Optimizing Intelligent Reflecting Surface-Base Station Association for Mobile Networks

Dongzi Jin*, Yong Xiao*§, Yingyu Li*, Guangming Shi†§, and Dusit Niyato‡
*School of Electronic Inform. & Commun., Huazhong University of Science & Technology, China
†School of Artificial Intelligence, Xidian University, Xi'an, China
§Pazhou Lab, Guangzhou, China
‡School of Computer Science and Engineering, Nanyang Technological University, Singapore

***Abstract*—This paper studies a multi-Intelligent Reflecting Surfaces (IRSs)-assisted wireless network consisting of multiple base stations (BSs) serving a set of mobile users. We focus on the IRS-BS association problem in which multiple BSs compete with each other for controlling the phase shifts of a limited number of IRSs to maximize the long-term downlink data rate for the associated users. We propose MDLBI, a Multi-agent Deep Reinforcement Learning-based BS-IRS association scheme that optimizes the BS-IRS association as well as the phase-shift of each IRS when being associated with different BSs. MDLBI does not require information exchanging among BSs. Simulation results show that MDLBI achieves significant performance improvement and is scalable for large networking systems.**

***Index Terms*—Multi-IRS, Transmit Beamforming, Reflect Beamforming, BS-IRS association, Reinforcement Learning;**

I. Introduction

Intelligent Reflecting Surface (IRS) is a programmable meta-surface consisting of a large number of low-cost and passive reflecting elements. The phase shifts of these elements can be controlled and optimized to enhance the signal reception at the receiver. Recent studies suggest that the IRS has the potential to establish a reflected link with performance that is comparable to the Line-of-Sight (LoS) link without consuming any extra energy for signal power amplification [1]. Due to its potential to improve the wireless communication performance with relatively low implementation and maintaining costs, IRS has been promoted by both industry and academia as the key enabling technology for next generation wireless communication systems [2].

Recent report suggests that deploying multiple IRSs has the potential to further improve the signal reflection performance and alleviate the co-channel interference between different signal sources [3]. However, enabling multiple IRSs to enhance the system performance introduces several challenges. First, it is known that, to maximize communication performance, an IRS needs to keep track of the channel state information (CSI) between itself and the signal source as well as the intended receivers. In the multi-IRS system, each IRS should not only coordinate with with its own associated sources and receivers but also carefully coordinate with other IRSs to avoid introducing the co-channel interference caused by reflecting signals toward unintended receivers. The total volume of coordination information among signal sources, IRSs and the receivers is expected to grow significantly with the number of IRSs and the number of elements of each IRS. Second, in a multi-user system consisting of multiple signal sources and receivers, how to improve the overall system performance by allocating different IRSs to serve different sources or receivers is still an open problem. This problem is further exacerbated by the fact that in mobile network system, users can constantly move from one location to another. In this case, dynamically evaluating and adjusting the BS-IRS association is critical for maximizing the long-term performance of IRS-assisted wireless system.

In this paper, we consider a mobile networking system consisting of multiple base stations (BSs), and each offers wireless services to users within an exclusive service area (e.g., cell). Multiple IRSs are deployed throughout the service areas of BSs to further enhance the downlink data communication performance from BSs to users. We consider a dynamic environment in which users can move between different service areas of BSs at different time. As mentioned earlier, optimizing the IRS-assisted data communication generally require constantly global coordination among all the BSs, IRSs, as well as the users. Inspired by recent success in learning-based methods [4]–[7], we propose MDLBI, a multi-agent deep reinforcement learning-based BS-IRS association scheme to allow each BS to compete for IRSs to serve their users. In MDLBI, each BS can neither know the IRS selection policy of others nor communicate with other BSs. Each BS, however, can learn and maintain a parameterized actor function which maps its locally observed states into its optimal decision. MDLBI is easy to implement and scalable to large networking systems. Extensive simulation has been conducted. Our results show that MDLBI converges well compared to existing benchmark methods and can be directly applied into networks with a large number of BSs and IRSs. To the best of our knowledge, this is the first work to study the BS-IRS association problem in a dynamic networking environment.

The reminder of this paper is organized as follows. Section II presents the related work. System model and problem for-

mulation are described in Section III. We present the proposed MDLBI in Section IV. The simulation results are presented in Section V and paper is concluded in Section VI.

II. Related Work

Single IRS-assited wireless network: Most existing works on IRS-assisted wireless networks focus on optimizing the single-IRS scenarios. Particularly, in [8], the authors applied IRS to minimized the total transmit power at the BS. A joint optimization problem of both transmit beamforming of active BS antennas and the reflect beamforming of passive phase shifters has been investigated. The authors in [4] applied deep reinforcement learning to optimize the overall energy efficiency for IRS powered by harvested energy. The authors in [5] adopted a deep reinforcement learning method to optimize the transmit beamforming and reflect beamforming in dynamic environment. The authors in [9] proposed a novel phase shift solution for IRS to minimize the co-channel interference for multiple receivers sharing the same spectrum. Recently, IRS is also utilized to maximize the secrecy rate of the legitimate communication link [10] and extend the wireless coverage with ultra-reliable low-latency communication services [11].

Multi-IRS-assisted wireless network: Multi-IRS-assisted network has attracted significant interest due to its potential to significantly improve the spectrum and energy efficiency [12]–[16]. For example, the authors in [12] studied the joint optimization problem of the transmit beamforming, reflect beamforming, and the set of active IRSs to minimize both the transmit power consumption of the BS and circuit power consumption of the IRSs. The authors in [13] investigated the joint design of transmit beamforming and artificial noise covariance matrix at an access point and the reflect beamforming at the IRSs to maximize the system sum rate while limiting the maximum information leakage at the potential eavesdroppers. The joint active and passive beamforming optimization problem for IRS-assisted simultaneous wireless information and power transfer (SWIPT) is studied in [14]. Different from the full knowledge channel state information (CIS) assumption in [12]–[14], the authors in [15] designed a transmit and reflect beamforming solution based on the imperfect location information of users. In [16], the authors performs an initial investigation on the IRS association problem, where the dynamics of environment is not considered.

III. System Model And Problem Formulation

A. System Model

We consider a multiple IRS-assisted network consisting of M BSs, each has N_b antennas, that provides services to K single-antenna users in the considered area, as shown in Fig. 1. Let $\mathbb{M} = \{b_1, b_2, \ldots, b_M\}$ and $\mathbb{K} = \{1, 2, \ldots, K\}$ be the sets of BSs and users, respectively. Each BS covers an exclusive sub-region in the service area. We consider a mobile network in which users can move from one sub-region to another. The user mobility can be regarded as a slotted process in which the set of users located in the sub-region of each BS can be considered as fixed within each time slot, i.e., we use $\mathbb{C}_{m,t}$ to denote as the set of users served by BS b_m during time slot t and $\cup_{m\in\mathbb{M}}\mathbb{C}_{m,t} = \mathbb{K}$. To simplify our description, we focus on the downlink communication and the main objective of each BS is to maximize the data rates from itself to the users. We assume a proper inter-cell interference cancellation mechanism has been adopted between BSs and thus the data transmission of each BS does not cause any noticeable interference to the users located in the coverage area of other BSs.

Suppose L IRSs are deployed in the service area that can be utilized by BSs to improve the downlink data communication performances of BSs. Let $\mathbb{L} = \{1, 2, \ldots, L\}$ be the set of all the IRSs and N_l be the number of passive reflecting elements of the lth IRS for $l \in \mathbb{L}$. BSs compete for the control of IRSs to serve their users. We assume each IRS can only serve a single BS in each time slot. Each BS, however, can take control of multiple IRSs. $\mathbb{N}_{m,t}$ denotes the IRS set controlled by BS b_m during time slot t and we have $\cup_{m\in\mathbb{M}}\mathbb{N}_{m,t} = \mathbb{L}$.

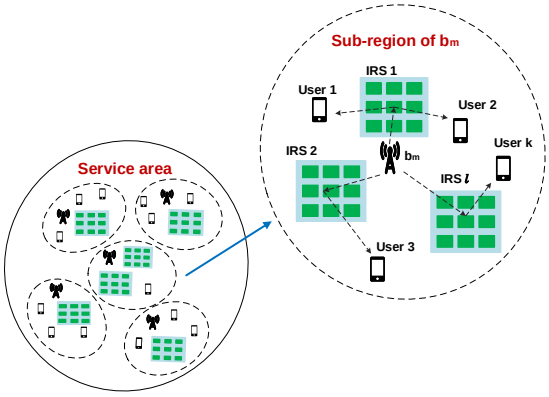

Fig. 1. System Model.

Let $\mathbf{G}_{m,l} \in \mathbb{C}^{N_l \times N_b}$ and $\mathbf{h}_{l,k}^H \in \mathbb{C}^{1\times N_l}$ to be the complex equivalent baseband channel vectors between each BS b_m and the lth IRS, and between the lth IRS and the kth user, respectively. The channel gain between BS b_m and the kth user is given by $\mathbf{H}_{m,k} \in \mathbb{C}^{1\times N_b}$. We assume $\mathbf{H}_{m,k}$, $\mathbf{G}_{m,l}$, and $\mathbf{h}_{l,k}^H$ can be regarded as constants during each transmission time slot [12]. Furthermore, the channel state information (CSI) is perfectly available at the BSs as assumed in [5], [8], [12], [13]. Let $\mathbf{\Phi}_l = \text{diag}\left\{\left[e^{j\theta_l^1}, \ldots, e^{j\theta_l^{N_l}}\right]\right\}$ be the phase shift matrix of the lth IRS, and $\theta_l^e \in [0, 2\pi]$ be the eth phase shift of the lth IRS, where $\text{diag}\{\cdot\}$ denotes diagonal matrix and $e \in \{1, \ldots, N_l\}$. We write $\mathbf{\Phi} = \text{blkdiag}\left(\mathbf{\Phi}_1, \mathbf{\Phi}_2, \ldots, \mathbf{\Phi}_l\right)$ as the phase shift matrices of all IRSs, where $\text{blkdiag}\{\cdot\}$ denotes block diagonal matrix. Let $n_k \sim \mathcal{CN}\left(0, \sigma_k^2\right)$ be the additive white Gaussian noise where σ_k^2 is the received noise power of user k.

Let $\mathbf{A} = \text{blkdiag}\left[\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_M\right]$ be the BS-IRS associ-

ation matrix, where $\mathbf{A}_m \in \mathbb{C}^{K\times L}$ is the BS-IRS association matrix of the mth BS. $[\mathbf{A}_m]_{k',l'}$ is a binary variable specifying the association relationship among the mth BS, the k'th user, and the l'th IRS. In other words, $[\mathbf{A}_m]_{k',l'} = 1$ means that the mth BS b_m allocates the l'th IRS to assist the downlink communication to the k'th user, where $l' \in \mathbb{N}_m$ and $k' \in \mathbb{C}_m$. When focusing on a specific time slot t, we write $\mathbb{N}_{m,t}$ and $\mathbb{C}_{m,t}$ as $\mathbb{N}_m$ and $\mathbb{C}_m$ for simplification. Since each user is assumed to be served by a single IRS, we have $\sum_{l'\in\mathbb{N}_m}[\mathbf{A}_m]_{k',l'} = 1, \forall k' \in \mathbb{C}_m$. We can use $\mathbf{\Phi}_{k'} = \sum_{l'\in\mathcal{L}} \mathbf{\Phi}'_l \mathbb{1}_{\{[\mathbf{A}_m]_{k',l'}=1\}}$ to represent the phase shift matrix of the single IRS allocated to the k'th user by the mth BS b_m, where $\mathbb{1}_{\{.\}}$ is the indicator function.

The received signal at the k'th user served by BS b_m can be written as

$$y_{m,k'} = \mathbf{H}_{m,k'}\mathbf{x} + \mathbf{h}_{l,k'}^H \mathbf{\Phi}_{k'} \mathbf{G}_{m,l}\mathbf{x} + n_{k'}, \tag{1}$$

where $(.)^H$ is conjugate transpose.

The transmitted signal of the BS b_m can be expressed as $\mathbf{x} = \sum_{k'\in\mathbb{N}_m} \mathbf{w}_{m,k'} s_{k'}$, where $s_{k'}$ denotes the desired signal of the k'th user with $s_{k'} \sim \mathcal{CN}(0,1)$, and $\mathbf{w}_{m,k'} \in \mathbb{C}^{N_b\times 1}$ is the transmit beamforming vector for the k'th user. Each BS has a maximum transmit power constraint:

$$\sum_{k'\in\mathbb{C}_m} \|\mathbf{w}_{m,k'}\|^2 \le P_{\max}, \forall m \in \mathbb{M}. \tag{2}$$

By substituting $\mathbf{x}$ into (1), we have

$$y_{m,k'} = \mathbf{H}_{m,k'} \sum_{i\in\mathbb{N}_m} \mathbf{w}_{m,i}s_i + \mathbf{h}_{l,k'}^H \mathbf{\Phi}_{k'} \mathbf{G}_{m,l} \sum_{i\in\mathbb{N}_m} \mathbf{w}_{m,i}s_i + n_{k'}. \tag{3}$$

Accordingly, the SINR at the kth user served by BS b_m is given by

$$r_{m,k'} = \log\left(1 + \frac{\left|\mathbf{H}_{m,k'}\mathbf{w}_{k'} + \mathbf{h}_{l,k'}^H \mathbf{\Phi}_{k'} \mathbf{G}_{m,k'} \mathbf{w}_{k'}\right|^2}{\sum_{i\neq k'}^{K} \left|\mathbf{h}_{l,k'}^H \mathbf{\Phi}_{k'} \mathbf{G}_{m,k'} \sum_{i\in\mathbb{N}_m} \mathbf{w}_i\right|^2 + \sigma_{k'}^2}\right). \tag{4}$$

The achievable sum rate of a multiple IRS-assisted wireless communication system is given by

$$R = \sum_{m\in\mathcal{M}} \sum_{k'\in\mathbb{C}_m} r_{m,k'}. \tag{5}$$

B. Problem Formulation

Given the defined system model, our goal is to maximize the sum rate of the multiple IRS-assisted communication system by jointly optimizing the user scheduling and association matrix $\mathbf{A}$, phase shifts matrix $\mathbf{\Phi}$ and transmit beamforming matrix $\mathbf{w}$, i.e., the joint optimizing problem can be written as follows:

$$(\mathcal{P}): \max_{\mathbf{w},\mathbf{\Phi},\mathbf{A}} R, \tag{6a}$$

$$\text{s.t.} \sum_{k'\in\mathbb{C}_m} \|\mathbf{w}_{m,k'}\|^2 \le P_{\max}, \forall m \in \mathcal{M}, \tag{6b}$$

$$\left|e^{\theta_{l'}^m}\right| = 1, \forall m \in \mathcal{M}, l' \in \mathbb{N}_m, \tag{6c}$$

$$[\mathbf{A}_m]_{k',l'} \in \{0,1\}, \forall m \in \mathcal{M}, l' \in \mathbb{N}_m, k' \in \mathbb{C}_m. \tag{6d}$$

We can observe that the objective function of problem $(\mathcal{P})$ is generally non concave and the three optimization variables $\mathbf{w}$, $\mathbf{\Phi}$, and $\mathbf{A}$ are coupled with each other which make the problem difficult to solve. Besides, the constraint (6c) is highly non-convex as the phase of each element is forced to have a unit magnitude. In the rest of the paper, we propose a Multi-agent Deep Deterministic Policy Gradient (MDDPG)-based solution, called MDLBI to jointly optimize transmit beamforming $\mathbf{w}$, reflect beamforming $\mathbf{\Phi}$, and BS-IRS association $\mathbf{A}$.

IV. Multi-agent DDPG for the Multiple IRSs-assisted Communication System

In this section, we introduce a MDDPG-based solution, called MDLBI for optimizing the multiple IRSs-assisted communication system.

In this method, each BS can observe the state including the channel information (i.e., $\mathbf{H}_{m,k'}$, $\mathbf{G}_{m,l'}$, $\mathbf{h}_{l',k'}^H$, and leakaged control signal, $\forall l' \in \mathbb{N}_m, \forall k' \in \mathbb{C}_m$) , output the actions (i.e., phase shifts $\mathbf{\Phi}_{l'}, \forall l' \in \mathbb{N}_m$, the transmit beamforming $\mathbf{w}_m, \forall m \in \mathbb{M}$, and the association matrix $\mathbf{A}_m, \forall m \in \mathbb{M}$), and obtain reward (i.e., sum rate $\sum_{k'\in\mathbb{C}_m} r_{k'}$) during each time slot t, as shown in Fig. 2.

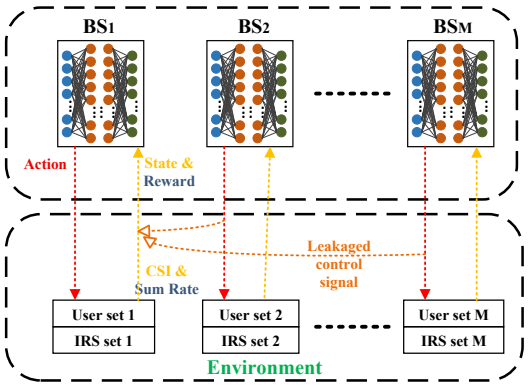

Fig. 2. MDDPG based optimization for the multiple IRSs-assisted communication system

One of the main advantages for adopting MDDPG-based method to address problem $(\mathcal{P})$ is that MDDPG is applicable to continuous and high-dimensional action spaces. This makes it suitable for our problem, in which the phase shifts $\mathbf{\Phi}$ are continuous and the association matrix $\mathbf{A}$ is high-dimensional.

DDPG is a deep reinforcement learning algorithm that could operate over continuous action space by maintaining an actor network to specify the current policy deterministically by mapping the observed state to a specific action. Besides, DDPG requires fewer steps of experiences than Deep Q Network (DQN) to find optimal solutions in the Atari domain [17]. Under the actor-critic architecture, DDPG maintains two networks, namely the actor network and the critic network. The critic network is trained to approximate the Q-table using neural networks without the curse of dimension, while the actor network is trained to generate a deterministic policy instead of policy gradient. Furthermore, target networks are also adopted to improve the stability [18]. However, DDPG is not specialized for the multi-agent environment since the environment is non-stationary from the perspective of each agent (i.e. BS). By utilizing of the control signal between the BSs and IRSs, we extend DDPG into a multi-agent version for our task. More specifically, there are four sub-nets, namely the critic network $Q^{\mu}_{\theta^{'}_m}(o_m, a_m)$, the target critic network $Q^{\mu^{'}}_{\theta^{'}}(o_m, a_m)$, the actor network $\mu_{\theta_m}(o_m)$ (abbreviated as μ_m), and the target actor network $\mu^{'}_{\theta^{'}}(o_m)$. We use θ_c, $\theta_{m'}$, θ_m, and $\theta^{'}$ to denote the parameters of the critic network, the actor network, and the target networks during certain time slots, respectively. Note that a_m is the action of b_m at the current time slot, and o_m is the current state of BS b_m. It contains two parts $o_m = (s_m, \rho_{-m})$, where s_m is the local state of physical environments (i.e., all channel state information in the sub-region of b_m) and ρ_{-m} represents the other BS agents' strategies to dominate the IRS sets. This observation of other BSs' strategies to choose certain IRS sets can be realized by capturing the leakage of control siganls between other BSs and IRSs, since the control of IRS are typically connected with the BS through wireless [2].

With the above notations, we formally describe the construction of the model in detail as follows.

State Space: $o_m \in \mathbb{O}_m$ is the state of b_m, which is determined by the physical environment (direct channel $\mathbf{H}_{m,k}$, channel between b_m and the lth IRS $\mathbf{G}_{m,l}$ and channel between the lth IRS and the kth user $\mathbf{h}^H_{l,k}, \forall l \in \mathbb{N}_m, \forall k \in \mathbb{C}_m$) and the other BS agents' strategies ρ_{-m}. Since neural network can only take real numbers and the channel information are complex, we separate the real and imaginary parts of the channel information as independent inputs. There are $2card(\mathbb{C}_m)N_b$, $2card(\mathbb{N}_m)N_bN_l$, $2card(\mathbb{N}_m)N_lcard(\mathbb{C}_m)$ and $(M-1)L$ entries respectively contributed by $\mathbf{H}_{m,k}$, $\mathbf{G}_{m,l}$, $\mathbf{h}^H_{l,k}$ and ρ_{-m}, where $card(\cdot)$ represents the cardinality of a set. Hence, the total number of entries for sate action is $D_s = 2card(\mathbb{C}_m)N_b + 2card(\mathbb{N}_m)N_bN_l + 2card(\mathbb{N}_m)N_lcard(\mathbb{C}_m) + (M-1)L$.

Action Space: $a_m \in \mathbb{A}_m$ is the action at current state, which is constructed by phase shifts $\mathbf{\Phi}_l, \forall l \in \mathbb{N}_m$, transmit beamforming vectors $\mathbf{w}_k, \forall k \in \mathbb{C}_m$ and association matrix $\mathbf{A}_m$. Likewise, there are $2N_bcard(\mathbb{C}_m)$, $card(\mathbb{N}_m)N_l$ and L entries of action contributed by $\mathbf{w}_k, \forall k \in \mathbb{C}_m$, $\mathbf{\Phi}_l, \forall l \in \mathbb{N}_m$ and $\mathbf{A}_m$, respectively. Hence, the total number of entries for action space is $D_a = 2N_bcard(\mathbb{C}_m) + card(\mathbb{N}_m)N_l + L$.

Reward: r_m represents the instant reward defined as the sum rate of users in the sub-region of b_m, which can be obtained by knowing the other BS agents' strategies ρ_{-m}, the instantaneous channel information $\mathbf{H}_{m,k}, \forall k$, $\mathbf{G}_{m,l} \forall l$ and $\mathbf{h}^H_{l,k'} \forall l, k'$ and the action (i.e., $\mathbf{w}_{m,k'}, \forall k'$, $\mathbf{\Phi}_l, \forall l \in \mathbb{N}_m$ and $\mathbf{A}_m$) obtained from the actor network.

The expected reward of each BS b_m is given by

$$J(\theta_m) = \mathbb{E}_{s\sim p^{\mu}, a_m \sim \mu_m}\left[\nabla_{\theta_m} \log \mu_m(a_m \mid o_m) Q^{\mu}_m(o_m, a_m)\right], \tag{7}$$

where p^{μ} is the state distribution.

Considering the continuous of action space, the gradient for the parameter θ_m of the deterministic policy μ_{θ_m} (abbreviated μ_m) is given as

$$\begin{aligned}\nabla_{\theta_m} J(\boldsymbol{\mu}_m) =& \mathbb{E}_{o\sim\mathcal{D}}\left[\nabla_{\theta_m} \boldsymbol{\mu}_m(a_m \mid o_m)\right. \\ & \left. \cdot \nabla_{a_m} Q^{\mu}_m(o_m)|_{a_m = \boldsymbol{\mu}_m(o_m)}\right]\end{aligned} \tag{8}$$

where $\mathcal{D}$ is the experience buffer contain trumples $(o_m, o^{'}_m, r_m)$, recording the experiences of all the agent BS.

Then the critic network is updated by minimizing the following loss function:

$$\mathcal{L}(\theta_m) = \mathbb{E}_{o_m, o^{'}_m, r_m}\left[\left(Q^{\mu}_m(o_m) - y\right)^2\right], \tag{9}$$

$$y = r_m + \gamma Q^{\boldsymbol{\mu}^{'}}_m(o_m)\Big|_{\rho_{-m}}, \tag{10}$$

where $\mu^{'}$ is the set of target policies.

Finally, DDPG softly updates the target networks with a small instant $\tau \ll 1$, i.e.,

$$\theta^{'} \leftarrow \tau\theta_m + (1-\tau)\theta^{'}. \tag{11}$$

This means the target networks are changed in a much slower speed than the actor and critic networks, which greatly improving the stability of the learning [17].

Given the set of IRS $\mathbb{N}_m$, the BS b_m interacts with the environment in a trial-and-error manner to optimize the sum rate of user set $\mathbb{C}_m$ in its sub-region. During each time step t of an episode, each BS (e.g. b_m) observes the current state o_m, applies a action a_m defined by policy μ_{θ_m} to the environment, and obtains the instant reward r_m.

The details of the proposed algorithm are presented in Algorithm 1. **At the beginning of the algorithm,** parameters of the BSs and environment are initialized. In this paper, the experience buffer $\mathcal{D}$, the other users' trategies ρ_{-m}, the actor network parameters θ_m, the target network parameters $\theta^{'}$, the critic network parameters $\theta^{'}_m$ and the association matrix $\mathbf{A}$ are randomly initialized. And the transmit beamforming $\mathbf{w}$, the phase shifts $\mathbf{\Phi}$ are simply initialized as identity matrix. **After initialization**, the new experience $(s_t, a_t, r_{t+1}, s_{t+1})$ are collected into the experience buffer $\mathcal{B}$ (i.e., the step 4-8). A minibatch of experience with size w is randomly sampled

from $\mathcal{B}$, i.e., step 9. The step 10 and 11 describe the update of the critic network and the actor network. Finally, the target networks are updated, i.e., the step 12. The algorithm run over N episodes and each episode with T time steps. During each episode, the algorithm terminates whenever it converges or finishes the T time steps.

Algorithm 1 Multi-agent DDPG based optimization with IRS-BS association b_m

Output: $a_m = \{\mathbf{w}_m, \mathbf{\Phi}_l \forall l \in \mathbb{N}_m, \mathbf{A}_k\}$, Q value function
Initialization: experience buffer $\mathcal{D}$ with size D_l, the actor network parameter θ_m, the target networks parameter $\theta_\prime$, the critic network parameter θ'_m, the transmit beamforming $\mathbf{w}_m$, the phase shifts $\mathbf{\Phi}_l, \forall l \in \mathbb{N}_m$ and the association matrix $\mathbf{A}_m$;

1: **for** $episode = 0, 1, 2, \ldots, N-1$ **do**
2: Collect the channel information $\mathbf{H}_{m,k}, \forall k \in \mathbb{C}_m$, $\mathbf{G}_{m,l}, \forall l \in \mathbb{N}_m$ and $\mathbf{h}_{l,k}^H, \forall l \in \mathbb{N}_m, \forall k \in \mathbb{C}_m$ for the nth episode to obtain the first state s_0;
3: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
4: Obtain the IRS set $\mathbb{C}_{m,t}$ dominate by the BS b_m
5: Observe action $a_m = \{\mathbf{w}_m, \mathbf{\Phi}_l \forall l \in \mathbb{N}_{m,t}, \mathbf{A}_k\}$= $\mu(\theta_m \mid o_m)$ from the actor network;
6: Observe the next state o'_m given action a_m;
7: Obtain the reward r_m;
8: Store the experience (o_m, a_m, r_m, o'_m) in the experience buffer $\mathcal{D}$;
9: Sample a random minibatch of W transitions from the experience buffer $\mathcal{D}$;
10: Update the critic network by minimizing the loss as described in Eq. (9);
11: Update the actor network using the policy gradient as described in Eq. (8);
12: Update the target networks using Eq. (11);
13: **end for**
14: **end for**

V. Stimulation and Results

In this section, we evaluate the performance of the proposed MDDPG-based algorithm. All the channels are assumed to suffer from both path loss and Rayleigh fading. As in [12], [19], the path loss exponents of the BS-IRS channel and the IRS-user channel are set as 2.5 and 2.4, respectively. The actor and critic network are both fully connected deep neural network with one input layer, two hidden layers and an output layer. The output layer of the actor network has the same dimension as the action and we use tanh function as the activation function. The output of the critic network is a scalar with one dimension. The activation function of all the hidden layers is Relu function.

In Fig.3, we investigate the convergence performance of MDLBI compared to the original DDPG-based solution. Since the batch sampling is utilized, we use time as the x axis instead of time steps. We can observe that the proposed solution converges much faster and can obtain a higher sum rate at around 8.5 bps/Hz. This means that the proposed MDLBI could improve the transmit beamforming compared to DDPG-based method. In practical system, each BS could utilize the leakage control signals of IRSs sent by other BSs to estimate their competition over the IRSs. In this way, the possibility of collisions between BSs when competing for the same IRS can be reduced.

TABLE I
Simulation Congiguration

Parameters	Description	Value
γ	Discount factor	0.99
μ_c	Learning rate of the actor	0.0001
μ_a	Learning rate of the critic	0.001
τ	Soft target update factor	0.001
D	Experience buffer Size	10000
N	Number of episodes	5000
T	Maximum time steps of each episode	800
W	Batch size	64

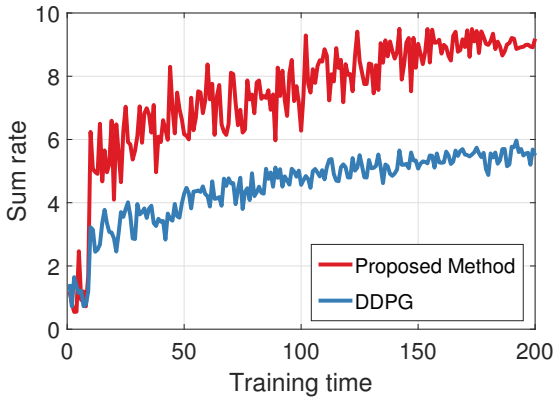

Fig. 3. Instant reward (dB) as a function of time

In Fig.4, we consider the instant reward (i.e. sum rate) under different system settings $L = \{1, 2, 4\}$,$K = \{10, 30\}$. The number of BS M is fixed to 2. It can be observed that the sum rate of the service area increases with the number of IRSs. By comparing cases with $\{K = 10, L = 1\}$ and $\{K = 30, L = 1\}$, we can observe that the performance enhancement induced by a single IRS increased a little with the number of served users. However, when more IRSs can be deployed in the service area, such as cased with $\{K = 30, L = 2\}$ and $\{K = 30, L = 4\}$, the sum rate could be further improved. This result verifies that deploying more number of IRSs even in a distributed manner could improve the system capacity.

Fig. 5 compares the sum rate of our proposed method to that of a fixed BS-IRS association solution. It can be observed

that although the optimization with a fixed association strategy may obtain higher sum rate at the beginning of the time of consideration, the overall system performance may fluctuate with time due to the user mobility.

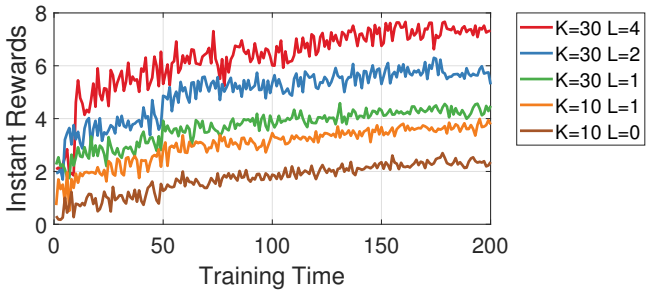

Fig. 4. Instant reward (dB) as a function of time under different system settings

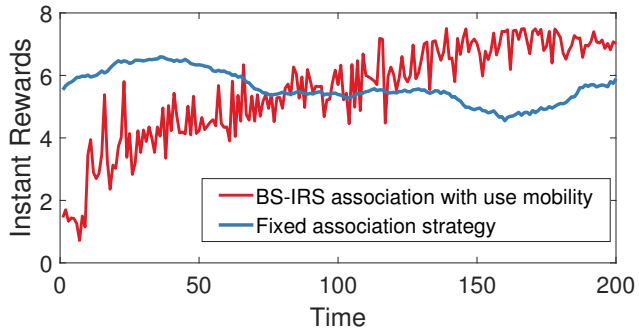

Fig. 5. Instant reward (dB) as a function of time under different association strategies

VI. Conclusion

In this paper, we investigate the IRS-BS association problem in a mobile network consisting of multiple BSs serving a set of mobile users assisted by multiple IRSs. In order to maximize the long-term data communication performance for the associated users located in their service coverage areas, the BSs compete with each other for controlling the phase shift of a limited number of IRSs. A multi-agent reinforcement learning-based solution, named as MDLBI, is proposed to optimize the BS-IRS association and the phase-shift of each IRS. The MDLBI achieves the maximum downlink communication sum rate without requiring any data exchange among BSs. Extensive simulations have been conducted to demonstrate that MDLBI achieves significant performance improvement even when being implemented in large networking systems.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grants 62071193 and 61632019, the Key R & D Program of Hubei Province of China under Grant 2020BAA002, China Postdoctoral Science Foundation under Grant 2020M672357.

References

[1] M. Dunna, C. Zhang, D. Sievenpiper, and D. Bharadia, "Scattermimo: Enabling virtual mimo with smart surfaces," in *International Conference on Mobile Computing and Networking (MobiCom)*, London, United Kingdom, Apr. 2020.

[2] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y. C. Liang, "Towards smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Communications Surveys and Tutorials*, vol. 22, no. 4, pp. 2283–2314, Jun. 2020.

[3] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.

[4] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *IEEE International Conference on Communications (ICC)*, Dublin, Ireland, Jul. 2020.

[5] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser miso systems exploiting deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1839–1850, Jun. 2020.

[6] Y. Xiao and M. Krunz, "Distributed optimization for energy-efficient fog computing in the tactile internet," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2390–2400, Nov. 2018.

[7] Y. Xiao, G. Shi, Y. Li, W. Saad, and H. Vincent Poor, "Toward self-learning edge intelligence in 6g," *IEEE Communications Magazine*, vol. 58, no. 12, pp. 34–40, Dec. 2020.

[8] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[9] X. Tan, Z. Sun, J. M. Jornet, and D. Pados, "Increasing indoor spectrum sharing capacity using smart reflect-array," in *IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, Jul. 2016.

[10] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Communications Letters*, vol. 8, no. 5, pp. 1410–1414, May 2019.

[11] L. Yang, Y. Yang, M. O. Hasna, and M. S. Alouini, "Coverage, probability of snr gain, and dor analysis of ris-aided communication systems," *IEEE Wireless Communications Letters*, vol. 9, no. 8, pp. 1268–1272, Apr. 2020.

[12] S. Sun, M. Fu, Y. Shi, and Y. Zhou, "Towards reconfigurable intelligent surfaces powered green wireless networks," in *IEEE Wireless Communications and Networking Conference (WCNC)*, Seoul, Korea, May 2020.

[13] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2637–2652, Jul. 2020.

[14] Q. Wu and R. Zhang, "Joint active and passive beamforming optimization for intelligent reflecting surface assisted swipt under qos constraints," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1735–1748, Jul. 2020.

[15] X. Hu, C. Zhong, Y. Zhang, X. Chen, and Z. Zhang, "Location information aided multiple intelligent reflecting surface systems," *IEEE Transactions on Communications*, vol. 68, no. 12, pp. 7948–7962, Aug. 2020.

[16] W. Mei and R. Zhang, "Joint base station-irs-user association in multi-irs-aided wireless network," in *IEEE Global Communications Conference*, Taipei, Taiwan, Dec. 2020.

[17] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *2015, arXiv:1509.02971*, [Online]. Availiable: https://arxiv.org/abs/1509.02971v1.

[18] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[19] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 4157–4170, Jun. 2019.