IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 11, 2019, accepted December 29, 2019, date of publication January 3, 2020, date of current version January 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2963751

Intelligent Resource Allocation for Train-to-Train Communication: A Multi-Agent Deep Reinforcement Learning Approach

JUNHUI ZHAO[1,2], (Senior Member, IEEE), YANG ZHANG[1], YIWEN NIE[1], AND JIN LIU[1]
[1]School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
[2]School of Information Engineering, East China Jiaotong University, Nanchang 330013, China

Corresponding author: Junhui Zhao (junhuizhao@hotmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61661021, in part by the National Science and Technology Major Project of the Ministry of Science and Technology of China under Grant 2016ZX03001014-006, in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University, under Grant 2017D14, in part by the Jiangxi Provincial Cultivation Program for Academic and Technical Leaders of Major Subjects under Grant 20172BCB22016, in part by the Key Technology Research and Development Program of Jiangxi Province under Grant 20171BBE50057, and in part by the Beijing Natural Science Foundation under Grant L182018.

ABSTRACT The application of train-to-train (T2T) communication in urban rail transit is expected to simplify system structure, reduce maintenance costs, and improve operational efficiency. In particular, train-to-wayside (T2W) communication coexist with T2T communication in the train control system based on T2T communication. To make full use of limited spectrum resources, frequency reuse is adopted as an efficient technique, but it brings the co-channel interference unfortunately, which affects the quality of service (QoS) for T2T and T2W users. In this paper, we propose a multi-agent deep reinforcement learning (MADRL) based autonomous channel selection and transmission power selection algorithm for T2T communication to reduce the co-channel interference. Specifically, each agent interacts with the environment and selects actions to implement a distributed resource allocation mechanism independently, adopting asynchronous updates to avoid different agents choosing the same sub-band. Simulation results show the superiority of our proposed algorithm: compared with the existing resource allocation schemes for T2T communication, the system throughput and the successful transmission probability of T2T links are greatly improved.

INDEX TERMS Train-to-train (T2T) communication, resource allocation, multi-agent deep reinforcement learning (MADRL), urban rail transit.

I. INTRODUCTION

With the continuous expansion of urban scale and the pressure of rail transit increasing, efficient and safe rail transit is highly valued [1]. In the past decade, the communication-based train control (CBTC) system has been widely used for its punctuality and higher operational efficiency [2], [3]. However, key functions such as train route and safety protection are based on bidirectional train-to-wayside (T2W) communication structure in CBTC systems, which bring many problems such as multiple configuration equipment and complicated system structure [4], [5]. Reliable direct train-to-train (T2T) communication can significantly improve efficiency and safety of train operation, and reduce wayside equipment in the train control system [6], [7]. Then, T2T technology was proposed and applied to the train control system, which is regarded as the next generation train control system [7]–[9].

Related research on T2T communication technology has been carried out widely. The channels for T2T communication at different frequencies were measured and modelled in [6], [10]–[13], which is the foundation for further research on T2T technology. The authors of [14] designed a novel CBTC system based on T2T communication, and proposed the local security certification and cooperative security check scheme to detect and against Sybil attacks. In [15], the CBTC data communication system based on T2T communication was proposed, and the reliability of the system was evaluated. In [16], the millimeter wave (mmWave) band was applied to

The associate editor coordinating the review of this manuscript and approving it for publication was Tony Thomas.

T2T communication, and the authors studied the alignment of narrow beams between trains in turning scenes. The authors of [17] studied the switch control function of the CBTC system based on T2T communication.

Although the T2T communication based CBTC system has many advantages, the wayside equipment in the system is still necessary. While the two adjacent trains acquire each other's position and status information through the T2T link, the train also needs to communicate with the wayside equipment. By multiplexing the frequency resources of the T2W uplink in T2T link, spectrum utilization can be improved effectively. However, it also produces co-channel interference in the system. Therefore, an effective resource allocation scheme is required to manage the interference [18]–[20]. The authors of [21] proposed a bio-inspired algorithm to achieve distributed channel allocation, which could effectively increase system throughput and reduce communication delay. To improve channel utilization and system performance, the authors of [22], [23] proposed a novel distributed channel allocation algorithm and a evolutionary scheme (named E-MAC) to achieve collision-free transmissions. The authors of [24] proposed a power control algorithm based on statistical-feature, which could reduce the average D2D transmit power and increase the energy efficiency of D2D communications in the cellular. The authors of [25] designed a mean-field game (MFG) theoretic framework and achieved a novel distributed power control scheme within the MFG framework. Notice that, none of the above works involved machine learning algorithms.

The reinforcement learning (RL) based resource allocation schemes have been applied to device-to-device (D2D) communication widely. In [26], a Q-learning based power control algorithm was proposed which decorrelated the actions selected by users and expand the solution space, and it had higher quality of service (QoS) than the schemes based on correlated Q-learning. In [27], two RL based power control methods were proposed, i.e., centralized method and distributed method. The simulation results showed that the distributed method had better system performance. In [28], a distributed learning based spectrum allocation scheme was proposed, which could maximize system throughput and spectral efficiency. However, in the above schemes, power control and channel selection were realized separately. In [29] and [30], new methods were proposed to solve this defect. In [29], an actor-critic RL based on policy gradient was proposed to improve D2D throughput and system throughput. In [30], a novel Bayesian (RL) model was proposed, and Bayesian RL-based coalition formation algorithms were implemented in a long-term evolution advanced network.

Recently, multi-agent RL has been gradually applied to wireless networks for its excellent performance and efficient implementation of distributed mechanisms. In [31], a collaborative multi-agent RL anti-jamming algorithm was proposed to solve the problem of external malicious jamming and mutual interference among users. An autonomous channel selection scheme based on multi-agent RL was proposed in [32], which could accelerate the convergence speed of the algorithm as well as improve the throughput of the system. In [33], a multi-agent deep reinforcement learning (MADRL) based distributed dynamic power allocation scheme was proposed, which achieved near-optimal power allocation. In [34], a MADRL method was adopted to realize cooperative spectrum sensing. Compared with traditional RL methods, the proposed algorithm had advantages in both the convergence speed and the reward performance.

However, the machine learning based resource allocation scheme for T2T communication is still scarce. In [35], Stackelberg game was proposed for power control, and weight factors based on proportional fairness were introduced for channel selection, which realized the resource allocation in the T2T scenario. The scheme can improve the throughput of the system and ensure the stability of the T2T communication. However, in this scheme, the system model has some disadvantages, e.g., the resource of one T2W uplink can only be multiplexed by one T2T link.

In this paper, we design a novel CBTC system structure based on T2T communication with Long Term Evolution for Metro (LTE-M), since Beijing Yanfang urban railway has already adopted LTE-M to transmit CBTC traffic [9]. Then, MADRL is adopted to the T2T scenario for the first time, and a novel distributed resource management scheme is realized. Specifically, in the proposed scheme, each T2T transmitter is regarded as an agent. Through interaction between agents and environment, each agent obtains state information, including resource block (RB) reuse and channel state, etc. According to the policy, each agent chooses actions, including power selection and RB selection. Compared with random allocation scheme and existing resource allocation scheme for T2T communication, the proposed scheme can effectively improve the throughput of the system, and improve the successful transmission probability of T2T links within the specified time.

The remainder of this paper is organized as follows. Section II briefly introduces the system model and formulates the resource allocation problem in the T2T scenario. In Section III, we describe the MADRL based resource allocation algorithm for T2T communication. The performances of the proposed schemes are simulated and compared in Section IV. Finally, we conclude the paper in Section V.

Notation: $\mathbf{a}$, a and $\mathcal{A}$ represent a vector, a scalar and a set, respectively; $|\mathcal{A}|$ denotes the size of set $\mathcal{A}$; $\mathbb{R}^n$ stands for the set of n-vector real numbers; $\mathbb{E}[\cdot]$ denotes the expectation.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. T2T COMMUNICATION BASED CBTC SYSTEM

A novel CBTC system based on T2T communication with LTE-M is shown in Fig. 1. The CBTC data communication system is mainly composed of Evolved Packet Core (EPC) of LTE-M systems, base station (BS), and terminal equipments. Compared with the traditional CBTC system, the functions of computer interlocking (CI) and zone controller (ZC) are

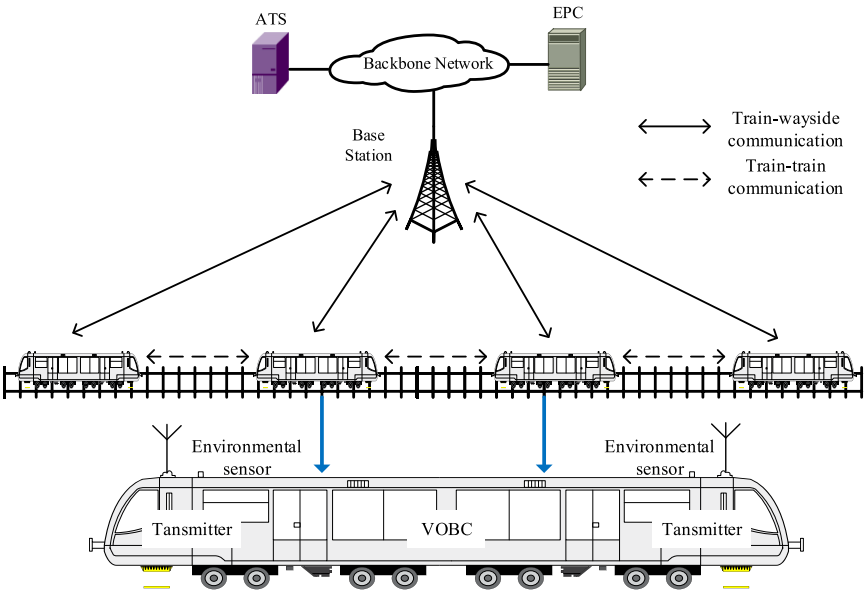

FIGURE 1. A novel CBTC system structure based on T2T communication with LTE-M.

integrated into trains and trackside controllers. In this novel CBTC system, the train becomes more "core" and "intelligent". Automatic train supervision (ATS) system sends the routing plan to vehicle on board controller (VOBC) by T2W communication, then VOBC can straightforwardly control the rotation and opening of the turnout according to the routing plan [3]. By adopting D2D communication technology, adjacent trains can directly perform T2T communication and exchange the key information such as train position and speed with each other. Based on the key information, the train can timely generate updated movement authority (MA) without the assistance of ZC or other equipments. There is a transmitter with the environmental sensor at the front and rear of the train to better support the T2T communication. The environmental sensor can obtain "environmental state information" such as instantaneous channel state information (CSI) and interference power, etc. The main function of the environmental sensor will be introduced in Section III.

The novel design can not only effectively reduce the communication processes between trains and improve the performance of the entire system, but also simplify the system structure.

B. PROBLEM FORMULATION

The T2T communication scenario in a single cell is shown in Fig. 2. In this scenario, we assume that the total number of RBs in the system is M, where M is the maximum number of trains in the area covered by the cell. The RBs are orthogonal to each other. Moreover, N trains require to establish N T2W uplinks to communicate from train to wayside equipment, and each link denotes as $n \in \mathcal{W} = \{1, 2, \ldots, N\}$. Each T2W link uses one RB, and the RBs are different from each

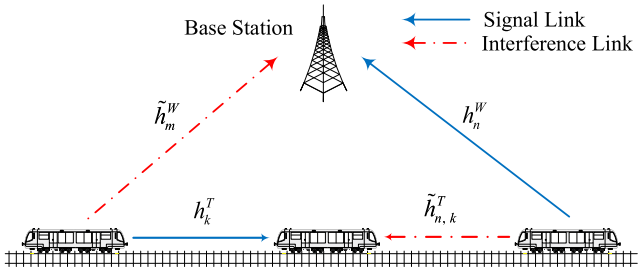

FIGURE 2. Co-channel interference caused by T2T communication and T2W communication.

other. Furthermore, each train requires to interact with the two adjacent trains to attain the location and state information of the adjacent train. So there are K T2T links denoted by $\mathcal{T} = \{1, 2, \ldots, K\}$, and K is twice as much as the number of trains. In particular, the first and last trains, only one train adjacent to them, and they still establish two T2T links with the adjacent train for redundant transmission. The anti-interference ability at the BS is stronger compared with that at the train, and the available spectrum resources for wireless communication are limited. Therefore, each T2T link reuses the orthogonal spectrum resource of the T2W uplink, and the same RB can be reused by multiple T2T links at the same time slot. When different links in the system use the same RB, co-channel interference (i.e., collision transmissions) occurs. The co-channel interference will affect the throughput and performance of the system.

The signal to interference plus noise ratio (SINR) of the nth T2W user at the BS can be expressed as

$$\gamma^W[n] = \frac{P_n^W h_n^W}{\sigma^2 + \sum_{k \in \mathcal{T}} \rho_k[n] P_k^T \tilde{h}_k^W}, \tag{1}$$

where P_n^W is the transmission power of the nth T2W user, h_n^W is the channel gain of the useful signal corresponding to the nth T2W user, and σ^2 is the noise power. $\rho_k[n] = 1$ when the kth T2T user has reused the frequency resource of the nth T2W user and $\rho_k[n] = 0$ otherwise. P_k^T, $\widetilde{h}_k^W$ is the transmission power of the kth T2T user and the channel gain of the interference to the BS, respectively. According to the Shannon theorem, the throughput of the nth T2W user can be formulated as

$$C^W[n] = Blog_2(1 + \gamma^W[n]), \tag{2}$$

where B is the bandwidth. The co-channel interference caused by reusing the same frequency resource between the T2T user and the T2W user is

$$I_W = \sum_{n \in \mathcal{W}} \rho_k[n] P_n^W \widetilde{h}_{n,k}^T, \tag{3}$$

and the co-channel interference among all T2T users which use the same RB is

$$I_T = \sum_{n \in \mathcal{W}} \sum_{k' \in \mathcal{T}, k' \neq k} \rho_k[n] \rho_{k'}[n] P_{k'}^T \widetilde{h}_{k',k}^T, \tag{4}$$

hence, the SINR of the kth T2T user can be expressed as

$$\gamma^T[k] = \frac{P_k^T h_k^T}{\sigma^2 + I_W + I_T}, \tag{5}$$

where, h_k^T is the channel gain of the useful signal corresponding to the kth T2T user, and $\widetilde{h}_{n,k}^T$ is the channel gain of the interference from the nth T2W or T2T user to the kth T2T user. The throughput of the kth T2T user can be expressed as

$$C^T[k] = Blog_2(1 + \gamma^T[k]). \tag{6}$$

In this system, each T2T transmitter is regarded as an agent. Each agent chooses transmission power and RB by interacting with the environment. By designing an appropriate reward function, our proposed scheme can maximize system throughput and improve the reliability of information transmission in each T2T link. In order to ensure the safe operation of trains, the position and status of trains need to be transmitted periodically between adjacent trains, so the reliability of the T2T link is particularly important. To evaluate the reliability of the T2T links, we define the successful transmission probability of T2T links. The information transmission is considered to be unsuccessful if the T2T link fails to transmit the required information within the specified time. More details will be discussed in Section III.

III. MULTI-AGENT DEEP REINFORCEMENT LEARNING FOR T2T RESOURCE ALLOCATION

MADRL can effectively implement a distributed resource allocation mechanism. Deep RL is a combination of deep learning and RL [36]. Deep learning is used to solve modelling problems between value function and policy, and RL is used to define problems and optimize goals. This section will be divided into two parts. The first part introduces RL. For the resource management in the T2T scenario, the basic elements of the RL model are designed, including state space $\mathcal{S}$, action space $\mathcal{A}$, policy π and reward function R. The second part introduces the deep Q-network (DQN) and multi-agent deep Q-network (MADQN) algorithm, which solve the mapping relationship between observation and value functions, and finds the optimization policy.

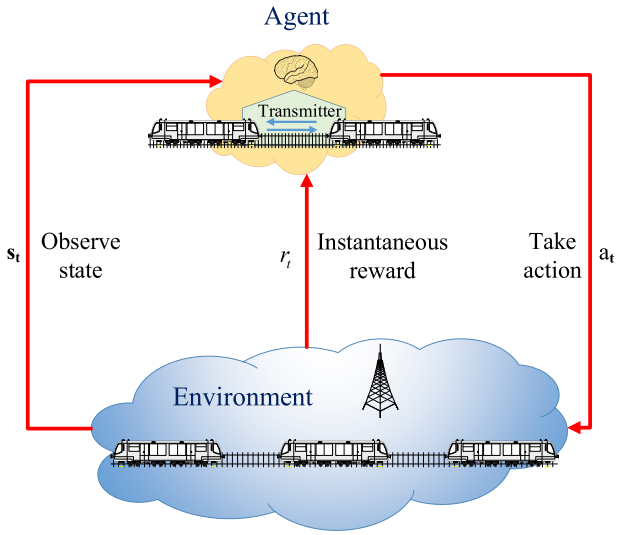

FIGURE 3. The framework of reinforcement learning.

A. REINFORCEMENT LEARNING

As shown in Fig. 3, the framework of RL is composed of two parts: agent and environment, which can interact with each other. In the process of interaction, the agent can continuously learn and ultimately complete the learning task. The key elements of the RL model are designed as follows:

- **States:** For the T2T resource allocation management, the agent can sense the external environment and generate its states $\mathbf{s}_t$ based on the onboard environment sensor [37], and $\mathbf{s}_t$ consists of six parts:

$$\mathbf{s}_t = \{G_t, H_t, I_{t-1}, D_{t-1}, E_t, F_t\}, \tag{7}$$

  where, $G_t \in \mathbb{R}^M$ and $H_m \in \mathbb{R}^M$ are channel gains of the T2T links and the T2W links at current time slot t respectively, $I_{t-1} \in \mathbb{R}^M$ and $D_{t-1} \in \mathbb{R}^M$ are interference power and times of the RBs being reused by adjacent agents at the previous time slot, respectively. E_t and F_t denote transmission duration and load quantity. At different time slots, the states observed by the agent will be different, all possible states constitute the state space $\mathcal{S}$.
- **Actions:** At time slot t, the agent observes the state $\mathbf{s}_t$ from the environment and selects the action $\mathbf{a}_t$, $\mathbf{a}_t \in \mathcal{A}$, according to the policy π. Policy π is a mapping function from state space $\mathcal{S}$ to action space $\mathcal{A}$, which determines the action selection in state $\mathbf{s}_t$. $\mathbf{a}_t$ includes the selection of the RBs to reuse and the transmit power level, which

can be expressed as

$$\mathbf{a}_t = \{RB_t, P_t\}. \tag{8}$$

As mentioned in Section II, the total number of RBs in the system is M, hence, $RB_t \in \{1, 2, \ldots, M\}$. Considering the complexity of the DQN network and T2T user requirements comprehensively, three levels of transmission power is adopted, and $P_t \in \{P_1, P_2, P_3\}$. Therefore, the size of the action space $|\mathcal{A}|$ (i.e., number of different actions) can be formulated as

$$|\mathcal{A}| = 3M. \tag{9}$$

After the agent takes action $\mathbf{a}_t$, it will act on the environment. The state of the environment becomes $\mathbf{s}_{t+1}$ from $\mathbf{s}_t$, and an instant reward r_{t+1} feeds back to the agent. Such interaction can go on like this:

$$\mathbf{s}_0, \mathbf{a}_0, r_1, \mathbf{s}_1, \mathbf{a}_1, \ldots, r_{t-1}, \mathbf{s}_{t-1}, \mathbf{a}_{t-1}, r_t, \mathbf{s}_t. \tag{10}$$

- **Reward Function:** To recognize the impact of the selected action on the system, we define the reward function as the weighted total throughput of the T2T and T2W links, rather than the throughput of the link related to the agent. The instant reward r_t is expressed as

$$r_t = \lambda \sum_{n \in \mathcal{W}} C^W[n] + (1-\lambda) \sum_{k \in \mathcal{T}} C^T[k], \tag{11}$$

where $\lambda \in [0, 1]$ is the weight factor. In RL, besides instant reward, a total reward should be considered to ensure the stability of long-term performance of the system. In the T2T scenario, the environment has no termination state, and the total reward will be infinite. To solve this problem, the discount rate γ is introduced to control the weight of the long-term reward, and the discount reward r'_t is defined as

$$r'_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i}, \tag{12}$$

where $\gamma \in [0, 1]$, the agent is more concerned about long-term rewards when γ approaches 1, and the current reward becomes more important when γ approaches 0. The target of RL is to learn a policy to maximize the expected discount reward, which can be defined as

$$R_t = \mathbb{E}[r'_t]. \tag{13}$$

The performance of the system is controlled by designing the reward function.

B. MULTI-AGENT DEEP Q-LEARNING

Many effective algorithms have been proposed to achieve the target of RL, Q-learning is one of the commonly used algorithms. For a policy π, Q-learning optimizes policy π by Q-value. The Q-value is closely related to the state $\mathbf{s}_t$ and the selected action $\mathbf{a}_t$, denoted as $Q(\mathbf{s}_t, \mathbf{a}_t)$. It can be approximated as the expected total reward of the agent selecting the action $\mathbf{a}_t$ in the state $\mathbf{s}_t$. The action with the highest Q-value is selected to update the policy π, then the Q-value is updated with the new policy, and repeat this process until Q-value converges to the optimal Q-value, Q^*. The optimal policy π^* can be found, once the Q^* is obtained. The iteration formula of Q-value is as follows

$$Q(\mathbf{s}, \mathbf{a}) \leftarrow Q(\mathbf{s}, \mathbf{a}) + \alpha \left(r + \gamma \max_{\mathbf{a}'} Q(\mathbf{s}', \mathbf{a}') - Q(\mathbf{s}, \mathbf{a}) \right), \tag{14}$$

where α is the learning rate. In Q-learning, the Q-value is stored in the Q-table, and the size of the Q-table is $|\mathcal{A}|^{|\mathcal{S}|}$. As the state-action space increases, the size of the Q-table will increase dramatically. In the resource allocation for the T2T communication, the state space $|\mathcal{S}|$ is large and uncertain, so the classic Q-learning cannot be applied. This problem can be solved well by using the neural network. As shown in Fig. 4, the observed state is regarded as the input of the neural network, and the neural network outputs the Q-value of each action. The Q-table is replaced by the neural network which can be called as Q-network.

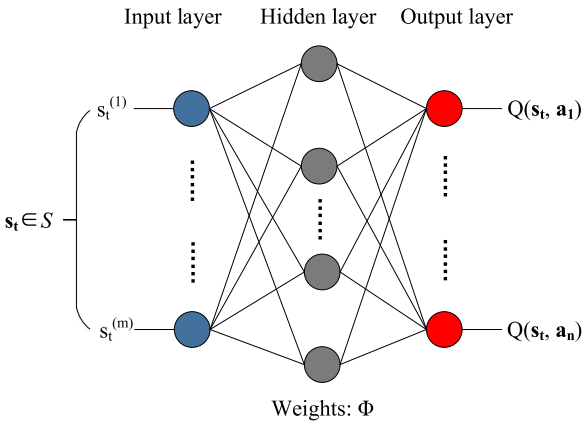

FIGURE 4. Structure of the deep Q-network.

In the resource allocation problem that we proposed, action $\mathbf{a}_t \in \mathcal{A}$ is discrete and finite. The output of Q-network can be expressed as:

$$Q_\phi(\mathbf{s}_t) = \begin{bmatrix} Q_\phi(\mathbf{s}_t, \mathbf{a}_1) \\ \vdots \\ Q_\phi(\mathbf{s}_t, \mathbf{a}_n) \end{bmatrix}, \tag{15}$$

where ϕ denote the weights in the Q-network and is learned to ensure $Q_\phi(\mathbf{s}_t)$ close to the real Q-value. There are two problems in the process of learning: one is that the target is unstable, and the goal of parameter learning depends on the parameter itself; the other is that there is a strong correlation between the samples. To solve the two problems, DQN was proposed. The DQN takes two measures: one is the freezing target network, i.e., the parameters in the target network are fixed in a period to stabilize the learning target, and the second is experience replay, an experience pool is built to remove data dependencies [38]. To solve the T2T resource allocation problem proposed in this paper, we adopt the MADQN algorithm, i.e., there are multiple agents which

Algorithm 1 MADQN for T2T Resource Allocation

Input: State space $\mathcal{S}$, action space $\mathcal{A}$, discount rate γ, learning rate α

Output: Multi-agent deep Q-network

Initialize replay memory D to capacity N;
Initialize $Q_k(\mathbf{s}, \mathbf{a})$ for each $k \in \mathcal{T}$;
Randomly initialize the weights ϕ of the Q-network;
Randomly initialize the weights of the target Q-network $\hat{\phi} = \phi$;
for episode $= 1 : j$ **do**
  Initialize state $\mathbf{s}_k$ for each $k \in \mathcal{T}$;
  for step $= 1 : i$ **do**
    for $k \in \mathcal{T}$ **do**
      In state $\mathbf{s}_k$, select action $\mathbf{a}_k$ with policy π;
      Take action $\mathbf{a}_k$, observe the reward r_k and a new state $\mathbf{s}'_k$;
      Save $\mathbf{s}_k, \mathbf{a}_k, r_k, \mathbf{s}'_k$ into D;
      Sample $\mathbf{ss}$, $\mathbf{aa}$, rr, $\mathbf{s}'\mathbf{s}'$ from D;
      $y = rr + \gamma \max_{\mathbf{a}'} Q_{\hat{\phi}}(\mathbf{s}'\mathbf{s}', \mathbf{a}')$;
      Train the multi-agent deep Q-network with the loss function
      $\text{Loss}(\phi) = (y - Q_{\phi}(\mathbf{ss}, \mathbf{aa}))^2$;
      $\mathbf{s}_k \leftarrow \mathbf{s}'_k$;
    Every C steps $\hat{\phi} \leftarrow \phi$;
return: Multi-agent deep Q-network with weights ϕ.

select actions with DQN independently. The learning process of the MADQN is described in Algorithm 1.

In the process of MADQN training, to make the agent explore the environment sufficiently, ϵ-greedy method is adopted, i.e., the agent selects the action which has the largest Q value with probability $1-\epsilon$ and randomly selects the action from $\mathcal{A}$ with probability ϵ.

With the completion of the training, the Q-value converges, and the learning effect of the MADQN will be tested. Different from the training process, the ϵ-greedy method is not adopted in the testing stage. The action with the largest Q-value is directly selected to maximize total reward and improve the performance of the system. In the distributed resource allocation scheme, each agent cannot know the actions selected by other agents at current time slot, and multiple agents may reuse the same RB, thereby generating large interference, reducing the reward and failing to obtain a higher system performance. To solve this problem, only a few agents update their actions in each time slot, synchronous update becomes asynchronous update. At a different time, the impact of the action selected by the agent on the environment can be observed by other agents. For higher rewards, the reuse of the same RB by adjacent or multiple agents at the same time will be reduced or even avoided.

In summary, the MADQN algorithm is proposed in this paper, which can solve the resource management problem in the T2T communication scenario. The specific performance analysis is given in Section IV.

IV. SIMULATION ANALYSIS

In this section, detailed simulation parameters are given, and the simulation results are conducted to evaluate the performance of our proposed scheme.

A. SIMULATION PARAMETERS

For the MADQN, considering the number of inputs and outputs, a three-layer fully connected neural network is adopted, which consists of an input layer, a hidden layer, and an output layer, where the number of neurons in the hidden layer is set as 90. MADQN input size n_{in} is $4M + 2$ according to Equation (7). The output size n_{out} is equal to $|\mathcal{A}|$, and it can be seen in Equation (9). The number of neurons in the input layer and the output layer can be set once n_{in} and n_{out} are determined. In the training stage, the ϵ-greedy method with variable ϵ is adopted. At the beginning, the agent randomly selects the action with a high probability to constantly explore the environment and accumulate experience. With the number of training steps increasing, ϵ gradually decreases, which can effectively balance exploration and exploitation. More specifically, ϵ can be expressed as

$$\epsilon = \frac{1}{\frac{x}{b} + 1}, \tag{16}$$

where, x denotes the number of training steps, and b is a constant. With the value of b increasing, the agent will spend longer time to explore the environment. In this paper, we take the value of b as 5500. From Fig. 5, the effect of the training process can be seen. Fig. 5 shows the variation of the average reward in the system with the number of training steps increasing. At the beginning, the agent is still in the stage of environmental exploration, i.e., the agent chooses action randomly with a high probability, so that the action with low reward will also be selected. Therefore, the average reward is constantly fluctuating. When training steps reach about 7700, due to the probability of selecting action with the largest Q-value is already greater than the probability of randomly selecting the action, the average reward begins to rise continuously. Hence, the exploration of the environment by agents begins to decrease.

The detail parameters are shown in Table 1. The path loss models of T2T links and T2W links are from the real information in Beijing Yanfang subway line [9].

B. PERFORMANCE ANALYSIS

To evaluate our proposed scheme, we compare it with the scheme proposed by [35] and the random allocation scheme. In the first scheme, the channel selection was based on the weighting factor of proportional fairness, and the power control was performed by Stackelberg game. For simplification, the scheme is called as scheme I. The other scheme is randomly choosing an action for each agent.

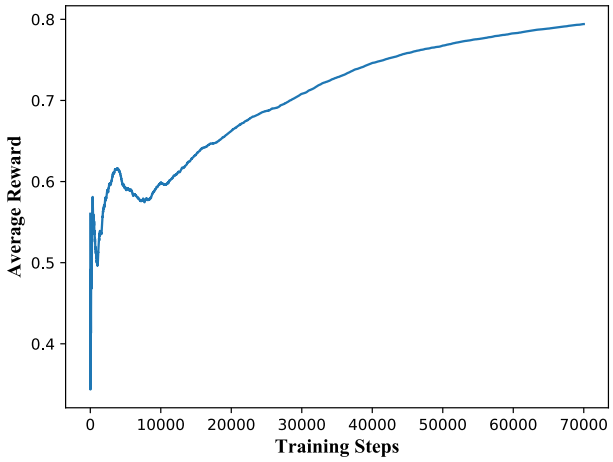

FIGURE 5. Average reward versus the training steps.

TABLE 1. Simulation parameters.

Parameter	Value
RB bandwidth B	1.5 MHz
Number of RBs M	7
BS coverage radius	3 km
BS antenna gain	7 dBi
BS receiver noise figure	5 dB
Train antenna gain	4 dBi
Train receiver noise figure	8 dB
Train speed	65 - 90 km/h
Distance between adjacent trains	600 - 900 m
T2W transmit power	43 dBm
T2T transmit power levels $\{P_1, P_2, P_3\}$	$\{8, 14, 23\}$ dBm
T2W link path loss model	37.6log10(d)+128.1 dB
T2T link path loss model	40log10(d) + 148 dB
Specified time of T2T links transmit	100 ms
Learning rate α	0.001
Discount rate γ	0.4
Weight factor λ	0.2

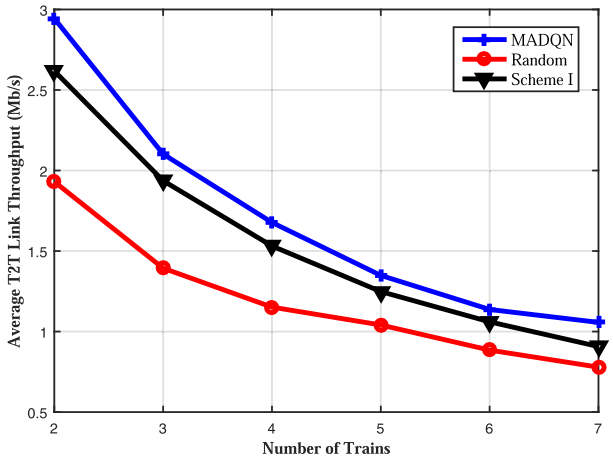

FIGURE 6. Average T2T link throughput as a function of the number of trains.

Fig. 6 shows the relationship between the average throughput of the T2T links and the number of trains. It can be seen that the proposed scheme effectively reduces the interference of the T2T link, and its throughput is higher than the other

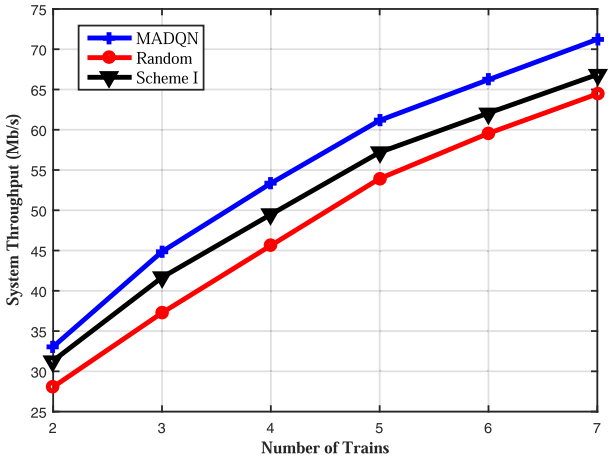

FIGURE 7. System throughput as a function of the number of trains.

two schemes in different train quantities. As the number of trains grows, more T2T links are established. Due to the limited quantity of available RBs, the interference from the T2W link to the T2T link, and among the different T2T links increases, which lead to a reduction in the average throughput of the T2T link. In detail, when the number of trains increases from 2 to 3, the average T2T link throughput decreases significantly. The specific reasons are as follows: Ideally, the co-channel interference could be eliminated when the number of train is 2, and the total number of links in the system is less than the number of available RBs. However, the co-channel interference is inevitable when we increase the number of trains to 3. Meanwhile, the total number of links is larger than the available RBs due to the increase number of trains.

Fig. 7 illustrates the total throughput of the system with respect to the number of trains in the system. From the simulation results, we can see that the proposed scheme can effectively increase the total throughput of the system, and the advantage of our scheme becomes more obvious as the number of trains increasing compared with the scheme I. As the number of trains increases, the total throughput of the system also increases. However, the increased T2T link and T2W link quantities bring more co-channel interference, which results in the slowdown of the increase rate of the system throughput.

Fig. 8 shows the successful transmission probability of T2T links as a function of the number of trains. It can be seen that our scheme has the highest transmission success rate, and with train quantities increasing, the transmission success rate decreases less, which can effectively guarantee the reliability of T2T links. It is because that the agent can attain the state of transmission during the process of interacting with the environment, which can increase the throughput of the T2T links while improving the successful transmission probability of T2T links.

Fig. 9 illustrates the probability for the agents to choose power levels. After training, the maximum transmission

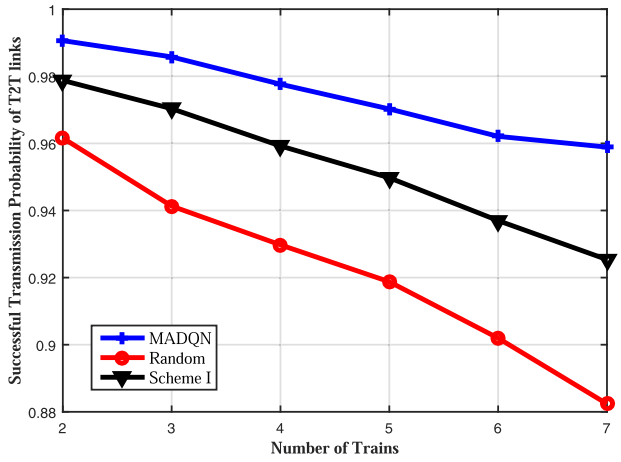

FIGURE 8. T2T information transmission success rate versus the number of trains.

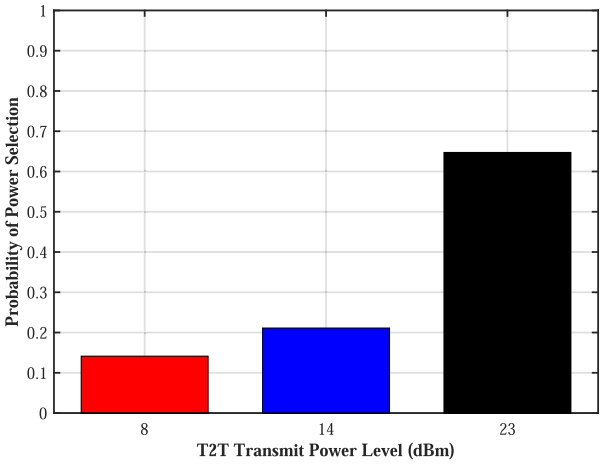

FIGURE 9. Probability of power selection.

power is selected with the highest probability. Combined with Fig. 6 and Fig. 7, it can be seen that in order to get more reward, the agent learns to select the maximum transmitting power to improve the throughput of the T2T link and learn to reduce the co-channel interference in the system effectively.

V. CONCLUSION

T2T communication is proposed in the next generation train control system, and the resource allocation problem is caused by the T2T links multiplexing the T2W uplinks spectrum resource. In this paper, we propose a distributed resource allocation scheme based on MADRL. Simulation results demonstrate that our scheme can effectively reduce the interference in the system. It can improve the throughput of T2T links and the system, and ensure the successful transmission probability of the T2T links within the specified time. Our scheme can play an important role in resource allocation for T2T communication.

REFERENCES

[1] J. Zhao, Y. Liu, Y. Gong, C. Wang, and L. Fan, "A dual–link soft handover scheme for C/U plane split network in high–speed railway," *IEEE Access*, vol. 6, pp. 12473–12482, 2018.
[2] L. Zhu, F. R. Yu, B. Ning, and T. Tang, "Communication–based train control (CBTC) systems with cooperative relaying: Design and performance analysis," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2162–2172, Jun. 2014.
[3] L. Zhu, D. Yao, and H. Zhao, "Reliability analysis of next–generation CBTC data communication systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2024–2034, Mar. 2019.
[4] H. Song and E. Schnieder, "Availability and performance analysis of train-to-train data communication system," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 7, pp. 2786–2795, Jul. 2019.
[5] J. Farooq and J. Soler, "Radio communication for communications–based train control (CBTC): A tutorial and survey," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1377–1402, 3rd Quart., 2017.
[6] A. Lehner, T. Strang, and P. Unterhuber, "Direct train-to-train communications at low UHF frequencies," *IET Microw., Antennas Propag.*, vol. 12, no. 4, pp. 486–491, Mar. 2017.
[7] J. Zhao, J. Liu, S. Ni, and Y. Gong, "Enhancing transmission on hybrid precoding based train-to-train communication," *Mobile Netw. Appl.*, no. 11, pp. 1–10, 2019.
[8] Y. Liu and L. Yuan, "Research on train control system based on train to train communication," in *Proc. Int. Conf. Intell. Rail Transp. (ICIRT)*, Dec. 2018, pp. 1–5.
[9] X. Wang, L. Liu, T. Tang, and W. Sun, "Enhancing communication–based train control systems through train-to-train communications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 4, pp. 1544–1561, Apr. 2019.
[10] A. Lehner, T. Strang, and P. Unterhuber, "Train-to-train propagation at 450 MHz," in *Proc. 11th Eur. Conf. Antennas Propag. (EUCAP)*, Mar. 2017, pp. 2875–2879.
[11] K. Guan, "Channel sounding and ray tracing for train-to-train communications at the THz band," in *Proc. 13th Eur. Conf. Antennas Propag. (EuCAP)*, Mar. 2019, pp. 1–5.
[12] C. Zhou and R. Jacksha, "Modeling and measurement of radio propagation in tunnel environments," *IEEE Antennas Wireless Propag. Lett.*, vol. 16, pp. 1431–1434, 2017.
[13] P. Unterhuber, S. Sand, U.-C. Fiebig, and B. Siebler, "Path loss models for train-to-train communications in typical high speed railway environments," *IET Microw., Antennas Propag.*, vol. 12, no. 4, pp. 492–500, Mar. 2018.
[14] X. Wang, L. Liu, L. Zhu, and T. Tang, "Joint security and QoS provisioning in train–centric CBTC systems under Sybil attacks," *IEEE Access*, vol. 7, pp. 91169–91182, 2019.
[15] D. Yao, L. Zhu, and H. Zhao, "Availability analysis of next generation CBTC data communication systems," in *Proc. Chin. Autom. Congr. (CAC)*, Oct. 2017pp. 3689–3694, .
[16] J. Zhao, J. Liu, Y. Nie, and S. Ni, "Location–assisted beam alignment for train-to-train communication in urban rail transit system," *IEEE Access*, vol. 7, pp. 80133–80145, 2019.
[17] X. Jikang, J. Sen, and X. Yan, "Switch control function of new CBTS system developed based on train-train communication," *Railway Signal. Commun. Eng.*, no. 3, p. 16, May 2017.
[18] J. Zhao, S. Ni, L. Yang, Z. Zhang, Y. Gong, and X. You, "Multiband cooperation for 5G hetnets: A promising network paradigm," *IEEE Veh. Technol. Mag.*, vol. 14, no. 4, pp. 85–93, Dec. 2019.
[19] Z. Junhui, Y. Tao, G. Yi, W. Jiao, and F. Lei, "Power control algorithm of cognitive radio based on non-cooperative game theory," *China Commun.*, vol. 10, no. 11, pp. 143–154, Nov. 2013.
[20] J. Zhao, X. Guan, and X. P. Li, "Power allocation based on genetic simulated annealing algorithm in cognitive radio networks," *Chin. J. Electron.*, vol. 22, no. 1, pp. 177–180, Jan. 2013.
[21] J. Li, H. Zhao, A. S. Hafid, J. Wei, H. Yin, and B. Ren, "A bio–inspired solution to cluster–based distributed spectrum allocation in high–density cognitive Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9294–9307, Dec. 2019.
[22] H. Zhao, K. Ding, N. I. Sarkar, J. Wei, and J. Xiong, "A simple distributed channel allocation algorithm for D2D communication Pairs," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10960–10969, Nov. 2018.
[23] H. Zhao, J. Wei, N. I. Sarkar, and S. Huang, "E–MAC: An evolutionary solution for collision avoidance in wireless ad hoc networks," *J. Netw. Comput. Appl.*, vol. 65, pp. 1–11, Apr. 2016.

[24] P. Sun, K. G. Shin, H. Zhang, and L. He, "Transmit power control for D2D-underlaid cellular networks based on statistical features," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4110–4119, May 2017.
[25] C. Yang, J. Li, P. Semasinghe, E. Hossain, S. M. Perlaza, and Z. Han, "Distributed interference and energy–aware power control for ultra–dense D2D networks: A mean field game," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1205–1217, Feb. 2017.
[26] S. Toumi, M. Hamdi, and M. Zaied, "An adaptive Q-learning approach to power control for D2D communications," in *Proc. Int. Conf. Adv. Syst. Electr. Technol. (IC_ASET)*, Mar. 2018, pp. 206–209.
[27] S. Nie, Z. Fan, M. Zhao, X. Gu, and L. Zhang, "Q-learning based power control algorithm for D2D communication," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–6.
[28] S. Sharma and B. Singh, "Weighted cooperative reinforcement learning-based energy-efficient autonomous resource selection strategy for underlay D2D communication," *IET Commun.*, vol. 13, no. 14, pp. 2078–2087, Aug. 2019.
[29] P. Khuntia and R. Hazra, "An actor–critic reinforcement learning for device-to-device communication underlaying cellular network," in *Proc. TENCON IEEE Region Conf.*, Oct. 2018, pp. 50–55.
[30] A. Asheralieva, "Bayesian reinforcement learning–based coalition formation for distributed resource sharing by device-to-device users in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5016–5032, Aug. 2017.
[31] F. Yao and L. Jia, "A collaborative multi–agent reinforcement learning anti–jamming algorithm in wireless networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1024–1027, Aug. 2019.
[32] K. Zia, N. Javed, M. N. Sial, S. Ahmed, and F. Pervez, "Multi–agent RL based user–centric spectrum allocation scheme in D2D enabled hetnets," in *Proc. IEEE 23rd Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Sep. 2018, pp. 1–6.
[33] Y. S. Nasir and D. Guo, "Multi–agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.
[34] Y. Zhang, P. Cai, C. Pan, and S. Zhang, "Multi–agent deep reinforcement learning–based cooperative spectrum sensing with upper confidence bound exploration," *IEEE Access*, vol. 7, pp. 118898–118906, 2019.
[35] Q. Zhou, X. Hu, J. Lin, and Z. Wu, "Train-to-train communication resource allocation scheme for train control system," in *Proc. 10th Int. Conf. Commun. Softw. Netw. (ICCSN)*, Jul. 2018, pp. 210–214.
[36] V. Mnih, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Jun. 2016, pp. 1928–1937.
[37] X. Wang, L. Liu, L. Zhu, and T. Tang, "Train-centric CBTC meets age of information in train-to-train communications," *IEEE Trans. Intell. Transp. Syst.*, to be published.
[38] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

JUNHUI ZHAO received the M.S. and Ph.D. degrees from Southeast University, Nanjing, China, in 1998 and 2004, respectively. From 1998 to 1999, he was with the Nanjing Institute of Engineers, ZTE Corporation. Then, he worked as an Assistant Professor with the Faculty of Information Technology, Macao University of Science and Technology, in 2004, where he was an Associate Professor, till 2007. In 2008, he joined Beijing Jiaotong University as an Associate Professor, where he is currently a Professor with the School of Electronics and Information Engineering. He was also a short-term Visiting Scholar with Yonsei University, South Korea, in 2004, and a Visiting Scholar with Nanyang Technological University, Singapore, from 2013 to 2014. Since 2016, he has been with the School of Information Engineering, East China Jiaotong University. His current research interests include wireless and mobile communications and the related applications, which contain 5G mobile communication technology, vehicle network communication, wireless localization, and cognitive radio.

YANG ZHANG received the B.Eng. degree in communication engineering from Beijing Jiaotong University, Beijing, China, in 2018, where he is currently pursuing the M.S. degree. His research interests include machine learning and rail transit.

YIWEN NIE received the B.Eng. degree in computer science from East China Jiaotong University, Nanchang, China, in 2017. He is currently pursuing the Ph.D. degree with Beijing Jiaotong University, Beijing. His research interests include machine learning and wireless communication.

JIN LIU received the B.Eng. degree in communication engineering from Beijing Jiaotong University, Beijing, China, in 2017, where she is currently pursuing the M.S. degree. Her research interests include LTE-R and rail transit.