Experience-Driven Power Allocation Using Multi-Agent Deep Reinforcement Learning for Millimeter-Wave High-Speed Railway Systems

Jianpeng Xu, *Graduate Student Member, IEEE*, and Bo Ai, *Senior Member, IEEE*

***Abstract*— Railway is stepping into the field of smart railway. Unfortunately, the challenge of obtaining accurate instantaneous channel state information in high-speed railway (HSR) scenario makes it difficult to apply conventional power allocation schemes. In this paper, to respond to the challenge, we propose an innovative experience-driven power allocation algorithm in the millimeter-wave (mmWave) HSR systems with hybrid beamforming, which is capable of learning power decisions from the past experience instead of the accurate mathematical model, just like one person learns one new skill, such as driving. To be specific, with the purpose of maximizing the achievable sum rate, we first characterize the power allocation problem of the mmWave HSR systems as a multi-agent deep reinforcement learning problem and then solve it by using emerging multi-agent deep deterministic policy gradient (MADDPG) approach, which enables the agent, i.e., the mobile relay onboard the train to learn the power decisions from the past experience in a distributed manner. The simulation results indicate that the spectral efficiency of proposed MADDPG algorithm significantly outperforms existing state-of-the-art schemes.**

***Index Terms*— High-speed railway (HSR), millimeter-wave communications, power allocation, multi-agent deep reinforcement learning.**

I. Introduction

HIGH-SPEED railway (HSR), as a green and convenient transport mode, is stepping into the field of smart railway, where more and more interconnection among infrastructures, passengers, and high-speed trains will be established [1], [2]. To realize this good wish, it is significant to provide high data transmission rate in HSR scenarios.

While there has been a lot of research to improve the transmission rate under the time-varying HSR channel conditions [3]–[8], they are actually idealistic and are challenging to be applied in practice. This is because all of them are model-based, which exhibit good performance with the following two conditions. Firstly, some key parameters are accurately acquired, such as exact instantaneous channel state information. Secondly, both HSR networks and user requirements can accurately be characterized in mathematical terms. However, it's hard to meet these two conditions in millimeter-wave (mmWave) HSR networks. Fortunately, artificial intelligence (AI), particularly deep reinforcement learning (DRL), has been proved to be an effective method to make clever decisions under the uncertain circumstances. The successful application of DRL technique in AlphaGo [9] has generated a lively interest in adopting DRL-based approach to handle problems in a proliferation of fields, such as mobile edge computing [10], [11], spectrum sharing [12], user association [13], multi-path TCP [14], cache [15] and so on. Nevertheless, there is little research to investigate a DRL-based algorithm to tackle power allocation problem in mmWave HSR systems. Therefore, this paper investigates the power allocation issue based on AI-assisted mmWave HSR network architecture shown in Fig. 1. An experience-driven method, i.e., multi-agent deep deterministic policy gradient (MADDPG) is developed, which makes the agent have the capacity of intelligently orchestrating power resource from the past experience to achieve maximum achievable sum rate (ASR), just like one person learns one new skill, e.g., driving. Note that a few relevant studies adopted the term "data-driven." However, compared to "data driven," the term "experience-driven" may be more deterministic. The reason is that a communication network includes three types of data, i.e., user data, control data as well as runtime statistics [14]. If the term "data-driven" is employed, we cannot obviously understand what type of data is referenced. In fact, runtime statistics denotes the past experience of the network. Consequently, "experience-driven" fundamentally represents that a network is able to learn the best control strategy based on its runtime statistics which are collected in the past.

In this paper, to the best of our knowledge, we are the first to address power allocation problem in the mmWave HSR

Manuscript received April 18, 2020; revised November 19, 2020; accepted January 15, 2021. Date of publication February 3, 2021; date of current version May 31, 2022. This work was supported in part by the National Key Research and Development Program under Grant 2016YFE0200900; in part by the Royal Society Newton Advanced Fellowship under Grant NA191006; in part by the NSFC under Grant 61725101, Grant 6196113039, and Grant U1834210; in part by the Major Projects of Beijing Municipal Science and Technology Commission under Grant Z181100003218010; in part by the State Key Lab of Rail Traffic Control and Safety under Grant RCS2018ZZ007, Grant RCS2020ZT010, and Grant RCS2019ZZ007; in part by the Fundamental Research Funds for the Central Universities under Grant 2020YJS201; in part by the Beijing Natural Haidian Joint Fund under Grant L172020; and in part by the Open Research Fund from Shenzhen Research Institute of Big Data under Grant 2019ORF01006. The Associate Editor for this article was V. Punzo. *(Corresponding author: Bo Ai.)*

The authors are with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing 100044, China, and also with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China (e-mail: 18111037@bjtu.edu.cn; boai@bjtu.edu.cn).

Digital Object Identifier 10.1109/TITS.2021.3054511

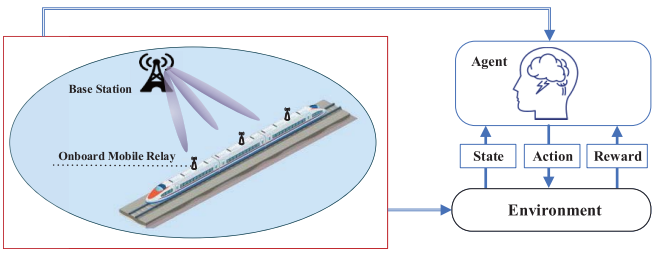


Fig. 1. AI-assisted mmWave HSR network architecture.

networks by leveraging DRL. Since the formulated problem is non-convex and it is difficult to deal with, we decompose the original optimization problem into a beamforming design sub-problem and a power allocation sub-problem. In the first sub-problem, we make the beamforming decisions at the transmitters (TXs) and receivers (RXs). Once the TX-RX beamforming design is given, the second sub-problem is solved by using the proposed multi-agent DRL power allocation algorithm. Specifically, our main contributions in this work can be summarized as follows.

- We investigate the joint beamforming and power allocation in the uplink mmWave HSR systems with low-resolution phase shifters (PSs). With the objective of maximizing ASR, we formulate the ASR maximization problem by jointly designing beamforming and power allocation.
- To address the formulated problem, we decompose the original optimization problem into two sub-problems as: 1) Hybrid beamforming (HBF) at RXs and analog beamforming (ABF) at TXs optimization; 2) Power allocation optimization. For the first sub-problem, after acquiring the optimal ABF design at TXs and RXs, it is easy to calculate the digital beamforming (DBF) at RXs according to the effective mmWave HSR channel.
- After the power allocation problem of the mmWave HSR systems being characterized as a multi-agent problem, we take advantage of recent progress of multi-agent DRL to propose an experience-driven power allocation algorithm to maximize ASR through meticulously designing state, action and reward.
- Simulation results demonstrate that by a meticulous training mechanism, the mobile relay (MR) can learn from the past power allocation experiences under various mmWave HSR channel environments and intelligently orchestrate a clever power allocation decision, while collaborating with each other in a distributed manner to improve mmWave HSR systems performance. Moreover, we also compare the performance of the proposed framework with three other algorithms, i.e., FP [Algorithm 3, 16], random power and maximal power allocation algorithms by carrying out extensive simulation experiments. Simulation results indicate that the proposed algorithm is capable of achieving higher spectral efficiency compared with the existing schemes.

The remainder of this paper is structured as follows. The related work is briefly reviewed in the next section. Both the system model of mmWave HSR systems and problem formulation are described in Section III. In Section IV, we discuss the beamforming design. We customize the power allocation approach by using multi-agent DRL to obtain the maximum ASR in Section V. Simulation results are shown in Section VI. Finally, the paper is summarized in Section VII.

Notation: We employ $\mathbf{A}$, $\mathbf{a}$, and a to indicate a matrice, a vector and a scalar, separately. $(\cdot)^T$, $(\cdot)^H$ and $(\cdot)^{-1}$ represent transpose, conjugate transpose, and inverse, whereas $|a|$ and $||\mathbf{a}||$ denote the magnitude of a scalar a and the Euclidean norm of vector $\mathbf{a}$, respectively. $[\mathbf{A}]_{p,q}$, $[\mathbf{A}]_{:,q}$ and $[\mathbf{A}]_{p,:}$ denote the entry in the p-th row and qth column, the q-th column, and the p-th row of matrix $\mathbf{A}$ respectively. $[\mathbf{a}]_p$ is the p-th entry of $\mathbf{a}$. $\mathcal{CN}\left(\mu, \sigma^2\right)$ is used to indicate a complex Gaussian random vector with mean μ and covariance σ^2. $\mathbf{I}$ represents the identity matrix. In the final, $\mathrm{Re}[\cdot]$ and $\mathbb{E}[\cdot]$ denotes the real part of a complex number and expectation, respectively.

II. Related Work

A. Resource Allocation for HSR Networks

How to optimize the allocation of resource, such as beamforming and power in HSR systems is a very interesting yet challenging topic worthy of study. There have been a few studies recently on resource allocation in HSR communications. A stable beamforming design containing the long-term TX beamforming as well as the short-term RX beamforming was presented in [4] to improve the transmission reliability and efficiency in a control/user-plane (C/U-plane) decoupled HSR wireless networks. In [6], the authors introduced an angle-domain channel tracking and HBF algorithm, which consists of Doppler frequency offset compensation, angular beamforming, and beam domain reception/precoding to overcome the performance loss due to the fast time-varying characteristic of HSR channel. Based on the practical mmWave HSR channel model, the authors of [7] investigated a productive HBF scheme with two stages to maximize the overall throughput in HSR scenarios. In the first stage, obtain the HBF at TX with weighted minimum mean square error (WMMSE) algorithm. In the second stage, decouple the DBF and the ABF at TX with the orthogonal matching pursuit (OMP) algorithm. In [17], with the aid of HSR location information, the authors proposed two open loop beamforming approaches to obtain high spectrum efficiency in HSR communications, which showed it's of great signification to specify the mapping relationships between the optimal transmission strategies and HSR location information with big data analysis.

There are also some works that focus on power allocation to enhance system performance in HSR communications. The power allocation problem with multiple MRs in orthogonal frequency division multiple access (OFDMA) HSR systems to obtain minimum total transmit power consumption was investigated in [8]. In [18], optimizing power allocation with antenna selection scheme in HSR communications was proposed for minimizing the average transmit power. The authors of [19] presented the power allocation algorithm to maximize energy efficiency in HSR networks with buffer constraint. The authors of [20] claimed that it's the first time that the power control problem was discussed in mmWave HSR systems. The above works work well with the assumptions of accurate

channel state information and mathematical model. However, it is scarcely possible to meet the assumptions in the HSR environments.

B. DRL for Wireless Networks

In the wireless communications fields, DRL has attracted numerous scholars' interest for the sake of enhancing system performance under uncertainty. With the goal of achieving the maximum weighted sum computation rate in wireless powered mobile edge computing networks, the authors of [10] jointly modelled the task offloading and resource allocation problem, which was decomposed into two sub-problems, i.e., offloading decision and resource allocation. In order to solve the first sub-problem, the online offloading algorithm based on DRL, i.e., DROO was proposed, which enabled the user equipment to learn the offloading strategy from the past experience. The second sub-problem was addressed with a one-dimensional bi-section search [23]. The authors in [12] proposed a multi-agent DRL based approach to solve the spectrum sharing problem in vehicular networks for the sake of enhancing performance of vehicle-to-infrastructure as well as vehicle-to-vehicle links. In [13], a multi-agent DRL based approach was introduced to jointly optimize user association and resource allocation problem for maximizing the long-term downlink utility, subject to guaranteeing each user's QoS requirement. The authors of [14] developed a new multi-path TCP by using DRL, i.e., DRL-CC to improve goodput In wireless networks. The authors in [21] investigated the joint power allocation and channel assignment problem in the multi-carrier non-orthogonal multiple access (NOMA) system with the aim of maximizing sum rate. The problem was solved through a two-stage approach method. In the first stage, with the given channel assignment, the authors derived the optimal power allocation based on [22]. Then channel assignment scheme was proposed with DRL in the second stage. In [24], the multi-agent DRL algorithm was applied to address power allocation problem in multi-cell networks with single antenna to achieve the maximum weighted sum-rate utility function. In [25], a distributed resource allocation algorithm via multi-agent DRL was presented for enhancing system throughput in train-to-train communications. However, there is no research to tackle the power allocation problem by leveraging emerging DRL in the mmWave HSR systems with low-resolution PSs.

III. System Model

In this paper, the uplink mmWave multiuser multiple input multiple output (MU-MIMO) HSR system is considered. Fig. 1 shows a potential mmWave HSR system model, which is recommended by 3rd Generation Partnership Project (3GPP) [26]. To mitigate penetration losses, the MR is mounted on top of each carriage, which keeps connection with the in-cabin wireless access point and forwards data between the passengers and BS. With the beamforming technique, the signals quality between the MRs and BSs can be enhanced. The power allocation problem between the MRs and the BS is the main study of the paper.

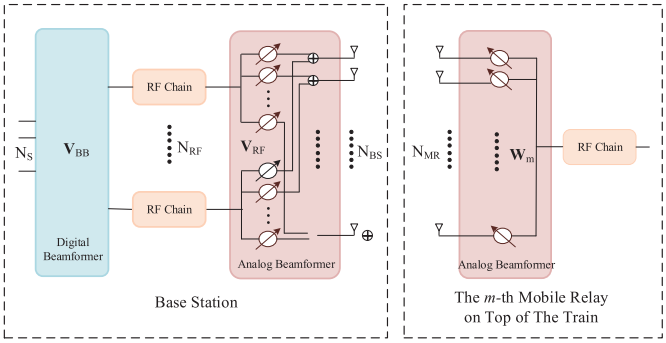


Fig. 2. The mmWave MU-MIMO HSR communication system, where the BS employs HBF to serve m-th MR with only ABF.

We consider the uplink of the multiuser mmWave MIMO HSR system, which is shown in Fig. 2. The BS equipped with N_{BS} antennas and N_{RF} radio frequency (RF) chains can be capable of simultaneously serving M MRs. Due to power consumption and hardware limitations, every MR with N_{MR} antennas transfers data to the BS with only one data stream. As a result, the sum number of data streams equals the number of MRs, i.e., $M = N_S$. Furthermore, in order to guarantee successful M data streams transmission, we assume $M \le N_{RF} \le \min(N_{BS}, N_{MR})$ [7].

In the uplink of HSR communication system, the m-th MR ($m = 1, 2, \ldots, M$) transmits the signal s_m to BS with $\mathbb{E}\{|s_m|^2\} = 1$ and transmission power p_m. Let $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_M]$ indicates the total MR ABF matrix. After the ABF process at the m-th MR, the transmitted signal of the m-th MR can be denoted by

$$x_m = \sqrt{p_m}\mathbf{w}_m s_m, \tag{1}$$

where $\mathbf{w}_m$ is the ABF vector corresponding to the m-th MR, whose elements also have the constant magnitude $\frac{1}{\sqrt{N_{MR}}}$ and low resolution discrete phases, i.e.,

$$[\mathbf{W}]_{p,q} \in \mathcal{W} \triangleq \left\{ \frac{1}{\sqrt{N_{MR}}} e^{\frac{2\pi b}{2^B}} | \mathrm{b} = 1, 2, \cdots, 2^B \right\}, \tag{2}$$

where B means quantization bits that dominate the resolution of PSs.

Define HBF matrix $\mathbf{V} = \mathbf{V}_{RF}\mathbf{V}_{BB} = [\mathbf{v}_1, \cdots, \mathbf{v}_m, \cdots, \mathbf{v}_M]$, in which $\mathbf{V}_{BB} = [\mathbf{v}_{BB}^1, \mathbf{v}_{BB}^2, \ldots, \mathbf{v}_{BB}^M]$, $\mathbf{v}_m = \mathbf{V}_{RF}\mathbf{v}_{BB}^m$, and let $\mathbf{H}_m \in \mathbb{C}^{N_{\mathrm{BS}} * N_{\mathrm{MR}}}$ be the HSR mmWave channel model from the m-th MR to the BS. Then the received signal at the BS is

$$\mathbf{r} = \sum_{m=1}^{M} \sqrt{p_m}\mathbf{H}_m \mathbf{w}_m s_m + \mathbf{u}, \tag{3}$$

where $\mathbf{u} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ represents the additive complex Gaussian white noise. After taking advantage of the HBF technology at the BS, we can model the estimated symbol of the M-th MR as

$$y_m = \underbrace{\mathbf{v}_m^{\mathrm{H}}\mathbf{H}_m\mathbf{w}_m\sqrt{p_m}s_m}_{\text{Desired signal}} + \underbrace{\mathbf{v}_m^{\mathrm{H}} \sum_{n=1, n\neq m}^{M} \mathbf{H}_n\mathbf{w}_n\sqrt{p_{\mathrm{n}}}s_n}_{\text{Multi-MR interference}} + \underbrace{\mathbf{v}_m^{\mathrm{H}}\mathbf{u}}_{\text{Noise}} \tag{4}$$

where $\mathbf{v}_{BB}^{m} \in \mathbb{C}^{N_{RF}\times 1}$ is the DBF vector for the m-th MR. $\mathbf{V}_{RF} \in \mathbb{C}^{N_{BS}\times N_{RF}}$ is the ABF matrix. For the fully-connected architecture, the ABF matrix $\mathbf{V}_{RF}$ is

$$\mathbf{V}_{RF} = \left[\mathbf{v}_{RF}^{1}, \ldots, \mathbf{v}_{RF}^{N_{RF}}\right], \tag{5}$$

where the elements of $\mathbf{V}_{RF}$, not only have the constant magnitude $\frac{1}{\sqrt{N_{BS}}}$, but also low resolution discrete phases, i.e.,

$$[\mathbf{V}_{RF}]_{p,q} \in \mathcal{V} \triangleq \left\{ \frac{1}{\sqrt{N_{BS}}} e^{j\frac{2\pi b}{2^B}} | \mathrm{b} = 1, 2, \cdots, 2^B \right\}. \tag{6}$$

A. mmWave Channel Model in HSR Scenarios

As is known to us, because of a finite number of scatters in the mmWave band, it is unlikely that mmWave channel is modelled as rich scattering model assumed within low frequencies. There are diverse angles of departure (AoDs) and angles of arrival (AoAs) in the limited multipath components (MPCs). For the sake of simplicity, it is assumed that every scattering contributes a single propagation path. Up to now, the Saleh-Valenzuela geometric channel model has been widely used to depict the mmWave channel in static or low mobility scenarios [27]. However, it's very necessary to take the effects of Doppler shift into account in extremely time-varying HSR scenarios. Consequently, the mmWave channel model in HSR scenarios can be characterized as [28]

$$\mathbf{H}_m = \gamma \sum_{l=1}^{L_m} \alpha_{m,l} \mathrm{a}_{BS}\left(\phi_{m,l}\right) \mathrm{a}_{MR}^{H}\left(\vartheta_{m,l}\right) e^{j2\pi \upsilon_l t}, \tag{7}$$

where $\gamma = \sqrt{\frac{N_{\mathrm{BS}} N_{\mathrm{MR}}}{L}}$ denotes a normalization factor, L_m is scatterers for the channel of m-th MR, $\alpha_{m,l}$ and υ_l represent the complex gain and the Doppler shift corresponding to the l-th path, respectively. $\phi_{m,l}$ and $\vartheta_{m,l}$ are the l-th path's AoAs and AoDs respectively. At last, $\mathrm{a}_{\mathrm{BS}}\left(\phi_{m,1}\right)$ and $\mathrm{a}_{\mathrm{MR}}\left(\vartheta_{m,l}\right)$ denote the antenna array response vectors associated with the BS and the m-th MR, respectively. In this paper, we assume that both BS and MR adopt uniform linear antenna arrays (ULAs). Then, we have

$$\mathrm{a}_{MR}(\vartheta) = \frac{1}{\sqrt{N_{MR}}}\left[1, e^{j\frac{2\pi}{\lambda}d\sin(\vartheta)}, \cdots, e^{j(N_{MR}-1)\frac{2\pi}{\lambda}d\sin(\vartheta)}\right]^T,$$

$$\mathrm{a}_{BS}(\phi) = \frac{1}{\sqrt{N_{BS}}}\left[1, e^{j\frac{2\pi}{\lambda}d\sin(\phi)}, \cdots, e^{j(N_{BS}-1)\frac{2\pi}{\lambda}d\sin(\phi)}\right]^T,$$

in which λ represents the signal wavelength, d denotes the antenna spacing.

B. Problem Formulation

The signal-to-interference-and-noise ratio (SINR) of the m-th MR is represented as

$$SINR_m = \frac{\left|\sqrt{p_m}\mathbf{v}_m^H \mathbf{H}_m \mathbf{w}_m\right|^2}{\zeta_{\mathrm{m}}}, \tag{8}$$

where $\zeta_m = \sum_{n=1,n\neq m}^{M}\left|\sqrt{p_n}\mathbf{v}_m^H \mathbf{H}_n \mathbf{w}_{\mathrm{n}}\right|^2 + \sigma^2$. The achievable rate of the m-th MR is

$$R_m = \log_2\left(1 + SINR_m\right). \tag{9}$$

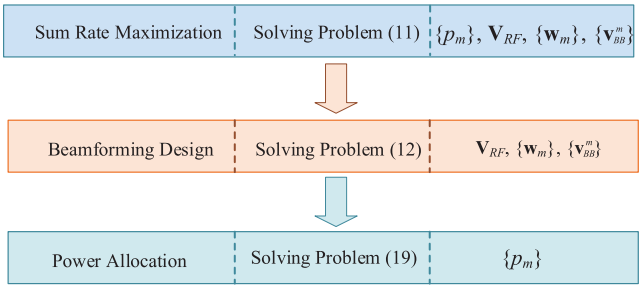


Fig. 3. The two-level architecture of addressing problem (11).

Finally, the ASR is given by

$$R_{\mathrm{sum}} = \sum_{m=1}^{M} R_m. \tag{10}$$

In the paper, we adopt the ASR as performance criteria. Then, the problem is to seek out $\left\{p_m^o\right\}_{m=1}^{M}$, $\mathbf{V}_{RF}^o$, $\left\{\mathbf{w}_m^o\right\}_{\mathrm{m=1}}^{M}$, $\left\{\mathbf{v}_{BB}^{m\ o}\right\}_{m=1}^{M}$ to achieve the maximum ASR, which can be formulated as

$$\begin{aligned} \max_{\{p_m\},\{\mathbf{V}_{RF}\},\{\mathbf{w}_m\},\{\mathbf{v}_{BB}^m\}} & \sum_{m=1}^{M} R_{sum} \\ \text{s.t. } & C_1: 0 \le p_m \le P_{max}, \quad \forall m, \\ & C_2: [\mathbf{V}_{RF}]_{p,q} \in \mathcal{V}, \quad \forall p, q, \\ & C_3: [\mathbf{W}]_{p,q} \in \mathcal{W}, \quad \forall p, q, \end{aligned} \tag{11}$$

in which the constraint C_1 denotes the power allocated to each MR should not be less than 0 and no more than P_{max} which is the maximum transmission power of MR. Both C_2 and C_3 are the constraints for the ABF matrix at BS and MR, respectively. The ASR in (11) can be increased by the excellent design on power allocation $\left\{p_m^o\right\}_{m=1}^{M}$, ABF matrix $\mathbf{V}_{RF}^o$, ABF $\left\{\mathbf{w}_m^o\right\}_{m=1}^{M}$ and DBF $\left\{\mathbf{v}_{BB}^{m\ o}\right\}_{m=1}^{M}$. It is extremely challenging to get the best solutions of all these design parameters at the same time. However, once the beamforming design is given, problem (11) can reduce to a power allocation problem. Say concretely, as shown in Fig. 3, problem (11) can be transformed into two sub-problems, namely, beamforming design and power allocation.

Beamforming design: To simplify the beamforming design, we decompose it into two independent optimizations. Firstly, we compute the ABF associated with the BS and MR, separately. Subsequently, after getting the effective baseband channel, the DBF matrix at the BS can be calculated.

Power allocation: Traditional power allocation algorithms can work well relying on accurate channel state information and mathematical model, which is fundamentally infeasible for time-varying mmWave HSR systems. To solve the problem, we develop a MADDPG-based power allocation algorithm, which can work well under uncertainty.

Accordingly, as shown in Fig. 3, we will first focus on designing ABF matrix $\mathbf{V}_{RF}^o$, ABF $\left\{\mathbf{w}_m^o\right\}_{\mathrm{m=1}}^{M}$ and DBF $\left\{\mathbf{v}_{BB}^{m\ o}\right\}_{m=1}^{M}$ and then introduce the power allocation proposed in this paper for mmWave HSR communication systems in Section V.

IV. Beamforming Design

In this paper, we consider the analog beamformers at the BS and the MRs are with two-bit resolution PSs, i.e., B = 2. In other words, the $\mathcal{V}$ as well as $\mathcal{W}$ can be rewritten as

$$\mathcal{V} = \frac{1}{\sqrt{N_{BS}}}\{\pm 1, \pm j\}^{N_{BS}},$$
$$\mathcal{W} = \frac{1}{\sqrt{N_{MR}}}\{\pm 1, \pm j\}^{N_{MR}}.$$

Then, with the given power allocation the BS and the m-th MR can determine the ${\mathbf{v}_{RF}^{m}}^{o}$ and $\mathbf{w}_m^o$ by dealing with the following problem

$$\begin{aligned}\left\{\mathbf{w}_m^o, {\mathbf{v}_{RF}^{m}}^{o}\right\} = \arg\max & \left\| {\mathbf{v}_{RF}^{m}}^{H} \mathbf{H}_m \mathbf{w}_m \right\| \\ \text{s.t. } & \mathbf{v}_{RF}^{m} \in \mathcal{V}, \quad \forall m, \\ & \mathbf{w}_m \in \mathcal{W}, \quad \forall m. \end{aligned} \tag{12}$$

After the Gram-Schmidt procedure [29], which is capable of suppressing the interference between MRs during the beamforming phase, the (m-1)-th MR's analog beamformer can be given by

$$\mathbf{g}_{m-1} = {\mathbf{v}_{RF}^{m-1}}^{o} - \sum_{n=1}^{m-2} \mathbf{e}_n^H {\mathbf{v}_{RF}^{m-1}}^{o} \mathbf{e}_n, \tag{13}$$

$$\mathbf{e}_{m-1} = \frac{\mathbf{g}_{m-1}}{\|\mathbf{g}_{m-1}\|}, \tag{14}$$

in which $\mathbf{e}_1 = \mathbf{v}_{RF}^{1^o}$, and $\mathbf{v}_{RF}^{1^o}$ represents the analog beamformer for the first MR. Afterwards, we remove the previous precoder components from the m-th MR's channel and obtain the modified channel

$$\tilde{\mathbf{H}}_m = \left(\mathbf{I}_{N_{BS}} - \sum_{n=1}^{m-1} \mathbf{e}_n \mathbf{e}_n^H\right) \mathbf{H}_m. \tag{15}$$

Subsequently, based on the modified channel $\tilde{\mathbf{H}}_m$, the analog beamformer pair for the m-th MR can be rewritten as [30]

$$\begin{aligned}\left\{\mathbf{w}_m^o, {\mathbf{v}_{RF}^{m}}^{o}\right\} = \arg\max & \left\| {\mathbf{v}_{RF}^{m}}^{H} \tilde{\mathbf{H}}_m \mathbf{w}_m \right\| \\ \text{s.t. } & \mathbf{v}_{RF}^{m} \in \mathcal{V}, \quad \forall m, \\ & \mathbf{w}_m \in \mathcal{W}, \quad \forall m. \end{aligned} \tag{16}$$

After determining the analog beamformers associated with the BS and MR, the equivalent channel for m-MR on the roof of the train can be given by

$$\bar{\mathbf{h}}_m^H = \left(\mathbf{V}_{RF}^o\right)^H \mathbf{H}_m \mathbf{w}_m^o. \tag{17}$$

In accordance with the effective channel, BS is able to compute the DBF $\bar{\mathbf{V}}_{BB}$, which can be generated using the zero-forcing precoding as

$$\bar{\mathbf{V}}_{BB} = \left[\, \bar{\mathbf{v}}_{BB}^1, \cdots, \bar{\mathbf{v}}_{BB}^M \,\right] = \bar{\mathbf{H}}^{\mathrm{H}} \left(\bar{\mathbf{H}} \bar{\mathbf{H}}^{\mathrm{H}}\right)^{-1}, \tag{18}$$

in which $\bar{\mathbf{H}} = \left[\bar{\mathbf{h}}_1, \bar{\mathbf{h}}_2, \ldots, \bar{\mathbf{h}}_M\right]^H$. So far, ABF matrix $\mathbf{V}_{RF}$, DBF $\left\{\mathbf{v}_{BB}^m\right\}_{m=1}^M$ and ABF $\{\mathbf{w}_m\}_{m=1}^M$ have been carefully chosen to provide both antenna array gains and multiplexing gains, improving the transmission dependability in HSR communication systems. Be aware that this paper primarily aims at studying the power allocation problem in the mmWave HSR systems. How to design an efficient and reliable beamforming scheme in mmWave HSR systems based on DRL is still an open issue and a very interesting direction worthy of further study. In the next section, aiming at maximizing the ASR of mmWave HSR systems, we propose a scheme of dynamic power allocation based on multi-agent DRL.

V. Power Allocation With Multi-Agent DRL

Game theory [31] and convex optimization [32] are two extensively applied methods to tackle the power allocation problems. Nevertheless, these methods exist a few shortcomings as follows.

Firstly, they assume that the critical factors, such as channel conditions can be accurately acquired. Nevertheless, it is very challenging in the realistic HSR environments.

Secondly, the high-speed mobility of trains causes complicated and dynamic HSR network topology. As a result, applying these approaches to tackle power allocation problems can hardly ensure efficient and dependable data transmission.

Thirdly, by using most of these approaches, optimal or near optimal solutions can be obtained only for one snapshot of the HSR communication systems. They ignore the long-term impact of current decision on power allocation.

Compared with these approaches, AI, especially DRL is a promising technique to realize resource allocation efficiently by interacting with environment [12]. Since the power variable is continuous, we develop an experience-driven power allocation algorithm shown in Fig. 4 based on MADDPG in this section. Different from these approaches above, the experience-driven power allocation algorithm is capable of making advisable power decisions with taking the real-time runtime state into consideration, instead of depending on any mathematical model.

Since we have successfully designed the ABF matrix $\mathbf{V}_{RF}$, DBF $\left\{\mathbf{v}_{BB}^m\right\}_{m=1}^M$ and ABF $\{\mathbf{w}_m\}_{m=1}^M$ in the previous section, the original problem (11) can be reduced to

$$\begin{aligned}\max_{\{p_m\}} & \sum_{m=1}^{M} R_{sum} \\ \text{s.t. } & C_1 : 0 \le p_m \le P_{max}, \quad \forall m. \end{aligned} \tag{19}$$

It is observed that the optimization problem (19) is non-convex. To deal with the problem (19), we propose an experience-driven power allocation algorithm with multi-agent DRL.

A. Overview of DDPG

Q-learning [33], a typical reinforcement learning (RL) algorithm, makes use of a Q-table to store the value of state-action pairs. Given the current state, the agent will take action according to its policy. Afterwards, the agent receives reward and then the state converts into to a new state. To go along with this, the value in the Q-table will be updated. By doing so, the agent can learn to take the best action within finite time steps. Let $s \in \mathcal{S}$ and $a \in \mathcal{A}$ represent the state and

action, respectively. Then, the expected value function at state s is defined as

$$\mathcal{F}_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \varsigma^t r_t(s_t, a_t)\,|s_0 = s\right] = \mathbb{E}_\pi\left[r_t(s_t, a_t) + \varsigma \mathcal{F}_\pi(s_{t+1})\,|s_0 = s\right], \quad (20)$$

in which $\varsigma \in [0, 1)$ denotes the discount factor influencing long-term reward. In order to obtain optimal policy π^o, we need to extract the optimal action with the following expression

$$\mathcal{F}^o(s) = \max_{a_t}\left\{\mathbb{E}_\pi\left[r_t(s_t, a_t) + \varsigma \mathcal{F}_\pi(s_{t+1})\right]\right\}. \quad (21)$$

Since the optimal Q-function is presented as

$$\mathcal{Q}^o(s, a) = r_t(s_t, a_t) + \varsigma \mathbb{E}_\pi\left[\mathcal{F}_\pi(s_{t+1})\right], \quad (22)$$

(21) is equivalent to

$$\mathcal{F}^o(s) = \max_a\left\{\mathcal{Q}^o(s, a)\right\}. \quad (23)$$

Now, the aim of agent can be transformed into finding the optimal value, namely $\mathcal{Q}^o(s, a)$. Specially, the Q-function is updated as

$$\mathcal{Q}(s_t, a_t) \leftarrow \mathcal{Q}(s_t, a_t) + \eta\left(r_t(s_t, a_t) + \varsigma \max_{a_{t+1}} \mathcal{Q}(s_{t+1}, a_{t+1}) - \mathcal{Q}(s_t, a_t)\right), \quad (24)$$

in which $\eta \in (0, 1)$ denotes the learning rate. $\mathcal{Q}(s_t, a_t)$ will have absolute convergence with the reasonable learning rate.

The classical Q-learning approach makes use of a Q-table to store the value of state-action pairs. Unfortunately, owing to the HSR environments' high time variability and complexity, it is challenging to store all values of state-action pairs, which is detrimental to achieve the optimal policy. Fortunately, emerging DRL combining RL and deep learning makes it possible to tackle the issue with infinite state spaces. Deep Q-network (DQN), which leverages deep neural network (DNN) to represent state-action pairs and approximates the Q-function is able to yield better performance than Q-learning. DNN is capable of outputting a corresponding action value $\mathcal{Q}(s, a|\varphi)$ based on the state s coming from HSR scenarios. For these values generated by DNN, ϵ-greedy policy [34], [35] is adopted for making decisions on the action.

The DQN is trained to minimize the gap between the $\mathcal{Q}(s, a|\varphi)$ and target Q-value by minimizing the loss function $\mathcal{L}(\varphi)$ at every time step, which is given by [36]

$$\mathcal{L}(\varphi) = \mathbb{E}\left[(y_t - \mathcal{Q}(s_t, a_t|\varphi))^2\right]. \quad (25)$$

Here, $\mathcal{Q}(s_t, a_t|\varphi)$ indicates evaluation Q-value generated by evaluation network with weights φ. $y_t = r_t + \varsigma \max_{a_{t+1}} \mathcal{Q}'\left(s_{t+1}, a_{t+1}|\varphi^-\right)$ represents the target Q-value given by target network with weights φ^-. The target network not only plays the role of providing fixed label, but also plays the role of increasing the stability as well as the convergence speed of training. To accomplish this, the following two effective and efficient techniques are introduced [37].

Experience replay: At time t, the agent stores the past experience, i.e., the sequence (s_t, a_t, r_t, s_{t+1}) into the relay memory $\mathcal{D}$ and update the Q-value $\mathcal{Q}(s_t, a_t|\varphi)$ by choosing mini-batch of samples from the relay memory.

Network cloning: The agent estimates target values y_t by leveraging a separate target network. Both evaluation network and target network have the same initial weights. However, φ is updated per time step, while φ^- is updated per B>1 time steps, which is conducive to improving learning stability.

As mentioned in [38], DON is targeted for solving discrete control problems with low-dimensional action spaces. Consequently, It can hardly be applied to continuous control tasks, such as power allocation. Fortunately, DDPG [38] was recently proposed, which showed the advantages in handling continuous control problems. DDPG realizes continuous control with the aid of deterministic policy gradient as well as deep neural network. As depicted in Fig. 4, the core idea of DDPG is that two functions are maintained at the same time. One is the function $\mu(s_t \mid \emptyset)$ derived by actor network responsible for generating actions. The other is the function $\mathcal{Q}(s_t, a_t \mid \varphi)$ derived by critic network responsible for evaluating actions. The critic network is updated by minimizing the loss function given by (25). The actor network function is updated by another DNN. Specifically, apply the chain rule to the anticipated cumulative reward $\mathcal{J}$ in relation to the actor network parameters to update the actor network [38], [39]

$$\nabla_\emptyset \mathcal{J} \approx \mathbb{E}\left[\nabla_\emptyset \mathcal{Q}(s, a \mid \varphi)|_{s=s_t, a=\mu(s_t|\emptyset)}\right] = \mathbb{E}\left[\nabla_a \mathcal{Q}(s, a \mid \varphi)|_{s=s_t, a=\mu(s_t)} \nabla_\emptyset \mu(s_t \mid \emptyset)\right]. \quad (26)$$

B. Modeling of Multi-Agent Scenarios

In this subsection, we transform the power allocation issue into the multi-agent DRL issue, which is shown in Fig. 5. Each MR m is an agent and accumulates adequate experiences to guide its own power allocation policy decisions through interacting with HSR environment. Concretely, at each time t, after observing the state s_t^m that includes the features of HSR environment, the m-th MR takes an action a_t^m based on its policy, forming joint actions a_t, and then receives a reward r_t^m. Afterwards, the new state s_{t+1} can be acquired.

The proposed experience-driven power allocation algorithm based on emerging multi-agent DRL technology proceeds in two stages called the centralized learning (or training) as well as the distributed implementation stages. Specifically, in the stage of training, every independent MR agent not only has convenient access to obtaining mmWave HSR system performance-oriented reward, but also learning to make best action according to continuous training. In the stage of implementation, after observing the state of HSR environment, every MR agent is quickly capable of taking best action based on its own past experience. Next, the key elements of the experience-driven power allocation algorithm will be detailedly discussed.

C. State Space

The first element of state is every MR agent's own signal channel, i.e., $\mathbf{H}_m$. What's more, the local observation information of each MR agent m also contains beamforming design,

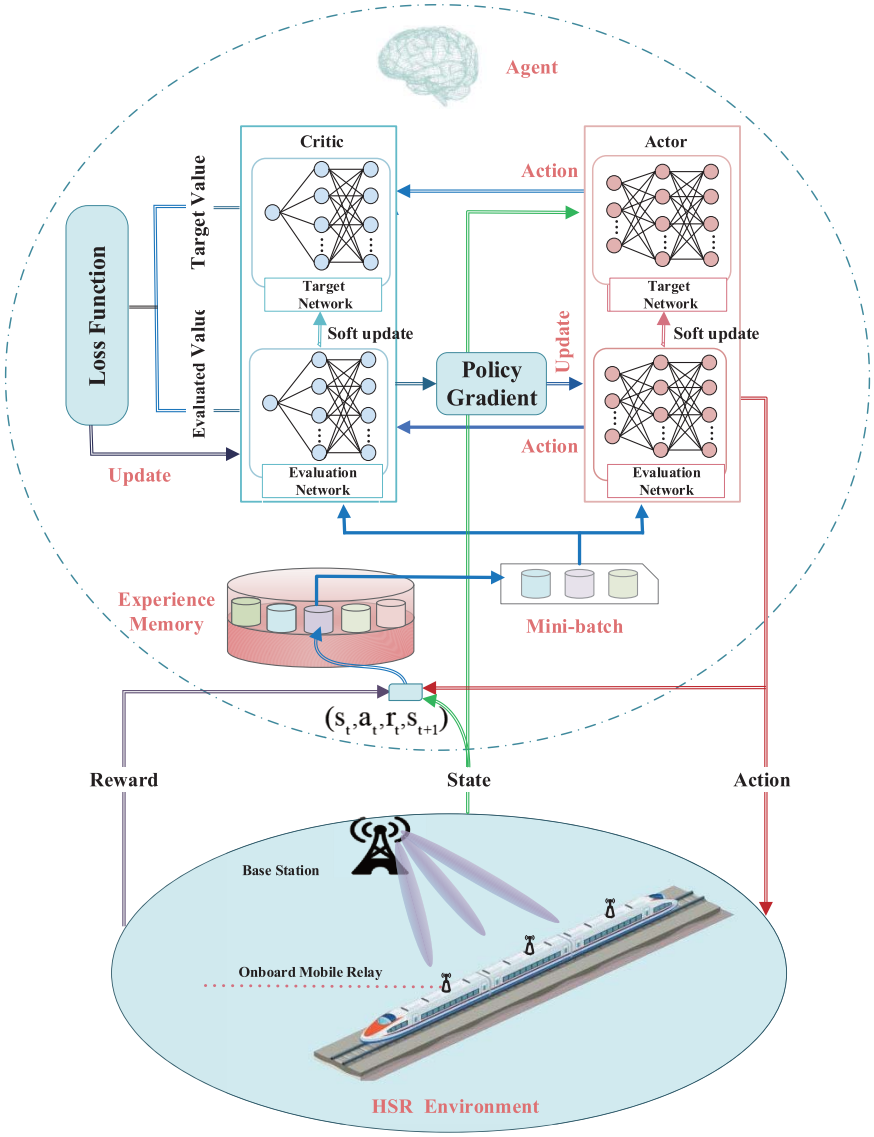


Fig. 4. The procedure of experience-driven power allocation in mmWave HSR systems.

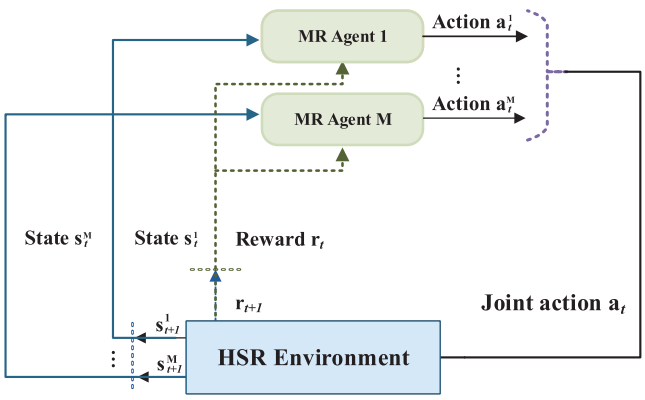


Fig. 5. The interaction between multi-agent and HSR environment.

i.e., $\{\mathbf{v}_{BB}^m\}$, $\{\mathbf{v}_{RF}^m\}$, and $\{\mathbf{w}_m\}$. Finally, in order to let the agent obtain the optimal power value more readily, two additional auxiliary features, i.e., R_{t-1}^m and p_{t-1}^m are introduced, which denote each agent's achievable rate and emitting power at last step, respectively. In a word, the states of all agents in mmWave HSR system can be defined as

$$\boldsymbol{s}_t = \left\{s_t^1, s_t^2, \ldots, s_t^m, \ldots, s_t^M\right\}, \tag{27}$$

where

$$s_t^m = \left\{\left\{\mathbf{v}_{BB}^m\right\}_t, \left\{\mathbf{v}_{RF}^m\right\}_t, \{\mathbf{w}_m\}_t, \{\mathbf{H}_m\}_t, R_{t-1}^m, p_{t-1}^m\right\}. \tag{28}$$

D. Action Space

Instead of making use of discrete power levels in [24], we employ continuous values from 0 to P_{max} as the transmit power of each agent. As a result, the action of MR agent m is expressed as

$$a_t^m = \{p_t^m\}. \tag{29}$$

In contrast with DQN, the robustness of learning based on DDPG is accordingly enhanced. If we solve the power

allocation problem with DQN, the action space must be discretized. In order to achieve finer grained discretization, the number of actions may be going to be an explosion. In addition, we should also pay attention to that it is harmful for the HSR networks performance to discretize the action space. Faced with this situation, it is significant to solve the problem with the action defined in (29).

E. Reward Design

As described in Section III, we aim to maximize the ASR in smart railway. As a result, (10) is acted as the reward function at each time step t, which is expressed as

$$r_t^m = \sum_{m=1}^{M} R_m. \tag{30}$$

Note that to meet the need of collaboration among the MRs, all MRs share the same reward function.

Algorithm 1: MADDPG-Based Power Allocation

Initialization:
  Initialize experience replay memory $\mathcal{D}$
  Initialize actor network $\mu(\cdot)$ and critic network $\mathcal{Q}(\cdot)$ with random weights $\emptyset$ and φ separately
  Initialize target networks $\mu'(\cdot)$, $\mathcal{Q}'(\cdot)$ with weights $\emptyset' = \emptyset$ and $\varphi' = \varphi$ separately
for each episode k **do**
  Initialize state s^m, for all $m \in M$
  for each step t **do**
    Obtain s_t^m
    Each MR agent orchestrates action $a_t^m = \mu(s_t^m)$ based on current policy
    Perform action $\boldsymbol{a}_t = \left\{a_t^1, \cdots, a_t^M\right\}$, obtain reward $\boldsymbol{r}_t$ and new state $\boldsymbol{s}_{t+1}$
    Save $(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{s}_{t+1})$ into $\mathcal{D}$
    for each MR agent m **do**
      Sample a random mini-batch of ϖ transitions from $\mathcal{D}$
      Update critic network by according to minimizing the loss: $\mathcal{L}(\varphi^m) = \frac{1}{\varpi}\sum_l \left(y_l - \mathcal{Q}^m\left(\mathbf{s}_l, a_l^1, \cdots, a_l^M\right)\right)^2$
      Update actor network by employing sampled policy gradient
    Update the target networks: $\emptyset' \leftarrow \zeta\emptyset + (1-\zeta)\emptyset'$; $\varphi' \leftarrow \zeta\varphi + (1-\zeta)\varphi'$
Perform Algorithm 2 to make power allocation strategic decisions.

F. Learning Algorithm

1) Training Process: As Algorithm 1 declares, once the three critical elements, namely, state, action as well as reward are defined, in the first place, experience replay memory, the actor and critic networks are initialized. At the beginning of each episode k, initialize the state. Afterwards, at each step t, each MR agent observes the state s_t^m followed by taking an action a_t^m and receiving reward r_t^m as well as new state s_{t+1}^m. Next, store the experience $(\boldsymbol{s}_t, \boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{s}_{t+1})$ in the replay memory $\mathcal{D}$. At last, sample random mini-batch of experience from $\mathcal{D}$ to update critic and actor networks. Here, we set ζ as 0.001.

Be aware that for the sake of gaining enough experience to make timely and effective decisions under similar circumstances, it is essential to collect sufficient number of samples for training the agents.

Algorithm 2: Implementation Process

Require:
  The trained actor network
All agents observe initial state $\boldsymbol{s}_o = \left\{s_o^1, s_o^2, \ldots, s_o^m, \ldots, s_o^M\right\}$
for each step t **do**
  All agents orchestrate action $\boldsymbol{a}_t = \left\{a_t^m = \mu^m\left(s_t^m\right), m \in M\right\}$ based on current policy
  Obtain reward as well as new state

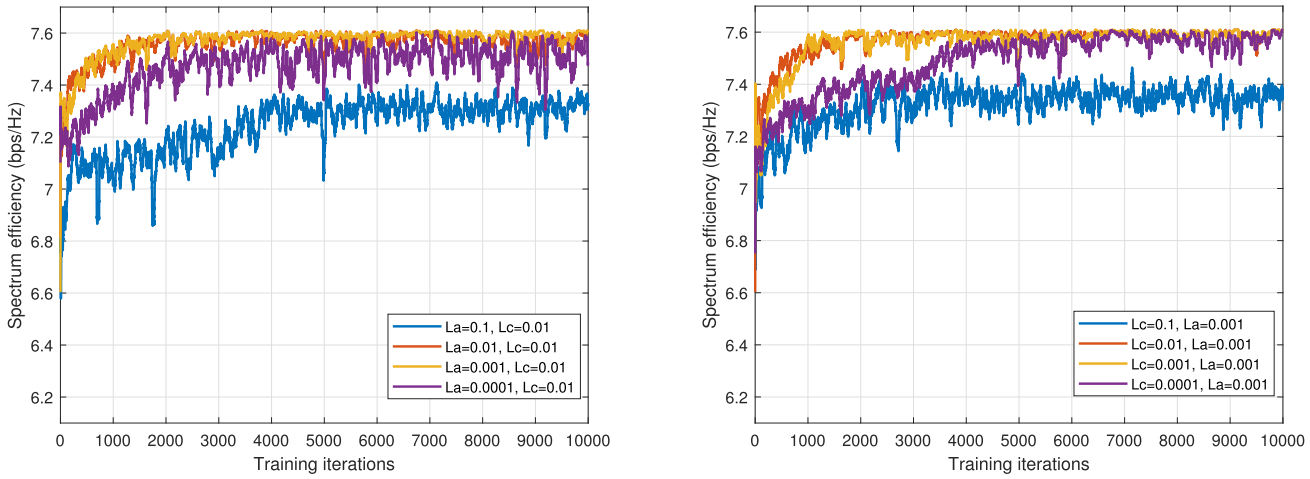
2) Distributed Implementation: The proposed experience-driven power allocation approach is implemented on DRL unit deployed on MR. As Algorithm 2 shows, in the beginning, each MR collects the state information of HSR environment, and updates the state. At per time step t, each MR takes the action $a_t^m = \mu(s_t^m)$ based on current policy. After that, all MRs begin transmitting data with the power level.

Note that the training process shown in Algorithm 1 is rather time-consuming, which can be performed offline. The low-cost implementation process is executed online in the deployment. Since the railway transportation line is fixed, it obeys regular space and time characteristics. Unless the HSR environment characteristics change very much, there is no need to retrain all agents in a short time, which is decided by environment dynamics [12]. In other words, it's possible that all agents need to be retrained once a week or even a month.

VI. Simulation Results

In this section, numerous simulation experiments are carried out to evaluate the performance of the proposed experience-driven power allocation scheme in the mmWave HSR system. In the simulations, we set the maximum transmission power of the MR P_{max} as 23 dBm. The pathloss is expressed as $PL(dB) = 61.4 + 34\ lg(d)$ [28], in which d indicates the distance. The speed of the train is 360 km/h. The MR and BS are equipped with $N_{MR} = 9$ and $N_{BS} = 64$ antennas, respectively.

In the stage of implementation, for performance evaluation, the following benchmarks where the same beamforming design is adopted, but different power allocation schemes are employed are included. The first benchmark is maximal power allocation, where each MR is assigned the maximal power, i.e., P_{max} for all time steps. The second benchmark is random



(a) (b)

Fig. 6. Spectrum efficiency under various learning rates when $M = 4$, $N_{MR} = 9$, and $N_{BS} = 64$: (a) learning rates of the actor network; (b) learning rates of the critic network.

power allocation, where each MR's power is randomly allocated from 0 to P_{max} at each time step. The third benchmark is fractional programming (FP) [Algorithm 3, 16]. It is assumed that FP knows full instantaneous channel state information. The last benchmark is DQN [24], where the action space is discrete.

Next, hyper-parameters used for training the MADDPG architecture of experience-driven power allocation method are shown. For the sake of enabling each MR agent to make power policy decision as soon as possible, we consider a relatively simple structure. The actor and the critic networks consist of three hidden layers, where the hidden layers contain N1 = 256, N2 = 128, and N3 = 128 neurons, respectively. The rectified linear unit (ReLU) is regarded as the activation function between hidden layers, The network parameters are updated by using Adam algorithm [41].

Fig. 6 (a) shows the effect of the various actor's learning rates, with the critic's learning rate set as the fixed value Lc = 0.01. Fig. 6 (b) shows the effect of the various critic's learning rates, with the actor's learning rate set as the fixed value La = 0.001. We can see that during the training of MADDPG algorithm, the learning rate has strong effect on the stability and quickness of the spectral efficiency.[1] That is because learning rate indicates the learning step to implement the convergence of spectral efficiency. If the learning rate is large, there is a high probability of missing the global optimum during training procedure. While if the learning rate is small, it would probably slow down the convergence rate. From Fig. 6, we also see that when actor's and critic's learning rates are set to 0.001 and 0.01, respectively, the spectral efficiency realized by MADDPG can quickly converge to a higher and more stable value. Note that the following simulation results are provided after the agents are trained to be stable.

Fig. 7 shows the spectral efficiency of different strategies against the number of MRs. Note that in the experiment,

[1]In this paper, we employ the ASR in (10) as the spectral efficiency [30], [42].

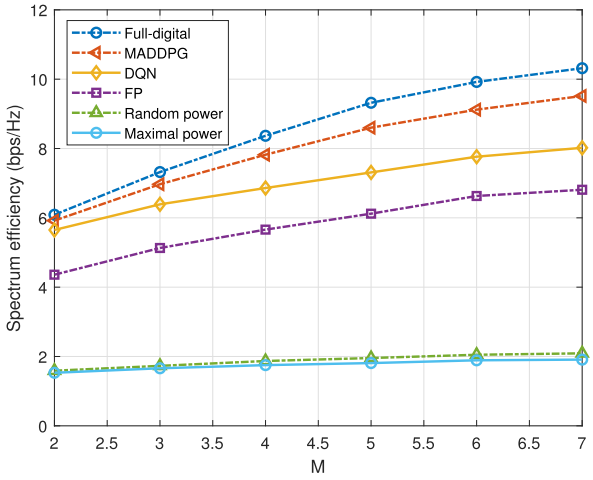


Fig. 7. Spectrum efficiency of different strategies against the number of MRs, when $N_{BS} = 64$ and $N_{MR} = 9$.

spectral efficiency realized by the optimal full-digital beamforming is taken as the upper bound. It's clear from the Fig. 7 that as the number of MRs increases, the spectral efficiency achieved by all algorithms increases. But the random power and maximal power scheme increase little improvement in the spectral efficiency, while FP, DQN and MADDPG promote the performance significantly. What's more, as the number of MRs increases in the mmWave HSR system, the spectral efficiency achieved by the proposed MADDPG scheme has a higher spectral efficiency than that of other four power allocation algorithms and is closest to the upper bound. Moreover, the performance of both the random power and maximal power scheme is always poorer than FP, DQN and MADDPG, which illustrates by the joint optimization of power allocation and beamforming, it can improve the spectrum efficiency of mmWave HSR systems.

Fig. 8 shows the spectrum efficiency of various power allocation algorithms for 1, 000 samples of the mmWave

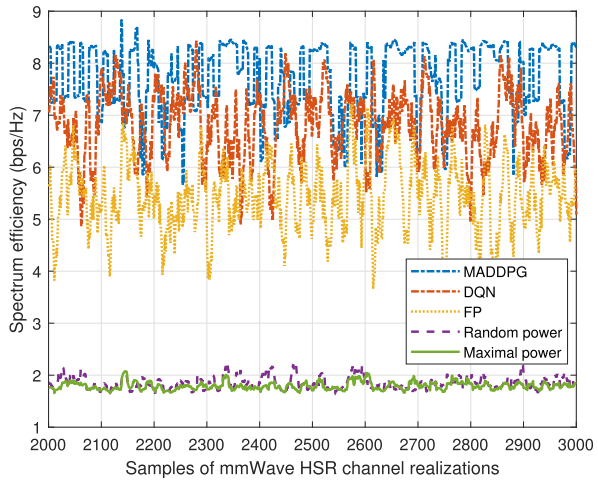


Fig. 8. Spectrum efficiency comparison of different algorithms over 1, 000 samples of the mmWave HSR channel realizations, when $M = 4$, $N_{MR} = 9$, and $N_{BS} = 64$.

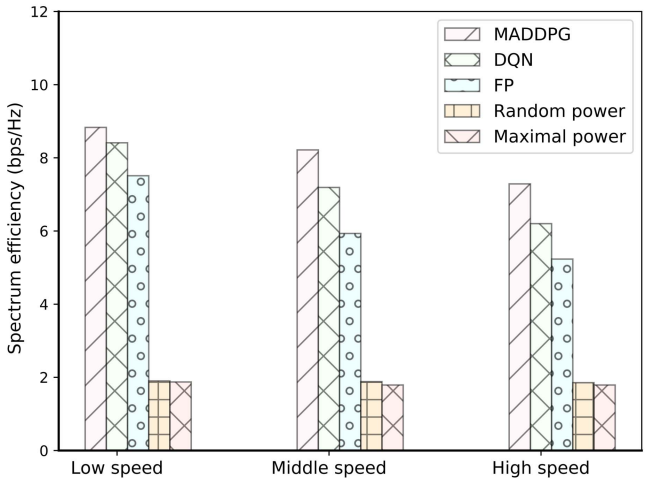


Fig. 9. Spectrum efficiency comparison of different algorithms against the speeds of train, when $M = 4$, $N_{MR} = 9$, and $N_{BS} = 64$.

HSR channel realizations. Again it is easy to see that in the time-varying mmWave HSR system, the proposed MADDPG scheme outperforms DQN, FP, random power and maximal power scheme in term of spectral efficiency. It is also observed that since the maximal power and random power algorithms do not take the time-varying characteristics of HSR channel into consideration, their curves are more stable than the ones of proposed MADDPG, DQN and FP methods. Besides, the performance of MADDPG shows the best, which demonstrates that MADDPG-based power allocation algorithm is capable of learning power allocation strategies efficiently from past experience, rather than relying on any pre-defined rule or any accurate mathematical model. In addition, we should note that DQN needs to discretize the action space, which leads to that in few time steps, the performance of DQN is slightly worse than FP.

Fig. 9 shows the spectrum efficiency of various algorithms against the speeds of train, i.e., high speed (360 km/h), middle speed (160 km/h) as well as low speed (50 km/h). We can observe that as the speed increases, the spectrum efficiency of all schemes significantly decreases. The reason is that with the increasing speed of train, it will cause the more serious inter-channel interference due to Doppler frequency offset.

VII. Conclusion

In this paper, we have developed an experience-driven power allocation method by leveraging multi-agent DRL, to maximize ASR for smart railway. The experience-driven power allocation approach, i.e., MADDPG with the ability of learning from the past power experiences to intelligently orchestrate power strategy in the mmWave HSR communication systems, was categorized into two processes: centralized training process as well as distributed implementation process. The simulation results demonstrated its effectiveness and advantage over the state-of-the-art schemes (including DQN, FP, random power and maximal power) in the mmWave HSR communication systems in terms of spectral efficiency.

References

[1] B. Ai, A. F. Molisch, M. Rupp, and Z.-D. Zhong, "5G key technologies for smart railways," *Proc. IEEE*, vol. 108, no. 6, pp. 856–893, Jun. 2020.

[2] B. Ai *et al.*, "Challenges toward wireless communications for high-speed railway," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2143–2158, Oct. 2014.

[3] J. Xu, B. Ai, L. Chen, L. Pei, Y. Li, and Y. Y. Nazaruddin, "When high-speed railway networks meet multipath TCP: Supporting dependable communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 202–205, Feb. 2020.

[4] L. Yan, X. Fang, and Y. Fang, "Stable beamforming with low overhead for C/U-plane decoupled HSR wireless networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6075–6086, Jul. 2018.

[5] J. Xu, B. Ai, G. Shi, Z. Zhong, S. Lukman, and B. Juliyanto, "Cross-layer assisted TCP for dependable communications in high-speed railway networks," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, pp. 1–6.

[6] K. Xu, Z. Shen, Y. Wang, and X. Xia, "Location-aided mMIMO channel tracking and hybrid beamforming for high-speed railway communications: An angle-domain approach," *IEEE Syst. J.*, vol. 14, no. 1, pp. 93–104, Mar. 2020.

[7] M. Gao *et al.*, "On hybrid beamforming of mmWave MU-MIMO system for high-speed railways," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.

[8] H. Ghazzai, T. Bouchoucha, A. Alsharoa, E. Yaacoub, M.-S. Alouini, and T. Y. Al-Naffouri, "Transmit power minimization and base station planning for high-speed trains with multiple moving relays in OFDMA systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 175–187, Jan. 2017.

[9] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[10] L. Huang, S. Bi, and Y.-J.-A. Zhang, "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 19, no. 11, pp. 2581–2593, Nov. 2020.

[11] L. Qian, Y. Wu, F. Jiang, N. Yu, W. Lu, and B. Lin, "NOMA assisted multi-task multi-access mobile edge computing via deep reinforcement learning for industrial Internet of Things," *IEEE Trans. Ind. Informat.*, early access, Jun. 10, 2020, doi: 10.1109/TII.2020.3001355.

[12] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.

[13] N. Zhao, Y. Liang, D. Niyato, Y. Pei, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6.

[14] Z. Xu, J. Tang, C. Yin, Y. Wang, and G. Xue, "Experience-driven congestion control: When multi-path TCP meets deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1325–1336, Jun. 2019.

[15] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. S. Quek, "Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 630–644, Jan. 2020.
[16] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, Mar. 2018.
[17] S. Han *et al.*, "Achieving high spectrum efficiency on high speed train for 5G new radio and beyond," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 62–69, Oct. 2019.
[18] J. Lu, K. Xiong, X. Chen, and P. Fan, "Toward traffic patterns in high-speed railway communication systems: Power allocation and access selection," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12273–12287, Dec. 2018.
[19] X. Wang, M. Yu, J. Hu, and Y. Xu, "Low-complexity energy-efficient power allocation with buffer constraint in HSR communications," *IEEE Access*, vol. 7, pp. 113867–113879, 2019.
[20] L. Wang, B. Ai, Y. Niu, X. Chen, and P. Hui, "Energy-efficient power control of train–ground mmWave communication for high-speed trains," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7704–7714, Aug. 2019.
[21] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.
[22] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2744–2757, Dec. 2017.
[23] S. Bi and Y. J. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, Jun. 2018.
[24] 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.
[25] J. Zhao, Y. Zhang, Y. Nie, and J. Liu, "Intelligent resource allocation for Train-to-Train communication: A multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 8, pp. 8032–8040, 2020.
[26] *Technical Specification Group Radio Access Network: Study on Scenarios and Requirements for Next Generation Access Technologies (Release 14)*, document 3GPP TR 38.913 V0.3.0, Jan. 2016.
[27] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Oct. 2014.
[28] M. Cheng, J.-B. Wang, J.-Y. Wang, M. Lin, Y. Wu, and H. Zhu, "A fast beam searching scheme in mmWave communications for high-speed trains," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, May 2019, pp. 1–6.
[29] F. Dong, W. Wang, and Z. Wei, "Low-complexity hybrid precoding for multi-user mmWave systems with low-resolution phase shifters," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 9774–9784, Oct. 2019.
[30] Z. Wang, M. Li, Q. Liu, and A. L. Swindlehurst, "Hybrid precoder and combiner design with low-resolution phase shifters in mmWave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 256–269, May 2018.
[31] Y. Liu, Z. Wen, N. C. Beaulieu, D. Liu, and X. Liu, "Power allocation for SWIPT in full-duplex AF relay interference channels using game theory," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 608–611, Mar. 2020.
[32] J. Huang, C.-C. Xing, and M. Guizani, "Power allocation for D2D communications with SWIPT," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2308–2320, Apr. 2020.
[33] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[34] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
[35] A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, and P. J. How, "A tutorial on linear function approximators for dynamic programming and reinforcement learning," *Found. Trends Mach. Learn.*, vol. 6, no. 4, 2013, Art. no. 375451.
[36] L. Chen, H. Qu, J. Zhao, B. Chen, and J. C. Principe, "Efficient and robust deep learning with correntropy-induced loss function," *Neural Comput. Appl.*, vol. 27, no. 4, pp. 1019–1031, May 2016.
[37] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, 7540, pp. 529–533, 2015.
[38] J. Xu, B. Ai, L. Wu, and L. Chen, "Handover-aware cross-layer aided TCP with deep reinforcement learning for high-speed railway networks," *IEEE Netw. Lett.*, early access, Dec. 21, 2020, doi: 10.1109/LNET.2020.3045967.
[39] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6379–6390.
[40] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. USENIX Symp. Operating Syst. Design Implement.*, 2016, pp. 265–283.
[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015, pp. 1–15.
[42] J. Xu, B. Ai, Y. Sun, and Y. Chen, "Power allocation for millimeter-wave high-speed railway networks with multi-agent deep reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 2020, pp. 1–6.

Jianpeng Xu (Graduate Student Member, IEEE) received the M.S. degree from Inner Mongolia University, Hohhot, China. He is currently pursuing the Ph.D. degree with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China. His current research interests include deep reinforcement learning-based wireless communications, and network architecture for high-mobility broadband wireless communications.

Bo Ai (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2002 and 2004, respectively. He was with Tsinghua University, Beijing, China, where he was an Excellent Post-Doctoral Research Fellow in 2007. He is currently a Professor and an Advisor of Ph.D. candidates with Beijing Jiaotong University, Beijing, where he is also the Deputy Director of the State Key Laboratory of Rail Traffic Control and Safety. He is currently with the Engineering College, Armed Police Force, Xi'an. He has authored/coauthored six books and 270 scientific research papers, and holds 26 invention patents in his research areas. His interests include the research and applications of orthogonal frequency-division multiplexing techniques, high-power amplifier linearization techniques, radio propagation and channel modeling, global systems for mobile communications for railway systems, and long-term evolution for railway systems. He is a fellow of the IET and an IEEE VTS Distinguished Lecturer. He has received many awards such as the Outstanding Youth Foundation from the National Natural Science Foundation of China, the Qiushi Outstanding Youth Award by the Hong Kong Qiushi Foundation, the New Century Talents by the Chinese Ministry of Education, the Zhan Tianyou Railway Science and Technology Award of the Chinese Ministry of Railways, and the Science and Technology New Star of the Beijing Municipal Science and Technology Commission.