

Collaborative Intelligent Reflecting Surface Networks with Multi-Agent Reinforcement Learning

Jie Zhang, Jun Li, Yijin Zhang, Qingqing Wu, Xiongwei Wu, Feng Shu, Shi Jin, Wen Chen

Abstract—Intelligent reflecting surface (IRS) is envisioned to be widely applied in future wireless networks. In this paper, we investigate a multi-user communication system assisted by cooperative IRS devices with the capability of energy harvesting. Aiming to maximize the long-term average achievable system rate, an optimization problem is formulated by jointly designing the transmit beamforming at the base station (BS) and discrete phase shift beamforming at the IRSs, with the constraints on transmit power, user data rate requirement and IRS energy buffer size. Considering time-varying channels and stochastic arrivals of energy harvested by the IRSs, we first formulate the problem as a Markov decision process (MDP) and then develop a novel multi-agent Q-mix (MAQ) framework with two layers to decouple the optimization parameters. The higher layer is for optimizing phase shift resolutions, and the lower one is for phase shift beamforming and power allocation. Since the phase shift optimization is an integer programming problem with a large-scale action space, we improve MAQ by incorporating the Wolpertinger method, namely, MAQ-WP algorithm to achieve a sub-optimality with reduced dimensions of action space. In addition, as MAQ-WP is still of high complexity to achieve good performance, we propose a policy gradient-based MAQ algorithm, namely, MAQ-PG, by mapping the discrete phase shift actions into a continuous space at the cost of a slight performance loss. Simulation results demonstrate that the proposed MAQ-WP and MAQ-PG algorithms can converge faster and achieve data rate improvements of 10.7% and 8.8% over the conventional multi-agent DDPG, respectively.

Index Terms—Intelligent reflecting surface, beamforming, energy harvesting, multi-agent reinforcement learning

I. INTRODUCTION

RECENTLY, advanced technologies have been developed in the fifth generation (5G), such as massive

multiple input multiple output (MIMO) and network densification [1, 2], to achieve high throughput, ultra low latency and high reliability. However, progressively dense deployment of MIMO base stations (BSs) with large-scale multi-antenna arrays suffers high cost and substantial power consumption.

To tackle this challenge, a promising paradigm called intelligent reflecting surface (IRS) has been proposed and has aroused considerable research enthusiasm due to its superiority of energy efficiency and low cost [3]. An IRS can be seen as a flat composed of many passive reflecting units, which is able to tune the phases and amplitudes of incident signals and then reflect them into the desired directions. By this way, it can significantly enhance the signal-to-interference-plus-noise ratio (SINR) [4]. Notably, an IRS can reflect the input signals passively by controlling the electronic devices without the need of utilizing radio-frequency chains. In addition, an IRS is generally portable and scalable, and thus can be easily deployed on the indoor furnitures and outdoor walls [5]. Owing to these nice features, the IRS technology is expected to be widely applied in various wireless communication scenarios to improve the network performance.

A. Related Work

Many recent studies have been devoted to investigating properties and challenges of the IRS-assisted conventional systems. Such a system with a single-antenna BS and multiple single-antenna users was studied in [6]. For IRS-assisted multi-antenna systems, [7–9] employed an alternating optimization (AO) approach and a semi-definite relaxation (SDR) algorithm to maximize the signal-to-noise ratio (SNR) with a guarantee of the user secrecy rate. Zhou *et al.* [10] investigated the robust beamforming at the BS for IRS-assisted multiple-input single-output (MISO) systems. For an IRS-assisted multi-user MIMO (MU-MIMO) system, the study in [11] used a two-step stochastic program to formulate the average received SNR maximization issue and utilized a minorization-maximization (MM) based algorithm to solve the passive beamforming and information transfer problem.

In addition, the IRS technology has been applied to some novel communication systems. For an IRS-assisted millimeter-wave system, Xiu *et al.* [12] designed the IRS beamforming to offer more feasible propagation paths, and proposed an alternating manifold optimization method to maximize the weighted sum rate. For an IRS-assisted unmanned aerial vehicle (UAV) system, the studies in [13–15] investigated how to increase the received signal strength by passive beamforming at each IRS. For a simultaneous wireless information

This work was supported in part by National Natural Science Foundation of China under Grants 61872184, 62071234, 62071236 and 62071296, in part by National Key Project under Grants 2018YFB1801102 and 2020YFB1807700, in part by Shanghai Fundamental Project under Grant 20JC1416502, in part by the Hainan Major Projects ZDKJ2021022, in part by the Scientific Research Fund Project of Hainan University under Grant KYQD(ZR)-21008, and in part by the Fundamental Research Funds for the Central Universities of China No. 30921013104 and No. 30920021127. (Corresponding author: Jun Li).

J. Zhang, J. Li and Y. Zhang are with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, 210094, China (e-mail: {zhangjie666; jun.li}@njjust.edu.cn; yijin.zhang@gmail.com).

Q. Wu is with the State Key Laboratory of IoT for Smart City, University of Macau, Macao 999078, China (email: qingqingwu@um.edu.mo).

X. Wu is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR of China (e-mail: xwwu@ee.cuhk.edu.hk).

F. Shu is with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China (email: shufeng0101@163.com).

S. Jin is with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China (e-mail: jinshi@seu.edu.cn).

W. Chen is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wenchen@sjtu.edu.cn).

and power transfer (SWIPT) system, the studies in [16–18] employed IRSs to serve energy harvesting receivers and information decoding receivers, and optimized the IRSs phase shifts to maximize the weighted sum-power. For an IRS-enhanced orthogonal frequency division multiplexing (OFDM) system, Yang *et al.* [19] maximized the achievable rate through reasonably allocating the transmit power and designing the passive beamforming.

Regarding the design of IRS passive beamforming, most previous studies concentrated on the continuous phase shift optimization at each IRS, which leads to excessively high resolution, and thus its computational burden is unacceptable in practice. When each IRS has a finite number of phase shifts, Wu *et al.* [20] demonstrated that the discrete phase shifts and continuous phase shifts can achieve the same power gain. Further, the study in [21] designed the IRS passive beamforming under discrete phase shifts and imperfect channel state information. However, the above mentioned works are not applicable to dynamic environments (e.g., varying channels, stochastic arrivals of mobile users) by utilizing traditional optimization approaches, e.g., AO, SDR, MM algorithms, to address the beamforming optimization problems.

Artificial intelligence (AI) has recently developed as a remarkably impressive technology to tackle dynamic optimization problems in large-scale systems [22–26]. Considering the superiority of AI, deep learning (DL) has been used to maximize the user received signal strength by formulating the IRSs online wireless configuration [27]. The studies in [28, 29] applied reinforcement learning (RL) to achieve the maximum SNR by optimizing the IRS passive phase shift. The work in [30, 31] developed deep RL (DRL) methods to improve the system secrecy rate and the energy efficiency by jointly optimizing the BS beamforming and the IRSs' reflecting beamforming. In [32, 33], the joint design of the BS digital beamforming and the IRSs' analog beamforming was formulated as an NP-hard optimization problem to improve the coverage range by leveraging DRL.

To the best of our knowledge, multi-agent RL (MARL) algorithm has not been developed in the existing works to cope with the joint beamforming optimization problems in multiple IRSs-assisted multi-user systems, under the condition of time-varying discrete phase shifts design and energy harvesting mechanism.

B. Contributions

In this paper, we aim to maximize the long-term average achievable system data rate by optimizing BS transmit beamforming and IRSs' discrete phase shift beamforming with transmit power limits, user data rate requirements and IRS energy storage buffer constraints, assuming that each IRS has adjustable phase shift resolutions and is equipped with energy harvesting devices. The main contributions of this paper are summarized as follows:

- 1) Considering a multi-user MISO system assisted by distributed IRSs and a central BS, we formulate the joint transmit beamforming and phase shift beamforming optimization with the objective of maximizing the long-term

average achievable system rate and propose a cooperative multi-agent Markov decision process (MDP) to model the distributed IRS-assisted system due to the time-varying channels and the stochastic harvested energy. The BS and all the IRSs are considered as agents that can interact with the system environment and learn through the historical interactive experience.

- 2) We develop a novel multi-agent Q-mix (MAQ) framework with two layers to decouple the optimization parameters. The high-level layer is for optimizing phase shift resolutions and the low-level layer is for phase shift beamforming and power allocation. To efficiently handle the exponentially large number of phase shift actions, we propose a MAQ with Wolpertinger method (MAQ-WP) algorithm to obtain sub-optimal beamforming policies. We further propose a MAQ with policy gradient (MAQ-PG) algorithm to overcome the weakness of high complexity in the MAQ-WP algorithm at the cost of a slight performance loss.
- 3) We generalize the MAQ-WP and MAQ-PG algorithms by segregating the high-dimensional and discrete phase shift actions into two hierarchical actions to significantly accelerate the learning process and reduce the computational complexity. We propose Wolpertinger and policy gradient methods to map proto-actions with high dimensionality into actual actions with low dimensionality, thereby improving the learning rate effectively.
- 4) Extensive simulation results demonstrate the effectiveness of the proposed algorithms in improving both the convergence value and learning speed under the constraints compared with benchmarks. In addition, the MAQ-WP and MAQ-PG algorithms can increase the long-term average achievable system rate by 10.7% and 8.8% compared with the multi-agent deep deterministic policy gradient (MADDPG) [34] based approach, respectively.

The rest of this paper is organized as follows. The system model and the problem formulation are provided in Section II. The MAQ framework based on an MDP formulation and the proposed MAQ-WP and MAQ-PG algorithms are presented in Section III. Section IV provides numerical results to evaluate the proposed algorithms. Section V concludes this paper.

Math notation: Vectors and matrices are denoted by bold-face lowercase letters \mathbf{a} and boldface capital letters \mathbf{A} , respectively. $(\cdot)^T$ and $(\cdot)^H$ are transpose and conjugate transpose operations, respectively. Let $|\cdot|$ and $\|\cdot\|$ denote the absolute value and the Euclidean norm operations, respectively. The operator $\text{diag}(\cdot)$ represents the diagonal matrix with argument of a vector. We use $x \sim \mathcal{CN}(0, 1)$ to indicate that the random variable x obeys the complex Gaussian distribution with zero-mean and unit variance.

II. SYSTEM MODEL

In this section we introduce the considered signal model and energy harvesting model. Based on these two models, under the objective of maximizing the long-term average system data, we formulate a constrained optimization problem. For ease of reference, Table I lists all the main notation of the system model.

Table I: Summary of main notation

Notation	Description
L, \mathcal{L}	Number of IRSs, set of IRSs
K, \mathcal{K}	Number of users, set of users
$\mathbf{h}_{l,k}^{\text{RU}}$	Channel state information between l -th IRS and k -th user
\mathbf{h}_k^{BU}	Channel state information between BS and the k -th user
\mathbf{H}_l^{BR}	Channel state information between BS and the l -th IRS
Φ_l	The reflection matrix of the l -th IRS
b_l	Bit resolution of the l -th IRS
$\rho_{l,n}(t)$	Working status of the l -th IRS n -th element at the t -th slot
$P^{\text{IRS}}(t)$	total power consumption of all the IRSs at the t -th slot
$E_l(t)$	Remaining energy of the l -th IRS at the t -th slot

A. Signal Model

As illustrated in Fig. 1, we consider a multi-user MISO communication system assisted by multiple IRSs, where a downlink BS equipped with M antennas communicates with K single-antenna users under L IRSs' cooperation. Let $\mathcal{K} \triangleq \{1, \dots, K\}$ and $\mathcal{L} \triangleq \{1, \dots, L\}$ denote indices of users and indices of IRSs, respectively. It is assumed that the IRS $l, \forall l \in \mathcal{L}$ is equipped with N_l reflecting elements or unit cells. We denote the l -th IRS reflecting elements set as $\mathcal{N}_l \triangleq \{1, \dots, N_l\}$.

The received signal at the k -th user can be given by

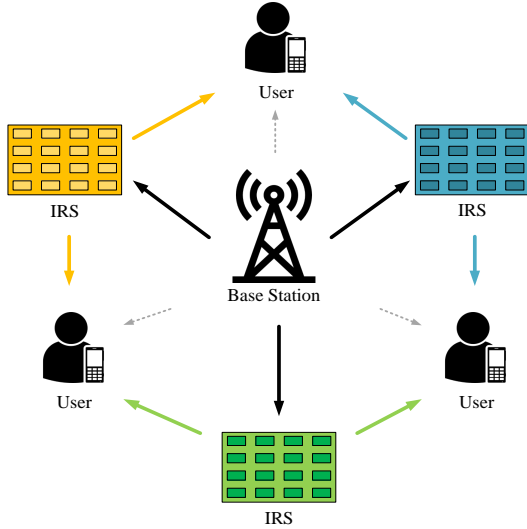


Figure 1: Downlink multi-user MISO system with multiple IRSs

$$y_k = \left(\sum_{l=1}^L \left(\mathbf{h}_{l,k}^{\text{RU}} \right)^H \Phi_l \mathbf{H}_l^{\text{BR}} + \left(\mathbf{h}_k^{\text{BU}} \right)^H \right) \sum_{j=1}^K \mathbf{v}_j s_j + n_k, \quad (1)$$

where $\mathbf{h}_{l,k}^{\text{RU}} \in \mathbb{C}^{N_l \times 1}$ denotes the channel coefficients vector between the l -th IRS and the k -th user, $\mathbf{H}_l^{\text{BR}} \in \mathbb{C}^{N_l \times M}$ denotes the channel coefficients matrix between the BS and the l -th IRS, $\mathbf{h}_k^{\text{BU}} \in \mathbb{C}^{M \times 1}$ denotes the channel coefficients vector between the BS and the k -th user. $\Phi_l \triangleq \text{diag}(\phi_l) \in \mathbb{C}^{N_l \times N_l}$ denotes reflection coefficients matrix of the l -th IRS, with the l -th IRS reflection coefficients vector ϕ_l being defined as

$$\phi_l \triangleq [\beta_{l,1} e^{j\theta_{l,1}}, \beta_{l,2} e^{j\theta_{l,2}}, \dots, \beta_{l,N_l} e^{j\theta_{l,N_l}}]^T, \quad (2)$$

where $\beta_{l,i} \in [0, 1], i \in \mathcal{N}_l$ denotes the i -th element amplitude reflection coefficient of the l -th IRS, and $\theta_{l,i} \in [0, 2\pi), i \in$

\mathcal{N}_l denotes the i -th element phase shift reflection coefficient of the l -th IRS. In this paper, we assume that the amplitude reflection coefficient of each IRS element is set to be one for maximizing the signal reflection, i.e., $\beta_{l,i} = 1, \forall i \in \mathcal{N}_l, l \in \mathcal{L}$. The transmit beamforming vector $\mathbf{v}_j \in \mathbb{C}^{M \times 1}, \forall j \in \mathcal{K}$ and information symbol $s_j \sim \mathcal{CN}(0, 1), \forall j \in \mathcal{K}$ are designed for the j -th user. Let $n_k \sim \mathcal{CN}(0, \sigma_n^2)$ denote the additive white Gaussian noise (AWGN) with zero mean and σ_n^2 variance. We separate the k -th user's received signal expressed in (1) into three parts: desired signal, inter-user interference signal and noise signal, i.e.,

$$y_k = \underbrace{\left(\sum_{l=1}^L \left(\mathbf{h}_{l,k}^{\text{RU}} \right)^H \Phi_l \mathbf{H}_l^{\text{BR}} + \left(\mathbf{h}_k^{\text{BU}} \right)^H \right) \mathbf{v}_k s_k}_{\text{desired signal}} + \underbrace{\sum_{j \neq k} \left(\sum_{l=1}^L \left(\mathbf{h}_{l,k}^{\text{RU}} \right)^H \Phi_l \mathbf{H}_l^{\text{BR}} + \left(\mathbf{h}_k^{\text{BU}} \right)^H \right) \mathbf{v}_j s_j}_{\text{inter-user interference signal}} + \underbrace{n_k}_{\text{noise signal}}. \quad (3)$$

Then the SINR at the k -th user can be obtained by

$$\gamma_k = \frac{\left| \left(\sum_{l=1}^L \left(\mathbf{h}_{l,k}^{\text{RU}} \right)^H \Phi_l \mathbf{H}_l^{\text{BR}} + \left(\mathbf{h}_k^{\text{BU}} \right)^H \right) \mathbf{v}_k \right|^2}{\sum_{j \neq k} \left| \left(\sum_{l=1}^L \left(\mathbf{h}_{l,k}^{\text{RU}} \right)^H \Phi_l \mathbf{H}_l^{\text{BR}} + \left(\mathbf{h}_k^{\text{BU}} \right)^H \right) \mathbf{v}_j \right|^2 + \sigma_n^2}, \quad (4)$$

and thus, the achievable system data rate can be obtained as

$$R = \sum_{k=1}^K \log_2 (1 + \gamma_k). \quad (5)$$

B. Energy Harvesting Model

An IRS is a passive component that reflects incident signal without amplification in theory. However, an IRS indeed consumes energy in practice due to the operations of IRS elements, IRS controller and circuit board. Hence, the IRS power consumption depends on the phase shift resolution and the operation status per IRS element.

Ideally, the phase shift of each IRS element can be tuned continuously. In practice, however, the phase shift is finite and discrete due to the complex hardware limitation. Let $\mathcal{B} \triangleq \{1, \dots, B\}$ denote the set of all possible bit resolution values. Then the set of all possible discrete phase shifts taking the b -bit resolution can be indicated as

$$\mathcal{F}_b \triangleq \{0, \Delta\theta, \dots, (2^b - 1)\Delta\theta\}, \forall b \in \mathcal{B}, \quad (6)$$

where $\Delta\theta = 2\pi/2^b$ and thus each element is within the range of $[0, 2\pi)$. We assume that each IRS can alter its phase shift bit resolution according to the actual requirement. For example, as shown in [35], power consumption per element is 1.5, 4.5 and 6mW for 3-, 4- and 5-bit resolution phase shifting, respectively. The higher the resolution of IRS phase shifts, the better its beamforming performance, but the power consumption will increase greatly. Therefore, there is a tradeoff between resolution and power consumption. Such power consumption

is much lower than that of an amplify-and-forward (AF) relay. However, in a large-scale multiple IRSs system, the IRS power consumption is still considerable.

In our work, a time-slotted system with a minimum slot length Δt is considered. Each IRS is assumed to carry an energy storage buffer and an energy harvesting device, which can gather solar energy with the usage of solar panels. It is assumed that the harvested energy obeys a certain statistical distribution, e.g., Poisson distribution. By fabricating the tunable reflecting element based on PIN diode, we can set each IRS element to an ON or OFF working status. If an IRS element works at the ON status, the incident signal will be reflected by the IRS element without amplitude attenuation and the signal phase will be changed. Otherwise, the incident signal will not be reflected and the signal power will not be absorbed into the IRS energy storage buffer. Thus, by adjusting the working status of the IRSs elements, the desired signal can be boosted and the interference signals can be weakened effectively.

The status of the l -th IRS n -th element at the t -th slot is denoted as $\rho_{l,n}(t)$. If $\rho_{l,n}(t) = 1$, the element operates at the ON status at the t -th slot. We assume that all elements in the same IRS have the same phase shift resolution. The ON-status power consumption of the l -th IRS element under $b_l \in \mathcal{B}$ bit resolution at the t -th slot is denoted as $\mu_l(t)$. As a result, the power consumption of the l -th IRS at the t -th slot can be expressed as

$$P_l^{\text{IRS}}(t) = \sum_{n=1}^{N_l} \rho_{l,n}(t) \mu_l(t), \forall l \in \mathcal{L}. \quad (7)$$

Then total power consumption of all the IRSs at the t -th slot can be obtained as

$$P^{\text{IRS}}(t) = \sum_{l=1}^L P_l^{\text{IRS}}(t). \quad (8)$$

The remaining energy stored in the l -th IRS energy storage buffer can be updated by

$$E_l(t+1) \triangleq \min \left\{ \max \left\{ E_l(t) - c_l(t), E_l^{\min} \right\} + a_l(t), E_l^{\max} \right\}, \quad (9)$$

where $E_l(t)$ is the remaining stored energy of the l -th IRS at the t -th slot, $c_l(t) = P_l^{\text{IRS}}(t) \Delta t$ is the energy consumption of the l -th IRS at the t -th slot, $a_l(t)$ is the harvested energy at the t -th slot, E_l^{\min} and E_l^{\max} are the minimum threshold and the maximum capacity of the energy storage buffer, respectively.

C. Problem Formulation

Our main objective is to jointly design the BS transmit beamforming matrix $\mathbf{V} \triangleq [\mathbf{v}_1, \dots, \mathbf{v}_K]$, the IRS element bit resolution vector $\mathbf{b} \triangleq [b_1, \dots, b_L]$, the IRS reflecting matrix $\Phi_l, \forall l \in \mathcal{L}$ and the ON/OFF status vector $\rho_l \triangleq [\rho_{l,1}, \dots, \rho_{l,N_l}], \forall l \in \mathcal{L}$ to maximize the long-term average achievable system data rate in the multiple IRSs-assisted communication system.

Let $\mathbf{X}(t) \triangleq [\mathbf{V}(t), \mathbf{b}(t), \{\Phi_l(t)\}_1^L, \{\rho_l(t)\}_1^L]$ denote the set consisting of all the variables at the t -th slot. Let $\{x_i(t)\}_1^L \triangleq \{x_1(t), \dots, x_L(t)\}$ for convenience. We aim to

maximize the long-term average achievable system data rate $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T R(t)$ subject to the data rate requirement $R_k^{\text{req}}(t)$ of the k -th user at the t -th slot, the transmit power limitation P^{\max} at the BS and the energy storage buffer constraints E_l^{\min}, E_l^{\max} at the l -th IRS. Accordingly, the optimization function is formulated as

$$\mathbf{P} : \max_{\mathbf{X}(t)} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T R(t), \quad (10)$$

$$\text{s.t. } C1: \sum_{k=1}^K P_k(t) \leq P^{\max}, \forall t, \quad (10a)$$

$$C2: R_k(t) \geq R_k^{\text{req}}(t), \forall k \in \mathcal{K}, \forall t, \quad (10b)$$

$$C3: E_l^{\min} \leq E_l(t) \leq E_l^{\max}, \forall l \in \mathcal{L}, \forall t, \quad (10c)$$

where $P_k(t) = \|\mathbf{v}_k\|^2$ is the transmit power of the k -th user signal, $R_k(t) = \log_2(1 + \gamma_k(t))$ ¹ and $R(t)$ are the achievable data rate of the k -th user and the achievable system data rate at the t -th slot, respectively. The optimization function (10) computes the long-term average achievable system data rate. The constraint (10a) indicates that the total transmit power is no more than the transmit power constraint P^{\max} . The constraint (10b) means that each user achievable data rate $R_k(t)$ should satisfy his data rate requirement $R_k^{\text{req}}(t)$. The constraint (10c) takes into account the IRS energy storage buffer size.

The formulated objective function in (10) with constraints in (10a), (10b) and (10c) is a non-convex problem because the transmit beamforming and phase shifts are coupled in (4). As conventional optimization techniques (e.g. convex optimization methods) are challenging to solve it, we aim to a DRL-based method to address this problem.

III. DRL-BASED SOLUTION

In this section, we construct an MDP formula for the proposed constrained optimization function and then propose an MAQ structure and two algorithms to solve the optimization problem.

A. MDP Formula

The optimization problem given in (10) is a complex non-convex problem with three types of constraint conditions. Besides, the channel state information, the user data rate requirements and the harvested energy are all time-varying. Thus, traditional optimization approaches that transform a dynamic system into a static system may achieve poor performance and have no guarantee of constraints. As such, we transform the dynamic system into a cooperative multi-agent MDP model by viewing the multiple IRSs-assisted multi-user MISO communication system in Section II as an interactive system and viewing the BS and all the IRSs as distributed learning agents. The main elements of the constructed multi-agent MDP are defined as follows:

¹Without further explanation, in the following sections, notation with subscript t will substitute the functional form notation at the t -th slot for convenience, e.g., $(R_k(t) \rightarrow R_{k,t})$.

1) *State*: The local state of the l -th IRS at the t -th slot should reflect the user data rate and the energy consumption, which is defined as

$$s_{l,t} \triangleq \left[\left\{ \exp\left(-\frac{d_{l,k}}{d_{l,0}}\right) \cdot f_k(R_{k,t-1}) \right\}_{k \in \mathcal{K}}, E_{l,t} \right], \quad (11)$$

where $\exp\left(-\frac{d_{l,k}}{d_{l,0}}\right)$ is a discount factor that embodies the impact of the k -th user on the l -th IRS, $d_{l,k}$ is the distance between the l -th IRS and the k -th user, $d_{l,0}$ is the reference distance of the l -th IRS. The indicator function $f_k(R_{k,t})$ is defined as

$$f_k(R_{k,t}) \triangleq \begin{cases} 1 & \text{if } R_{k,t} > R_{k,t}^{\text{req}}, \\ 0 & \text{else.} \end{cases} \quad (12)$$

The local state $s_{l,t}$ indicates that if the user is adjacent to the IRS and its data rate requirement is satisfied, the user will be significant to the IRS. The local state of the BS at the t -th slot is defined as the transmit power, i.e.,

$$s_{B,t} \triangleq \left[\{P_{k,t}\}_{k \in \mathcal{K}} \right]. \quad (13)$$

The local states of all the agents constitute the global state, i.e.,

$$s_t \triangleq \left[\{s_{l,t}\}_{l \in \mathcal{L}}, s_{B,t} \right]. \quad (14)$$

Let \mathcal{S}_l be the state space of the l -th IRS. \mathcal{S}_B and \mathcal{S} are the BS agent state space and the global state space, respectively. We have $s_{l,t} \in \mathcal{S}_l$, $s_{B,t} \in \mathcal{S}_B$, $s_t \in \mathcal{S}$ and $\mathcal{S} \triangleq \left\{ \bigcup_{l \in \mathcal{L}} \{\mathcal{S}_l\} \right\} \cup \mathcal{S}_B$.

2) *Action*: The local action of the BS is defined as the transmit beamforming vector at the t -th slot:

$$\mathbf{a}_{B,t} \triangleq \left[\{\mathbf{v}_{k,t}\}_{k \in \mathcal{K}} \right]. \quad (15)$$

The local action of each IRS implies the chosen bit resolution, the phase shift vector and the ON/OFF status vector at the t -th slot as

$$\mathbf{a}_{l,t} \triangleq [b_l, \boldsymbol{\theta}_{l,t}, \boldsymbol{\rho}_{l,t}]. \quad (16)$$

The optimization argument $\boldsymbol{\Phi}_l, \forall l \in \mathcal{L}$ in (10) is a matrix while matrix processing is more complex than vector processing. For simplicity, we assume that each IRS element reflects the signal independently without signal coupling. Therefore, the optimization variables $\boldsymbol{\Phi}_l$ is sparse and equivalent to the phase shift vector $\boldsymbol{\theta}_l = [\theta_{l,1}, \dots, \theta_{l,N_l}]$.

The local action space of the l -th IRS agent is expressed as

$$\mathcal{A}_l \triangleq \{(b_l, \boldsymbol{\theta}_l, \boldsymbol{\rho}_l) | b_l \in \mathcal{B}, \theta_{l,n} \in \mathcal{F}_{b_l}, \rho_{l,n} \in \{0, 1\}, \forall n \in \mathcal{N}_l\}. \quad (17)$$

The local actions of all the agents make up the joint action at the t -th slot defined as

$$\mathbf{a}_t \triangleq [\{\mathbf{a}_{l,t}\}_{l \in \mathcal{L}}, \mathbf{a}_{B,t}]. \quad (18)$$

Let \mathcal{A}_B , $\mathcal{A} \triangleq \left\{ \bigcup_{l \in \mathcal{L}} \{\mathcal{A}_l\} \right\} \cup \mathcal{A}_B$ denote the BS action space and the joint action space, respectively.

3) *Reward*: Our objective is to maximize system data rate under constraints of the user data rate request, the transmit power and the IRS energy storage buffer. The reward represents the optimization objective with constraints, thus, the instant reward at the t -th slot is defined as

$$\begin{aligned} r_t \triangleq & \underbrace{\sum_{k=1}^K R_{k,t}}_{\text{part 1}} + \xi_1 \cdot \underbrace{\sum_{k=1}^K \min\{R_{k,t} - R_{k,t}^{\text{req}}, 0\}}_{\text{part 2}} \\ & + \xi_2 \cdot \underbrace{\min\left\{P^{\text{max}} - \sum_{k=1}^K P_{k,t}, 0\right\}}_{\text{part 3}} \\ & + \xi_3 \cdot \underbrace{\sum_{l=1}^L \min\{E_l(t) - E_l^{\text{min}}, 0\}}_{\text{part 4}}, \end{aligned} \quad (19)$$

where the part 1 is the achievable system data rate, the part 2, the part 3 and the part 4 are penalty which are defined as the user data rate request satisfaction level, the transmit power constraint and the IRS energy consumption degree, respectively. ξ_1, ξ_2 and ξ_3 are the trade-off coefficients to balance the rate and the penalty. The return is defined as the cumulative discounted future reward as follows $G_t \triangleq \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, wherer γ denotes the discount factor. Under the premise of satisfying the constraints, we aim to obtain an optimal policy that maximizes long-term expected return, which is defined as

$$\mathbb{E}_{\pi} [G_t | s_t = s]. \quad (20)$$

It is conspicuous to see that such a maximization is equivalent to solve the aforementioned optimization problem (10). Note that the optimal policy indicates the optimal BS transmit beamforming vector and the appropriate resolution, the phase shifting vector and the working status vector on each IRS. The state value function $V(s)$ and the state-action value function $Q(s, \mathbf{a})$ can be given as $V(s) = \mathbb{E}_{\pi} [G_t | s_t = s]$ and $Q(s, \mathbf{a}) = \mathbb{E}_{\pi} [G_t | s_t = s, \mathbf{a}_t = \mathbf{a}]$, respectively. The value functions satisfy the Bellman equation [36], and thus can be expressed as

$$\begin{aligned} V(s) &= \sum_{\mathbf{a}} \pi(\mathbf{a}|s) \sum_{s'} P_{ss'}^{\mathbf{a}} [r + \gamma V(s')], \\ Q(s, \mathbf{a}) &= \sum_{s'} P_{ss'}^{\mathbf{a}} \left[r + \gamma \sum_{\mathbf{a}'} \pi(\mathbf{a}'|s') Q(s', \mathbf{a}') \right], \end{aligned} \quad (21)$$

where $P_{ss'}^{\mathbf{a}} \triangleq \Pr(s_{t+1} = s' | s_t = s, \mathbf{a}_t = \mathbf{a})$ is the state transition probability from the current state s to the next state s' given the current action \mathbf{a} , policy $\pi(\mathbf{a}|s)$ denotes the conditional probability of taking action \mathbf{a} on the state s .

B. Multi-Agent Q-mix Networks

Considering the fact that the action space in (17) has high-dimensional and discrete characteristics, the training of the neural network would make the the computation resources overloaded. Moreover, value-based methods (e.g. Q-learning [36]) suffer from slow convergence and policy-based methods (e.g. DDPG [37]) may be trapped in local

optimal solutions. In addition, they are problematic to solve the discrete action space with high dimensionality. MADDPG [34] is a typical MARL algorithm, which works in terms of centralized training and distributed execution (CTDE) framework briefly. More specifically, it uses global information to update the centralized Q-value and renew each agent's policy through a distributed policy gradient method. However, it shows high computational complexity in large-scale action space. To handle with hybrid action space, the work in [38] proposed a framework called parameterized deep Q-networks (PDQN) to combine the algorithm DQN for discrete action space with the algorithm DDPG for continuous action space. More specifically, PDQN first chooses the continuous action x based on the discrete action b , utilizes a neural network $Q(s, b, x|\omega)$ parameterized by ω to approximate the state-action value function and updates the network parameters by minimizing the loss function defined as

$$l(\omega) = \mathbb{E} \left[(y - Q(s, b, x|\omega))^2 \right], \quad (22)$$

$$y = r + \gamma \max_{b'} Q(s', b', x'|\omega),$$

where y is the one-step discount reward value, s', b', x' are the global state, the discrete action vector and the continuous action vector in the next step, respectively. However, PDQN only serves the situation of the large-scale discrete-continuous hybrid action space, and thus *cannot* be used in the systems with discrete-discrete hybrid action space.

In our model, we have two levels of discrete actions of each IRS agent: the high-level action: the bit resolution b_l ; the low-level action: the pair of the phase shift vector and the ON/OFF state vector $x_l = (\theta_l, \rho_l)$. As for the l -th IRS, we first choose the high-level action b_l , which confirms the phase shift value set \mathcal{F}_{b_l} and power consumption per element μ_l , and then select the low-level action x_l , which determines the reflection matrix Φ_l and the power consumption P_l^{IRS} of the l -th IRS.

The high-level action space with B elements is low-dimensional, whereas the low-level action space with the dimension increasing exponentially with the growth of the number of IRS elements. A common method to handle with such a situation is to use ergodic methods (e.g. the ε -greedy method [36]), but would lead to the execution complexity growing linearly with $|\mathcal{A}_l|$, which is quite intractable.

To cope with the aforementioned issues in this section, we propose a novel multi-agent Q-mix framework, which, as illustrated in Fig. 2, is made up of three parts: a high-level Q-mix network, policy networks and a low-level Q-mix network.

1) *High-level Q-mix network*: It consists of L agent networks and a high-level mixing network. We assume that each IRS agent employs the same structure of the agent network. The agent network performs individual value function calculations for the corresponding IRS agent. The high-level mixing network can be viewed as a monotonic function with the input of all the individual value functions.

For the IRS agent $i, \forall i \in \mathcal{L}$, the local state $s_{i,t}$ and the bit resolution $b_{i,t-1}$ at the previous slot are utilized as the inputs of its agent network; moreover, in order to distinguish from other agents, the agent index i is also utilized as the input after one-hot encoding. Using the ε -greedy method, with probability

ε , the high-level action b_i is uniformly and randomly chosen from \mathcal{B} ; with probability $1 - \varepsilon$, b_i is selected using

$$b_i = \arg \max_{b_i \in \mathcal{B}} Q_i^{\text{high}}(s_i, b_i | \omega_i^{\text{high}}), \quad (23)$$

where we combine the local state $s_{i,t}$, the bit resolution $b_{i,t-1}$ and the agent number index i as the new local state s_i ; the individual value function Q_i^{high} is parameterized by ω_i^{high} . Here, we ignore the time subscript for simplicity.

The high-level mixing network at the BS takes the individual value functions of all the IRS agents as its input and the system information consisting of the global state s and joint low-level action $x \triangleq [x_1, \dots, x_L]$ as its auxiliary message input, since the global system information is critical for the distributed agent decision. It is assumed that the global information collection and the mixing network training are executed by the BS agent.

To guarantee high stability of the learning process and low variance of the value function, we utilize the double-network method [39] and construct two networks including a target network $Q^{\text{high,tar}}$ parameterized by $\omega^{\text{high,tar}}$, and an evaluated network $Q^{\text{high,eval}}$ parameterized by $\omega^{\text{high,eval}}$. The evaluated network evaluates the value function and updates the parameters in real time, while the target network copies the evaluated network parameters every a certain number of iterations. The parameters $\omega^{\text{high}} = \{\omega_1^{\text{high}}, \dots, \omega_L^{\text{high}}, \omega^{\text{high,eval}}\}$ of the high-level network can be updated by minimizing the loss function as

$$l(\omega^{\text{high}}) \triangleq \mathbb{E} \left[\left(y^{\text{high}} - Q^{\text{high,eval}}(s, b, x) \right)^2 \right], \quad (24)$$

$$y^{\text{high}} = r + \gamma \max_{b'} Q^{\text{high,tar}}(s', b', x'),$$

where y^{high} is the one-step target value derived from the target network, s', b', x' are the global state, the resolution action vector and joint low-level action in the next step derived from the experience buffer, respectively.

2) *Policy network*: Each IRS agent $i, \forall i \in \mathcal{L}$ can obtain the low-level action x_i through its policy network taking the IRS local state and bit resolution as input, while the BS agent can get its local action taking the BS local state as input. The IRS policy network utilizes policy methods, i.e., the Wolpertinger method and the proto-action policy gradient method, which will be discussed later, to map the continuous proto-actions with high dimensionality into actual discrete actions with low dimensionality.

3) *Low-level Q-mix network*: It is made up of $L + 1$ agent networks, which calculate the individual value function for all the agents, a low-level mixing network, which takes the individual value functions of all the agents as its input, and takes the global state s and the phase resolution vector b as its auxiliary information input.

Each IRS agent $i, \forall i \in \mathcal{L}$ can obtain its low-level individual value function $Q_i^{\text{low}}(s_i, b_i, x_i | \omega_i^{\text{low}})$ parameterized by ω_i^{low} and the BS agent maintain its individual value function $Q_B^{\text{low}}(s_B, a_B | \omega_B^{\text{low}})$ parameterized by ω_B^{low} . We also utilize the double-network method, that is, jointly utilize a target network $Q^{\text{low,tar}}$ parameterized by $\omega^{\text{low,tar}}$ and an evaluated network

$Q^{\text{low,eval}}$ parameterized by $\omega^{\text{low,eval}}$. The parameters $\omega^{\text{low}} = \{\omega_1^{\text{low}}, \dots, \omega_L^{\text{low}}, \omega_B^{\text{low}}, \omega^{\text{low,eval}}\}$ of the low-level network can be updated by minimizing the loss function as

$$l(\omega^{\text{low}}) \triangleq \mathbb{E} \left[\left(y^{\text{low}} - Q^{\text{low,eval}}(s, \mathbf{b}, \mathbf{x}) \right)^2 \right], \quad (25)$$

$$y^{\text{low}} = r + \gamma \max_{\mathbf{x}'} Q^{\text{low,tar}}(s', \mathbf{b}', \mathbf{x}'),$$

where y^{low} is the one-step target value obtained from the low-level mixing target network.

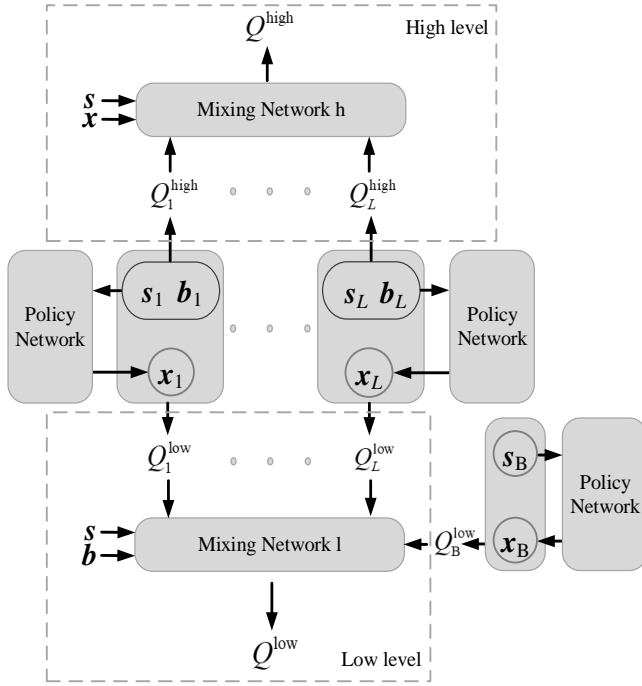


Figure 2: Multi-agent Q-mix framework

Based on the proposed framework, we summarize a multi-agent Q-mix algorithm in Algorithm 1. In the training process, each agent initializes its network parameters and observes its local state, while the BS agent collects all the local states as the global state. Then, each IRS agent $i, \forall i \in \mathcal{L}$ selects the bit resolution action b_i using the ε -greedy method and obtains the low-level action x_i through its policy network, while the BS agent gets the local action a_B . After executing a joint action, which consists of all the local actions, all the agents receive a reward from the system and observe their next local states. Then the transition experience, which consists of the global state, the joint action, the next global state and the reward, is stored in the experience replay buffer \mathcal{D} . We minimize the loss function and update the parameters of all the networks through experiences sampling from the experience buffer \mathcal{D} . The training process is completed when the reward converges. Finally, we can obtain an optimal setting of the beamforming matrix \mathbf{V} , the IRS bit resolution vector \mathbf{b} , the IRS phase shift matrix $\Phi_i, \forall i \in \mathcal{L}$ and the ON/OFF status vector $\rho_i, \forall i \in \mathcal{L}$.

Consider the fact that each IRS policy network has a continuous output space, while each IRS agent employs a discrete action space. To solve this problem, we propose a multi-agent Q-mix with Wolpertinger policy (MAQ-WP)

Algorithm 1: Multi-agent Q-mix

Input: The channel matrix $\mathbf{h}_{l,k}^{\text{RU}}, \mathbf{H}_l^{\text{BR}}, \mathbf{h}_k^{\text{BU}}, \forall l \in \mathcal{L}, k \in \mathcal{K}$

Output: The optimal transmit beamforming matrix \mathbf{V} , the optimal phase shift matrix Φ_l , the optimal ON/OFF status vector $\rho_l, \forall l \in \mathcal{L}$

- 1 Initialize: The high-level Q-mix network parameters $\omega^{\text{high}}, \omega^{\text{high,tar}}$; The low-level Q-mix network parameters $\omega^{\text{low}}, \omega^{\text{low,tar}}$; The learning rate of high-level Q-mix network, policy network and low-level Q-mix network α_h, α_p and α_l ; The replay buffer \mathcal{D} ;
- 2 **for** epoch = 1, 2, ..., J **do**
- 3 initialize the state s_0 ;
- 4 **for** $t = 0, 1, \dots, T-1$ **do**
- 5 **for** $i \in \mathcal{L}$ **do**
- 6 Select high-level action b_i by the ε -greedy method in (23);
- 7 **for** $i \in \mathcal{L}$ **do**
- 8 Get the action x_i through policy network;
- 9 Get the BS action $a_{B,t}$ and the IRS action $a_{i,t}, \forall i \in \mathcal{L}$;
- 10 Execute joint action \mathbf{a}_t and observe the next state s_{t+1} and reward r_t ;
- 11 Store the experience $(s_t, \mathbf{a}_t, s_{t+1}, r_t)$ in \mathcal{D} ;
- 12 Sample a random mini-batch of $N_{\mathcal{D}}$ experiences $(s_i, \mathbf{a}_i, s_{i+1}, r_i)$ from \mathcal{D} ;
- 13 **for** $i \in \mathcal{L}$ **do**
- 14 Get the high-level $Q_i^{\text{high}}(s_i, b_i)$;
- 15 **for** $i \in \mathcal{L}$ **do**
- 16 Get the low-level $Q_i^{\text{low}}(s_i, b_i, x_i)$;
- 17 Get the BS individual value function $Q_B^{\text{low}}(s_B, a_B)$;
- 18 Update the high-level Q-mix network parameters by minimizing the loss using (24);
- 19 Update the policy network parameters;
- 20 Update the low-level Q-mix network parameters by minimizing the loss using (25);
- 21 Update the target networks by copying parameters from the evaluated networks;

algorithm by utilizing the Wolpertinger policy method [38]. We assume that each IRS agent maintains a policy network $\mu_i^{\xi_i}(s_i, b_i)$ parameterized by ξ_i , while the BS agent manages a policy network $\mu_B^{\xi_B}(s_B)$ parameterized by ξ_B . As shown in Algorithm 2, each i -th IRS agent first receives the continuous-valued proto-actions from the Wolpertinger policy given its local state and bit resolution, and then retrieves k closest discrete-valued actions using the K-nearest neighbor (KNN) method [40].

As the low-level actions are discrete-valued, the parameters of the policy network *cannot* be updated with the gradient ascent method directly. Hence, the i -th IRS policy network is updated approximately with continuous-valued proto-actions

Algorithm 2: Wolpertinger Policy**Input:** The local state s_i , the high-level action $b_i, \forall i \in \mathcal{L}$ and the Wolpertinger factor k **Output:** The low-level action $x_i = (\theta_i, \rho_i)$

- 1 Get the proto-actions $\hat{x}_i = \mu_i^{\xi_i}(s_i, b_i)$;
- 2 Get the discrete action space $\mathcal{A}_{i,b_i} \subset \mathcal{A}_i$ given b_i ;
- 3 Use KNN to get the set of k closest actions

$$\mathcal{A}_{i,b_i,k} = \arg \min_{x_i \in \mathcal{A}_{i,b_i}}^k \|x_i - \hat{x}_i\|_2;$$

- 4 Get the discrete-valued low-level action

$$x_i = \arg \max_{x_i \in \mathcal{A}_{i,b_i,k}} Q_i^{\text{low}}(s_i, b_i, x_i)$$

using the following gradient

$$\begin{aligned} \nabla_{\xi_i} l(\xi_i) &\approx \nabla_{\xi_i} Q_i^{\text{low}}(s_i, b_i, \hat{x}_i \mid \hat{x}_i = \mu_i^{\xi_i}(s_i, b_i)) \\ &= \nabla_{\hat{x}_i} Q_i^{\text{low}}(s_i, b_i, \mu_i^{\xi_i}(s_i, b_i)) \nabla_{\xi_i} \mu_i^{\xi_i}(s_i, b_i). \end{aligned} \quad (26)$$

As the BS agent employs a continuous action space, we can update the parameters of the network by the gradient ascent method directly as

$$\begin{aligned} \nabla_{\xi_B} l(\xi_B) &= \nabla_{\xi_B} Q_B^{\text{low}}(s_B, a_B \mid a_B = \xi_B^{\xi_B}(s_B)) \\ &= \nabla_{a_B} Q_B^{\text{low}}(s_B, \mu_B^{\xi_B}(s_B)) \nabla_{\xi_B} \mu_B^{\xi_B}(s_B). \end{aligned} \quad (27)$$

However, the proposed MAQ-WP algorithm is still computationally demanding as the search of the closest action is linear related to the Wolpertinger factor k and the number of the IRS agents L . In addition, the approximation in (26) may increase the variance of the individual value function $Q_i^{\text{low}}(s_i, b_i, x_i), \forall i \in \mathcal{L}$. To overcome these weaknesses, we develop a multi-agent Q-mix with policy gradient (MAQ-PG) algorithm, which allows each agent to maintain an policy network μ_{θ_i} parameterized by θ_i and a mapping function $\phi(e) : \mathcal{E} \rightarrow \mathcal{A}_{i,b_i}$ for mapping the proto-actions $\mu_{\theta_i}(e|s_i)$ based on a given state s_i to the actual low-level actions. It is assumed that, given the range set of the proto-actions \mathcal{E}_a , the mapping function $\phi(e)$ will deterministically generate an actual low-level action, i.e., $\forall e \in \mathcal{E}_a$ such that $\phi(e) = a$. Here, the input of the $\mu_{\theta_i}(e|s_i)$, s_i , is defined by combining the i -th IRS agent local state and its high-level action.

In addition, the local long-term discounted function is defined as

$$J_i^{\theta_i} \triangleq \mathbb{E}_{s_i} [V_i^{\theta_i}(s_i)] = \int_{s_i} d_i(s_i) V_i^{\theta_i}(s_i) ds_i, \quad (28)$$

where $d_i(s_i)$ is the original probability distribution of the state s_i . We use the Bellman Equation to rewrite the equation (28) as

$$J_i^{\theta_i} = \int_{s_i} d_i(s_i) \sum_{x_i \in \mathcal{A}_{i,b_i}} \pi_{\theta_i}(x_i|s_i) Q_i^{\theta_i}(s_i, x_i) ds_i. \quad (29)$$

With a slight abuse of notation, denote the low-level action policy of the i -th IRS agent by π_{θ_i} . We can get the action

representation form of π_{θ_i} by accumulating the proto-action probabilities

$$\pi_{\theta_i}(x_i|s_i) = \int_{\mathcal{E}_{x_i}} \mu_{\theta_i}(e|s_i) de. \quad (30)$$

Lemma 1 Given the proto-actions $\mu_{\theta_i}(e|s_i)$ and the mapping function $\phi(e)$, the policy gradient can be calculated as

$$\nabla_{\theta_i} J_i^{\theta_i} = \mathbb{E}_{s_i, e} \left[\nabla_{\theta_i} \log \mu_{\theta_i}(e|s_i) Q_i^{\theta_i}(s_i, \phi(e)) \right]. \quad (31)$$

Proof: See Appendix A. ■

$\phi(e)$ can be used to obtain an action with the probability 1, while it may not be a prior knowledge. Thus, we can construct an estimator $\hat{\phi}(x_i|e)$ to approximate the action selection probability using the Kullback-Leibler (KL) divergence. As we assume that the action x_i is conditionally independent with the state s_i given the proto-action e , we denote the true probability from state s_i to action x_i by $p(x_i|s_i) = p(e|s_i)p(x_i|e)$ and denote the estimator by $\hat{p}(x_i|s_i) = p(e|s_i)\hat{\phi}(x_i|e)$. Thus, the KL divergence between p and \hat{p} is expressed as

$$D_{\text{KL}}(p||\hat{p}) = \mathbb{E}_{x_i} \left[\log \left(\frac{p(x_i|s_i)}{\hat{p}(x_i|s_i)} \right) \right] = -\mathbb{E}_{x_i} \left[\log \left(\frac{\hat{\phi}(x_i|s_i)}{p(x_i|e)} \right) \right]. \quad (32)$$

As the denominator in (32) is irrelevant to the parameters of the estimator, we can neglect the denominator and update the estimator parameters by minimizing the loss function as

$$l(\hat{\phi}) \triangleq -\mathbb{E}_{x_i} [\log \hat{\phi}(x_i|e)]. \quad (33)$$

C. Computational Complexity Analysis

The computational complexity is primarily determined by the network architectures of the high-level agent networks, the low-level agent networks, the high-level mixing network, the low-level mixing network and the policy networks.

As for the high-level agent network, the number of neurons in the input layer is specified by the dimension of the local state and the bit resolution, which is $K + 2$. The number of neurons in the output layer is 1. It is assumed that the agent network utilizes a total of L^{ha} fully connected neural networks, where the l -th ($2 \leq l \leq L^{\text{ha}} - 1$) hidden layer contains n_l^{ha} neurons. Then, the number of the weights in the input layer, the l -th hidden layer, and the final hidden layer are $(K + 2)n_1^{\text{ha}}, n_{l-1}^{\text{ha}}n_l^{\text{ha}}$ and $n_{L^{\text{ha}}-1}^{\text{ha}}$, respectively. Similar to the high-level agent network, we assume each that low-level agent network possess L^{la} hidden fully connected layers, where the l -th ($2 \leq l \leq L^{\text{la}} - 1$) hidden layer contains n_l^{la} neurons. Therefore, the number of the weights in the input layer, the l -th hidden layer, and the final hidden layer are $(K + 2 + 2N)n_1^{\text{la}}, n_{l-1}^{\text{la}}n_l^{\text{la}}$ and $n_{L^{\text{la}}-1}^{\text{la}}$, respectively. For simplicity, here we assume that each IRS agent has identical number of reflecting elements N .

Moreover, we adopt L^{ip} hidden fully connected for IRS agent policy network and L^{bp} hidden fully connected for BS agent policy network, where the corresponding l -th ($2 \leq l \leq L^{\text{ip}} - 1, 2 \leq l \leq L^{\text{bp}} - 1$) hidden layer contains n_l^{ip} and n_l^{bp} neurons, respectively. Thus for IRS policy network, the number of the weights in the input layer, the l -th hidden layer, and the final hidden layer are $(K + 2)n_1^{\text{ip}}, n_{l-1}^{\text{ip}}n_l^{\text{ip}}$ and $n_{L^{\text{ip}}-1}^{\text{ip}}$,

respectively. For BS policy network, the number of the weights in the input layer, the l -th hidden layer, and the final hidden layer in turn are $Kn_1^{\text{bp}}, n_{l-1}^{\text{bp}}n_l^{\text{bp}}$ and n_{L-1}^{bp} . For high-level mixing network and low-level mixing network, the number of the total weights are given by $n^{\text{hmix}}(KL+K+2L+1)(L+1)$ and $n^{\text{lmix}}(KL+K+2L+2NL)(L+1)$, respectively, where n^{hmix} and n^{lmix} is the total number of the high-level and low-level mixing network neurons.

Suppose that the computational complexity to train a single weight is W . Finally, the computational complexity of the proposed MAQ-WP is $O\left(W\left[\left((K+2)n_1^{\text{ha}} + \sum_{l=2}^{L-1} n_{l-1}^{\text{ha}}n_l^{\text{ha}} + n_{L-1}^{\text{ha}}\right)L + \left((K+2+2N)n_1^{\text{la}} + \sum_{l=2}^{L-1} n_{l-1}^{\text{la}}n_l^{\text{la}} + n_{L-1}^{\text{la}}\right)L + \left((K+2)n_1^{\text{ip}} + \sum_{l=2}^{L-1} n_{l-1}^{\text{ip}}n_l^{\text{ip}} + n_{L-1}^{\text{ip}}\right)L + Kn_1^{\text{bp}} + \sum_{l=2}^{L-1} n_{l-1}^{\text{bp}}n_l^{\text{bp}} + n_{L-1}^{\text{bp}} + n^{\text{hmix}}(KL+K+2L+1)(L+1) + n^{\text{lmix}}(KL+K+2L+2NL)(L+1)\right]\right)$, which mainly increases quadratically with the IRS number L and increases linearly with the number of users K . Regarding centralized policy network, it is necessary to traverse under various resolution \mathbf{b} to choose the joint low-level action \mathbf{x} . The computational complexity is thus grows exponentially with the number of IRS L .

IV. SIMULATION RESULTS

In this section, we evaluate the performance of the proposed MAQ-WP and MAQ-PG algorithms numerically. In our simulations, we consider the multi-IRSs-assisted down-link communication system described in Section II and assume that each agent has a full knowledge of perfect channel information.

1) *Simulation settings*: The BS equipped with 4 antennas is located at the coordinate (0,0). The users are uniformly distributed in the ring area centered on the BS with the inner radius 100m and the outer radius 120m. The IRSs are uniformly distributed on the circle with the center of the BS and the radius of 100m. In addition, the maximum BS transmit power P^{max} is set to be varying from 5dBm to 30dBm, and the system noise power is set to be -80dBm. We assume that the channel \mathbf{h}_k^{BU} follows Rayleigh fading and the IRS-assisted channels $\mathbf{H}_l^{\text{BR}}, \mathbf{h}_{l,k}^{\text{RU}}, \forall l \in \mathcal{L}$ follow Rician fading, which can be modeled as

$$\begin{aligned} \mathbf{H}_l^{\text{BR}} &= \text{PL}_{1,l} \cdot \left(\sqrt{\frac{\epsilon_1}{1+\epsilon_1}} \mathbf{a}_{N_l}(\vartheta_l) \mathbf{a}_M(\varphi)^H + \sqrt{\frac{1}{1+\epsilon_1}} \bar{\mathbf{H}}_l^{\text{BR}} \right), \\ \mathbf{h}_{l,k}^{\text{RU}} &= \text{PL}_{2,l,k} \cdot \left(\sqrt{\frac{\epsilon_2}{1+\epsilon_2}} \mathbf{a}_{N_l}(\chi_{l,k}) + \sqrt{\frac{1}{1+\epsilon_2}} \bar{\mathbf{h}}_{l,k}^{\text{RU}} \right), \end{aligned} \quad (34)$$

where ϵ_1, ϵ_2 are Rician factors, $\mathbf{a}(\cdot)$ is the steering vector that is defined as $\mathbf{a}_N(x) = \left[1, e^{-j2\pi \frac{d}{\lambda} \sin x}, \dots, e^{-j2(N-1)\pi \frac{d}{\lambda} \sin x} \right]^T$, $\vartheta_l, \varphi, \chi_{l,k}$ are angular settings, $\bar{\mathbf{H}}_l^{\text{BR}}$ and $\bar{\mathbf{h}}_{l,k}^{\text{RU}}$ denote the non-line-of-sight (NLOS) components whose elements both follow the symmetric complex Gaussian distribution $\mathcal{CN}(0, 1)$, $\text{PL}_{1,l}$ denotes the path loss between the BS and the l -th IRS, $\text{PL}_{2,l,k}$ denotes the path loss between the l -th IRS and the k -th user. These two types of path loss are defined by

$$\begin{aligned} \text{PL}_{1,l} &= (\text{PL}_0 - 10\kappa_l \lg(d_l/d_0)), \\ \text{PL}_{2,l,k} &= (\text{PL}_0 - 10\kappa_{l,k} \lg(d_{l,k}/d_0)), \end{aligned} \quad (35)$$

where d_l is the distance between the BS and the l -th IRS, $d_{l,k}$ is the distance between the l -th IRS and the k -th user. The harvested energy obeys Poisson distribution with the probability function $\Pr(X = k) = \frac{\varsigma^k}{k!} e^{-\varsigma}, k = 0, 1, \dots$, where $\varsigma = 2.2$. The descriptions and values of the parameters $\text{PL}_0, \kappa_l, \kappa_{l,k}, \epsilon_1, \epsilon_2$ in (34) and (35) are listed in Table II.

Table II: System Model Parameters

Parameters	Description	Values
PL_0	The path loss at the reference distance $d_0 = 1\text{m}$	30dB
κ_l	the path loss factor between BS and IRS	2
$\kappa_{l,k}$	the path loss factor between user and IRS	2.8
ϵ_1, ϵ_2	the Rician factor	10

The network parameters settings of the proposed algorithms are summarized in Table III

Table III: Network Parameters

J	Total training epochs	20000
T	Training steps per epoch	100
D	Experience replay buffer size	100000
$N_{\mathcal{D}}$	Sample mini-batch	128
γ	Discount factor	0.99
α_h	Learning rate of the high-level Q-mix network	0.0001
α_p	Learning rate of the policy networks	0.0001
α_l	Learning rate of the low-level Q-mix network	0.0001
ϵ	ϵ -greedy policy factor	[0.02, 0.2]

2) *Comparisons with benchmarks*: We consider the number of the IRSs $L = 2$, the number of the users $K = 4$, the number of the elements per IRS $N_l = 10$, the Wolpertinger policy factor $k = 50$ and the maximum BS transmit power $P^{\text{max}} = 5\text{dBm}$. Fig. 3 compares the proposed MAQ-WP and MAQ-PG algorithms with two benchmarks: MADDPG and independent learning (IL). IL means that in multi-agent environment, each agent only cares about itself and learns independently without considering the impact of other agents' actions or policies to it. It is obvious that the proposed algorithms both show appealing performance. The MAQ-WP algorithm finally achieves similar performance compared with the MAQ-PG algorithm at a larger variance as we make an approximation in (26). The rewards of the proposed algorithms converge faster compared with MADDPG. The MAQ-PG algorithm converges at about the 5000-th epoch and the MAQ-WP algorithm converges at about the 7500-th epoch, while MADDPG converges at the 10000-th epoch. This is because that our algorithms develop the hierarchical actions and calculate the gradient in (26) and (31) based on the low-level action, while MADDPG updates the gradient by the joint action with extra computational complexity. Fig. 4 compares the user average data rates under different algorithms with the growth of IRS number L . It is observed that, the average data rates increase with the number of IRS L , resulting from the increase in the number of beamforming policies as L increases. The MAQ-PG and MAQ-WP algorithms enjoy an data rate improvement of 4.9% ~ 8.8% and 6.1% ~ 10.7% compared with MADDPG, respectively. In addition, the proposed algorithms both significantly outperform the IL. The comparisons in Figs. 3-4 jointly indicate that our proposed algorithms enjoy both excellent network performance and fast convergence in learning.

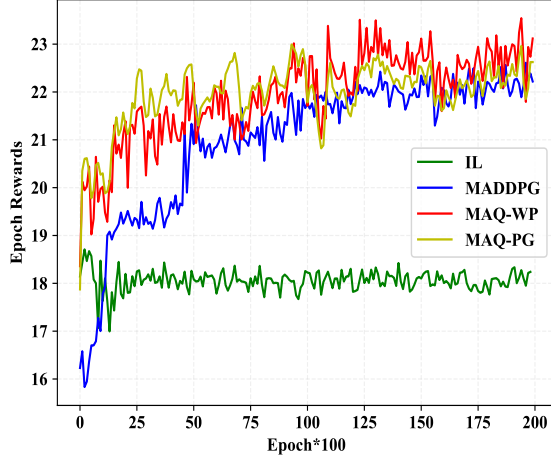


Figure 3: Convergence comparisons of different algorithms

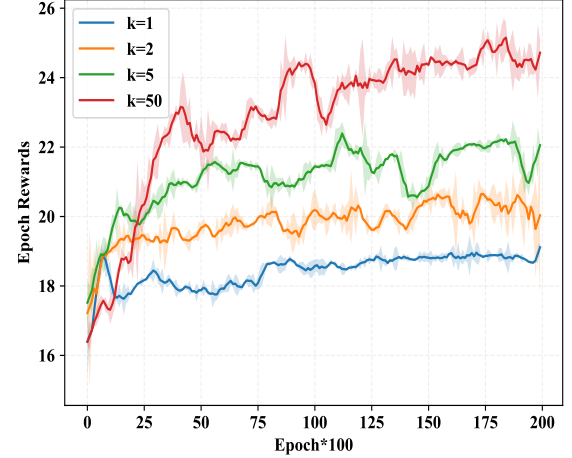


Figure 5: Performance comparisons versus epochs under different Wolpertinger policy factors

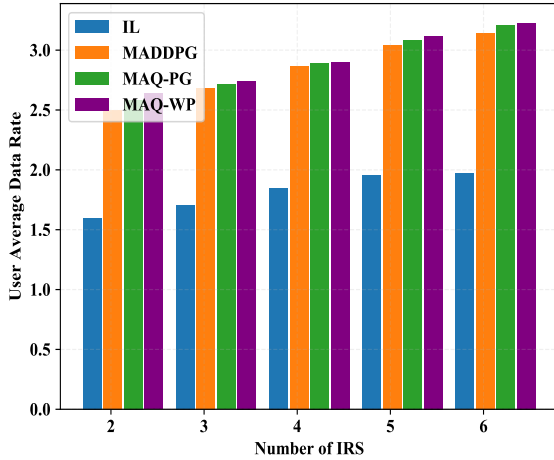


Figure 4: User average data rate of different algorithms

3) *Impact of Wolpertinger policy factor:* Fig. 5 shows the performance of the proposed MAQ-WP algorithm under different Wolpertinger policy factors, i.e., $k = 1, 2, 5, 50$. It is worthy noting that, the solid line represents the average reward curve, and the shaded region around each learning curve shows the reward variance. It can be observed that the convergence value grows with the Wolpertinger factor k . The learning processes with the factor $k = 1, 2$ both achieve convergence after 2500 epochs, whereas the processes with the factor $k = 5, 50$ converge much slower as the Wolpertinger policy selects an optimal action from k actions. The results imply that if the factor k is large enough, the performance can reach a considerable convergence value at cost of the training rate.

4) *Impact of maximum BS transmit power:* Fig. 6 presents the system rate of the MAQ-PG algorithm as a function of the maximum transmit power P^{\max} . It can be observed that the system rate increases steadily and tends to be stable gradually. The reason behind is that the achievable system rate

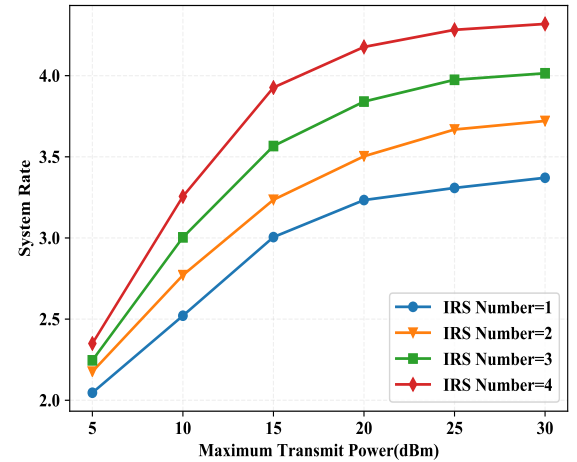


Figure 6: Performance of system rate comparisons versus different transmit power for MAQ-PG

increases with the transmitting signal maximum power limit. Furthermore, as the channel interferences *cannot* be ignored under large P^{\max} , the system rate will eventually reach the convergence.

Figs. 7-8 show the user data rate satisfaction rate and the ratios of final convergence under different P^{\max} and different numbers of IRS L , respectively. As expected, the user data rate satisfaction rate increases monotonically with P^{\max} and L . The reason is that when P^{\max} and L increase, the received SINR in (4) increases, resulting in the improvement of the user data rate satisfaction rate.

5) *Impact of IRS element:* Fig. 9 shows the epoch reward of the MAQ-PG algorithm when $N_l = 10, 20, 30, 40$, $K = 4$ and $P^{\max} = 5\text{dBm}$. It is apparent that as N_l increases, the convergence value of the epoch reward grows. The reason is that the IRS with more elements can be used to provide more accurate beamforming policies and more reasonable energy

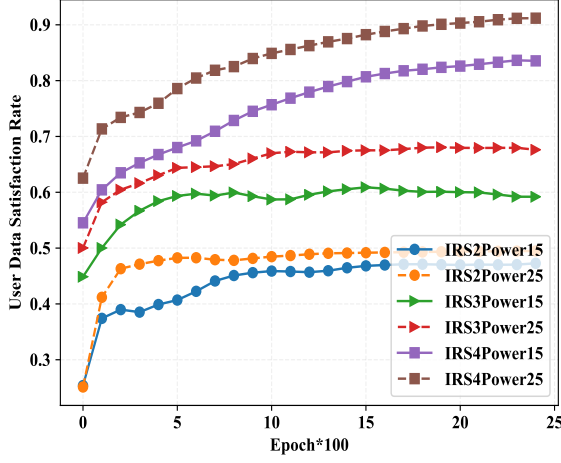


Figure 7: Performance of data satisfaction rate comparisons curve versus epochs under different transmit power for MAQ-PG

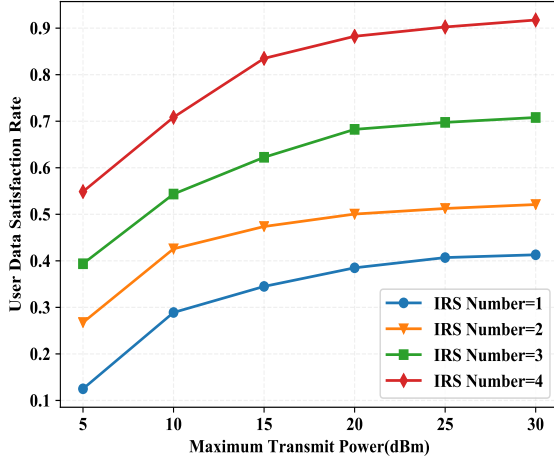


Figure 8: Performance of data satisfaction rate comparisons versus different transmit power for MAQ-PG

consumption methods, which effectively improves the instant reward and diminishes the penalties in (19).

Fig. 10 shows the user data rate satisfaction rate of the MAQ-PG algorithm under different numbers of the IRS elements N_l . The results indicate that such a satisfaction rate increases with N_l , and achieves 91.75% when $N_l = 40$, $K = 5$, $L = 4$ and $P^{\max} = 5\text{dBm}$, which implies that most of the users meet the individual requirement of their data rates. Fig. 11 illustrates the user data rate satisfaction rate for different numbers of IRS elements N_l and different numbers of IRSs L . It is also found that the user data satisfaction rate increases with N_l and L . This is because that more IRSs and IRS elements, more signal paths and signal power can be reflected by the IRSs to improve the received SINR in (4). The results indicate that an appropriate N_l is beneficial to improve

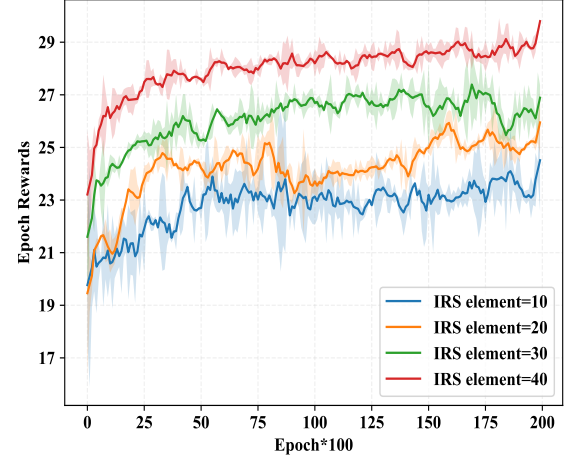


Figure 9: Performance comparisons versus epochs under different IRS elements for MAQ-PG

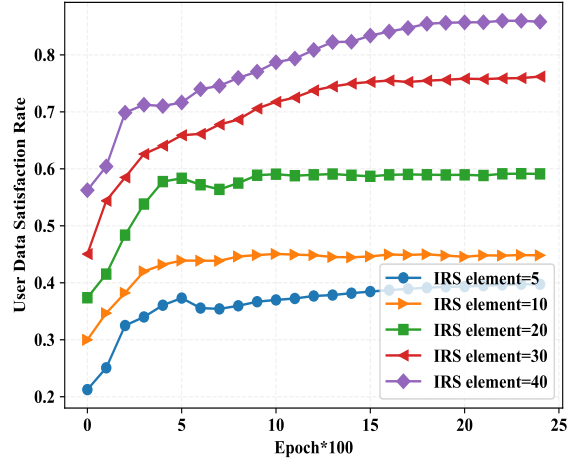


Figure 10: Performance of data satisfaction rate curve comparisons versus epochs under different IRS elements for MAQ-PG

the user data rate satisfaction rate; otherwise would cause the dissatisfaction of the user data rate constraints and the waste of the hardware resources.

V. CONCLUSION

In this paper, a multiple IRS-assisted multi-user communication system with an energy harvesting mechanism was considered. We formulated a BS transmit beamforming and IRSs phase shift beamforming joint optimization problem with transmit power limits, user data rate requirements and energy storage buffer constraints. We further converted this complicated non-convex optimization problem into an MDP model and designed a multi-agent Q-mix framework to decouple the optimization parameters. Moreover, the MAQ-WP and MAQ-PG algorithms were proposed to handle the high-dimensional

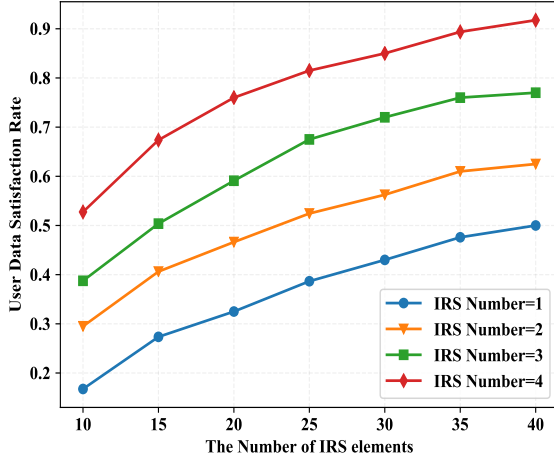


Figure 11: Performance of data satisfaction rate comparisons versus different IRS elements for MAQ-PG

and hybrid action space. The proposed algorithms separate the phase shift actions into two hierarchical actions, and thus are able to significantly reduce the overload of computational cost and accelerate the convergence of the learning process on the premise of a high achievable system rate. Simulation results confirmed the performance advantage of the proposed algorithms over other algorithms.

REFERENCES

- [1] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.
- [2] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, 2017.
- [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] O. Ozdogan, E. Bjornson, and E. G. Larsson, "Intelligent reflecting surfaces: Physics, propagation, and pathloss modeling," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 581–585, 2020.
- [5] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3313–3351, 2021.
- [6] 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, 2019.
- [7] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Communications Letters*, vol. 23, no. 9, pp. 1488–1492, 2019.
- [8] 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, 2019.
- [9] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Transactions on Communications*, vol. 68, no. 9, pp. 5849–5863, 2020.
- [10] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "A framework of robust transmission design for irs-aided miso communications with imperfect cascaded channels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 5092–5106, 2020.
- [11] W. Yan, X. Yuan, and X. Kuai, "Passive beamforming and information transfer via large intelligent surface," *IEEE Wireless Communications Letters*, vol. 9, no. 4, pp. 533–537, 2020.
- [12] Y. Xiu, Y. Zhao, Y. Liu, J. Zhao, O. Yagan, and N. Wei, "IRS-assisted millimeter wave communications: Joint power allocation and beamforming design," in *2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2021, pp. 1–6.
- [13] D. Ma, M. Ding, and M. Hassan, "Enhancing cellular communications for UAVs via intelligent reflective surface," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–6.
- [14] M. Hua, L. Yang, Q. Wu, C. Pan, C. Li, and A. Lee Swindlehurst, "UAV-assisted intelligent reflecting surface symbiotic radio system," *IEEE Transactions on Wireless Communications*, Early Access, 2021.
- [15] L. Ge, P. Dong, H. Zhang, J.-B. Wang, and X. You, "Joint beamforming and trajectory optimization for intelligent reflecting surfaces-assisted UAV communications," *IEEE Access*, vol. 8, pp. 78 702–78 712, 2020.
- [16] Q. Wu and R. Zhang, "Weighted sum power maximization for intelligent reflecting surface aided SWIPT," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 586–590, 2020.
- [17] Q. Wu and R. Rui, "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, 2020.
- [18] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1719–1734, 2020.
- [19] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4522–4535, 2020.
- [20] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1838–1851, 2020.
- [21] C. You, B. Zheng, and R. Zhang, "Channel estimation and passive beamforming for intelligent reflecting surface: Discrete phase shift and progressive refinement," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2604–2620, 2020.
- [22] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.
- [23] X. You, C. Zhang, X. Tan, S. Jin, and H. Wu, "AI for 5G: Research directions and paradigms," *Sci. China Inf. Sci.*, vol. 62, no. 2, pp. 21 301:1–21 301:13, 2019.
- [24] X. Wu, X. Li, J. Li, P. C. Ching, V. C. M. Leung, and H. V. Poor, "Caching transient content for IoT sensing: Multi-agent soft actor-critic," *IEEE Transactions on Communications*, pp. 1–1, 2021.
- [25] Q. Liu, L. Shi, L. Sun, J. Li, M. Ding, and F. Shu, "Path planning for UAV-mounted mobile edge computing with deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 5723–5728, 2020.
- [26] C. Ma, J. Li, M. Ding, H. H. Yang, F. Shu, T. Q. S. Quek, and H. V. Poor, "On safeguarding privacy and security in the framework of federated learning," *IEEE Network*, vol. 34, no. 4, pp. 242–248, 2020.
- [27] C. Huang, G. C. Alexandropoulos, C. Yuen, and M. Debbah, "Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019, pp. 1–5.
- [28] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 745–749, 2020.
- [29] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, "Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1–5.
- [30] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 375–388, 2021.
- [31] G. Lee, M. Jung, A. T. Z. Kargari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

- [32] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, and M. Debbah, "Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1663–1677, 2021.
- [33] 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, 2020.
- [34] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Annual Conference on Neural Information Processing Systems*, 2017, pp. 6379–6390.
- [35] 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, 2019.
- [36] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.
- [37] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. A. Riedmiller, "Deterministic policy gradient algorithms," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, vol. 32, 2014, pp. 387–395.
- [38] J. Xiong, Q. Wang, Z. Yang, P. Sun, L. Han, Y. Zheng, H. Fu, T. Zhang, J. Liu, and H. Liu, "Parametrized deep Q-networks learning: Reinforcement learning with discrete-continuous hybrid action space," *arXiv:1810.06394*, 2018.
- [39] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nat.*, vol. 518, no. 7540, pp. 529–533, 2015.
- [40] T. M. Mitchell, *Machine Learning*. Machine Learning, 2003.
- [41] X. Wu, J. Li, M. Xiao, P. C. Ching, and H. Vincent Poor, "Multi-agent reinforcement learning for cooperative coded caching via homotopy optimization," *IEEE Transactions on Wireless Communications*, Early Access, 2021.

APPENDIX

A. Proof of Lemma 1

Using Lemma 1, we operate the gradient of the equation (29) and unfold it as

$$\begin{aligned}
 \nabla_{\theta_i} J_i^{\theta_i} &= \nabla_{\theta_i} \int_{s_i} d(s_i) \sum_{x_i \in \mathcal{A}_{i,b_i}} \pi_{\theta_i}(x_i|s_i) Q_i^{\theta_i}(s_i, x_i) ds_i \\
 &= \int_{s_i} d(s_i) \left(\sum_{x_i \in \mathcal{A}_{i,b_i}} \nabla_{\theta_i} \pi_{\theta_i}(x_i|s_i) Q_i^{\theta_i}(s_i, x_i) \right. \\
 &\quad \left. + \sum_{x_i \in \mathcal{A}_{i,b_i}} \pi_{\theta_i}(x_i|s_i) \nabla_{\theta_i} Q_i^{\theta_i}(s_i, x_i) \right) ds_i. \quad (36)
 \end{aligned}$$

Then the proof follows the procedures in [37, 41]. We can obtain

$$\begin{aligned}
 \nabla_{\theta_i} J_i^{\theta_i} &= \int_{s_i} d(s_i) \sum_{k=0}^{\infty} \int_{s_{i'}} \gamma^k p(s_i \rightarrow s_{i'}, k) \\
 &\quad \cdot \sum_{x_i \in \mathcal{A}_{i,b_i}} \nabla_{\theta_i} \pi_{\theta_i}(x_i|s_i) Q_i^{\theta_i}(s_i, x_i) ds_{i'}, \quad (37)
 \end{aligned}$$

where $p(s_i \rightarrow s_{i'}, k)$ denotes the probability of transition from state s_i to state $s_{i'}$ in k steps. Let

$$\rho(s_i) = \sum_{k=0}^{\infty} \int_{s_{i'}} d(s_i) \gamma^k p(s_i \rightarrow s_{i'}, k) ds_{i'} \quad (38)$$

indicate the stationary state distribution of the Markov chain starting from the state s_i . Substitute the equation (38) into the

gradient and using action representation (30) to replace the policy, resulting in

$$\nabla_{\theta_i} J_i^{\theta_i} = \int_{s_i} \rho(s_i) \sum_{x_i \in \mathcal{A}_{i,b_i}} \int_{\mathcal{E}_{x_i}} \nabla_{\theta_i} \mu_{\theta_i}(e|s_i) \cdot Q_i^{\theta_i}(s_i, x_i) de ds_i. \quad (39)$$

In the range of \mathcal{E}_{x_i} , each e maps a unique action x_i , $\phi(e) = x_i$. The summation of action over the action space can be substituted by the entire domain of e , i.e., \mathcal{E} . As a result, the final gradient formula can be calculated as follows

$$\begin{aligned}
 \nabla_{\theta_i} J_i^{\theta_i} &= \int_{s_i} \rho(s_i) \int_{\mathcal{E}} \nabla_{\theta_i} \mu_{\theta_i}(e|s_i) \cdot Q_i^{\theta_i}(s_i, \phi(e)) de ds_i \\
 &= \int_{s_i} \rho(s_i) \int_{\mathcal{E}} \mu_{\theta_i}(e|s_i) \\
 &\quad \cdot \nabla_{\theta_i} \log \mu_{\theta_i}(e|s_i) \cdot Q_i^{\theta_i}(s_i, \phi(e)) de ds_i \\
 &= \mathbb{E}_{s_i, e} \left[\nabla_{\theta_i} \log \mu_{\theta_i}(e|s_i) \cdot Q_i^{\theta_i}(s_i, \phi(e)) \right]. \quad (40)
 \end{aligned}$$



Jie Zhang received the B.S. degree from the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2019, where he is pursuing the M.S. degree currently. His research interests include reinforcement learning, deep learning and intelligent reflecting surface.



Jun Li (M'09-SM'16) received Ph. D degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, P. R. China in 2009. From January 2009 to June 2009, he worked in the Department of Research and Innovation, Alcatel Lucent Shanghai Bell as a Research Scientist. From June 2009 to April 2012, he was a Postdoctoral Fellow at the School of Electrical Engineering and Telecommunications, the University of New South Wales, Australia. From April 2012 to June 2015, he was a Research Fellow at the School of Electrical

Engineering, the University of Sydney, Australia. From June 2015 to now, he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China. He was a visiting professor at Princeton University from 2018 to 2019. His research interests include network information theory, game theory, distributed intelligence, multiple agent reinforcement learning, and their applications in ultra-dense wireless networks, mobile edge computing, network privacy and security, and industrial Internet of things. He has co-authored more than 200 papers in IEEE journals and conferences, and holds 1 US patents and more than 10 Chinese patents in these areas. He was serving as an editor of IEEE Communication Letters and TPC member for several flagship IEEE conferences. He received Exemplary Reviewer of IEEE Transactions on Communications in 2018, and best paper award from IEEE International Conference on 5G for Future Wireless Networks in 2017.



Yijin Zhang (M'14–SM'18) received the Ph.D. degree in information engineering from The Chinese University of Hong Kong, in 2010. He joined the Nanjing University of Science and Technology, China, in 2011, where he is currently a Professor with the School of Electronic and Optical Engineering. His research interests include sequence design and resource allocation for communication networks.



Qingqing Wu (S'13–M'16–SM'21) received the B.Eng. and the Ph.D. degrees in Electronic Engineering from South China University of Technology and Shanghai Jiao Tong University (SJTU) in 2012 and 2016, respectively. He is currently an assistant professor with the State key laboratory of Internet of Things for Smart City, University of Macau. From 2016 to 2020, he was a Research Fellow in the Department of Electrical and Computer Engineering at National University of Singapore. His current

research interest includes intelligent reflecting surface (IRS), unmanned aerial vehicle (UAV) communications, and MIMO transceiver design. He has coauthored more than 150 IEEE papers with 24 ESI highly cited papers and 8 ESI hot papers, which have received more than 10,000 Google citations. He was listed as Clarivate ESI Highly Cited Researcher in 2021 and World's Top 2% Scientist by Stanford University in 2020.

He was the recipient of the IEEE Communications Society Young Author Best Paper Award in 2021, the Outstanding Ph.D. Thesis Award of China Institute of Communications in 2017, the Outstanding Ph.D. Thesis Funding in SJTU in 2016, the IEEE ICC Best Paper Award in 2021, and IEEE WCSP Best Paper Award in 2015. He was the Exemplary Editor of IEEE Communications Letters in 2019 and the Exemplary Reviewer of several IEEE journals. He serves as an Associate Editor for IEEE Transactions on Communications, IEEE Communications Letters, IEEE Wireless Communications Letters, IEEE Open Journal of Communications Society (OJ-COMS), and IEEE Open Journal of Vehicular Technology (OJVT). He is the Lead Guest Editor for IEEE Journal on Selected Areas in Communications on "UAV Communications in 5G and Beyond Networks", and the Guest Editor for IEEE OJVT on "6G Intelligent Communications" and IEEE OJ-COMS on "Reconfigurable Intelligent Surface Based Communications for 6G Wireless Networks". He is the workshop co-chair for IEEE ICC 2019–2022 workshop on "Integrating UAVs into 5G and Beyond", and the workshop co-chair for IEEE GLOBECOM 2020 and ICC 2021 workshop on "Reconfigurable Intelligent Surfaces for Wireless Communication for Beyond 5G". He serves as the Workshops and Symposia Officer of Reconfigurable Intelligent Surfaces Emerging Technology Initiative and Research Blog Officer of Aerial Communications Emerging Technology Initiative. He is the IEEE Communications Society Young Professional Chair in Asia Pacific Region.



Xiongwei Wu received the the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong (CUHK), Hong Kong SAR, China, in 2021. From August 2018 to December 2018, he was a Visiting International Research Student with The University of British Columbia (UBC), Vancouver, BC, Canada. From July 2019 to January 2020, he was a Visiting Student Research Collaborator with Princeton University, Princeton, NJ, USA. His research interests include signal processing and resource allocation, mobile edge computing, big data

analytics, and machine learning.



Feng Shu (M'2016) was born in 1973. He received the Ph.D., M.S., and B.S. degrees from the Southeast University, Nanjing, in 2002, XiDian University, Xi'an, China, in 1997, and Fuyang teaching College, Fuyang, China, in 1994, respectively. From September 2009 to September 2010, he is a visiting post-doctor at the University of Texas at Dallas. From October 2005 to November 2020, he was with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China, where he was promoted from associate professor to a full professor of supervising Ph.D students in 2013. Since December 2020, he has been with the School of Information and Communication Engineering, Hainan University, Haikou, where he is currently a Professor and supervisor of Ph.D and graduate students. He is awarded with the Leading-talent Plan of Hainan Province in 2020, Fujian hundred-talent plan of Fujian Province in 2018, and Mingjian Scholar Chair Professor in 2015. His research interests include wireless networks, wireless location, and array signal processing. He was an IEEE Trans on Communications exemplary reviewer for 2020. Now, he is an editor for the journals IEEE Wireless Communications Letters and IEEE Systems Journal. He has published more than 300 in archival journals with more than 120 papers on IEEE Journals and 170 SCI-indexed papers. His citations are 4020. He holds sixteen Chinese patents and also are PI or CoPI for six national projects.



Shi Jin received the B.S. degree in communications engineering from the Guilin University of Electronic Technology, Guilin, China, in 1996, the M.S. degree from the Nanjing University of Posts and Telecommunications, Nanjing, China, in 2003, and the Ph.D. degree in information and communications engineering from Southeast University, Nanjing, in 2007. From June 2007 to October 2009, he was a Research Fellow with the Adastral Park Research Campus, University College London, London, U.K. He is currently with the Faculty of the National

Mobile Communications Research Laboratory, Southeast University. His research interests include space time wireless communications, random matrix theory, and information theory. He and his coauthors have been awarded the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communication theory and the 2010 Young Author Best Paper Award by the IEEE Signal Processing Society. He serves as an Associate Editor for the IEEE Transactions on Communications, IEEE Transactions on Wireless Communications, the IEEE Communications Letters, and IET Communications.



Wen Chen (M'03–SM'11) is a tenured Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, China, where he is the director of Broadband Access Network Laboratory. He is a fellow of Chinese Institute of Electronics and the distinguished lecturers of IEEE Communications Society and IEEE Vehicular Technology Society. He is the Shanghai Chapter Chair of IEEE Vehicular Technology Society, an Editors of IEEE Transactions on Wireless Communications, IEEE Transactions on Communications, IEEE Access and IEEE Open Journal of Vehicular Technology. His research interests include multiple access, wireless AI and meta-surface communications. He has published more than 110 papers in IEEE journals and more than 120 papers in IEEE Conferences, with citations more than 7000 in google scholar.