2021 IEEE International Conference on Communications Workshops (ICC Workshops) | 978-1-7281-9441-7/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCWorkshops50388.2021.9473585

Multi-IRS-assisted Multi-Cell Uplink MIMO Communications under Imperfect CSI: A Deep Reinforcement Learning Approach

Junghoon Kim*, Seyyedali Hosseinalipour*, Taejoon Kim†, David J. Love* and Christopher G. Brinton*
*Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
†Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA
*{kim3220, hosseina, djlove, cgb}@purdue.edu, †taejoonkim@ku.edu

***Abstract*—Applications of intelligent reflecting surfaces (IRSs) in wireless networks have attracted significant attention recently. Most of the relevant literature is focused on the single cell setting where a single IRS is deployed and perfect channel state information (CSI) is assumed. In this work, we develop a novel methodology for *multi-IRS-assisted multi-cell networks* in the uplink. We consider the scenario in which (i) channels are dynamic and (ii) only partial CSI is available at each base station (BS); specifically, scalar effective channel powers from only a subset of user equipments (UE). We formulate the sum-rate maximization problem aiming to jointly optimize the IRS reflect beamformers, BS combiners, and UE transmit powers. In casting this as a sequential decision making problem, we propose a multi-agent deep reinforcement learning algorithm to solve it, where each BS acts as an independent agent in charge of tuning the local UE transmit powers, the local IRS reflect beamformer, and its combiners. We introduce an efficient information-sharing scheme that requires limited information exchange among neighboring BSs to cope with the non-stationarity caused by the coupling of actions taken by multiple BSs. Our numerical results show that our method obtains substantial improvement in average data rate compared to baseline approaches, e.g., fixed UE transmit power and maximum ratio combining.**

I. Introduction

Intelligent reflecting surfaces (IRSs) are one of the innovative technologies for 6G and beyond [1], [2]. An IRS is an array of passive reflecting elements with a control unit. It manipulates the propagation of an incident signal by providing an abrupt phase shift, which can control the communication channel. IRSs are utilized to provide enhanced communication efficiency without building extra infrastructure [3]–[9]. In this paper, we study a scenario where multiple IRSs are deployed in a multi-cell cellular setting to provide enhanced data rates to the users.

A. Related Work

Exploiting IRSs in cellular networks initiated with applications of this technology in the downlink (DL). Studying IRS use cases in the uplink (UL) is thus comparably more recent.

1) Utilizing IRSs in the DL: Most of the relevant literature has considered a *single cell* system with a *single IRS* [3], [4]. Specific investigations have included quality of service (QoS)-constrained transmit power minimization [3] and weighted sum-rate maximization [4] to obtain the base station (BS) beamformer and IRS reflect beamformer/precoder in the DL. Unlike the prior approaches, the work in [5] considers a *multi-cell* scenario with a single IRS, where the BS precoders and IRS reflect beamformer are designed to maximize sum-rate.

2) Utilizing IRSs in the UL: Most of the works in UL design are also focused on single cell systems with a single IRS [6]–[8]. Several of these works have studied IRS reflect beamformer design and uplink user equipment (UE) power control problems, where the impact of quantized IRS phase values [6] and compressed sensing-based user detection [8] on the uplink throughput have also been investigated. The concept of IRS resembles analog beamforming in millimeter-wave (mmWave)-based systems [7]. Recently, systems with two IRSs have been considered focusing on SINR fairness [9].

Despite the potential benefit of improving multi-cell-wide performance, multi-IRS deployment in multi-cell UL scenarios has not been thoroughly modeled and studied due to the added optimization complexity involved in controlling multiple IRSs.

B. Overview of Methodology and Contributions

In this work, we develop an architecture for multi-IRS-assisted multi-cell UL networks. Our methodology explicitly considers multi-order reflections among IRSs, which is rarely done in existing literature. We address the scenario where (i) channels are time-varying, and (ii) only partial/imperfect CSI is available, in which each BS only has knowledge of scalar effective channel powers from a subset of UEs. This is more practical and realistic as compared to the prior approaches [6]–[9] that assume perfect knowledge of all channel matrices. We formulate the sum-rate maximization problem aiming to jointly optimize UE transmit powers, IRS reflect beamformers, and BS combiners across cells.

Given the interdependencies between the design variables across different cells, we cast the problem as one of sequential decision making and tailor a multi-agent deep reinforcement learning (DRL) algorithm to solve it. We consider each BS as an independent learning agent that controls the local UE transmit powers, the local IRS reflect beamformer, and its combiners via only index gradient variables. We design the state, action, and reward function for each BS to capture the interdependencies among the design choices made at different BSs. We further develop an information-sharing scheme where only limited information among neighboring BSs is exchanged

978-1-7281-9441-7/21/$31.00 ©2021 IEEE

to cope with the non-stationarity issue caused by the coupling between the actions at other BSs. Through numerical simulations, we show that our proposed scheme outperforms the conventional baselines for data rate maximization.

II. Multi-cell Systems with Multiple IRSs

In this section, we first introduce the signal model under consideration (Sec. II-A). Then, we formulate the optimization and discuss the challenges associated with solving it (Sec. II-B).

A. Signal Model

We consider a multi-cell system with multiple IRSs for the uplink (UL) as depicted in Fig. 1. The system is comprised of a set of L cells $\mathcal{L} = \{1, ..., L\}$ and R IRSs $\mathcal{R} = \{1, ..., R\}$. For simplicity we assume that each cell has one IRS, i.e., $R = L$, though our method can be readily generalized to the case where $R \neq L$. The IRSs are indexed such that cell ℓ contains IRS ℓ.

Each cell $\ell \in \mathcal{L}$ contains (i) K_ℓ UEs with single antenna, denoted by $\mathcal{K}_\ell = \{1, ..., K_\ell\}$, (ii) an IRS with N_ℓ reflecting elements, denoted by $\mathcal{N}_\ell = \{1, ..., N_\ell\}$, and (iii) a BS with M_ℓ antennas denoted by $\mathcal{M}_\ell = \{1, ..., M_\ell\}$. We let UE (i, j) refer to UE j in cell i. The received signal vector at BS $\ell \in \mathcal{L}$ at the tth channel instance is given by

$$\mathbf{y}_\ell[t] = \sum_{i\in\mathcal{L}} \sum_{j\in\mathcal{K}_i} \Bigg\{ \Bigg(\mathbf{h}^{\mathrm{UB}}_{(i,j),\ell}[t] + \sum_{r\in\mathcal{R}} \mathbf{G}^{\mathrm{IB}}_{r,\ell}[t] \mathbf{\Phi}_r[t] \mathbf{h}^{\mathrm{UI}}_{(i,j),r}[t] + \sum_{r_2\in\mathcal{R}} \sum_{r_1\in\mathcal{R}\setminus\{r_2\}} \mathbf{G}^{\mathrm{IB}}_{r_2,\ell}[t] \mathbf{\Phi}_{r_2}[t] \mathbf{G}^{\mathrm{II}}_{r_1,r_2}[t] \mathbf{\Phi}_{r_1}[t] \mathbf{h}^{\mathrm{UI}}_{(i,j),r_1}[t] \Bigg) \sqrt{p_{i,j}[t]} s_{i,j}[t] \Bigg\} + \mathbf{n}_\ell[t], \tag{1}$$

where $\mathbf{h}^{\mathrm{UB}}_{(i,j),\ell}[t] \in \mathbb{C}^{M_\ell \times 1}$ is the direct channel from UE (i, j) to BS ℓ, $\mathbf{h}^{\mathrm{UI}}_{(i,j),r}[t] \in \mathbb{C}^{N_r \times 1}$ is the channel from UE (i, j) to IRS $r \in \mathcal{R}$, $\mathbf{G}^{\mathrm{IB}}_{r,\ell}[t] \in \mathbb{C}^{M_\ell \times N_r}$ is the channel from IRS r to BS ℓ, and $\mathbf{G}^{\mathrm{II}}_{r_1,r_2}[t] \in \mathbb{C}^{N_{r_2} \times N_{r_1}}$ is the channel from IRS r_1 to IRS r_2, $r_1 \neq r_2$. Also, $p_{i,j}[t] \in \mathbb{R}^+$ is the transmit power and $s_{i,j}[t] \in \mathbb{C}$ is the transmit symbol of UE (i, j), where $\mathbb{E}[|s_{i,j}[t]|^2] = 1$. The noise vector $\mathbf{n}_\ell[t] \in \mathbb{C}^{M_\ell \times 1}$ at BS ℓ is assumed to be distributed according to zero mean complex Gaussian with covariance matrix $\sigma^2\mathbf{I}$, i.e., $\mathbf{n}_\ell[t] \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ denotes the identity matrix and σ^2 is the noise variance. Finally, $\mathbf{\Phi}_r[t] = \mathrm{diag}(\phi_{r,1}[t], \phi_{r,2}[t], ..., \phi_{r,N_r}[t]) \in \mathbb{C}^{N_r \times N_r}$ is a diagonal matrix with its diagonal entries representing the beamforming vector of IRS $r \in \mathcal{R}$. $\phi_{r,n}[t]$, $n \in \mathcal{N}_r$, is modeled as $\phi_{r,n}[t] = a_{r,n}[t]e^{j2\pi\theta_{r,n}[t]} \in \mathbb{C}$, incurring the signal attenuation $a_{r,n}[t] \in [0, 1]$ and phase shift $\theta_{r,n}[t] \in [0, 2\pi)$.

In (1), we consider the channels with three different paths from UE (i, j) to BS ℓ: (i) the direct channel, (ii) the channel after one reflection from the IRSs (the sum over r), and (iii) the channel after two reflections from the IRSs (the sum over r_1, r_2). Higher order reflections can also be incorporated in (1), i.e., signals reflected from more than two IRSs; we focus on up to the second-order reflections due to a large attenuation induced by multiple reflections between IRSs.

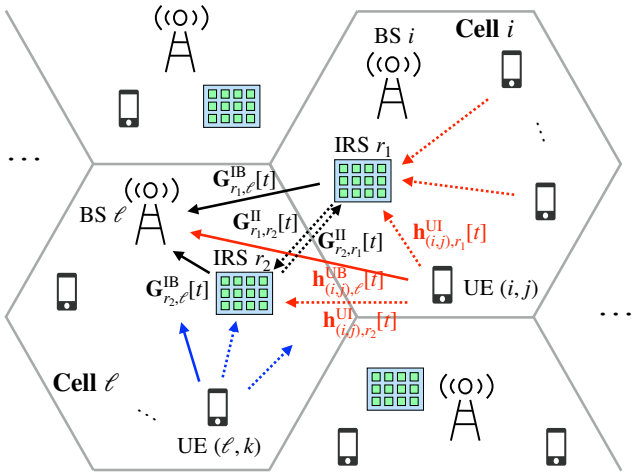


Fig. 1: Depiction of a multi-IRS-aided multi-cell system in the UL.

We assume that a linear combiner $\mathbf{z}_{\ell,k}[t] \in \mathbb{C}^{M_\ell \times 1}$ is employed at BS ℓ to restore $s_{\ell,k}[t]$ from $\mathbf{y}_\ell[t]$, which yields

$$\hat{y}_{\ell,k}[t] = \mathbf{z}^H_{\ell,k}[t]\mathbf{y}_\ell[t], \tag{2}$$

where superscript H denotes the conjugate transpose.

B. Problem Formulation and Challenges

We aim to maximize the sum-rate over all the UEs in the network through design of the UE powers $\{p_{\ell,k}[t]\}_{\ell,k}$, BS combiners $\{\mathbf{z}_{\ell,k}[t]\}_{\ell,k}$, and IRS beamformers $\{\boldsymbol{\phi}_r[t]\}_r$, where $\boldsymbol{\phi}_r[t] = [\phi_{r,1}[t], \phi_{r,2}[t], ..., \phi_{r,N_r}[t]]^\top \in \mathbb{C}^{N_r \times 1}$ is the IRS beamforming vector on the diagonal of $\mathbf{\Phi}_r[t]$, i.e., $\mathbf{\Phi}_r[t] = \mathrm{diag}(\boldsymbol{\phi}_r[t])$. With $\mathrm{SINR}_{\ell,k}[t]$ as the signal-to-interference ratio (SINR) of UE (ℓ, k), we propose the following optimization problem:

$$\begin{aligned} \text{maximize} \quad & \sum_{\ell\in\mathcal{L}} \sum_{k\in\mathcal{K}_\ell} \log_2(1 + \mathrm{SINR}_{\ell,k}[t]) \\ \text{subject to} \quad & p_{\ell,k}[t] \in \mathcal{P}, \ \ \mathbf{z}_{\ell,k}[t] \in \mathcal{Z}, \ \ \boldsymbol{\phi}_r[t] \in \mathcal{Q}, \ \ \forall \ell, \forall k, \forall r, \\ \text{variables} \quad & \{p_{\ell,k}[t]\}_{\ell,k}, \ \{\mathbf{z}_{\ell,k}[t]\}_{\ell,k}, \ \{\boldsymbol{\phi}_r[t]\}_r, \end{aligned} \tag{3}$$

where $\mathcal{P}$ is the set of power values, $\mathcal{Z}$ is the codebook for BS combiners, and $\mathcal{Q}$ is the codebook for IRS beamformers.[1]

The problem in (3) is an optimization problem at time t, where $t \in \mathcal{T} = \{0, T, 2T, ...\}$, i.e., the optimization of variables is performed once every T time instances. If the instantaneous channels $\mathbf{h}^{\mathrm{UB}}_{(i,j),\ell}[t]$, $\mathbf{h}^{\mathrm{UI}}_{(i,j),r}[t]$, $\mathbf{G}^{\mathrm{IB}}_{r,\ell}[t]$ and $\mathbf{G}^{\mathrm{II}}_{r_1,r_2}[t]$ in (1) are all known, then conventional optimization methods, e.g., successive convex approximation or integer programming, could be applied, since $\mathrm{SINR}_{\ell,k}[t]$ in (3) can be formulated as (4) (shown at the top of the next page) with

[1] A codebook structure can be employed for IRS because IRS is in practice controlled by a field-programmable gate array (FPGA) circuit where FPGA stores a set of coding sequences [10].

$$\mathrm{SINR}_{\ell,k}[t] = \frac{p_{\ell,k}[t]\left|\mathbf{z}_{\ell,k}^H[t]\left(\mathbf{h}^{\mathrm{UB}}_{(\ell,k),\ell}[t]+\sum\limits_{r\in\mathcal{R}}\mathbf{G}^{\mathrm{IB}}_{r,\ell}[t]\mathbf{\Phi}_r[t]\mathbf{h}^{\mathrm{UI}}_{(\ell,k),r}[t]+\sum\limits_{r_2\in\mathcal{R}}\sum\limits_{r_1\in\mathcal{R}\backslash\{r_2\}}\mathbf{G}^{\mathrm{IB}}_{r_2,\ell}[t]\mathbf{\Phi}_{r_2}[t]\mathbf{G}^{\mathrm{II}}_{r_1,r_2}[t]\mathbf{\Phi}_{r_1}[t]\mathbf{h}^{\mathrm{UI}}_{(\ell,k),r_1}[t]\right)\right|^2}{\sum\limits_{(i,j)\neq(\ell,k)}p_{i,j}[t]\left|\mathbf{z}_{\ell,k}^H[t]\left(\mathbf{h}^{\mathrm{UB}}_{(i,j),\ell}[t]+\sum\limits_{r\in\mathcal{R}}\mathbf{G}^{\mathrm{IB}}_{r,\ell}[t]\mathbf{\Phi}_r[t]\mathbf{h}^{\mathrm{UI}}_{(i,j),r}[t]+\sum\limits_{r_2\in\mathcal{R}}\sum\limits_{r_1\in\mathcal{R}\backslash\{r_2\}}\mathbf{G}^{\mathrm{IB}}_{r_2,\ell}[t]\mathbf{\Phi}_{r_2}[t]\mathbf{G}^{\mathrm{II}}_{r_1,r_2}[t]\mathbf{\Phi}_{r_1}[t]\mathbf{h}^{\mathrm{UI}}_{(i,j),r_1}[t]\right)\right|^2+\sigma^2} \tag{4}$$

the known channels. However, IRS-assisted wireless networks face the following challenges in practice:

- *IRS channel acquisition:* Although most of the works, e.g., [6]–[9], assume that channels are perfectly known, this assumption is impractical because an IRS is passive and often does not have sensing capabilities. While special IRS hardware with the ability to estimate the concatenated channels does exist [11], the time overhead could easily overwhelm the coherent channel resources especially when there are multiple IRSs.
- *Dynamic channels:* Channel dynamics in wireless environments adds another degree of difficulty to channel acquisition and estimation. This makes solving the optimization in (3) impossible with conventional model-based optimization approaches, due to dynamic and unknown channels.
- *Centralization:* A centralized implementation to solve (3) would require gathering all the information at a central point, which is impractical in our setting. Given the interdependencies among the design variables taken by different cells and their impact on the overall objective function, distributed optimization of the variables in (3) is challenging.

To address these challenges, we convert (3) into *a sequential decision making problem*, where the variables are designed via successive interactions with the environment through *deep reinforcement learning* (DRL). While conventional DRL assumes a centralized implementation, we develop a *multi-agent DRL* approach, where each BS acts as an independent agent in charge of tuning its local UEs transmit powers, local IRS beamformer, and combiners. To cope with the *non-stationarity* issue of multi-agent DRL [12], we carry out the learning through limited information-sharing among neighbouring BSs.

III. Multi-agent DRL Framework Design

In this section, we first introduce the information collection process at the BSs and design an information-sharing scheme (Sec. III-A). We then formulate a Markov decision process (MDP) (Sec. III-B) and propose a dynamic control scheme (Sec. III-C) to solve our optimization from Sec. II-B.

A. Local Observations and Information Exchange

We consider a setting where each BS only acquires *scalar effective channel powers* from *a subset of UEs*. When UE (i,j) transmits a pilot symbol with power $p_{i,j}[t]$, BS ℓ measures the scalar effective channel power $|\hat{h}_{(i,j),\ell,k}[t]|^2 \in \mathbb{R}$ (after combining with $\mathbf{z}_{\ell,k}[t]$), $k \in \mathcal{K}_\ell$, which is given by

$$|\hat{h}_{(i,j),\ell,k}[t]|^2 = |\mathbf{z}_{\ell,k}^H[t]\hat{\mathbf{h}}_{(i,j),\ell}[t]|^2, \tag{5}$$

where $\hat{h}_{(i,j),\ell,k}[t] \in \mathbb{C}$ is the *scalar effective channel*. The vector $\hat{\mathbf{h}}_{(i,j),\ell}[t] \in \mathbb{C}^{M_\ell \times 1}$ is the *effective channel* from UE (i,j) to BS ℓ (before combining), which is expressed as follows:

$$\hat{\mathbf{h}}_{(i,j),\ell}[t] = \sqrt{p_{i,j}[t]}\Bigg(\mathbf{h}^{\mathrm{UB}}_{(i,j),\ell}[t]+\sum_{r\in\mathcal{R}}\mathbf{G}^{\mathrm{IB}}_{r,\ell}[t]\mathbf{\Phi}_r[t]\mathbf{h}^{\mathrm{UI}}_{(i,j),r}[t] + \sum_{r_2\in\mathcal{R}}\sum_{r_1\in\mathcal{R}\backslash\{r_2\}}\mathbf{G}^{\mathrm{IB}}_{r_2,\ell}[t]\mathbf{\Phi}_{r_2}[t]\mathbf{G}^{\mathrm{II}}_{r_1,r_2}[t]\mathbf{\Phi}_{r_1}[t]\mathbf{h}^{\mathrm{UI}}_{(i,j),r_1}[t]\Bigg). \tag{6}$$

BS ℓ collects the scalar effective channel powers of the links (i) from local UEs (in cell ℓ) to BS ℓ, (ii) from neighbouring UEs (not in cell ℓ) to BS ℓ, and (iii) from local UEs to neighbouring BSs. BS ℓ measures (i) and (ii) as local observations, but needs to receive (iii), which cannot be measured by BS ℓ, from neighbouring BSs. Additionally, BS ℓ receives a *penalty value* from neighbouring BSs, where the penalty value is used for designing the reward function and will be formalized in Sec. III-B3. Note that concurrent estimation of the scalar effective channel powers of multiple UEs can be performed by UE-specific reference signals in the Long-Term Evolution (LTE) standard [13]. Acquiring scalar effective channel powers from only a subset of UEs lowers the CSI acquisition overhead compared to the conventional method of acquiring large-dimensional vector or matrix CSI from individual UE for each IRS.

To clarify which neighbouring UEs are included in (ii) and which neighbouring BSs are included in (iii), we define two sets of cell indices. First, we define the set of indices of *dominantly interfering* neighboring cells, $\mathcal{B}_\ell^{(1)}[t]$. UEs in cell $i \in \mathcal{B}_\ell^{(1)}[t]$ are dominantly interfering with the data link of local UEs (in cell ℓ). Formally, $\forall i \in \mathcal{B}_\ell^{(1)}[t],\ \forall i' \in \mathcal{L}\backslash\mathcal{B}_\ell^{(1)}[t]\backslash\{\ell\}$, we have

$$\sum_{j\in\mathcal{K}_i}\|\hat{\mathbf{h}}_{(i,j),\ell}[t]\|_2^2 \geq \sum_{j\in\mathcal{K}_{i'}}\|\hat{\mathbf{h}}_{(i',j),\ell}[t]\|_2^2. \tag{7}$$

The size of this set is a control variable $B^{(1)} = |\mathcal{B}_\ell^{(1)}[t]|$. For (ii), then, we include neighbouring UEs in cell $i \in \mathcal{B}_\ell^{(1)}[t]$.

Second, we define the set of indices of *dominantly interfered* neighboring cells, $\mathcal{B}_\ell^{(2)}[t]$. The data links of UEs in cell $i \in \mathcal{B}_\ell^{(2)}[t]$ are dominantly interfered by local UEs (in cell ℓ). Formally, $\forall i \in \mathcal{B}_\ell^{(2)}[t],\ \forall i' \in \mathcal{L}\backslash\mathcal{B}_\ell^{(2)}[t]\backslash\{\ell\}$, we have

$$\sum_{k\in\mathcal{K}_\ell}\|\hat{\mathbf{h}}_{(\ell,k),i}[t]\|_2^2 \geq \sum_{k\in\mathcal{K}_\ell}\|\hat{\mathbf{h}}_{(\ell,k),i'}[t]\|_2^2. \tag{8}$$

The size of this set is a control variable $B^{(2)} = |\mathcal{B}_\ell^{(2)}[t]|$. For (iii), then, we include neighbouring BSs of cell $i \in \mathcal{B}_\ell^{(2)}[t]$.

The effective channel gain, used in defining $\mathcal{B}_\ell^{(1)}[t]$ and $\mathcal{B}_\ell^{(2)}[t]$, can be acquired by the antenna circuit before digital processing (e.g., from the automatic gain control (AGC) circuit [14]), without the explicit effective channel vector or combiner implementation. BS ℓ also measures $\mathrm{SINR}_{\ell,k}[t]$ of all local UEs, by measuring the received signal strength indicator (RSSI) and the reference signal received power (RSRP), which are the conventional measures to evaluate the signal quality in LTE standards [13]. Using the SINRs, BS ℓ then calculates the achievable data rate of UE (ℓ, k) as $R_{\ell,k}[t] = \log_2(1 + \mathrm{SINR}_{\ell,k}[t])$. Here, we omit the bandwidth parameter, assuming the same bandwidth for all the data links.

B. Markov Decision Process Model

We formulate the decision making process of each BS as an MDP with states, actions, and rewards:

1) State: We define the state space of BS ℓ as

$$\mathcal{S}_\ell[t] = \mathcal{S}_{\ell,1}[t] \bigcup \mathcal{S}_{\ell,2}[t] \bigcup \mathcal{S}_{\ell,3}[t] \bigcup \mathcal{S}_{\ell,4}[t], \tag{9}$$

where each constituent set is described below.

(i) Local channel information. $\mathcal{S}_{\ell,1}[t]$ consists of the scalar effective channel powers from local UEs observed at two consecutive times $t - T$ and t, given by

$$\mathcal{S}_{\ell,1}[t] = \{|\hat{h}_{(\ell,j),\ell,k}[t-T]|^2,\ |\tilde{h}_{(\ell,j),\ell,k}[t]|^2\}_{j\in\mathcal{K}_\ell, k\in\mathcal{K}_\ell}.$$

Here, $|\hat{h}_{(\ell,j),\ell,k}[t-T]|^2$ can be obtained from (5) at time $t-T$, and $|\tilde{h}_{(\ell,j),\ell,k}[t]|^2$ is a version of (5) obtained at time t using previous-time variables $p_{\ell,k}[t-T]$, $\mathbf{z}_{\ell,k}[t-T]$, and $\{\boldsymbol{\phi}_r[t-T]\}_{r\in\mathcal{R}}$. Having them enables us to capture the effect of channel variation over time.

(ii) From-neighbor channel information. $\mathcal{S}_{\ell,2}[t]$ contains the scalar effective channel powers from UE (i, j) in neighboring cell i, and the index i, for $i \in \mathcal{B}_\ell^{(1)}[t]$, $j \in \mathcal{K}_i$. Formally,

$$\mathcal{S}_{\ell,2}[t] = \{|\hat{h}_{(i,j),\ell,k}[t-T]|^2\}_{j\in\mathcal{K}_i, k\in\mathcal{K}_\ell, i\in\mathcal{B}_\ell^{(1)}[t]} \bigcup \{i\}_{i\in\mathcal{B}_\ell^{(1)}[t]}.$$

This set captures the interference from neighbor UEs to cell ℓ.

(iii) To-neighbor channel information. $\mathcal{S}_{\ell,3}[t]$ contains the scalar effective channel powers from local UE (ℓ, k) to BS i, and the index i, for $i \in \mathcal{B}_\ell^{(2)}[t]$, $k \in \mathcal{K}_\ell$. Formally,

$$\mathcal{S}_{\ell,3}[t] = \{|\hat{h}_{(\ell,k),i,j}[t-T]|^2\}_{j\in\mathcal{K}_i, k\in\mathcal{K}_\ell, i\in\mathcal{B}_\ell^{(2)}[t]} \bigcup \{i\}_{i\in\mathcal{B}_\ell^{(2)}[t]}.$$

This set captures the amount of interference that local UEs in cell ℓ inflict on neighboring cells. This information enables BS ℓ to adjust the transmit powers of local UEs to reduce interference to the neighboring cells.

(iv) Previous local variables and local sum-rate. $\mathcal{S}_{\ell,4}[t]$ consists of previous local variables, i.e., $\{p_{\ell,k}[t-T]\}_{k\in\mathcal{K}_\ell}$, $\{\mathbf{z}_{\ell,k}[t-T]\}_{k\in\mathcal{K}_\ell}$, and $\boldsymbol{\phi}_\ell[t-T]$, and the local sum-rate $R_\ell[t-T] = \sum_{k\in\mathcal{K}_\ell} R_{\ell,k}[t-T]$. Formally,

$$\mathcal{S}_{\ell,4}[t] = \{p_{\ell,k}[t-T], \mathbf{z}_{\ell,k}[t-T]\}_{k\in\mathcal{K}_\ell} \bigcup \{\boldsymbol{\phi}_\ell[t-T], R_\ell[t-T]\}.$$

2) Action: The action space is defined as

$$\mathcal{A}_\ell[t] = \{b^p_{\ell,1}[t], ..., b^p_{\ell,K_\ell}[t], b^z_{\ell,1}[t], ..., b^z_{\ell,K_\ell}[t], b^\phi_\ell[t]\}, \tag{10}$$

where $b^p_{\ell,k}[t]$, $b^z_{\ell,k}[t]$, $b^\phi_\ell[t]$ are the *index gradient variables* used for updating the local UE (ℓ, k) transmit power, combiner k of BS ℓ, and local IRS ℓ reflect beamformer. These index gradient variables are defined over a binary $\{-1, 1\}$, or ternary $\{-1, 0, 1\}$ alphabet as we will describe in Sec. IV.

Once BS ℓ determines the action in (10), the BS feeds forward $b^p_{\ell,k}[t]$ to UE (ℓ, k), which then updates its power index as

$$i^p_{\ell,k}[t] = i^p_{\ell,k}[t-T] + b^p_{\ell,k}[t]. \tag{11}$$

The power of UE (ℓ, k) is set to $p_{\ell,k}[t] = \mathcal{P}(i^p_{\ell,k}[t])$, $k \in \mathcal{K}_\ell$, where $\mathcal{P}(i)$ denotes i-th element of the power set $\mathcal{P}$ in (3). The BS also feeds forward $b^\phi_\ell[t]$ to IRS ℓ, which then updates its beamformer index as

$$i^\phi_\ell[t] = i^\phi_\ell[t-T] + b^\phi_\ell[t]. \tag{12}$$

The beamformer of IRS ℓ is set to $\boldsymbol{\phi}_\ell[t] = \mathcal{Q}(i^\phi_\ell[t])$ where $\mathcal{Q}(i)$ is the i-th vector in the codebook $\mathcal{Q}$ in (3). Finally, the combiner index is updated as

$$i^z_{\ell,k}[t] = i^z_{\ell,k}[t-T] + b^z_{\ell,k}[t]. \tag{13}$$

The combiner k of BS ℓ is set to $\mathbf{z}_{\ell,k}[t] = \mathcal{Z}(i^z_{\ell,k}[t])$.

3) Reward: Aiming to only maximize the local sum-rate at each BS could increase the interference to the neighboring cells. To incorporate the entire system performance, we design the reward $r_\ell[t]$ including penalty terms as

$$r_\ell[t] = \sum_{k\in\mathcal{K}_\ell} R_{\ell,k}[t] - \sum_{i\in\mathcal{B}_\ell^{(2)}[t]} P_{\ell,i}[t], \tag{14}$$

where the first term is the sum-rate of cell ℓ and the second term is the sum of penalties. The penalty $P_{\ell,i}[t]$ is the rate loss of the dominantly interfered cell i caused by the interference of local UEs (in cell ℓ), which is calculated at BS i as

$$P_{\ell,i}[t] = \sum_{j\in\mathcal{K}_i} P_{\ell,(i,j)}[t] = \sum_{j\in\mathcal{K}_i} \Bigg[-R_{i,j}[t] + \log_2\left(1 + \frac{|\hat{h}_{(i,j),i,j}[t]|^2}{\sum_{(i',j')\neq(i,j), i'\neq\ell} |\hat{h}_{(i',j'),i,j}[t]|^2 + \sigma^2}\right)\Bigg], \tag{15}$$

where $P_{\ell,(i,j)}[t]$ is the rate loss of UE (i, j) caused by the interference of local UEs in cell ℓ. A similar reward function was found to be effective for multi-agent DRL-based beamforming [15]. The $\log(\cdot)$ term in (15) denotes the data rate of UE (i, j) without the interference of the local UEs in cell ℓ, while $R_{i,j}[t]$ is the data rate including the interference. If there is no interference, the two terms cancel with each other, leading to zero penalty. Otherwise, the penalty is positive.

Algorithm 1 Dynamic control based on multi-agent DRL.

1: Establish a train DQN with random weights $\mathbf{w}_\ell$, a target DQN with random weights $\mathbf{w}_\ell^-$, an empty experience pool $\mathcal{Y}_\ell$ with $|\mathcal{Y}_\ell| = 0$, and a pool size M_ℓ^{pool}. Set the discount factor γ_ℓ, initial ϵ-greedy value $\epsilon_\ell(0)$, mini-batch size M_ℓ^{batch}, and DQN-aligning period T_{align}, $\forall \ell \in \mathcal{L}$.
2: Agent ℓ (BS ℓ) randomly initializes the design variables $\{p_{\ell,k}[0]\}_{k\in\mathcal{K}_\ell}$, $\{\mathbf{z}_{\ell,k}[0]\}_{k\in\mathcal{K}_\ell}$, and $\boldsymbol{\phi}_\ell[0]$, and informs local UEs and local IRS of the initial variables, $\forall \ell \in \mathcal{L}$.
3: Agent ℓ selects its action $a_\ell \in \mathcal{A}$ randomly and executes it, $\forall \ell \in \mathcal{L}$.
4: $t \leftarrow T$. Agent ℓ observes the next state s'_ℓ, $\forall \ell \in \mathcal{L}$.
5: **repeat**
6: $s_\ell \leftarrow s'_\ell$
7: Agent ℓ selects its action a_ℓ at time t based on ϵ-greedy policy, $\forall \ell \in \mathcal{L}$: With probability $\epsilon_\ell(t)$, agent ℓ selects random action a_ℓ, and with probability $1-\epsilon_\ell(t)$, agent ℓ selects $a_\ell = \arg\max_{a\in\mathcal{A}} Q(s_\ell, a, \mathbf{w}_\ell)$.
8: Agent ℓ executes its action, $\forall \ell \in \mathcal{L}$.
9: $t \leftarrow t+T$. Agent ℓ observes the next state s'_ℓ and gets the reward r_ℓ, $\forall \ell \in \mathcal{L}$.
10: Agent ℓ stores the new experience $< s_\ell, a_\ell, r_\ell, s'_\ell >$ in its own experience pool $\mathcal{Y}_\ell$, $\forall \ell \in \mathcal{L}$.
11: **if** $|\mathcal{Y}_\ell| \geq M_\ell^{\text{batch}}$ **then**
12: Agent ℓ samples a mini-batch consisting of M_ℓ^{batch} experiences from its experience pool $\mathcal{Y}_\ell$, $\forall \ell \in \mathcal{L}$.
13: Agent ℓ updates the weights $\mathbf{w}_\ell$ of its train DQN using back propagation, $\forall \ell \in \mathcal{L}$.
14: Agent ℓ updates the weights of its target DQN $\mathbf{w}_\ell^- \leftarrow \mathbf{w}_\ell$ every T_{align}, $\forall \ell \in \mathcal{L}$.
15: **end if**
16: **until** Process terminates

C. Dynamic Control Scheme based on Multi-agent DRL

In the proposed MDP, the channel values used as states are continuous variables, which makes conventional RL, i.e., Q-learning based on Q-table, not applicable. We thus adopt deep Q-networks (DQN) [16]. BS ℓ possesses its own train DQN, $Q(s, a, \mathbf{w}_\ell)$, with weights $\mathbf{w}_\ell$, and target DQN, $Q(s, a, \mathbf{w}_\ell^-)$, with weights $\mathbf{w}_\ell^-$, where the state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ are defined in Sec. III-B. The pseudocode of the proposed dynamic control scheme based on multi-agent DRL is provided in Algorithm 1. Our algorithm follows a decentralized training with decentralized execution (DTDE) framework, where both training and execution are independently carried out at each agent. Therefore, our algorithm is independent of the UEs in other surrounding agents (BSs). Further, our algorithm incorporates the *index gradient* approach for codebook-based BS combining and IRS beamforming, which is independent of the number of antennas/elements and the size of the codebook.

IV. Numerical Evaluation and Discussion

In this section, we first describe the simulation setup (Sec. IV-A) and evaluation scenarios (Sec. IV-B). Then, we present and discuss the results (Sec. IV-C).

A. Simulation Setup

1) Parameter settings: We consider a cellular network with $L = 7$ hexagonal cells, as shown in Fig. 2. We assume $K_\ell = 3$, $M_\ell = 5$, and $N_r = 5$, $\forall \ell, r$, similar to [5]. The BSs are located at the center of each cell with 10 m height, and the distance between adjacent BSs is 100 m. Each IRS is deployed nearby the BS, and UEs are randomly placed in the cells. The set $\mathcal{P}$ for UE power control is given by $\mathcal{P} = \{p_{\min}, p_{\min}e^{\Delta_{\text{p}}}, p_{\min}e^{2\Delta_{\text{p}}}, ..., p_{\max}\}$, where $p_{\min} = 10$ dBm and $p_{\max} = 30$ dBm are the minimum and maximum transmit powers, and $\Delta_{\text{p}} = (\log p_{\max} - \log p_{\min})/(|\mathcal{P}| - 1)$. For BS combiner and IRS beamformer codebooks, we use a random vector quantization (RVQ) [17] codebook with size $|\mathcal{Z}| = |\mathcal{Q}| = 30$. We set $\sigma^2 = -114$ dBm, $B^{(1)} = B^{(2)} = 2$.

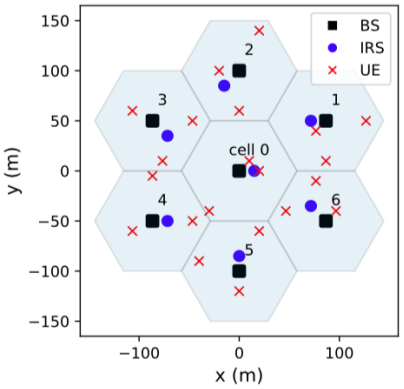


Fig. 2: The cellular network with $L = 7$ hexagonal cells and 100 m distance between adjacent BSs used in our simulations.

2) Channel modeling: We consider a single frequency band with flat fading and adopt a temporally correlated block fading channel model. Following a common cellular standard [18], we assume coherence time $T = 5$ ms and center frequency $f_c = 2.5$ GHz. The channel vector $\mathbf{h}^{\text{BS}}_{(i,j),\ell}[t]$ is modeled as

$$\mathbf{h}^{\text{UB}}_{(i,j),\ell}[t] = \sqrt{\beta^{\text{UB}}_{(i,j),\ell}}\mathbf{u}^{\text{UB}}_{(i,j),\ell}[t], \tag{16}$$

where $\beta^{\text{UB}}_{(i,j),\ell}$ denotes the large-scale fading coefficient from UE (i,j) to BS ℓ, modeled as

$$\beta^{\text{UB}}_{(i,j),\ell} = \beta_0 - 10\alpha^{\text{UB}}_{(i,j),\ell}\log_{10}(d^{\text{UB}}_{(i,j),\ell}/d_0). \tag{17}$$

Here, β_0 is the path-loss at the reference distance d_0, $d^{\text{UB}}_{(i,j),\ell}$ is the distance between UE (i,j) and BS ℓ, and $\alpha^{\text{UB}}_{(i,j),\ell}$ is the path-loss exponent between them. We set $\beta_0 = -30$ dB and $d_0 = 1$ m. $\mathbf{u}^{\text{UB}}_{(i,j),\ell}[t]$ denotes the Rayleigh fading vector, modeled by a first-order Gauss-Markov process [19]:

$$\mathbf{u}^{\text{UB}}_{(i,j),\ell}[t] = \rho^{\text{UB}}_{(i,j),\ell}\mathbf{u}^{\text{UB}}_{(i,j),\ell}[t-T] + \sqrt{1-(\rho^{\text{UB}}_{(i,j),\ell})^2}\mathbf{n}^{\text{UB}}_{(i,j),\ell}[t], \tag{18}$$

where $\mathbf{n}^{\text{UB}}_{(i,j),\ell}[t] \in \mathbb{C}^{M_\ell \times 1}$, $\mathbf{n}^{\text{UB}}_{(i,j),\ell}[t] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$, and $\mathbf{u}^{\text{UB}}_{(i,j),\ell}[0] \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$. The time correlation coefficient obeys the Jakes model [19], i.e., $\rho^{\text{UB}}_{(i,j),\ell} = J_0(2\pi\tilde{f}^{\text{UB}}_{(i,j)}T)$, where $J_0(\cdot)$ is the zeroth order Bessel function of the first kind, and $\tilde{f}^{\text{UB}}_{(i,j),\ell} = v^{\text{UB}}_{(i,j),\ell}f_c/c$ is the maximum Doppler frequency, with velocity $v^{\text{UB}}_{(i,j),\ell}$ of UE (i,j) and $c = 3\times10^8$ m/s. The same modeling for $\mathbf{h}^{\text{UB}}_{(i,j),\ell}[t]$ is applied for the channels between the UEs and the IRSs, i.e., $\mathbf{h}^{\text{UI}}_{(i,j),r}[t]$, $\forall i,j,r$, with path-loss exponent $\alpha^{\text{UI}}_{(i,j),r}$. Since IRSs are placed at

the desired locations to have less variations of IRS-BS/IRS-IRS channels as compared to UE-BS/UE-IRS channels [5], $\mathbf{G}^{\mathrm{IB}}_{r,\ell}[t]$ and $\mathbf{G}^{\mathrm{II}}_{r_1,r_2}[t]$ are assumed to be stationary. Each entry for the channels is distributed according to $\mathcal{CN}(0,\beta^{\mathrm{IB}}_{r,\ell})$ and $\mathcal{CN}(0,\beta^{\mathrm{II}}_{r_1,r_2})$, respectively. $\beta^{\mathrm{IB}}_{r,\ell}$ and $\beta^{\mathrm{II}}_{r_1,r_2}$ denote the large-scale fading coefficients with path loss exponents $\alpha^{\mathrm{IB}}_{r,\ell}$ and $\alpha^{\mathrm{II}}_{r_1,r_2}$, respectively.

We assume $\alpha^{\mathrm{UB}}_{(i,j),\ell}=\alpha^{\mathrm{UB}}$, $\forall i,j,\ell$, $\alpha^{\mathrm{UI}}_{(i,j),r}=\alpha^{\mathrm{UI}}$, $\forall i,j,r$, $\alpha^{\mathrm{IB}}_{r,\ell}=\alpha^{\mathrm{IB}}$, $\forall r,\ell$, and $\alpha^{\mathrm{II}}_{r_1,r_2}=\alpha^{\mathrm{II}}$, $\forall r_1,r_2$. To model the presence of extensive obstacles and scatterers, the path-loss exponent between the UEs and BS is taken to be $\alpha^{\mathrm{UB}}=3.75$. Because the IRS-aided link can have less path loss than that of direct UE-BS channel by properly choosing the location of the IRS, we set the path-loss exponents of the UE-IRS link, of the IRS-BS link, and of the IRS-IRS link to $\alpha^{\mathrm{UI}}=2.2$, $\alpha^{\mathrm{IB}}=1$, and $\alpha^{\mathrm{II}}=2$, respectively [5]. We assume $\rho^{\mathrm{UB}}_{(i,j),\ell}=\rho^{\mathrm{UI}}_{(i,j),r}=\rho$, $\forall i,j,\ell,r$ and adopt $\rho=0.999$ ($v\approx 1$ km/h), 0.99 ($v\approx 3$ km/h), and 0.9 ($v\approx 9$ km/h), where v is the UE speed.

B. Evaluation Scenarios

1) Scenario 1. The effective channels from local UEs are not known: In this scenario, each BS measures the scalar effective channel powers directly from received signals without explicitly obtaining the effective channels as a vector form in (6). We introduce two baselines in this scenario: RRR=(random, random, random) and MRR=(maximum, random, random). The name of each baseline is indicating how it selects its (UE power, IRS beamformer, BS combiner) variables as a tuple. We propose DQN1, where the action space consists of $2K+1$ elements for K UE powers, the IRS beamformer, and K BS combiners. The index gradient variables are binary, i.e., $\{-1,1\}$.

2) Scenario 2. The effective channels from local UEs are known: In this scenario, each BS measures the effective channels from local UEs as the vector form in (6). Each BS is assumed to adopt a maximum ratio combiner (MRC) by finding the index as

$$i^{\star}=\arg\max_i |\mathcal{Z}(i)^H\mathbf{h}[t]|^2, \tag{19}$$

where $\mathbf{h}[t]$ is the effective channel from local UE. We introduce several baselines: MRM=(maximum, random, MRC), FRM=(25% of maximum, random, MRC), RRM=(random, random, MRC), and MM with no IRS=(maximum, N/A, MRC). MM with no IRS assumes the IRSs to be turned off. In this scenario, we propose DQN2 and DQN3. In DQN2, the action space consists of $K+1$ elements for K UE powers and the IRS beamformer (the action space does not have the elements $b^z_{\ell,k}[t]$, $\forall k$ in (10)). The BS combiner is designed as MRC and the index gradient variable is binary, i.e., $\{-1,1\}$. The action space is DQN3 is the same as DQN2, except it uses a tenary index gradient variable, i.e., $\{-1,0,1\}$.

In both scenarios, the DQNs[2] are composed of an input layer, an output layer, and two fully-connected hidden layers. The input size is $6K^2+2K+6=66$. The output size is $2^{2K+1}=128$, $2^{K+1}=16$, and $3^{K+1}=81$ for DQN1, DQN2, and DQN3, respectively. For DQN1, the number of neurons in the two hidden layers is 70 and 100; for DQN2, 40 and 30; and for DQN3, 70 and 70. The rectified linear unit (ReLU) activation function is employed. In Algorithm 1, we adopt the ϵ-greedy method with $\epsilon_\ell(t)=\max\{\epsilon_{\min},(1-10^{-3.5})\epsilon_\ell(t-T)\}$, where $\epsilon_\ell(0)=0.6$ and $\epsilon_{\min}=0.005$, $\forall\ell$. We consider $M^{\mathrm{batch}}_\ell=10$, $M^{\mathrm{pool}}_\ell=300$, and $\gamma_\ell=0.7$, $\forall\ell$. We set $T_{\mathrm{align}}=50T$, i.e., the target DQN is updated with the weights of train DQN after a time of $50T$. We employ the RMSProp optimizer for training.

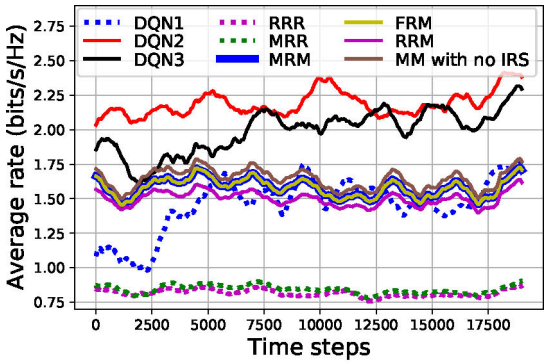
C. Simulation Results and Discussion

Fig. 3 depicts the average achievable data rate over all 21 UEs with different values of time correlation coefficient ρ: $\rho=0.999$ in (a), $\rho=0.99$ in (b), and $\rho=0.9$ in (c). The dotted lines show the performance of the schemes in Scenario 1. With varying channels, RRR and MRR select random or fixed indices for variables, and therefore have low average data rates over time. On the other hand, DQN1 learns and adapts to the varying channels over time by exploiting the local observations and information-sharing in our sequential decision making.

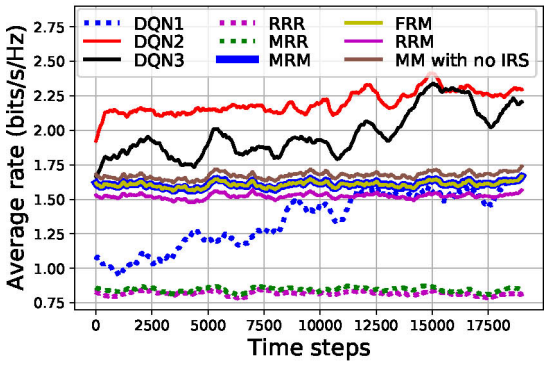
The solid lines represent the performances of schemes in Scenario 2. The MM with no IRS gives better performance than the baselines using IRS, implying that random IRS beamforming is worse than not deploying it at all. This also reveals the vulnerability of IRS-assisted systems to adversarial IRS utilization. Our DQN2 and DQN3 methods outperform the baselines, which emphasizes the benefit of carefully optimizing the IRS configuration with the rest of the cellular network. DQN2 yields slightly better performance and converges faster than DQN3: the faster convergence is due to neural networks training faster with a smaller number of outputs, and the better overall performance is consistent with the observation [16] that DQNs are more successful with smaller action spaces.

Comparing Scenario 1 with 2, i.e., the dotted lines with the solid lines in Fig. 3, we note that the performance of DQN1, which only uses scalar effective channel powers, is comparable with the baselines in Scenario 2, which use vectorized local effective channels for MRC. Also, with higher ρ values, the DQNs experience faster convergence, which is particularly noticeable in DQN1. The fluctuation of the DQN plots occurs due to the ϵ-greedy policy, which explores random action selection occasionally to avoid getting trapped in local optima. Overall, in each case, we see that our MDP-based algorithms obtain significant performance improvements, emphasizing the benefit of our multi-agent DRL method.

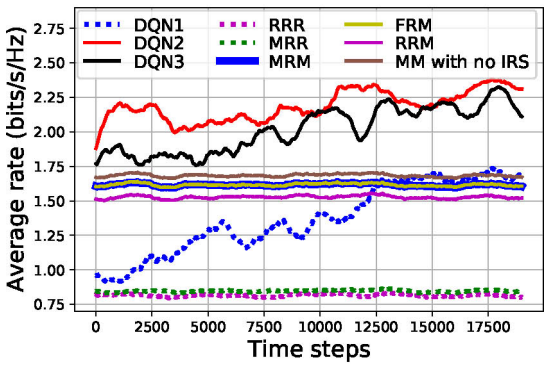
[2]All DQNs establish the same state space and reward function given in Sec. III-B. For the state information group (iv) in Sec. III-B1, the indices of previous local variables are stored in the state.



(a) $\rho = 0.999$



(b) $\rho = 0.99$



(c) $\rho = 0.9$

Fig. 3: Average achievable data rates over all 21 UEs obtained by each method with different values of ρ: $\rho = 0.999$ in (a), $\rho = 0.99$ in (b), and $\rho = 0.9$ in (c). The dotted-lines and solid lines show the performance of schemes in Scenario 1 and Scenario 2, respectively. Each data point in the plots is a moving average over the previous 1000 time slots.

V. Conclusion

We developed a novel methodology for uplink multi-IRS-assisted multi-cell systems. Due to temporal channel variations and difficulties of channel acquisition, we considered that BSs only acquire scalar effective channel powers from a subset of UEs. We developed an information-sharing scheme among neighboring BSs and proposed a dynamic control scheme based on multi-agent DRL, in which each BS acts as an agent and adaptively designs its local UE powers, local IRS beamformer, and its combiners. Through numerical simulations, we verified that our algorithm outperforms conventional baselines.

Acknowledgment

D.J. Love was supported in part by the National Science Foundation (NSF) under grants CNS1642982 and CCF1816013. C. G. Brinton was supported in part by the NSF under grants AST2037864. T. Kim was supported by the NSF under grants CNS1955561.

References

[1] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, 2020.

[2] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated to fog learning: Distributed machine learning over heterogeneous wireless networks," *IEEE Commun. Mag.*, 2020.

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

[4] H. Guo, Y. Liang, J. Chen, and E. G. Larsson, "Weighted sum-rate maximization for intelligent reflecting surface enhanced wireless networks," in *Proc. IEEE Glob. Commun. Conf.*, 2019, pp. 1–6.

[5] C. Pan, H. Ren, K. Wang, W. Xu, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Multicell MIMO communications relying on intelligent reflecting surfaces," *IEEE Trans. Wireless Commun.*, 2020.

[6] H. Zhang, B. Di, L. Song, and Z. Han, "Reconfigurable intelligent surfaces assisted communications with limited phase shifts: How many phase shifts are enough?" *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4498–4502.

[7] K. Zhi, C. Pan, H. Ren, and K. Wang, "Uplink achievable rate of intelligent reflecting surface-aided millimeter-wave communications with low-resolution ADC and phase noise," *arXiv:2008.00437*, 2020.

[8] L. Feng, X. Que, P. Yu, W. Li, and X. Qiu, "IRS assisted multiple user detection for uplink URLLC non-orthogonal multiple access," in *IEEE Conf. Comput. Commun. Workshop*, 2020, pp. 1314–1315.

[9] S. Zhang and R. Zhang, "Intelligent reflecting surface aided multi-user communication: Capacity region and deployment strategy," *arXiv:2009.02324*, 2020.

[10] T. J. Cui, M. Q. Qi, X. Wan, J. Zhao, and Q. Cheng, "Coding metamaterials, digital metamaterials and programmable metamaterials," *Light: Science & Applications*, vol. 3, no. 10, p. e218, 2014.

[11] Z. Wang, L. Liu, and S. Cui, "Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis," *IEEE Trans. Wireless Commun.*, 2020.

[12] A. Marinescu, I. Dusparic, and S. Clarke, "Prediction-based multi-agent reinforcement learning in inherently non-stationary environments," *ACM Trans. Auton. Adapt. Syst.*, vol. 12, no. 2, pp. 1–23, 2017.

[13] 3GPP TS 36.211, "LTE: Evolved universal terrestrial radio access (e-utra): Physical channels and modulation," vol. V14.2.0 Release 14, 2017.

[14] J. Mo, P. Schniter, and R. W. Heath, "Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1141–1154, 2017.

[15] J. Ge, Y.-C. Liang, J. Joung, and S. Sun, "Deep reinforcement learning for distributed dynamic MISO downlink-beamforming coordination," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6070–6085, 2020.

[16] V. Mnih, K. Kavukcuoglu *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[17] C. K. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 458–462, 2007.

[18] "IEEE P802.16m-2008 draft standard for local and metropolitan area network," *IEEE Standard 802.16m*, 2008.

[19] B. Sklar *et al.*, *Digital communications: fundamentals and applications*, 2001.