learning factor $\alpha = 0.2$ converges faster than $\alpha = 0.1$. For the proposed algorithm DDQL, $\alpha = 0.2$, it converges after about 50 iterations, and for $\alpha = 0.1$, it converges after about 80 iterations. The independent Q-learning algorithm IQL, for $\alpha = 0.2$, converges after approximately 170 iterations, and for $\alpha = 0.1$, the algorithm converges after approximately 200 iterations.

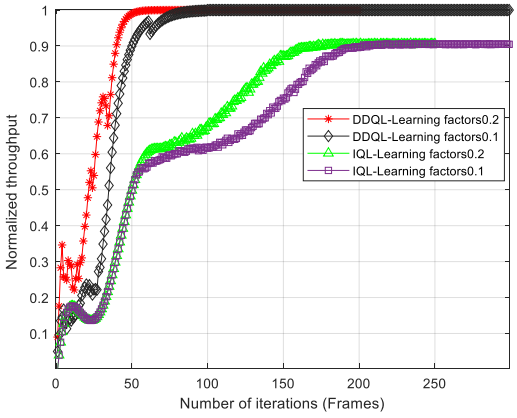


Figure 4. Relationship between normalized throughput and learning factor.

We verify the convergence of the proposed algorithm through simulation. Figure 5 shows that the normalized throughput of the proposed DDQL can converge to the optimal

value of 1 at the 50th iteration, which indicates that users can still communicate normally in the environment of multi-channel intelligent blocking jamming. Independent of binary Q-learning (IQL) at about 160 times convergence, convergence to 0.9, compared with the proposed DDQL gap in the convergence and reliability of the algorithm, occurs mainly because the classical Q-learning algorithm can interact with the environment, and trial and error to effectively avoid multi-channel jamming, but because each user is an independent subject, it does not interact with other users, so there is no way to avoid interference between users; the algorithm is based on perception and random channel effect is worse. Throughput has nothing to do with the number of iterations and convergence, the main reason being that users cannot perceive the transmission signals of other users; at the same time it will not be able to eliminate mutual interference due to the user frequency: random selection after the encountered jamming in channel transmission strategy cannot be efficient due to multichannel random disturbance. Therefore, it has the worst anti-interference effect and cannot guarantee the normal communication of users.

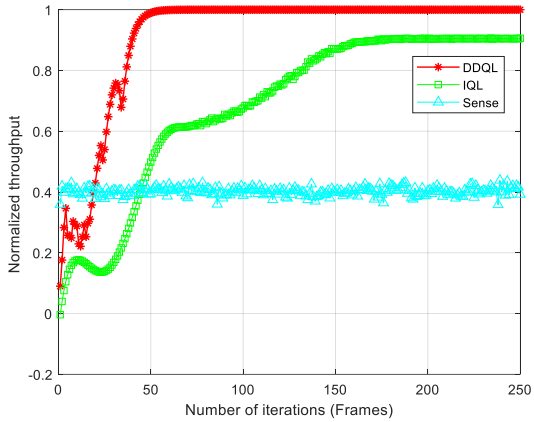


Figure 5. Relationship between normalized throughput and iteration times.

5. Conclusions

In this paper, in order to solve the problem of channel conflicts and external malicious jamming caused by random access in wireless communication networks, we derive a throughput optimization problem to find the optimal selection of subframes and channel policies for each user. In order to solve this optimization problem, a distributed double Q learning algorithm is proposed. Each user learns the subframe and channel selection according to the received congestion level using two Q-learning, respectively. Compared with the existing algorithms, the convergence performance and robustness of the proposed scheme are improved.

Author Contributions: Conceptualization, G.Z. and PL.; methodology, Y.N.; software, Y.N.; validation, Y.L. and Y.N.; formal analysis, G.Z.; resources, G.Z.; data curation, Y.L.; writing—original draft preparation, Y.N.; writing—review and editing, PL. and Y.L.; visualization, Y.N.; supervision, G.Z. and L.Z.; project administration, L.Z.; funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by National Science Foundation of China under contract No. U19B2014.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available from the author, G.Z., upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alsabah, M.; Naser, M.A.; Mahmmod, B.M. 6G Wireless Communications Networks: A Comprehensive Survey. *IEEE Access* **2021**, *9*, 2169–3536. [CrossRef]
2. Klaus, W.; Puttnam, B.J.; Luis, R.S.; Sakaguchi, J.; Mendinueta, J.D.; Awaji, Y.; Wada, N. Advanced space division multiplexing technologies for optical networks [Invited]. *J. Opt. Commun. Netw.* **2017**, *9*, C1–C11. [CrossRef]
3. Li, S.; Hou, Y.T.; Lou, W.; Jalaian, B.A.; Russell, S. Maximizing Energy Efficiency with Channel Uncertainty under Mutual Interference. *IEEE Trans. Wirel. Commun.* **2022**, *21*, 1276–1536. [CrossRef]
4. Lin, J.; Tian, B.; Wu, J.; He, J. Spectrum Resource Trading and Radio Management Data Sharing Based on Blockchain. In Proceedings of the 2020 IEEE 3rd International Conference on Information Systems and Computer Aided Education (ICISCAE), Dalian, China, 27–29 September 2020; pp. 83–87.
5. Zou, Y.; Zhu, J.; Wang, X.; Hanzo, L. A Survey on Wireless Security: Technical Challenges, Recent Advances, and Future Trends. *Proc. IEEE* **2016**, *104*, 1727–1765. [CrossRef]
6. Adil, M.; Khan, R.; Ali, J. An Energy Proficient Load Balancing Routing Scheme for Wireless Sensor Networks to Maximize Their Lifespan in an Operational Environment. *IEEE Access* **2020**, *8*, 163209–163224. [CrossRef]
7. Ram, S.S.; Ghatak, G. Optimization of Network Throughput of Joint Radar Communication System Using Stochastic Geometry. *Front. Sig. Proc.* **2022**, *2*, 835743. [CrossRef]
8. Zhang, S.; Li, M.; Jian, M. AIRIS:Artificial Intelligence Enhanced Signal Processing in Reconfigurable Intelligent Surface Communications. *China Commun.* **2021**, *18*, 1276–1536. [CrossRef]
9. Noels, N.; Moeneclaey, M. Performance of advanced telecommand frame synchronizer under pulsed jamming conditions. In Proceedings of the IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017.
10. Hall, M.; Silvennoinen, A.; Haggman, S. Effect of pulse jamming on IEEE 802.11 wireless LAN performance. In Proceedings of the MILCOM 2005—2005 IEEE Military Communications Conference, Atlantic City, NJ, USA, 17–20 October 2005; pp. 2301–2306.
11. Hwang, K.; Chen, M.; Gharavi, H. Artificial Intelligence for Cognitive Wireless Communications (Editorial). *IEEE Wirel. Commun.* **2019**, *26*, 1284–1536. [CrossRef]
12. Wang, J.; Jiang, C.; Zhang, H.; Ren, Y.; Chen, K.; Hanzo, L. Pareto-Optimal Wireless Networks. *IEEE Commun. Surv. Tutor.* **2020**, *22*, 1472–1514. [CrossRef]
13. Zhou, Z. Machine Learning. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 10–12 June 2008; pp. 1247–1250.
14. Sutton, R.S.; Barto, A.G. *Reinforcement Learning*; Electronic Industry Press: Beijing, China, 2019.
15. Galindo-Serrano, A.; Giupponi, L. Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks. *IEEE Trans. Veh. Technol.* **2010**, *59*, 1823–1834. [CrossRef]
16. Sharma, S.K.; Wang, X. Collaborative Distributed Q-Learning for RACH Congestion Minimization in Cellular IoT Networks. *IEEE Commun. Lett.* **2019**, *23*, 600–603. [CrossRef]
17. Aref, M.A.; Jayaweera, S.K. A novel cognitive anti-jamming stochastic game. In Proceedings of the Cognitive Communications for Aerospace Applications Workshop (CCAA), Cleveland, OH, USA, 27–28 June 2017; pp. 1–4.
18. Zhou, Q.; Li, Y.; Niu, Y.; Qin, Z.; Zhao, L.; Wang, J. "One Plus One Is Greater Than Two": Defeating Intelligent Dynamic Jamming with Collaborative Multi-Agent Reinforcement Learning. In Proceedings of the International Conference on Computer and Communications (ICCC), Chengdu, China, 11–14 December 2020.
19. Littman, M.L. Value-function reinforcement learning in Markov games. *Cogn. Syst. Res.* **2001**, *2*, 55–66. [CrossRef]
20. Aref, M.A.; Jayaweera, S.K.; Machuzak, S. Multi-agent reinforcement learning based cognitive anti-jamming. In Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, USA, 19–22 March 2017; pp. 1–6.
21. Machuzak, S.; Jayaweera, S.K. Reinforcement learning based anti-jamming with wideband autonomous cognitive radios. In Proceedings of the IEEE International Conference on Communications in China (ICCC), Chengdu, China, 27–29 July 2016; pp. 1–5.
22. Su, J.; Ren, G. An SCMA-Based Decoupled Distributed Q-Learning Random Access Scheme for Machine-Type Communication. *IEEE Commun. Lett.* **2021**, *10*, 1737–1741. [CrossRef]
23. Zhou, Q.; Li, Y.; Niu, Y. Intelligent Anti-Jamming Communication for Wireless Sensor Networks: A Multi-Agent Reinforcement Learning Approach. *IEEE Commun. Lett.* **2021**, *2*, 775–784. [CrossRef]
24. Wang, J. *Optimization Theory and Methods*; Beijing University of Technology Press: Beijing, China, 2018.
25. Vlassis, N. *A Concise Introduction to Multiagent Systems and Distributed Artificial Intelligence*; Morgan and Claypool Publishers: Williston, VT, USA, 2007.
26. Zhong, A. Preamble Design and Collision Resolution in a Massive Access IoT System. *Sensors* **2021**, *21*, 250. [CrossRef] [PubMed]