CHINA COMMUNICATIONS

Joint Active and Passive Beamforming Design in Intelligent Reflecting Surface (IRS)-assisted Covert Communications: A Multi-Agent DRL Approach

Gao Ang[1,*], Ren Xiaoyu[1], Deng Bin[2], Sun Xinshun[1], Zhang Jiankang[3]

[1] School of Electronic Information, Northwestern Polytechnical University, Xi'an, 710072, China
[2] Key Laboratory of Near Ground Detection and Perception Technology, Wuxi, 214035, China
[3] Department of Computing and Informatics, Bournemouth University, UK
[*] The corresponding author, email: gaoang@nwpu.edu.cn

Cite as: A. Gao, X. Ren, *et al*., "Joint active and passive beamforming design in intelligent reflecting surface (irs)-assisted covert communications: A multi-agent drl approach," *China Communications*, 2024, vol. 0, no. 0, pp. 1-16. **DOI**: 10.23919/JCC.YYYY.MM.NNN

Abstract: Intelligent Reflecting Surface (IRS), with the potential capability to reconstruct the electromagnetic propagation environment, evolves a new IRS-assisted covert communications paradigm to eliminate the negligible detection of malicious eavesdroppers by coherently beaming the scattered signals and suppressing the signals leakage. However, when multiple IRSs are involved, accurate channel estimation is still a challenge due to the extra hardware complexity and communication overhead. Besides the cross-interference caused by massive reflecting paths, it is hard to obtain the close-formed solution for the optimization of covert commutations. On this basis, the paper improves a heterogeneous multi-agent deep deterministic policy gradient (MADDPG) approach for the joint active and passive beamforming (Joint A&P BF) optimization without the channel estimation, where the base station (BS) and multiple IRSs are taken as different types of agents and learn to enhance the covert spectrum efficiency (CSE) cooperatively. Thanks to the 'centralized training and distributed execution' feature of MADDPG, each agent can execute the active or passive beamforming independently based on its partial observation without referring to others. Numeral results demonstrate that the proposed deep reinforcement learning (DRL) approach could not only obtain a preferable CSE of legitimate users and a low detection of probability (LPD) of warden, but also alleviate the communication overhead and simplify the IRSs deployment.

Keywords: Covert Communications; Intelligent Reflecting Surface; Deep Reinforcement Learning

Received:
Revised:
Editor:

I. INTRODUCTION

With the rapid growth of wireless communications, protecting the privacy of legitimate users becomes progressively challenging. To conquer the extra communication overhead incurred by the encryption in higher layers, secrecy communications at the physical layer have drawn significant research attention to achieve a positive secrecy rate thus the information can be conveyed confidentially [1]. However, in some specific military or financial scenarios, the communication entity further seeks to shield the existence of itself. Therefore, covert communications [2, 3], also referred to as 'low probability of detection (LPD) communications' or 'undetectable communications', is proposed as a methodology to shelter the presence of transmis-

sions from vigilant adversaries while guaranteeing a certain covert rate of legitimate users. Different from secrecy communications which tend to protect legitimate transmissions from being wiretapped by malicious eavesdroppers, covert communications focus on a negligible detection probability of observant adversaries.

Recently, intelligent reflecting surface (IRS), composed by abundant of low-cost passive reflecting metasurfaces, has been envisaged as a promising technology for reconfiguring the wireless propagation environment by passive beamforming (PBF) [4].

With the electromagnetic reconstruct capability of IRS, reflected and non-reflected signals can be added coherently at the covert users (CUs) while destructively suppressing the signaling leakage below the noise (steer a mull) at the warden, which is practically appealing in improving the performance of covert communications [3, 5–8].

In specific, the research in [3] insights into the fundamental of how an IRS can be integrated to benefit covert communications, and evaluates the impact of the transmission power of base station (BS) and the elements number of IRS on the system performance. The research in [5] aims to maximize the covert rate at both uplink and downlink of IRS-assisted non-orthogonal multiple access (NOMA) systems by the transmit power of BS and passive beamforming optimization of IRS. The research in [8] investigates the multi-input multi-output (MIMO) covert communications aided by IRS against a multi-antenna warden. It is worth noticing that the researches above assume that the instantaneous channel state information (CSI) including the transmitter-IRS and IRS-receiver channels can be estimated at warden according to the pre-known pilot signals, which is crucial to fully unleash the various performance gain brought by IRS with covertness constraints [9]. However, obtaining the instantaneous CSI is not always possible for the following reasons:

- IRS generally does not have to transmit radio frequency (RF) chain (receiving RF chain is still necessary) to send pilot signals actively for channel estimation.
- Although researches in [10, 11] suppose that IRS only has a few active elements equipped transmit RF chain to estimate channels by actively sending pilot signals to simply the procedure, such pilot-based channel estimation and CSI sharing scheme yields huge communication overhead.
- A more crucial issue is that to hide the existence, warden will not collude with either the BS or IRSs, which makes it impossible to obtain the instantaneous CSI.

With this in mind, in the research [12], the passive beamformers at IRSs are optimized through the statistical CSI, and the transmit beamformers at BSs are based on the instantaneous CSI of effective channels. Furthermore, researchers in [6, 13] take the joint power control and phase shift design by exploiting statistical CSI. The research in [7] also illustrates that even when only the statistical CSI of warden is available, the IRS-assisted system can significantly outperform the system without an IRS in the context of covert communications. Despite of tremendous achievements of existing researches, only the simple scenario is considered yet without the cooperation of multi-IRS and the active beamforming of BS. When multiple distributed IRSs are involved in the system, convex and other alternating optimization techniques may be no longer feasible, because the variables are coupled with each other and the optimization problem is hard to be decoupled into the form of linear-program (LP) or quadratically constrained quadratic program (QCQP).

Recently, machine learning (ML) has been widely used for the optimization in IRS-assisted communications for their powerful non-linear approximation capability, which makes it easier to solve the model-free and complex problems by training neural networks, especially in uncertain dynamic scenarios [14–18].

Inspired by this, the paper takes a more common scenario into consideration that the propagation environment is jointly shaped by multiple IRSs with the existence of multiple CUs. Besides, facing the challenge of obtaining instantaneous CSI covertly without using the traditional pilot-based channel estimation methods, the machine learning based approach which allows to estimate channels without explicit feedback/detection is worth exploring to devise feasible solutions [3, 7]. The main contributions of the paper can be summarized as follows:

- A challenging scenario is considered for IRS-assisted covert communications in that the ac-

Table 1. *Comparison with existing literature.*

Reference	Objective	Optimization Variables	Constrain	User NUM.	CSI	Methods
[3]	Covert SE	Passive beamforming	Covert constrain †	Single	Perfect CSI	Monte Carlo method
[5]	Downlink/Uplink covert SE	Transmission Power & Passive beamforming	Covert constrain † & QoS rate	NOMA pair	Perfect CSI	AO & Convex
[6]	Covert outage	Passive beamforming ‡	Covert constrain †	Single	Statistical CSI	Convex
[7]	Covert SNR	Transmission power & Passive beamforming ‡	Covert constrain †	Single	Statistical CSI	Convex approximation
[8]	Covert rate	Active * and Passive beamforming	Covert constrain †	Singe	Perfect CSI	AO & Convex ⋆
Our work	Covert SE	Active * and Passive beamforming	Covert constrain †	Multi-IRS Multi-CU	Statistical CSI	MADRL

†The summation of false alarm probability and miss-detection probability at warden.

‡Both the reflect phase shift and amplitude are jointly justified.

∗MIMO beamfroming for active transmission.

⋆SDR alternative solving for active and passive beamforming, and a low-complexity KKT solution is futher proposed.

tive MIMO beamforming of BS and the passive reflecting beamforming of IRSs are jointly optimized by the collaboration of multi-IRS to enhance the covert spectrum efficiency (CSE) while suppressing the signal leakage at warden.

- To maximize CSE by the joint active and passive beamforming (Joint A&P BF) optimization which is a typical non-linear non-convex programming problem, a heterogeneous multi-agent deep deterministic policy gradient (MADDPG) approach is adopted where BS and IRSs are taken as different types of agents. Due to the 'centralized training and distributed execution' feature of MADDPG, each agent is driven to learn to map its partial observation to a proper action independently, which greatly reduces the communication overhead caused by the information sharing.
- Each agent can learn from its historical action by simple scalar reward feedback from CUs, which avoids the complex CSI estimation and simplifies the hardware design. Numeral results demonstrate that the proposed deep reinforcement learning (DRL) approach not only obtains a preferable CSE with low detection probability at warden, but also helps to facilitate distributive infrastructures in IRS-assisted communication networks.

The differences between our work and the existing literature are summarized in Table 1. The rest of the paper is organized as follows. Sec. II presents the system model for IRS-assisted covert communications. The proposed multi-agent DRL (MADRL) approach for the joint active and passive beamforming optimization is detailed in Sec. III. The algorithm complexity is described in Sec. IV. The numerical results are demonstrated in Sec. V, and the paper is concluded in Sec. VI.

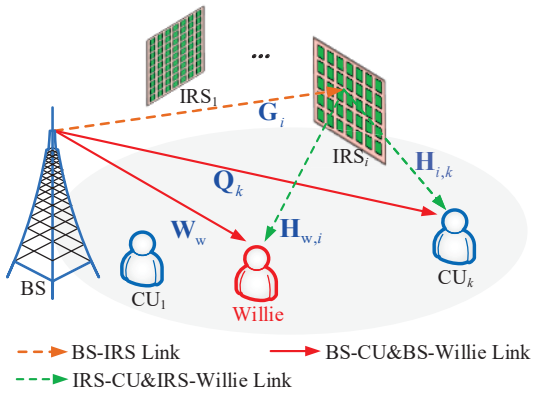


Figure 1. *System model for IRS-assisted covert communications.*

Notation: $\mathbb{C}^{x\times y}$ denotes the space of $x \times y$ complex-valued matrix. $\|\cdot\|$ denotes the Euclidean norm of a complex vector. $\mathbf{G}^{\mathrm{T}}, \mathbf{G}^{\mathrm{H}}$ denote the transpose and conjugate transpose of $\mathbf{G}$, respectively. $\mathbb{E}\{\cdot\}$ denotes the statistical expectation. $\otimes$ denotes the Kronecker product.

II. SYSTEM MODEL

As shown in Figure 1, an IRS-assisted covert communications system that there are I numbers of IRSs mounted on the exterior walls of buildings to assist the covert communications between BS and K CUs in the downlink which may be detected by a warden denoted as 'Willie' is considered. Assume that the BS is equipped with an uniform linear array (ULA) with N antennas, and each CU is equipped with a single omni-antenna. Each IRS has an uniform planar array (UPA) with M_{h} and M_{v} in horizontally and vertically

respectively, i.e. $M = M_{\mathrm{h}} \times M_{\mathrm{v}}$ elements in total.

The reflecting matrix of the i^{th} IRS can be denoted by $\Theta_i = \mathrm{diag}[\beta_{i,1}e^{j\theta_{i,1}}, \beta_{i,2}e^{j\theta_{i.2}} \cdots \beta_{i,M}e^{j\theta_{i,M}}] \in \mathbb{C}^{M\times M}$, where $\theta_{i,m} \in [0, 2\pi)$ is the phase shift and $\beta_{i,m} \in [0, 1]$ is the reflecting coefficient. In practice, it is costly to implement independent controllers for the amplitude and phase shift simultaneously [19]. The phase shift is usually selected from a finite number of discrete values from 0 to 2π for the ease of circuit implement [20]. Assume that all channels experience the quasi-static flat-fading and IRSs can be real-time reconfigured [1, 20]. The equivalent channels of BS-i^{th} IRS, i^{th} IRS-k^{th} CU and BS-k^{th} CU are denoted by $\mathbf{G}_i \in \mathbb{C}^{M\times N}$, $\mathbf{H}_{i,k} \in \mathbb{C}^{M\times 1}$ and $\mathbf{Q}_k \in \mathbb{C}^{1\times N}$, respectively. Let $\mathbf{s}_k \in \mathbb{C}^{1\times N}$ be the transmitting data to the k^{th} CU, which is an identically distributed (i.i.d.) random variable with zero mean and unit variance. Then the received signals at the k^{th} CU can be expressed as:

$$y_k = \underbrace{\left(\sum_{i=1}^{I} \mathbf{H}_{i,k}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{Q}_k\right)\mathbf{p}_k\mathbf{s}_k +}_{\text{covert signals}} \underbrace{\left(\sum_{i=1}^{I} \mathbf{H}_{i,k}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{Q}_k\right)\sum_{\substack{j=1\\ j\neq k}}^{K}\mathbf{p}_j\mathbf{s}_j}_{\text{interference from other CUs}} + \omega_k\ , \quad (1)$$

where $\mathbf{p}_k \in \mathbb{C}^{N\times 1}$ is the beamforming vector of BS used to transmit the original signals $\mathbf{s}_k$, and $\omega_k \sim \mathcal{CN}(0, \delta_\omega^2)$ is the noise following circularly symmetric complex Gaussian (CSCG) distribution.

The total transmission power of BS should be limited by the maximum value $P^{\max}$, which leads to the constraint:

$$\sum_{k=1}^{K} \| \mathbf{p}_k \|^2 \leq P^{\max}\ . \quad \text{(C1)}$$

For the k^{th} CU, signals from others are taken as the interference [21]. So the signal to interference plus noise ratio (SINR) at the k^{th} CU is:

$$\mathrm{SINR}_k = \frac{|\left(\sum_{i=1}^{I} \mathbf{H}_{i,k}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{Q}_k\right)\mathbf{p}_k|^2}{\sum_{\substack{j=1\\ j\neq k}}^{K}|\left(\sum_{i=1}^{I} \mathbf{H}_{i,k}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{Q}_k\right)\mathbf{p}_j|^2 + \delta_\omega^2}\ . \quad (2)$$

2.1 Binary Hypothesis Testing

In order to detect the possible covert communications, Willie is required to distinguish the following two hypotheses:

- The null hypothesis $\mathcal{H}_0$ indicating that there is no transmission of CUs.
- The alternative hypothesis $\mathcal{H}_1$ indicating that there is an ongoing covert transmission from BS to one specific CU.

In general, Willie takes power detection to determinate the existence of covert communications. In specific, a radiometer is equipped as a detector and takes an infinite number of signal samples for such binary detection, which implies that the uncertainties of additive white gaussian noises (AWGNs) can be partially eliminated [22].

Suppose BS-Willie and i^{th} IRS-Willie channels be $\mathbf{W}_{\mathrm{w}} \in \mathbb{C}^{N\times 1}$ and $\mathbf{H}_{\mathrm{w},i} \in \mathbb{C}^{M\times 1}$, respectively. According to Eq. (1), the received power at Willie with respect to the k^{th} CU would be:

$$\mathbb{T}_k = \begin{cases} \left|\left(\sum_{i=1}^{I} \mathbf{H}_{\mathrm{w},i}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{W}_{\mathrm{w}}\right)\sum_{\substack{j=1,\\ j\neq k}}^{K}\mathbf{p}_j\right|^2 + \delta_\omega^2\ , & \mathcal{H}_0\ . \\ \left|\left(\sum_{i=1}^{I} \mathbf{H}_{\mathrm{w},i}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{W}_{\mathrm{w}}\right)\sum_{\substack{j=1,\\ j\neq k}}^{K}\mathbf{p}_j\right|^2 & \\ \quad + \left|\left(\sum_{i=1}^{I} \mathbf{H}_{\mathrm{w},i}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{W}_{\mathrm{w}}\right)\mathbf{p}_k\right|^2 + \delta_\omega^2\ , & \mathcal{H}_1\ . \end{cases} \quad (3)$$

Define $\mathcal{D}_1$ and $\mathcal{D}_0$ be the binary decisions in favor of $\mathcal{H}_1$ and $\mathcal{H}_0$, respectively. Then the decision strategy embedded in the detector of Willie for the k^{th} CU can be expressed by:

$$\mathbb{T}_k \underset{\mathcal{D}_1}{\overset{\mathcal{D}_0}{\gtrless}} \tau\ , \quad (4)$$

where τ is the detection threshold for $\mathbb{T}_k$.

The detection performance of Willie can be further normalized by the detect error probability, which is defined as [6]:

$$\xi = \mathbb{P}_{\mathrm{FA}} + \mathbb{P}_{\mathrm{MD}}\ , \quad (5)$$

where $\mathbb{P}_{\mathrm{FA}} \triangleq \mathcal{P}\{\mathcal{D}_1|\mathcal{H}_0\}$ and $\mathbb{P}_{\mathrm{MD}} \triangleq \mathcal{P}\{\mathcal{D}_0|\mathcal{H}_1\}$ are defined as the false alarm probability and the miss detection probability, respectively.

Willie is expected to optimize the detection threshold τ to achieve the minimum value of ξ denoted as

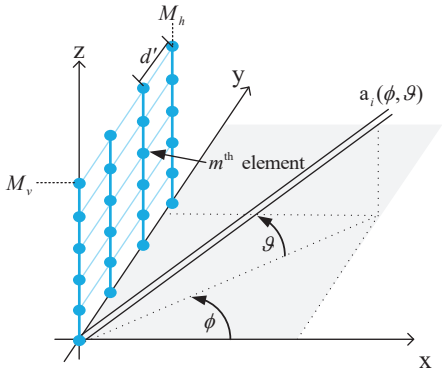


Figure 2. *Definition of azimuth and elevation angles of IRS.*

ξ^*. Thus, the covertness constraint can be written as:

$$\xi^* \geq 1-\epsilon,\ \epsilon \in [0,1]\ , \tag{C2}$$

where ϵ is a small value which implies the receiver operation characteristic (ROC) curve lies very close to the line of no-discrimination [23]. Note that to ensure LPD for a given covert communication, the key idea is to artificially create the phase-shift Θ_i to induce complex-valued random variable (RV) uncertainties at Willie. As a result, Willie can not tell the change of received power either due to the transmission of a specific CU or the interference from others.

2.2 Channel Model

In general, the position of BS and IRSs are pre-fixed, and the mobile position of the k^{th} CU in 3D can be denoted by $C_k = [c_{xk}, c_{yk}, c_{zk}]$. Define the Euclidean distance of BS-i^{th} IRS, i^{th} IRS-k^{th} CU and BS-k^{th} CU be $d_{0,i}$, $d_{i,k}$ and $d_{0,k}$, respectively. Then the corresponding path loss can be calculated as $\upsilon_{0,i} = \sqrt{\rho_0 d_{0,i}^{-\alpha_{\text{BI}}}}$, $\upsilon_{i,k} = \sqrt{\rho_0 d_{i,k}^{-\alpha_{\text{IU}}}}$ and $\upsilon_{0,k} = \sqrt{\rho_0 d_{0,k}^{-\alpha_{\text{BU}}}}$, where α_{BI}, α_{IU} and α_{BU} are the corresponding path loss exponents, and ρ_0 is the path loss at 1 meter reference distance.

As shown in Figure 2, one IRS is deployed in the $y-z$ plane, and the array steering vector at the direction of (ϕ, ϑ) is denoted by $\mathbf{a}(\phi, \vartheta)$, where $\phi_{0,k}$ and $\vartheta_{0,k}$ are the azimuth and elevation angles to the normal vector of UPA itself. For brevity, the azimuth and elevation angles are defined in the local coordinate system of each IRS or BS.

According to the 3D Saleh-Valenzuela channel model, the line-of-sight (LoS) component of BS-i^{th} IRS, i^{th} IRS-k^{th} CU and BS-k^{th} CU link can be expressed as:

$$\mathbf{G}_i^{\text{LoS}} = \upsilon_{0,i}\mathbf{a}_{0,i}(\phi_{0,i}^{\text{AoD}})\mathbf{a}_i^{\text{H}}(\phi_{0,i}^{\text{AoA}}, \vartheta_{0,i}^{\text{AoA}})\ , \tag{6}$$

$$\mathbf{H}_{i,k}^{\text{LoS}} = \upsilon_{i,k}\mathbf{a}_i(\phi_{i,k}^{\text{AoD}}, \vartheta_{i,k}^{\text{AoD}})\ , \tag{7}$$

$$\mathbf{Q}_k^{\text{LoS}} = \upsilon_{0,k}\mathbf{a}_{0,k}(\phi_{0,k}^{\text{AoA}})\ , \tag{8}$$

where $\mathbf{a}_{0,i} = \left[1, e^{j\frac{2\pi d'}{\lambda}\sin\phi_{0,i}^{\text{AoD}}}, \cdots, e^{j\frac{2\pi d'}{\lambda}(N-1)\sin\phi_{0,i}^{\text{AoD}}}\right]$ is the steering vector of BS to the direction of the i^{th} IRS, $\phi_{0,i}^{\text{AoD}}$ is the azimuth angle-of-departure (AoD) of BS to the i^{th} IRS. There are similarly definitions for $\mathbf{a}_{0,k}$ and $\phi_{0,k}^{\text{AoD}}$. Besides, λ is the carrier wavelength and d' is the antenna space at BS.

Furthermore, $\phi_{0,i}^{\text{AoA}}$ and $\vartheta_{0,i}^{\text{AoA}}$ in Eq. (6) are the azimuth and elevation angle-of-arrival (AoA) from BS to the i^{th} IRS, and $\mathbf{a}_i(\phi_{0,i}^{\text{AoA}}, \vartheta_{0,i}^{\text{AoA}}) = \mathbf{a}_i^{\text{v}}(\phi_{0,i}^{\text{AoA}}, \vartheta_{0,i}^{\text{AoA}}) \otimes \mathbf{a}_i^{\text{h}}(\phi_{0,i}^{\text{AoA}}, \vartheta_{0,i}^{\text{AoA}})$ is the steering vector of the i^{th} IRS, where $\mathbf{a}_i^{\text{v}} = [a_{i,m}^{\text{v}}] \in \mathbb{C}^{M_{\text{v}}\times 1}$ and $\mathbf{a}_i^{\text{h}} = [a_{i,m}^{\text{h}}] \in \mathbb{C}^{M_{\text{h}}\times 1}$ are the steering vector in vertical (z direction) and horizontal (y direction). In Eq. (7), $\mathbf{a}_i(\phi_{i,k}^{\text{AoD}}, \vartheta_{i,k}^{\text{AoD}})$ is defined similarly.

$$a_{i,m}^{\text{v}} = e^{j\frac{2\pi d'}{\lambda}(m-1)\sin\vartheta_{0,i}^{\text{AoA}}}\ ,\ \ m = \{1, \cdots, M_{\text{v}}\}\ . \tag{9}$$

$$a_{i,m}^{\text{h}} = e^{j\frac{2\pi d'}{\lambda}(m-1)\cos\vartheta_{0,i}^{\text{AoA}}\sin\vartheta_{0,i}^{\text{AoA}}}\ ,\ \ m = \{1, \cdots, M_{\text{h}}\}\ . \tag{10}$$

In general, mmWave channels consist of only a small number of dominant paths with a LoS component [19]. Due to the severe path loss, the transmit power of twice or more reflections can be ignored. In general, since IRSs are densely distributed in the hotspot space, only LoS model is reasonable [20, 24]. Therefore, BS-i^{th} IRS, i^{th} IRS-k^{th} CU and BS-k^{th} CU channels are modeled by Rician fading:

$$\mathbf{G}_i = \sqrt{\tfrac{K_1}{1+K_1}}\mathbf{G}_i^{\text{LoS}} + \sqrt{\tfrac{1}{1+K_1}}\mathbf{G}_i^{\text{NLoS}}\ , \tag{11}$$

$$\mathbf{H}_{i,k} = \sqrt{\tfrac{K_2}{1+K_2}}\mathbf{H}_{i,k}^{\text{LoS}} + \sqrt{\tfrac{1}{1+K_2}}\mathbf{H}_{i,k}^{\text{NLoS}}\ , \tag{12}$$

$$\mathbf{Q}_k = \sqrt{\tfrac{K_3}{1+K_3}}\mathbf{Q}_k^{\text{LoS}} + \sqrt{\tfrac{1}{1+K_3}}\mathbf{Q}_k^{\text{NLoS}}\ , \tag{13}$$

where K_1, K_2 and K_3 are Rician factors, and each non-line-of-sight (NLoS) element in $\mathbf{G}_i^{\text{NLoS}} \in \mathbb{C}^{M\times N}$, $\mathbf{H}_{i,k}^{\text{NLoS}} \in \mathbb{C}^{M\times 1}$ and $\mathbf{Q}_k^{\text{NLoS}} \in \mathbb{C}^{1\times N}$ is independent and identically distributed (i.i.d.) with a zero-mean

and unit variance.

Similarly, there are:

$$\mathbf{H}_{\mathrm{w},i} = \sqrt{\frac{K_4}{1+K_4}}\mathbf{H}_{\mathrm{w},i}^{\mathrm{LoS}} + \sqrt{\frac{1}{1+K_4}}\mathbf{H}_{\mathrm{w},i}^{\mathrm{NLoS}}\,, \quad (14)$$

$$\mathbf{W}_{\mathrm{w}} = \sqrt{\frac{K_5}{1+K_5}}\mathbf{W}_{\mathrm{w}}^{\mathrm{LoS}} + \sqrt{\frac{1}{1+K_5}}\mathbf{W}_{\mathrm{w}}^{\mathrm{NLoS}}\,. \quad (15)$$

2.3 Problem Formulation

The problem can be formulated to maximize CSE at downlink by the joint optimization of active beamforming of BS and passive beamforming of IRSs with subject to the maximum transmission power of BS and the detection error probability of Willie.

$$\mathcal{P}1: \max_{\mathbf{p}_k,\Theta_i} \min_{k\in\mathcal{K}} \ \log(1+\mathrm{SINR}_k)\,, \quad (16)$$

$$\text{s.t.} \ \sum\nolimits_{k=1}^{K} \parallel \mathbf{p}_k \parallel^2 \leq P^{\max}\,, \quad (C1)$$

$$\xi^* \geq 1-\epsilon,\ \epsilon\in[0,1]\,, \quad (C2)$$

$$\theta_{i,m} \in [0,2\pi)\,. \quad (C3)$$

Since the detection threshold τ has the effect on both the false alarm and miss detection probability, the optimal value to minimize ξ should be the likelihood ratio test to make them equal, which is commonly adopted in the existing researches of covert communications [23, 25]. However, the resultant expression for ξ^* involves incomplete gamma functions, which is not tractable for subsequent analysis and design. To overcome the difficulty, according to Pinsker's inequality, a lower bound of ξ^* is given as [25]:

$$\xi^* \geq 1-\sqrt{\frac{1}{2}\mathcal{D}(\mathbb{P}_{\mathrm{FA}}|\mathbb{P}_{\mathrm{MD}})}\,, \quad (17)$$

where $\mathcal{D}(\mathbb{P}_{\mathrm{FA}}|\mathbb{P}_{\mathrm{MD}})$ is the Kullback-Leibler (KL) divergence (also known as 'relative entropy') from $\mathbb{P}_{\mathrm{FA}}$ to $\mathbb{P}_{\mathrm{MD}}$:

$$\mathcal{D}(\mathbb{P}_{\mathrm{FA}}|\mathbb{P}_{\mathrm{MD}}) = \left[\ln\left(\frac{A+B}{A}\right) - \frac{B}{A} + B\right]\,, \quad (18)$$

where $A = |\left(\sum_{i=1}^{I}\mathbf{H}_{\mathrm{w},i}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{W}_{\mathrm{w}}\right)\sum_{j=1,j\neq k}^{K}\mathbf{p}_j|^2 + \delta_\omega^2$ and $B = |\left(\sum_{i=1}^{I}\mathbf{H}_{\mathrm{w},i}^{\mathrm{H}}\Theta_i\mathbf{G}_i + \mathbf{W}_{\mathrm{w}}\right)\mathbf{p}_k|^2$.

Therefore, with algebraic manipulations of (C2) and Eq. (17), $\mathcal{P}1$ can be rewritten with a more stringent constraint:

$$\mathcal{P}2: \max_{\mathbf{p}_k,\Theta_i} \min_{k\in\mathcal{K}} \ \log(1+\mathrm{SINR}_k)\,,$$

$$\text{s.t.} \ \sum\nolimits_{k=1}^{K} \parallel \mathbf{p}_k \parallel^2 \leq P^{\max}\,, \quad (C1)$$

$$\mathcal{D}(\mathbb{P}_{\mathrm{FA}}|\mathbb{P}_{\mathrm{MD}}) \leq 2\epsilon^2\,, \quad (C2.a)$$

$$\theta_{i,m} \in [0,2\pi)\,. \quad (C3)$$

It should be noticed that the objective function of $\mathcal{P}2$ is still non-convex with respect to Θ_i and $\mathbf{p}_k$ [26], and (C2.a) is also a non-convex constraint. Besides the impossibility of obtaining the instantaneous CSI, it is hard to solve the close-formed solution of $\mathcal{P}2$ by conventional convex methods in polynomial time.

To this end, a heterogeneous MADRL approach is a novelty proposed to solve the Joint A&P BF optimization problem where both BS and IRS agents can be trained with a simple scalar reward feedback from CUs, and learn to seek a strategy to maximize CSE spontaneously. Even without the instantaneous CSI, the model-free nature of DRL can resist the random small-fast fading. Thus that IRSs can be freed from the complex channel estimation and the information sharing with BS can be greatly reduced.

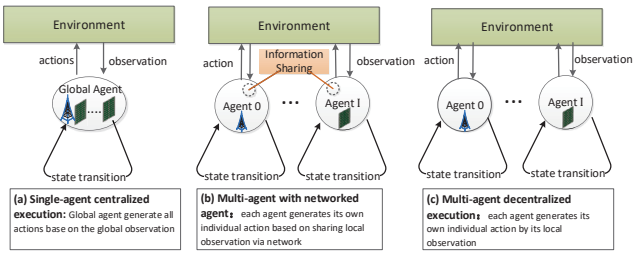


Figure 3. *Different execution schemes of DRL approaches.*

III. MULTI-AGENT DRL BASED OPTIMIZATION OF TRANSMIT BEAMFORMING AND PHASE SHIFTS

3.1 Motivation of MADRL

As shown in Figure 3, there are three different execution schemes of DRL approaches [27].

- Figure 3 (a) known as the single-agent centralized execution, takes the whole system as a single global agent and embeds a central controller to aggregate all the related information for training. Although such centralized DRL scheme

has achieved great success in computation vision [28], it may not be practicable for distributed communication systems.

- Although there are multiple agents in the scheme of Figure 3 (b), all the agents still need to share their local information spreading across the network, which will lead to a great amount of communication overhead. So it belongs to a semi-decentralized approach.
- The proposed learning approach for the Joint A&P BF optimization takes the fully-decentralized scheme shown as Figure 3 (c). Once well trained, each agent can work individually by its local observation without any extra information exchanging with each other.

It should be further noticed that there are two types of agents with tailored action and state space in the proposed heterogeneous MADRL approach, i.e., BS agent and IRS agent. Each agent is trained to learn the cooperation with others by a simple SINR reward feedback from CUs. By introducing the distributive architecture and heterogeneous agents to MADRL, the complex channel estimation can be avoided and the huge compunction overhead can be reduced.

3.2 Heterogeneous MADRL for Joint A&P BF

As shown in Figure 4, there are $I+1$ agents (i.e., I IRSs and one BS) involved in the system. Let a_t^i, s_t^i, r_t^i be the action, state, and reward of the i^{th} agent at the t^{th} slot, respectively, where $i \in \mathcal{I}^+$ and $\mathcal{I}^+ = 0 \cup \mathcal{I}$ ($i = 0$ represents BS). To interact with the environment, each agent incrementally updates the reflecting matrix Θ_i or the beamforming vector $\mathbf{p}_k$ to pursue a better reward. For notion brevity, the paper takes superscript and subscript to denote the agent number and the training episode, respectively.

- *Action*: When $i = \{1, \cdots, I\}$, action $a_t^i = \Theta_i[t]$ is defined as the reflecting matrix of each IRS agent. While when $i = 0$, action $a_t^0 = \mathbf{P}[t] = \{\mathbf{p_1}, \cdots, \mathbf{p}_K\}$ is the active beamforming vector of BS agent. Action can be either continuous or discrete, which depends on the quantization of elements at IRS and RF chains at BS.
- *State*: State $s_t^i = \{C_k[t], C_{\text{w}}\}(k \in \mathcal{K})$ is the position of all CUs and Willie. State for all agents including IRSs and BS is the same.

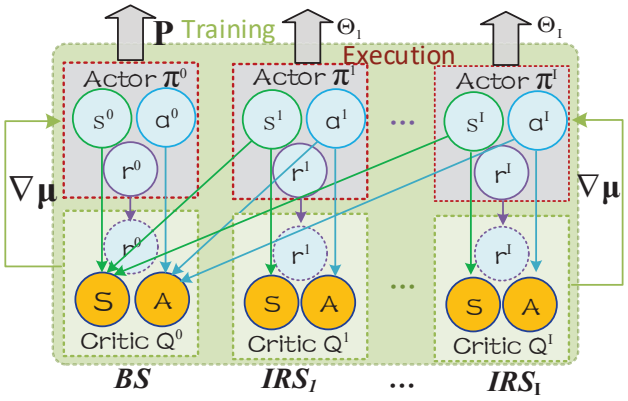


Figure 4. *Learning architecture of heterogeneous MADDPG.*

- *Reward*: Since it is difficult for $\mathcal{P}2$ to distinguish the contributions of each agent, the reward for all agents is uniformly defined as:

$$r_t^i = \min_{k \in \mathcal{K}} \text{SINR}_k + \eta(2\epsilon^2 - \mathcal{D}(\mathbb{P}_{\text{FA}} | \mathbb{P}_{\text{MD}})) \,, \quad (19)$$

where η is the penalty factor that squeezes the network to avoid violating constraint (C2.a). Since constraints (C1) and (C3) are related to action, they are easy to be satisfied by discarding the violated action during training.

By leveraging the non-linear approximation capability of neural networks, BS and each IRS in MADRL can be treated as a single agent and learn to map its observation into a proper action by training network parameters. As a result, they can make the immediate passive and active beamforming without perfect CSI.

3.3 Learning of MADDPG

By extending deep deterministic policy gradient (DDPG) into the multi-agent domain, MADDPG not only eliminates the drawback of traditional Q-learning or policy gradient methods that are inadequate for multi-agent environments, but also reserves the great advantage of DDPG that the action and state space can be continuous rather than discrete. Thanks to the feature of 'centralized training and decentralized execution', each BS or IRS in MADDPG is trained as an individual agent and learned by interacting with the environment.

The global state and global action are defined as the concatenation of each single agent, i.e. $\mathcal{A}_t = [a_t^0 :$

$a_t^i], \mathcal{S}_t = [s_t^i]$ where $i \in \mathcal{I}^+$ and $\mathcal{A}_t, \mathcal{S}_t \in \mathbb{C}^{N\times K+M\times I}$. Reward is a simple scalar defined in Eq. (19). Each agent in MADDPG has two networks, i.e., an actor network and a critic network.

- Actor i, a policy-network denoted by π^i, takes the current state s_t^i as the input with parameters μ^i and outputs the action $a_t^i = \pi(s_t^i|\mu^i)$. It is updated by $\mu_{t+1}^i = \mu_t^i + \alpha_\mu \nabla_\mu J(\mu)$, where α_μ is the step size for parameters updating. Each agent aims to learn a policy to maximize the cumulative historical reward, i.e., $J_i(\mu) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t^i]$, where γ is the discount factor.
- Critic i, a value-network denoted by Q^i, outputs the policy gradient $\nabla_\mu J(\mu^i)$ for actor π^i to evaluate the fitness of action a_t^i by a comprehensive consideration of the global state and action. The network parameters ρ^i are updated by minimizing the temporal difference (TD) error with Q-value of each global action-state pair $Q^i(\mathcal{S}_t, \mathcal{A}_t|\rho^i)$, where $\rho_{t+1}^i = \rho_t^i + \alpha_\rho \nabla_\rho Q^i$, and α_ρ is the step size for parameters updating.

During training, the deep coupling among agents (BS and multi-IRS have to work together to enhance CSE at downlink) can also be investigated by the network. Each agent collects the action and state of other agents to construct the global pair $(\mathcal{S}_t, \mathcal{A}_t)$, and takes SINR_k measured at CU receivers to calculate the reward according to Eq. (19).

During execution, the well trained agents can output the best action $\mathcal{A}^i$ independently without resorting to others. It is a great advantage to significantly reduce the interaction between BS and IRS controllers.

Both the actor and critic network further have two sub-networks, i.e., an on-line network and a target network for soft updating to overcome the overestimation [29]. Besides, there are $\mathbb{M}$ transitions $\{\mathcal{S}_t, \mathcal{A}_t, r_t^i, \mathcal{S}_{t+1}\}$ randomly selected as the mini-batch for training to avoid highly correlated action for successive updating. The details of heterogeneous MADDPG algorithm for Joint A&P BF optimization are shown in Algorithm 1.

Algorithm 1. *Heterogeneous MADDPG for Joint A&P BF Optimization*

Input: $\mathbf{G}_i$, $\mathbf{H}_{i,k}$, $\mathbf{H}_{\text{w},i}$, $\mathbf{Q}_k$ and $\mathbf{W}_\text{w}$, the positions of CUs, IRSs and Willie

Output: optimal action $a_i = \{\mathbf{p}_k, \Theta_i\}$

Initialization : experience replay memory O, training actor network parameter μ^i, training critic network parameter ρ^i, transmit beamforming matrix $\mathbf{p}_k$, phase matrix Θ_i
for $episode = 1 \rightarrow max - episode$ **do**
  Initialize a random process
  Receive initial state $\mathcal{S}_0$
  for $t = 1 \rightarrow T$ **do**
    For each agent i, select action $\mathcal{A}_t$ w.r.t. the current policy and exploration from the actor network
    Execute actions $\mathcal{A}_t = [a_t^0 : a_t^i]$ and observe reward r_t^i and new state $\mathcal{S}_{t+1}$
    Store $(\mathcal{S}_t, \mathcal{A}_t, r_t^i, \mathcal{S}_{t+1})$ in replay buffer O
    $\mathcal{S}_t \leftarrow \mathcal{S}_{t+1}$
    for agent $i = 0 \rightarrow I$ **do**
      Sample a random minibatch of $\mathbb{M}$ samples from O
      Set $y_i = r_t^i + \gamma Q^{i'}(\mathcal{S}_t, \mathcal{A}_t'|\rho^{i'})$
      Update critic by minimizing the loss
      $\ell_m = \frac{1}{\mathbb{M}}\mathbb{E}\left[(y_i - Q^i(\mathcal{S}_t, \mathcal{A}_t|\rho^m))^2\right]$
      Update actor network parameter by policy gradient:
      $\mu_{t+1}^i = \mu_t^i + \alpha_\mu \nabla_\mu J(\mu)$
      where α_μ is the stride for parameter updating and $\nabla_\mu J(\mu)$ is the gradient of μ^i
      Update target network parameters:
      $\rho_{t+1}^{i'} \leftarrow \tau\rho_t^i + (1-\tau)\rho_t^{i'}$
      $\mu_{t+1}^{i'} \leftarrow \tau\mu_t^m + (1-\tau)\mu_t^{i'}$
      $t = t + 1$
    end for
  end for
end for

3.4 MADDPG Implementation on IRSs

Although passive IRSs do not have an RF chain, it is not contradictory for them to be equipped with the computing capability. The embedded controller of IRS can be wirelessly connected with BS through an independent control channel, which is a general setup of IRS and enough to enforce the neural network training and execution [30, 31].

- The embedded controller of the IRS randomly generates the phase shift as an action. As a consequence, the reward related to the SINR of CUs and the detection probability of warden can be in-

ferred according to Eq. (19).

- The location of the warden can be detected from the local oscillator power which is unintentionally leaked from its radio frequency front end [32]. Therefore, the detection probability of warden can be calculated by Eq. (18), and the SINR of CUs can be feedback to BS through the independent control channel of BS.

With the global action, state and reward, each IRS can be taken as an independent agent to interact with the environment and then update phase shift matrix.

IV. COMPLEXITY ANALYSIS

4.1 Computation Complexity

Since the heterogeneous MADDPG proposed in the paper is distributive, each agent (no matter BS or IRSs) has its own neuron networks which can be trained and executed in parallel. In general, the complexity can be estimated by the neural network configuration listed in Table 2 in terms of the floating-point operations per second (FLOPS).

Let $\ell_{\mathrm{A},j}$ be the number of neurons in the j^{th} layer of actor, and $\ell_{\mathrm{C},l}$ be that in the l^{th} layer of critic, where $j \in \{0, 1, \cdots, J\}$, $l \in \{0, 1, \cdots, L\}$ and J, L are the number of layers for the actor and critic networks, respectively. For a fully connected layer with ℓ_j neurons as the input and ℓ_{j+1} neurons as the output, the dot product of FLOPS from the j^{th} to the $(j+1)^{\text{th}}$ layer should be $(2\ell_j - 1) \times \ell_{i+1}$.

Table 2. *MADDPG neural networks configuration*

Agent	Type	Layer	Neuron Number	Activation Function
BS/ IRSs	Actor	Input	3*(K+E)†	Relu
		Hidden	2 layers with 128	Relu
		Output	NK+IM‡	Sigmod
	Critic	Input	3*(K+E)+NK+IM*	Relu
		Hidden	2 layers with 128	Relu
		Output	1⋆	/

†Actor input is $\mathcal{S}_t = \{C_k[t], C_{\mathrm{w}}\}$ with the dimension of $3^*(K+E)$.

‡Actor output is $\mathcal{A}_t = [a_t^0 : a_t^i]$ with the dimension of $NK + IM$.

∗Critic input is $\{\mathcal{S}_t + \mathcal{A}_t\}$ with the dimension of $3^*(K+E) + NK + IM$.

⋆Critic output is the policy gradient value scalar. So the dimension of IRSs and BS are 1.

The neurons output is passed by a specific activation function. For example, Sigmod function has $\kappa_{\mathrm{sigmod}} = 4$ FLOPS because the function $\delta(z) = 1/(1+e^{-z})$ has four mathematical operations, i.e., division, summation, exponentiation and substation, and each of them needs one FLOPS. Similarity, rectified linear unit (ReLU) function has $\kappa_{\mathrm{ReLU}} = 1$ FLOPS.

Therefore, the time complexity for training is:

$$\begin{aligned} &2\sum_{j=0}^{J-1}((2\ell_{\mathrm{A},j}-1)\ell_{\mathrm{A},j+1} + \kappa_j \ell_{\mathrm{A},j+1}) \\ &+2\sum_{l=0}^{L-1}((2\ell_{\mathrm{C},l}-1)\ell_{\mathrm{C},l+1} + \kappa_l \ell_{\mathrm{C},l+1}) \\ &= \mathcal{O}(\sum_{j=0}^{J-1}\ell_{\mathrm{A},j}\ell_{\mathrm{A},j+1} + \sum_{l=0}^{L-1}\ell_{\mathrm{C},l}\ell_{\mathrm{C},l+1}) \,. \end{aligned}$$

The time complexity is reduced to $\mathcal{O}(\sum_{j=0}^{J-1}\ell_{\mathrm{A},j}\ell_{\mathrm{A},j+1})$ during the distributive execution.

Space is needed to store the learning transition. The memory for one fully connected network is a $(\ell_j \times \ell_{j+1})$ matrix and a ℓ_j bias vector. So the space complexity is:

$$\begin{aligned} &2\sum_{j=0}^{J-1}(\ell_{\mathrm{A},j}+1)\ell_{\mathrm{A},j+1} + \sum_{l=0}^{L-1}(\ell_{\mathrm{C},l}+1)\ell_{\mathrm{C},l+1} + \mathbb{M} \\ &= \mathcal{O}(\sum_{j=0}^{J-1}\ell_{\mathrm{A},j}\ell_{\mathrm{A},j+1} + \sum_{j=0}^{L-1}\ell_{\mathrm{C},l}\ell_{\mathrm{C},l+1} + \mathbb{M}) \,, \end{aligned}$$

where $\mathbb{M}$ is the size of mini-batch.

The space complexion is declined to $\mathcal{O}(\sum_{i=0}^{J-1}\ell_{\mathrm{A},j}\ell_{\mathrm{A},j+1})$ for distributive execution.

4.2 Communication Overhead

As defined in Sec. 3.2, the global state $\mathcal{S} = [s_t^i]$ and the global action $\mathcal{A} = [a_t^0 : a_t^i]$ should be shared among each agent, where $s_t \in \mathbb{C}^{3(K+E)}$, $a_t^0 = \mathbf{P}[t] \in \mathbb{C}^{N\times K}$ and $a_t^i = \Theta_i[t] \in \mathbb{C}^M$. Thus the size of $\mathcal{S}$ and $\mathcal{A}$ should be $3(K+E)$ and $NK + IM$, respectively.

Besides, a scalar reward will be feedback to each agent during training, which leads to the communication overhead be $3(K+E)+(NK+IM)+K$. Similar to the computation complexity, the communication overhead is declined to $3(K+E)$ for both BS agent and IRS agents during execution.

Table 3. *Simulation parameters*

Parameter	Description	Value
I	Number of IRS	2
K	Number of CU	2
N	Antenna number of BS	32
M_{h}	Number of horizontal elements at IRS	5
M_{v}	Number of vertical elements at IRS	5
$M = M_{\mathrm{h}} \times M_{\mathrm{v}}$	Number of an IRS elements	5×5
λ	Carrier wavelength	0.125m
$P^{\max}$	Maximum transmission power of BS	30dBm
ϵ	Covertness constraint	0.001
δ_ω^2	Variance of noise power	-80dBm
$\alpha_{\mathrm{BI}}, \alpha_{\mathrm{IU}}, \alpha_{\mathrm{BU}}$	BS-IRS, IRS-CU, BS-CU link path loss exponential, respectively [33]	2.2, 3, 6
$\alpha_{\mathrm{IW}}, \alpha_{\mathrm{BW}}$	IRS-Willie, BS-Willie link path loss exponential, respectively	3, 6
K_1, K_2, K_3	BS-IRS, IRS-CU and BS-CU link Rician factors, respectively[34]	6, 6, 6
K_4, K_5	IRS-Willie and BS-Willie link Rician factors, respectively	6, 6
ρ_0	Channel gain at reference of 1 meter distance	-30dB
γ	Discount factor	0.95
η	Penalty factor	5
O	Size of Replay Buffer	20000
$\mathbb{M}$	Size of minibatch	256

V. SIMULATION RESULTS

Consider an IRS-assisted covert communications system that there are $I = 2$ IRSs equipped with $M = 5 \times 5$ elements each between BS with $N = 32$ antennas and $K = 2$ CUs detected by a warden. As shown in Figure 6 (a), the position of BS, CUs, as well as Willie is set as (0, 0), $\{(50, 0), (85, 10)\}$ and (60, 20) in meter of a two-dimensional plane, respectively. IRSs are mounted at $\{(70, 30), (100, 30)\}$ with the height of 20m. On the basis, the path loss exponents of BS-IRS, IRS-CU, and BS-CU links denoted as $\alpha_{\mathrm{BI}}, \alpha_{\mathrm{IU}}$ and α_{BU} are set as 2.2, 3 and 6 respectively. The detailed simulation parameters are provided in Table 3.

5.1 Algorithm Convergence

Figure 5 depicts the convergence of different machine learning algorithms in terms of CSE, i.e., Q-learning, deep Q network (DQN), DDPG, multi-agent deep Q network (MADQN) and MADDPG. The weighted minimum mean square error (WMMSE) algorithm [35] which is generally used for utility maximization problems is taken as the benchmark for comparison.

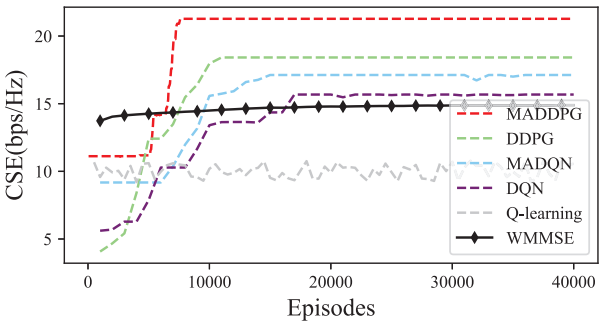


Figure 5. *Convergence of different algorithms.*

- Although the weight matrix is able to transform the original problem to be a convex optimization which can converge in limited iterations, the object function will be monotone decreased by each transformation. As a result, the covert spectrum efficiency finally converges to a local optimum which is far less than the value learned by DRL based methods.
- Q-learning performs poorly in this scenario (actually it works well when there is only one IRS equipped with 3*3 elements.). Since the lack of global observation, channels state is time-varying which makes the environment non-stationary. Along with the system scale, the fast growing size of Q-table makes the training difficult to ergodic all possible statuses, and agents are prevented to use past experience to replay in straight forward [36].
- All the other four DRL based algorithms can be converged along with the training episodes. However, since DQN and MADQN are value-based learning methods like Q-learning, they are defeated by the policy-based DDPG and MADDPG. Without the critic network, DQN and MADQN can only generate the action from their experience in a discrete space. On the contrary, DDPG and MADDPG learn to generate a deterministic action from the distribution of state, which implies that each agent is able to infer a better action even

when one state has not been experienced during training.

- Compared with centralized learning approaches, i.e., DQN and DDPG, the multi-agent learning scheme, i.e., MADQN and MADDPG can further stimulate agents to obtain a better reward due to the fact that each agent can be trained distributively to maximize its own reward rather than the global value.

5.2 Visualization of SINR Map

Table 4. *Algorithm execution time comparison*†

IRSs size	Time	2IRSs,2CUs	3IRSs,3CUs	4IRSs,3CUs
3×3 UPA	training	1359.942s	1510.467s	1544.883s
	execution	0.000256s	0.000346s	0.000367s
5×5 UPA	training	1476.827s	1708.033s	1927.322s
	execution	0.000292s	0.000352s	0.000388s
7×7 UPA	training	1702.583s	2016.760s	2203.481s
	execution	0.000328s	0.000391s	0.000404s
8×8 UPA	training	1768.496s	2082.822s	2208.843s
	execution	0.000354s	0.000576s	0.000631s

†The algorithm training and execution time are calculated on the platform with Intel Core(TM) i5-11400 2.60GHz CPU and 8GB memory.

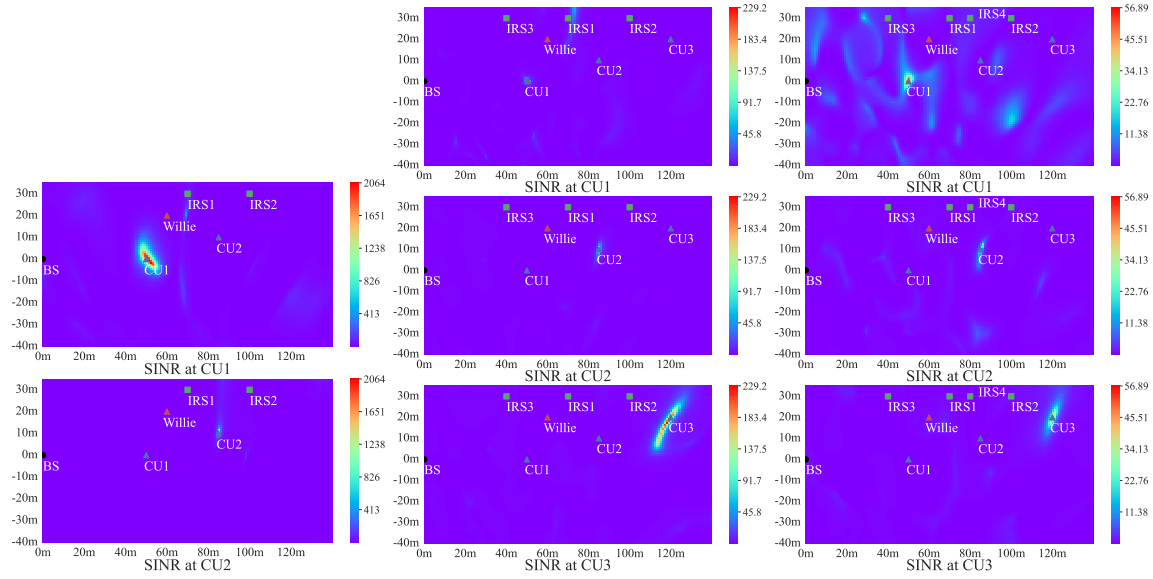
Figure 6 further demonstrates the scalability of MADDPG based Joint A&P BF optimization in more complex scenarios by visualizing the SINR map of each individual CU, i.e., 2 IRSs with 2 CUs, 3 IRSs with 3 CUs and 4 IRSs with 3 CUs in (a)-(c) respectively.

- SINR map provides insight into the quality of received signals based on the position of CUs, by which the level of interference and noise relative to the desired signal strength can be observed. The higher the SINR value, the better the communication quality.
- Obviously, SINR of each CU can only be significantly enhanced in a directed manner at a dedicated position. While the extremely low SINR value indicates that the system is robust against unauthorized access and ensures the confidentiality of legitimate communications. Those phenomenons serve as a strong validation of the effectiveness of the proposed approach in achieving targeted signal enhancement for individual user.
- Table 4 presents the training and execution time of the proposed heterogeneous MADDPG algorithm for the Joint A&P beamforming optimization. Since the global status and action sharing are inevitable to evaluate the fitness of actors which map the local state to a proper action, the training process is time-consuming. However, thanks to the 'centralized training and distributed execution' feature of MADDPG, once well-trained, each agent can work independently and immediately without interacting with others.
- Even though the different scenarios, the system scale has little effect on the effectiveness of the algorithm and the conclusion of all simulations. To save space and avoid too many curves in each figure, we just take the scenario of 2 IRSs with 2 CUs as an example for the following simulations without discussing the scalability anymore.

5.3 Impact of IRSs Size

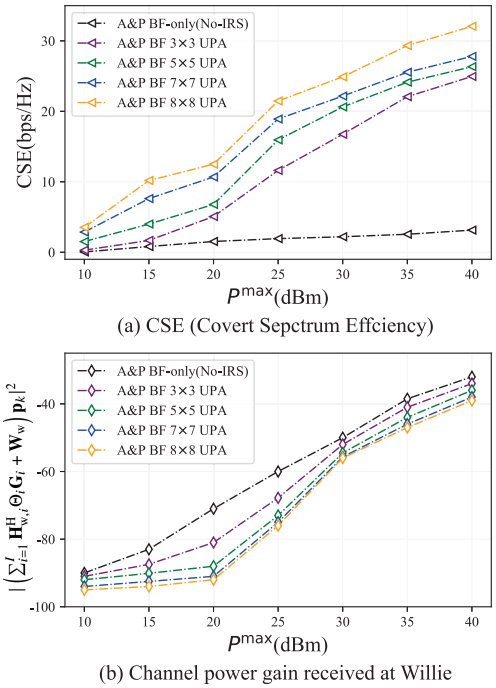
Figure 7 (a) and (b) demonstrate the impact of IRSs size on both the spectrum efficiency of CUs and the channel power gain received at Willie, respectively. The IRSs size is changing at the set of $\{3 \times 3, 5 \times 5, 7 \times 7, 8 \times 8\}$. The A&P BF-only (No-IRS) scenario is taken as the benchmark that there is no IRS working between BS and CUs.

- As the increment of IRSs size, CSE will be gradually enhanced while the channel power gain received at Willie is suppressed simultaneously, which reveals the potential of IRSs in covert communications. In specific, the KL divergence is a monotonically increasing function of $|\left(\sum_{i=1}^{I} \mathbf{H}_{\mathrm{w},i}^{\mathrm{H}} \Theta_i \mathbf{G}_i + \mathbf{W}_{\mathrm{w}}\right) \mathbf{p}_k|^2$, which reduces along with the IRSs size. As a result, CUs are allowed to take a higher transmission power without violating the covertness constraint.
- It is reasonable that CSE also improves when BS enhances the transmission power. Although the larger channel power gain received at Willie limits the transmission power of CUs to satisfy the more rigorous covertness constraint, the strengthened signals quality of BS helps to effectively overcome the signal attenuation and interference during the transmission which finally makes a positive effect on the system performance.



(a) *2 IRSs, 2 CUs* **(b)** *3 IRSs, 3 CUs* **(c)** *4 IRSs, 3 CUs*

Figure 6. *SINR map with more IRSs and CUs in the system.*



(a) CSE (Covert Sepctrum Effciency)

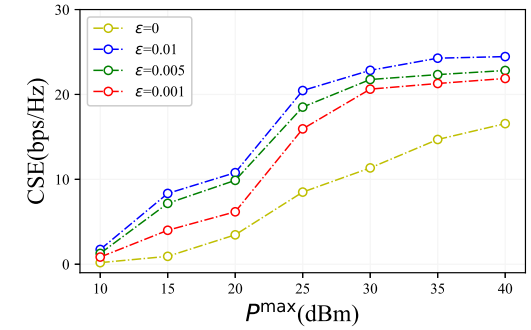
(b) Channel power gain received at Willie

Figure 7. *System performance along with IRS size.*

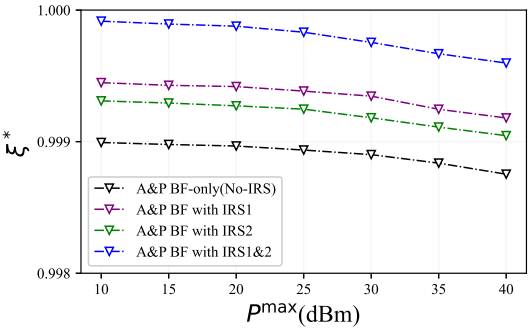
5.4 Impact of Covertness Constraint

The covertness constraint also plays an important role on the system performance.

- Figure 8(a) compares CSE under different levels of covertness constraint, i.e., $\epsilon \in \{0, 0.01, 0.005, 0.001\}$, along with the transmission power of BS. According to (C2) (i.e., $\xi^* \geq 1-\epsilon$), ϵ is a small value used to restrict the information received by Willie. When $\epsilon = 0$, CUs intend to achieve the perfect concealment, which implies a 100% probability of false detection and missed detection (i.e., $\xi^* = 1$). IRSs must continuously adjust the phase shift to prioritize the security and privacy of CUs and minimize the amount of information leaked to Willie. As the value of ϵ increases, the more relaxed requirement for concealment (e.g., when $\epsilon = 0.01$, $\xi^* = 0.99$; when $\epsilon = 0.001$, $\xi^* = 0.999$) allows each agent in MADDPG to obtain a better reward.
- Figure 8(b) further details the minimum value of ξ^* in the present of A&P BF-only, IRS1-only, IRS2-only and IRS1&2 respectively when $\epsilon = 0.001$. Without the action of IRS, the covertness constraint can not be satisfied anymore when the transmission power of BS is larger than 20dBm. Easily understood, there will be a more optimal value of ξ^* when two IRSs work together. Inter-



(a) CSE (Covert Sepctrum Effciency)



(b) When ε is 0.001

Figure 8. *System performance along with covertness constraint.*

estingly, the value of ξ^* by IRS1-only is better than that of IRS2-only. That is because IRS2 is more far away from the base station. Thus, signals passing through IRS2 will become relatively weak due to the free space path loss.

5.5 Impact of NLoS Component

Most existing literature related to IRS-assisted covert communications holds a common assumption that the small-scale fading is the same in current environments, i.e., BS-IRS, IRS-CU, and BS-CU channels share the same Rician K-factor [1, 21, 34] .

Since the K-factor represents the ratio of LoS signals power to the scattered multi-path signals power in Rician fading channels, Figure 9 tends to discuss the covert spectral efficiency when the K-factor is set at different values. The larger the K-factor, the greater the dominant component of LoS. In particular, the channel model degrades to LoS channel when 'K-factor $\rightarrow \infty$' or Rayleigh fading channel when 'K-factor $= 0$' [20].

- Along with the value of K-factor increases, LoS gradually becomes the dominant component

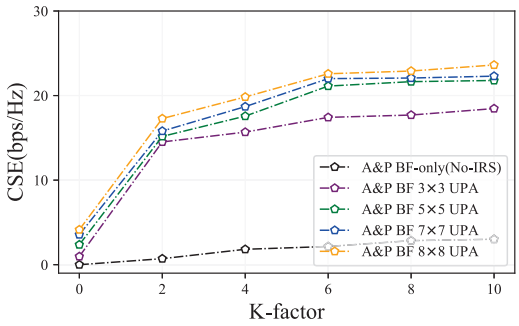


Figure 9. *System performance along with K-factor.*

which results in stronger and more reliable signal reception and thus a better spectrum efficiency. On the other hand, the multi-path components become more significant when K-factor is at a small value. The severe signal dispersion and potential interference will inevitably degrade the channels condition and the system performance.

- Note that even when more NLoS components are in the propagation, CSE is still satisfying. The reason can be interpreted that MADDPG works depending on the actual feedback reward rather than the channel estimation which is generally difficult to be obtained due to the time-varying multi-path delay spread and the resultant convolutions time-domain channel responses.

5.6 Performance Comparison with PBF-only Scenario

Figure 10 compares the system performance with PBF-only (i.e., the transmission power of BS is set at a fixed value) and Joint A&P BF schemes.

- No matter PBF-only or Joint A&P BF scheme, the spectrum efficiency of each CU as well as the total value can be greatly enhanced when both IRS 1&2 are deployed together. As aforementioned, since IRS1 is located geographically close to BS, it reflects more energy in a larger range, which leads to less radio spatial dissipation and better CSE.
- Compared with PBF-only, not only the spectrum efficiency of each CU can be further improved, the fairness among CUs can also be guaranteed by Joint A&P BF. That is because when BS works as an independent agent by Joint A&P BF, it is

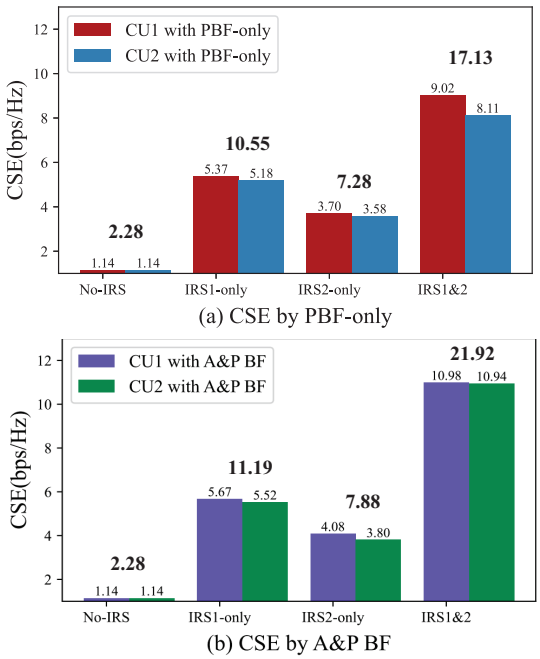


Figure 10. *CSE comparison of PBF-only and Joint A&P BF.*

driven to balance the transmitting power to compensate different path loss caused by attenuation. Besides, multiple IRSs can also cooperate beamforming for remote CUs.

VI. CONCLUSION

Different from the existing researches that do not consider the cooperation of multi-IRS and the active beamforming of BS, the paper focuses on tackling the Joint A&P BF optimization problem for IRS-assisted covert communications with low detection probability at warden. Facing the challenge of obtaining instantaneous CSI covertly by traditional pilot-based channel estimation methods, the machine learning based approach is adopted where each agent can learn from its historical action by a simple scalar reward feedback from CUs, which helps to avoid the complex CSI estimation and hardware design. Besides, the distributed execution feature of MADDPG further reduces the communication overhead caused by information sharing.

Although the paper only discusses the joint beamforming for covert communications in downlink, such approach can also be applied in uplink for IRS-assisted time-division duplexing (TDD) MIMO systems due to the uplink-downlink channel reciprocity. With respect to the frequency-division duplexing (FDD) systems, the approach can also be taken into action with the scalar reward feedback from BS instead.

ACKNOWLEDGEMENT

The work was supported by the Key Laboratory of Near Ground Detection and Perception Technology (No. 6142414220406 and 6142414210101), Shaanxi and Taicang Keypoint Research and Invention Program (No. 2021GXLH-01-15 and TC2019SF03).

REFERENCES

[1] CUI M, ZHANG G, ZHANG R. Secure wireless communication via intelligent reflecting surface[J]. IEEE Wireless Communications Letters, 2019, 8(5): 1410-1414.

[2] BASH B A, GOECKEL D, TOWSLEY D, et al. Hiding information in noise: Fundamental limits of covert wireless communication[J]. IEEE Communications Magazine, 2015, 53(12): 26-31.

[3] LU X, HOSSAIN E, SHAFIQUE T, et al. Intelligent reflecting surface enabled covert communications in wireless networks[J]. IEEE Network, 2020, 34(5): 148-155.

[4] TONG X, ZHANG Z, WANG J, et al. Joint multi-user communication and sensing exploiting both signal and environment sparsity[J]. IEEE Journal of Selected Topics in Signal Processing, 2021, 15(6): 1409-1422.

[5] LV L, WU Q, LI Z, et al. Covert communication in intelligent reflecting surface-assisted noma systems: Design, analysis, and optimization[J]. IEEE Transactions on Wireless Communications, 2021, 21(3): 1735-1750.

[6] WU C, YAN S, ZHOU X, et al. Intelligent reflecting surface (irs)-aided covert communication with warden's statistical csi[J]. IEEE Wireless Communications Letters, 2021, 10(7): 1449-1453.

[7] ZHOU X, YAN S, WU Q, et al. Intelligent reflecting surface (irs)-aided covert wireless communications with delay constraint[J].

IEEE Transactions on Wireless Communications, 2021, 21(1): 532-547.
[8] CHEN X, ZHENG T X, DONG L, et al. Enhancing mimo covert communications via intelligent reflecting surface[J]. IEEE Wireless Communications Letters, 2021, 11(1): 33-37.
[9] WU Q, ZHANG S, ZHENG B, et al. Intelligent reflecting surface-aided wireless communications: A tutorial[J]. IEEE Transactions on Communications, 2021, 69(5): 3313-3351.
[10] TAHA A, ALRABEIAH M, ALKHATEEB A. Deep learning for large intelligent surfaces in millimeter wave and massive mimo systems[C]// 2019 IEEE Global communications conference (GLOBECOM). IEEE, 2019: 1-6.
[11] HE Z Q, YUAN X. Cascaded channel estimation for large intelligent metasurface assisted massive mimo[J]. IEEE Wireless Communications Letters, 2019, 9(2): 210-214.
[12] GAN X, ZHONG C, HUANG C, et al. Multiple riss assisted cell-free networks with two-timescale csi: Performance analysis and system design[J]. IEEE Transactions on Communications, 2022, 70(11): 7696-7710.
[13] HAN Y, TANG W, JIN S, et al. Large intelligent surface-assisted wireless communication exploiting statistical csi[J]. IEEE Transactions on Vehicular Technology, 2019, 68(8): 8238-8242.
[14] GAN X, ZHONG C, HUANG C, et al. Ris-assisted multi-user miso communications exploiting statistical csi[J]. IEEE Transactions on Communications, 2021, 69(10): 6781-6792.
[15] HUANG C, MO R, YUEN C. Reconfigurable intelligent surface assisted multiuser miso systems exploiting deep reinforcement learning[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(8): 1839-1850.
[16] HUANG C, YANG Z, ALEXANDROPOULOS G C, et al. Multi-hop ris-empowered terahertz communications: A drl-based hybrid beamforming design[J]. IEEE Journal on Selected Areas in Communications, 2021, 39(6): 1663-1677.
[17] LIU X, LIU Y, CHEN Y. Machine learning empowered trajectory and passive beamforming design in uav-ris wireless networks[J]. IEEE Journal on Selected Areas in Communications, 2020, 39(7): 2042-2055.
[18] LIU X, LIU Y, CHEN Y, et al. Ris enhanced massive non-orthogonal multiple access networks: Deployment and passive beamforming design[J]. IEEE Journal on Selected Areas in Communications, 2020, 39(4): 1057-1071.
[19] WU Q, ZHANG R. Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design[C]//2018 IEEE Global Communications Conference (GLOBECOM). IEEE, 2018: 1-6.
[20] WU Q, ZHANG R. Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming[J]. IEEE transactions on wireless communications, 2019, 18(11): 5394-5409.
[21] GUO H, LIANG Y C, CHEN J, et al. Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks[J]. IEEE transactions on wireless communications, 2020, 19(5): 3064-3076.
[22] ZHENG T X, WANG H M, NG D W K, et al. Multi-antenna covert communications in random wireless networks[J]. IEEE Transactions on Wireless Communications, 2019, 18(3): 1974-1987.
[23] BASH B A, GOECKEL D, TOWSLEY D. Limits of reliable communication with low probability of detection on awgn channels[J]. IEEE journal on selected areas in communications, 2013, 31(9): 1921-1930.
[24] CAO Y, LV T, NI W. Intelligent reflecting surface aided multi-user mmwave communications for coverage enhancement[C]//2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. IEEE, 2020: 1-6.
[25] YAN S, HE B, ZHOU X, et al. Delay-intolerant covert communications with either fixed or random transmit power[J]. IEEE Transactions on Information Forensics and Security, 2018, 14(1): 129-140.
[26] GAO Y, YONG C, XIONG Z, et al. Reconfigurable intelligent surface for miso systems with proportional rate constraints[C]//ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 2020: 1-7.
[27] ZHANG K, YANG Z, BAŞAR T. Multi-agent reinforcement learning: A selective overview of theories and algorithms[J]. Handbook of rein-

forcement learning and control, 2021: 321-384.
[28] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. nature, 2015, 518(7540): 529-533.
[29] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[A]. 2015.
[30] WU Q, ZHANG R. Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network[J]. IEEE communications magazine, 2019, 58(1): 106-112.
[31] ZHANG J, BJÖRNSON E, MATTHAIOU M, et al. Prospective multiple antenna technologies for beyond 5g[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(8): 1637-1660.
[32] MUKHERJEE A, SWINDLEHURST A L. Detecting passive eavesdroppers in the mimo wiretap channel[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012: 2809-2812.
[33] GOLDSMITH A. Wireless communications[M]. Cambridge university press, 2005.
[34] ZHANG S, ZHANG R. Capacity characterization for intelligent reflecting surface aided mimo communication[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(8): 1823-1838.
[35] SHI Q, RAZAVIYAYN M, LUO Z Q, et al. An iteratively weighted mmse approach to distributed sum-utility maximization for a mimo interfering broadcast channel[J]. IEEE Transactions on Signal Processing, 2011, 59(9): 4331-4340.
[36] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. Advances in neural information processing systems, 2017, 30.

BIOGRAPHIES

Gao Ang received his Ph.D. degree in control theory and engineering from Northwestern Polytechnical University in 2011. He currently serves as an Associate Professor at the School of Electronics and Information, Northwestern Polytechnical University. His interests include resource management and reinforcement learning in wireless communications.

Ren Xiaoyu is currently a master student under the supervision of Prof. A. Gao with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. Her research interests include covert communication and deep reinforcement learning in wireless communication networks.

Deng Bin is currently serves as an engineer of the Key Laboratory of Near Ground Detection and Perception Technology, Wuxi, China. He is interested in the pattern recognition and intelligent systems.

Sun Xinshun is currently a master student under the supervision of Prof. A. Gao with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. His research interests include intelligent reflecting surface, beamforming and channel estimation.

Zhang Jiankang is a Senior Lecturer at Bournemouth University. Prior to joining in Bournemouth University, he was a senior research fellow at University of Southampton, UK. His research interests are in the areas of aeronautical communications and networks, evolutionary algorithms, machine learning algorithms and edge computing.