Sum Throughput Optimization of Wireless Powered IRS-Assisted Multi-User MISO System

Jing Xu[a], Jiarun Tang[a], Yuze Zou[b], Ruikai Wen[a], Wei Liu[a], Jianhua He[c]

[a]School of Electronics Information and Communications
Huazhong University of Science and Technology, Wuhan, China, 430074
[b]Wuhan Maritime Communication Research Institute, Wuhan, China, 430000
[c] School of Computer Science and Electronic Engineering
University of Essex, Colchester, U.K., CO4 3SQ

Abstract

Intelligent reflecting surface (IRS) is a promising technology for beyond-5G wireless communication systems. However, the energy demand of IRS is often overlooked in existing works, leading to performance issues in practical scenarios. To address this issue, this paper proposes an operating model based on time switching (TS) protocol for an IRS-assisted multi-user multiple-input single-output (MISO) system, which can provide energy for IRS through wireless power transfer (WPT) technology. The system throughput maximization problem is addressed to improve performance. Specifically, a two-stage algorithm combined with alternating optimization, denoted as TAO, is proposed. To further improve the optimization process in large-size IRS scenarios, an improved deep deterministic policy gradient (DDPG) method combined with TAO, denoted as TAO-DDPG, is also proposed. Numerical results demonstrate that the proposed TAO-DDPG algorithm achieves similar performance to TAO while greatly reducing the optimization time.

Keywords:
Intelligent reflecting surface, wireless communication, time switching, deep reinforcement learning, MU-MISO system.

1. Introduction

Intelligent reflecting surface (IRS) has emerged as a highly promising technology in recent years to enhance the performance of beyond-5G wireless communication systems. The IRS is an artificial surface that is composed of electromagnetic materials and contains an array of reflective elements. Each of these elements can be software-defined and self-regulate properties such as the phase shift of the electromagnetic waves reflected on it. By controlling the entire array of reflective elements jointly, the IRS can enhance the reflected signals, thereby improving the transmission of wireless communication systems.

In comparison to other advanced technologies, such as MIMO, Millimeter Wave, and Ultra-Dense Network, etc.[1] IRS exhibits numerous advantages, including ease of deployment, low cost, energy efficiency[2]-[3], and strong compatibility[4]. By leveraging these advantages, IRS has been introduced in various wireless communication fields. Recent studies have focused on optimizing wireless communication systems that utilize IRS, as evidenced by [5]-[6]. In [7], the authors addressed the transmit power minimization problem in an IRS-assisted MISO system for non-orthogonal multiple access (NOMA) transmission. To minimize power consumption, the authors in [8] proposed an improved quasi-degraded conditional transformation method, while in [9], the authors highlighted the importance of user pairing in NOMA networks with IRS, specifically in a discrete phase shift IRS-assisted single-input single-output (SISO) system. While several studies have focused on transmit power minimization, data rate optimization is

[*]Corresponding author: Wei Liu

Email addresses: xujing@hust.edu.cn (Jing Xu), tangjiarun@hust.edu.cn (Jiarun Tang), zouyuze@hust.edu.cn (Yuze Zou), wenruikai@hust.edu.cn (Ruikai Wen), liuwei@hust.edu.cn (Wei Liu), j.he@essex.ac.uk (Jianhua He)

also a key consideration. In [10] and [11], the authors discuss maximizing the data rate of a multi-user MISO system, utilizing an alternate optimization (AO) method. In [12] and [13], the focus was on optimizing the weighted sum of each user's data rate, with the former jointly adjusting the IRS phase shift scheme and user scheduling to balance between maximizing the sum data rates and ensuring fairness of data rate among users, and the latter pre-optimizing the IRS and combining it with current channel conditions to maximize the sum data rate of the wireless communication system. However, these simplifications may result in a loss of user fairness. In [14] and [6], the authors addressed the maximization of the minimum data rate of users to ensure that the system can provide communication services that meet the needs of all users.

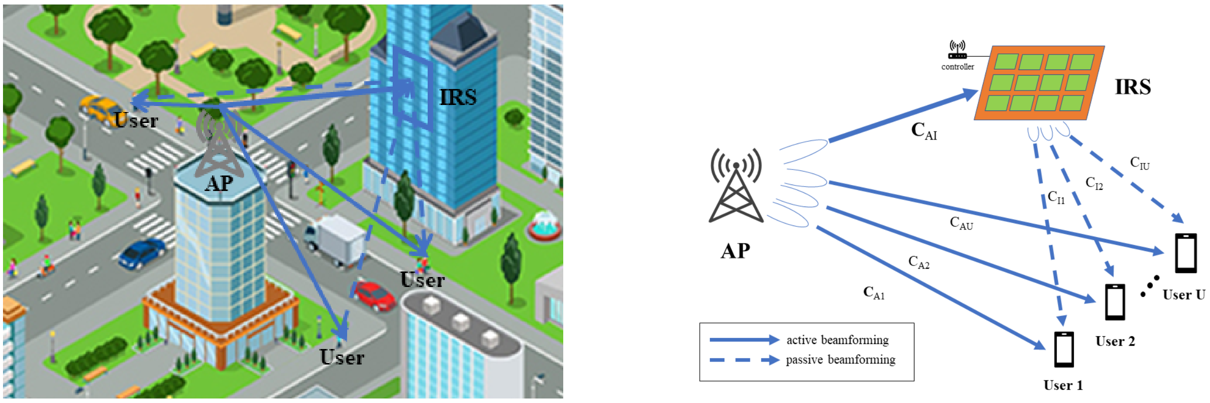
Despite the popularity of IRS-assisted wireless communication systems, all above studies assume that IRS is an ideal device with no energy consumption, ignoring the practical energy requirements of IRS. In [15]-[16], researchers introduced wireless power transfer (WPT) into IRS-assisted wireless communication systems, taking energy constraints into account. WPT has been extensively researched and discussed in communication-related studies [17]-[18], as it allows wireless devices to obtain energy from communication signals, thus enabling long-term maintenance [16]. In [19], two typical WPT protocols, time switching (TS) and power splitting (PS), were proposed, both of which have been shown to be effective when combined with WPT.

The introduction of WPT has made charging the IRS extremely challenging when considering capacity optimization. Researchers in [15] studied the optimization of IRS passive beamforming and energy allocation in a full-duplex wireless powered communication network (WPCN) that uses a hybrid access point (HAP) to transmit both energy and information signals. In [20], the maximum total energy received by the energy receiver was obtained through the joint optimization of AP active beamforming and IRS passive beamforming. Moreover, [21] and [22] maximized the minimum power that users receive in WPT network and the sum data rate in a WPT network, respectively. Additionally, the scenario was extended to multiple IRSs, and the requirements for quality of service (QoS) of different signal and energy receivers were used as constraints to minimize the AP transmit power.

The studies mentioned above mainly focused on scenarios in which WPT is combined with IRS, but there are still some limitations. For example, the work in [15] focused on the passive beamforming optimization of the IRS, but not considering the active beamforming of the HAP, implying that the system was not fully optimized. Similarly, the authors of [20] and [21] only considered the application of IRS in WPT systems. In our previous works[23]-[24], we considered the performance optimization of wireless-powered IRS-assisted single-user systems based on PS and TS protocols. Numerical results showed that PS outperforms TS when the IRS is closer to the AP, while TS outperforms PS when the IRS is closer to the users. However, only single-user situation was discussed in our work.

The utilization of artificial intelligence (AI) algorithms in communication systems has been explored in recent researches. Deep reinforcement learning (DRL) has emerged as a promising tool for optimizing IRS-assisted wireless communication systems, due to its unique advantages such as the ability to learn from the environment, make autonomous decisions, and significant learning speed[25]. Two typical DRL methods, deep Q network (DQN) and deep deterministic policy gradient (DDPG), have been proposed to address problems with different features of action and state spaces, namely discrete and continuous spaces respectively. It is commonly observed that the former method excels in solving continuous space problems, whereas the latter outperforms in solving problems with discrete space. For instance, the authors of [26] utilized DRL to predict and optimally tune the IRS phase shift matrices in order to solve the non-convex sum data rate maximization problem in an IRS-assisted downlink NOMA system. The results show that the IRS-assisted NOMA system performs better than an OMA-based one under the DRL scheme. In [27], the maximization of the sum data rate in IRS-assisted MISO interference channels is investigated, and a DRL-empowered algorithm is proposed to configure the digital beamforming and analog phase shift matrices. Furthermore, the work in [28] proposed a DRL-based algorithm to jointly optimize user scheduling, phase shift control, and beamforming in IRS-aided systems. Specifically, curriculum learning with DDPG (CL-DDPG) is used to optimize the phase shift control and beamforming vectors, which outperforms AO-based algorithms in terms of runtime when the number of reflective elements is large, indicating that DRL-based methods could significantly reduce optimization time in large size IRS scenarios.

Motivated by the aforementioned discussions, we consider a novel approach to optimize a multi-user IRS-assisted system with wireless power supply, where the TS protocol is employed to provide energy for the IRS. To maximize the system throughput, we propose a two-stage algorithm based on AO, denoted as TAO, which jointly optimizes the active beamforming of the AP with the passive beamforming of the IRS. Furthermore, to reduce the optimization time in case of dynamic channel changes, we introduce DRL to solve the problem. The major contributions of this paper



(a) scenario diagram

(b) principle diagram

Figure 1: Wireless powered IRS-assisted multi-user MISO system

are summarized as follows:

- We propose a multi-user system that optimizes the system with consideration of powering IRS. The proposed optimization algorithm is highly compatible with common multi-user systems in reality.
- TAO is introduced to improve the convergence of DDPG in the replay mechanism, which greatly accelerates the optimization of the system. Compared with TAO, TAO-DDPG significantly reduces the optimization time while maintaining similar performance.

The rest of this paper is organized as follows. Section II presents the system model and problem formulation. Section III develops TAO to solve the system throughput maximization problem. In Section IV, we propose TAO-DDPG to address the capacity optimization problem, which greatly reduces the optimization time. Section V shows numerical results of system simulations under different optimization algorithms. Finally, Section VI concludes this paper and discusses future work.

2. System model and problem formulation

The scenario and principle diagrams of the wireless powered IRS-assisted multi-user MISO system are illustrated in Fig. 1. The system mainly consists of three parts: AP, IRS, and multiple users. The AP serves as the transmitter for multiple users, while also supplying power to the IRS. The IRS obtains energy from the AP and assists in transmitting signals to the users. The system performs signal transmission based on TS, and the communication process is divided into two stages: energy harvesting (EH) and IRS-assisted transmission (IAT). During the EH stage, the IRS harvests energy from the active beamforming signal transmitted by the AP, which is stored and used in the IAT stage. In the IAT stage, the IRS, now fully charged, reflects signals to each user in the system along with the active beamforming signal from the AP. The two stages are dynamically switched according to the remaining power of the IRS and the operating time of the system. In this paper, the proportion of the working time of the first stage to the unit operating time of the system is denoted as t_{EH}, and the proportion of the second stage is recorded as t_{IAT}. The relation between t_{EH} and t_{IAT} is

$$t_{\mathrm{EH}} + t_{\mathrm{IAT}} = 1. \tag{1}$$

Thus, the working model of the wireless powered IRS-assisted multi-user MISO system based on TS can be obtained, as shown in Fig. 2.

Based on the aforementioned model, we formulate the capacity optimization problem of a wireless powered IRS-assisted multi-user MISO system as maximizing the system throughput for a given AP transmit power. The system throughput is defined as the sum of the throughput during the EH stage and the IAT stage. However, optimizing all the coupled variables simultaneously may result in an infeasible solution to the throughput maximization problem. To address this issue, we leverage the WPT model presented in [29] and assume that during the EH stage, the AP only transmits power signals. This simplifies the optimization problem for the AP, which no longer needs to perform

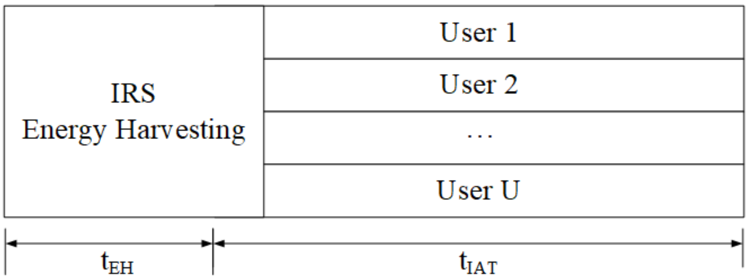


Figure 2: Working model of TS-based wireless powered IRS-assisted multi-user MISO system

signal encoding and beamforming to meet the various demands of users, thereby facilitating the optimization of the system's communication performance. Consequently, the overall system throughput in a single operating time unit can be approximated as the throughput of the IAT stage.

For the AP, the number of its antennas is denoted as M, and its transmit power is denoted as p_0. For the IRS, the number of reflective elements is denoted as N, the energy consumption of each reflective element is denoted as μ, and the offset phase corresponding to the reflective element is denoted as $\varphi_n \in [0, 2\pi), \forall n \in \mathcal{N} = \{1, \ldots, N\}$. The phase shift of each reflective element is uniformly expressed as the overall phase shift vector of IRS $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_N]^{\mathrm{T}} \in \mathbb{C}^{N\times 1}, \theta_n = e^{j\varphi_n}, |\theta_n|^2 = 1$, the corresponding phase shift matrix is written as $\mathbf{\Theta} = \mathrm{diag}(\boldsymbol{\theta}) \in \mathbb{C}^{N\times N}$. The efficiency of energy harvesting by IRS is denoted as η. For users, the number of users is denoted as U, and all users are unified into the set $\mathcal{U}$, that is $\forall u \in \mathcal{U} = \{1, \ldots, U\}$. For each group of channels in the system, the channel between the AP and the IRS is denoted as $\mathrm{C}_{\mathrm{AI}} \in \mathbb{C}^{M\times N}$. The channel between AP and each user is denoted as $\mathrm{C}_{\mathrm{A}u} \in \mathbb{C}^{M\times 1}$ $(u \in \mathcal{U})$. The channel between IRS and each user is denoted as $\mathrm{C}_{\mathrm{I}u} \in \mathbb{C}^{N\times 1}$ $(u \in \mathcal{U})$. The signal that user u receives could be calculated as

$$y_u = (\mathbf{C}_{\mathrm{A}u} + \mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u})^{\mathrm{H}}\mathbf{B}\mathbf{s} + n_u, \tag{2}$$

where $\mathbf{B}$ is the active beamforming matrix of AP, $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_U] \in \mathbb{C}^{M\times U}, \mathrm{b}_u \in \mathbb{C}^{M\times 1}$ is the beamforming vector for the transmission signal of user u; $\mathbf{s} = [s_1, \ldots, s_U]^{\mathrm{T}} \in \mathbb{C}^{U\times 1}$ is the signal matrix transmitted by the AP to each user, and $\mathbb{E}\left[\mathbf{s}\mathbf{s}^{\mathrm{H}}\right] = \mathbf{I}$; n_u is the additive white Gaussian noise in the transmission to user u, which has the mean value as 0, and the variance as σ_u^2.

In a multi-user scenario, the AP must simultaneously serve multiple users during the IAT stage. To represent the system throughput during this stage, we use SINR to calculate the data rate of each user by the AP. The SINR of user u is given by

$$\gamma_u = \frac{|(\mathbf{C}_{\mathrm{A}u} + \mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u})^{\mathrm{H}}\mathbf{b}_u|^2}{\sigma_u^2 + |(\mathbf{C}_{\mathrm{A}u} + \mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u})^{\mathrm{H}}\mathbf{B}_{-u}|^2}, \tag{3}$$

where $(\mathbf{C}_{\mathrm{A}u}+\mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u})$ represents the enhanced transmission channel between the AP and user u due to the introduction of the IAT. $\mathbf{b}_u$ refers to the AP active beamforming for user u, $\mathbf{B}_{-u}$ refers to the AP beamforming matrix without $\mathbf{b}_u$. The data rate of the overall system is

$$\begin{aligned} s_{\mathrm{IAT}}(\mathbf{B}, \Theta) &= \sum_{u\in\mathcal{U}} \log_2(1 + \gamma_u) \\ &= \sum_{u\in\mathcal{U}} \log_2\left(1 + \frac{\left|(\mathbf{C}_{Au} + \mathbf{C}_{\mathrm{AI}}\Theta\mathbf{C}_{\mathrm{I}})^{\mathrm{H}}\,\mathbf{b}_u\right|^2}{\sigma_u^2 + \left|(\mathrm{C}_{\mathrm{Au}} + \mathrm{C}_{\mathrm{AI}}\Theta\mathrm{C}_{\mathrm{I}u})^{\mathrm{H}}\,\mathbf{B}_{-u}\right|^2}\right) \end{aligned} \tag{4}$$

Given that the operating time of the IAT stage is denoted as t_{IAT}, the overall throughput of the system can be derived as follows

$$th = t_{\mathrm{IAT}} \cdot s_{\mathrm{IAT}}(\mathbf{B}, \Theta) \tag{5}$$

Then, the throughput maximization problem of wireless powered IRS-assisted multi-user MISO system can be expressed as

$$
\begin{aligned}
&(\mathbf{P1}) \max_{0\le t_{\mathrm{IAT}}\le 1,\mathbf{B},\mathbf{\Theta}} && t_{\mathrm{IAT}}\cdot\sum_{u=1}^{U} s_u && \text{(6a)}\\
&s.t. && \|\mathbf{B}\|_F^2 \le p_0 && \text{(6b)}\\
& && \eta t_{\mathrm{EH}}\gamma_{\mathrm{EH}} \ge t_{\mathrm{IAT}} N\mu && \text{(6c)}
\end{aligned}
$$

where the $\|\mathbf{B}\|_F^2$ is the F-norm of $\mathbf{B}$, meaning the power that AP active beamforming matrix $\mathbf{B}$ requires; p_0 is the AP transmit power; γ_{EH} is the rate of the IRS harvesting energy from the signals transmitted by the AP. (6b) and (6c) are the power constraints of AP active beamforming and IRS passive beamforming.

3. Capacity Optimization with CVX

In this section, we adopt a conventional approach to address the problem **P1** of maximizing system throughput. Specifically, we transform and decouple the **P1** into convex optimization subproblems, which can be solved with CVX.

Before solving **P1**, we could determine γ_{EH} with the EH stage assumption. For γ_{EH}, with the AP-IRS channel $\mathbf{C}_{\mathrm{AI}}$ and the beamforming of AP for powering IRS $\mathbf{b}_{AI}$, it could be implied as

$$\gamma_{\mathrm{EH}} = \|\mathbf{C}_{\mathrm{AI}}^{\mathrm{H}}\mathbf{b}_{\mathrm{AI}}\|_2^2 \tag{7}$$

According to [29], with AP only transmitting power signals in EH stage, the optimized $\mathbf{b}_{\mathrm{AI}}$ is the eigenvector of $\mathbf{C}_{\mathrm{AI}}$, and its modulus is proportional to p_0. Therefore, γ_{EH} could be reformulated as

$$\gamma_{\mathrm{EH}} = \|\mathbf{C}_{\mathrm{AI}}\|_2^2 * p_0 \tag{8}$$

In (8), the right side of the equal sign is full of known quantities, hence γ_{EH} could be determined.

With determined γ_{EH}, we could obtain the upper bound of t_{IAT} with the constraint (6c) and the EH stage assumption. We substitute (1) into the constraint (6c), and (6c) could be transformed into the following inequality

$$t_{\mathrm{IAT}} \le \frac{\eta\gamma_{\mathrm{EH}}}{N\mu + \eta\gamma_{\mathrm{EH}}} \tag{9}$$

With known γ_{EH}, the upper bound of t_{IAT} is determined in (9). To achieve the maximum throughput in **P1**, t_{IAT} must take its upper bound, hence t_{IAT} is determined. It is clear that the optimized t_{IAT} is independent of AP active beamforming $\mathbf{B}$ and IRS passive beamforming $\mathbf{\Theta}$, which proves that the subproblem of determining t_{IAT} is decoupled with maximizing s_{IAT}.

With determined upper bound of t_{IAT}, **P1** is tranformed in to a subproblem of maximizing s_{IAT}, which could be formulated as

$$
\begin{aligned}
&\max_{\mathbf{B},\mathbf{\Theta}} \sum_{u\in\mathcal{U}} \log_2\Big(1+\frac{|(\mathbf{C}_{\mathrm{A}u}+\mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u})^{\mathrm{H}}\mathbf{b}_u|^2}{\sigma_u^2+|(\mathbf{C}_{\mathrm{A}u}+\mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u})^{\mathrm{H}}\mathbf{B}_{-u}|^2}\Big) && \text{(10a)}\\
&\text{s.t. } \|\mathbf{B}\|_F^2 \le p_0 && \text{(10b)}
\end{aligned}
$$

To address this problem in a more convenient fashion, we firstly express the AP-IRS-User cascade channel $\mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u}$ in another form. $\mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u}$ could be expressed as

$$\mathbf{E}_{\mathrm{A}u}\boldsymbol{\theta} = \mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u}, \tag{11}$$

where $\mathbf{E}_{\mathrm{A}u} = \mathbf{C}_{\mathrm{AI}}\cdot\mathrm{diag}(\mathbf{C}_{\mathrm{A}u})$ denotes the equivalent channel of the AP-IRS-User cascade channel, and we will continue to use it in the subsequent derivation.

Based on [11], we solve **P2** with alternating optimization (AO) and block coordinate descent (BCD). With introduction of relaxing variables $\boldsymbol{\tau} = [\tau_1,\ldots,\tau_U]^{\mathrm{T}} \ge 0$ and $\boldsymbol{\upsilon} = [\upsilon_1,\ldots,\upsilon_U]^{\mathrm{T}} \ge 0$, **P2** could be transformed as

$$
\begin{aligned}
\max_{\boldsymbol{\tau},\boldsymbol{\upsilon},\mathbf{B},\boldsymbol{\theta}} f(\boldsymbol{\tau},\boldsymbol{\upsilon},\mathbf{B},\boldsymbol{\theta}) = & \sum_{u\in\mathcal{U}} log(1+\tau_u) - \tau_u \\
& + \sum_{u\in\mathcal{U}} 2\sqrt{1+\tau_u}\mathrm{Re}(\upsilon_u^*(\mathbf{C}_{\mathrm{A}u}+\mathbf{E}_{Au}\boldsymbol{\theta})^{\mathrm{H}}\mathbf{b}_u) \\
& - \sum_{u\in\mathcal{U}} |\upsilon_u|^2(|(\mathbf{C}_{\mathrm{A}u}+\mathbf{E}_{\mathrm{A}u}\boldsymbol{\theta})^{\mathrm{H}}\mathbf{B}|^2+\sigma_u^2)
\end{aligned} \tag{12}
$$

$\boldsymbol{\tau}, \boldsymbol{\upsilon}, \mathbf{B}, \boldsymbol{\theta}$ need to be updated in every iteration while solving **P2**. For $\boldsymbol{\tau}$ and $\boldsymbol{\upsilon}$, it could be calculated by extremal conditions of Lagrange Multipliers

$$
\tau_u = \frac{\overline{\phi}_u^2 + \overline{\phi}_u\sqrt{\overline{\phi}_u^2+4}}{2} \tag{13}
$$

$$
\upsilon_u = \frac{\sqrt{1+\tau_u}(\mathbf{C}_{\mathrm{A}u}+\mathbf{E}_{\mathrm{A}u}\boldsymbol{\theta})^{\mathrm{H}}\mathbf{b}_u}{|(\mathbf{C}_{\mathrm{A}u}+\mathbf{E}_{\mathrm{A}u}\boldsymbol{\theta})^{\mathrm{H}}\mathbf{B}|^2+\sigma_u^2} \tag{14}
$$

where $\overline{\phi}_u^2 = \mathrm{Re}(\upsilon_u^*(\mathbf{C}_{\mathrm{A}u}+\mathbf{E}_{\mathrm{A}u}\boldsymbol{\theta})^{\mathrm{H}}\mathbf{b}_u)$. Furthermore, based on the explanation in [11] and [30], $\boldsymbol{\tau}$ could be updated as

$$
\tau_u = \frac{|(\mathbf{C}_{\mathrm{A}u}+\mathbf{E}_{\mathrm{A}u}\boldsymbol{\theta})^{\mathrm{H}}\mathbf{b}_u|^2}{\sigma_u^2+|(\mathbf{C}_{\mathrm{A}u}+\mathbf{E}_{\mathrm{A}u}\boldsymbol{\theta})^{\mathrm{H}}\mathbf{B}_{-u}|^2} \tag{15}
$$

Then the AP active beamforming $\mathbf{B}$ could be updated as

$$
\begin{aligned}
\mathbf{b}_u = & \sqrt{1+\tau_u}\upsilon_u \\
& *(\sum_{v\in\mathcal{U}} |\upsilon_u|^2(\mathbf{C}_{Av}+\mathbf{E}_{Av}\boldsymbol{\theta})(\mathbf{C}_{Av}+\mathbf{E}_{Av}\boldsymbol{\theta})^H + \epsilon\mathbf{I}_M)^{-1} \\
& *(\mathbf{C}_{Au}+\mathbf{E}_{Au}\boldsymbol{\theta})
\end{aligned} \tag{16}
$$

where $\mathbf{I}_M$ is an identity matrix with the size of $M\times M$; ϵ is the dual optimization variable given by the power constraint. ϵ could be obtained by dichotomy. First, ϵ is initialized and substituted in (16) to calculate $\mathbf{B}$. Then, if $\mathbf{B}$ satisfies the constraint that $||\mathbf{B}||_F^2 < p_0$, ϵ could be updated to a larger value, otherwise ϵ is reduced, and $\mathbf{B}$ is calculated again. When ϵ is determined, $\mathbf{B}$ is updated.

The updating of $\boldsymbol{\theta}$ could be transformed as a quadratic constraint quadratic programming (QCQP) problem

$$
\min\ \boldsymbol{\theta}^H\mathbf{T}\boldsymbol{\theta} - 2\mathrm{Re}(\boldsymbol{\theta}^H\boldsymbol{\zeta}) \tag{17a}
$$

$$
\text{s.t. } |\theta_n| = 1, \tag{17b}
$$

The matrix $\mathbf{T}$ and vector $\boldsymbol{\zeta}$ are

$$
\mathbf{T} = \sum_{u\in\mathcal{U}} |\upsilon_u|^2 \sum_{v\in\mathcal{U}} \mathbf{i}_{v,u}\mathbf{i}_{v,u}^{\mathrm{H}} \tag{18}
$$

$$
\boldsymbol{\zeta} = \sum_{u\in\mathcal{U}}(\sqrt{1+\tau_u}\upsilon_u^*\mathbf{i}_{u,u} - |\upsilon_u|^2\sum_{v\in\mathcal{U}} j_{v,u}^*\mathbf{i}_{v,u}) \tag{19}
$$

where $\mathbf{i}_{v,u} = (\mathbf{E}_{\mathrm{A}u})^{\mathrm{H}}\mathbf{b}_v$, $j_{v,u} = \mathbf{C}_{\mathrm{A}u}^{\mathrm{H}}\mathbf{b}_v$. By introducing relaxing variable ι, the QCQP problem could be transformed as

$$
\min \begin{bmatrix}\boldsymbol{\theta}\\ \iota\end{bmatrix}^{\mathrm{H}} \begin{bmatrix}\mathbf{T} & -\boldsymbol{\zeta}\\ -\boldsymbol{\zeta}^{\mathrm{H}} & \mathbf{0}\end{bmatrix} \begin{bmatrix}\boldsymbol{\theta}\\ \iota\end{bmatrix} \tag{20}
$$

Then, consider $\begin{bmatrix}\boldsymbol{\theta}\\ \iota\end{bmatrix}$ as the optimization variable, (20) could be transformed as the following semi-definite programming

Algorithm 1 TAO: System Throughput Optimization with AO and BCD

```
Input: $M$, $p_0$, $N$, $\mu$, $U$, number of iterations $T$, current iteration $t = 0$
Output: maximum system throughput $th$
Initialize $\mathbf{C}_{\mathrm{AI}}$, $\mathbf{C}_{\mathrm{A}u}$ and $\mathbf{C}_{\mathrm{I}u}$
for $\forall$ random seed do
    Initialize $\boldsymbol{\theta}$ with current random seed, initialize current best data rate $s_{\max} = 0$
    while $t < T$ do
        update $\tau$ via (15)
        update $\upsilon$ via (14)
        update $\mathbf{B}_u$ via (16)
        update $\upsilon$ again via (14)
        update $\boldsymbol{\theta}$ via solving (21a)
        compare the current optimized data rate with $s_{\max}$, and update $s_{\max}$
        $t = t + 1$
    end while
    $s_{\mathrm{IAT}} = s_{\max}$
end for
determine $t_{\mathrm{IAT}}$ via (8) and (9) with initialized channels
obtain maximum system throughput $th = t_{\mathrm{IAT}} * s_{\mathrm{IAT}}$
```

(SDP) problem

$$\min_{\overline{\boldsymbol{\Theta}}} \quad \mathrm{Tr}(\mathbf{V}\overline{\boldsymbol{\Theta}}) \tag{21a}$$

$$\text{s.t.} \quad \overline{\boldsymbol{\Theta}} \geq 0, \tag{21b}$$

$$\overline{\boldsymbol{\Theta}}_{n,n} = 1, n = 1, ..., N+1 \tag{21c}$$

where $\mathbf{V} = \begin{bmatrix} \mathbf{T} & -\zeta \\ -\zeta^H & 0 \end{bmatrix}$, $\overline{\boldsymbol{\Theta}} \geq \begin{bmatrix} \theta \\ \iota \end{bmatrix} \begin{bmatrix} \theta \\ \iota \end{bmatrix}^{\mathrm{H}}$. Problem (21a) could be solved with CVX tool, and $\boldsymbol{\theta}$ could be recovered with Gaussian random method. $\boldsymbol{\tau}, \boldsymbol{\upsilon}, \mathbf{B}, \boldsymbol{\theta}$ are updated in an alternating way. The order of updating is demonstrated as follows

$$\cdots \rightarrow \tau \rightarrow \upsilon \rightarrow \mathbf{B} \rightarrow \upsilon \rightarrow \boldsymbol{\theta} \rightarrow \tau \rightarrow \cdots$$

In simulation, we discover that the initial selection of $\boldsymbol{\theta}$ significantly affects the optimization result. Therefore, in order to achieve the best optimization result, we utilize a series of different random seeds to initialize $\boldsymbol{\theta}$, and select the best from all the results of the optimization process.

By transforming and decoupling the optimization problem, we are able to maximize the system throughput. The complete algorithm is presented in Algorithm 1.

4. Capacity Optimization with DRL

In Section III, we propose a method to optimize the capacity of a wireless powered IRS-assisted multi-user MISO system by transforming the problem into solvable subproblems. However, this method involves multiple high-dimensional matrix computations, leading to high computational complexity. For large size IRS with a significant number of reflective elements, the method in Section III may require a considerable amount of time to optimize the system, making it challenging to adapt to dynamic channel conditions. Therefore, in this section, we propose a DDPG-based method with lower computational complexity to optimize the capacity of the system in a shorter time.

4.1. Problem Modeling under DRL Framework

4.1.1. State Modeling

Typically, a wireless communication system comprises one or more sets of transmitters and receivers, as well as the channels therebetween. However, in this paper, we only consider the transmitters and receivers as the source and destination of signals, and their operation is relatively irrelevant. Our focus is primarily on the channels themselves. Once the channels are determined, the state of the system is determined accordingly. Therefore, we employ the current channel information to represent the current state s_t, i.e. $s_t = \{\mathbf{C}_{\mathrm{AI}}, \mathbf{C}_{\mathrm{A}u}, \mathbf{C}_{\mathrm{I}u}\}(u \in \mathcal{U})$.

4.1.2. Action Modeling

Regarding action modeling, we have selected AP active beamforming and IRS passive beamforming as the actions taken by the system at every moment. During operation, the system transmits signals to each user via direct channel transmission and auxiliary transmission of the IRS passive beamforming using the AP's active beamforming. When considering the system as an agent, its active and passive beamforming align with the assumption of the agent's action. At the same time, in the capacity optimization problem, the active and passive beamforming are also served as variables for resolving the problem. Therefore, it is reasonable to consider the active beamforming of the AP and the passive beamforming of the IRS as the actions taken by the system under the DDPG framework. Thus, we employ the current AP active beamforming $\mathbf{B}$ and IRS passive beamforming $\mathbf{\Theta}$ to represent the current action a_t, i.e. $a_t = \{\mathbf{B}, \mathbf{\Theta}\}$.

4.1.3. Reward Modeling

Our optimization goal is to maximize the system throughput, plus the system throughput could be calculated with the parameters in the aforementioned action and state modeling, therefore we choose the system throughput as the reward function. The current reward of system is

$$\begin{aligned} r_t(s_t, a_t) &= t_{\mathrm{IAT}} \cdot s_{\mathrm{IAT}}(\mathbf{B},\mathbf{\Theta}) \\ &= t_{\mathrm{IAT}} \sum_{u \in \mathcal{U}} \log_2\Big(1 + \frac{|(\mathbf{C}_{\mathrm{A}u} + \mathbf{C}_{\mathrm{AI}}\mathbf{\Theta}\mathbf{C}_{\mathrm{I}u})^{\mathrm{H}}\mathbf{b}_u|^2}{\sigma_u^2 + |(\mathbf{C}_{Au} + \mathbf{C}_{AI}\mathbf{\Theta}\mathbf{C}_{Iu})^{H}\mathbf{B}_{-u}|^2}\Big) \end{aligned} \tag{22}$$

where t_{IAT} could still be determined with (8) and (9).

4.1.4. State Transition

In state modeling, we choose channel information to describe the state of the system. Under the ideal circumstance, channel parameters will not change during the operating time of the system. Therefore, the system has the same state at all time slots, i.e. $s_1 = ... = s_t = s_{t+1}, t = 1, 2, ...$ Hence, the state transition process under the DRL framework in this paper (s_t, a_t, r_t, s_{t+1}) is equal to (s, a_t, r_t, s), where $s = \{\mathbf{C}_{\mathrm{AI}}, \mathbf{C}_{\mathrm{A}u}, \mathbf{C}_{\mathrm{I}u}\}$.

On the other hand, for the time slot t, we consider that $t = 1$ always holds under the DRL framework, i.e. there only exists one time slot during the process that the system makes decisions to optimize the reward. Based on the previous analysis, the transmission time t_{IAT} of the system is independent of the AP active beamforming and IRS passive beamforming. Moreover, maximizing the system throughput equates to maximizing the sum data rate of the system at each time slot. Consequently, the state transition only occurs once, and the reward is obtained only once, which is the maximum system throughput.

4.2. Implementation and Improvement of DDPG

To address the issue of maximizing system throughput, we implement the DDPG method while adhering to the set parameters and constraints in the problem. Additionally, we improve the stability and performance of DDPG by integrating an alternate replay mechanism and TAO.

The state and action initialization are in consistent with the initialization of channels and beamforming outlined in Section III. Considering that deep neural networks (DNNs) only accept real input, we extract the real and imaginary parts of the complex matrices of channels and beamforming, subsequently converting them into one-dimensional vectors to comply with the input requirements.

Algorithm 2 TAO-DDPG: Improved DDPG with Alternate Replay Mechanism and TAO

```
Initialize $Q(s, a|\theta^Q)$ and $\mu(s|\theta^\mu)$ with parameters $\theta^Q, \theta^\mu$
Initialize target network $Q', \mu'$ with parameters $\theta^{Q'} \leftarrow \theta^Q, \theta^{\mu'} \leftarrow \theta^\mu$
Initialize channels and convert them to initial state $s$, initial the stochastic process of action $\mathcal{N}$
Use TAO to obtain the reference state transition tuple $(s, a_{ref}, r_{ref}, s)$
Initialize memory pool $R$
for episode = 1, M do
    According to current policy and exploration, obtain current action $a_t = \mu(s|\theta^\mu) + \mathcal{N}_t$
    obtain $r_t$ with current $a_t$ and $s$
    if episode achieves the set step length of reference then
        compare current reward $r_t$ with the reference reward $r_{ref}$, if $r_{ref}$ is higher, replace current tuple $(s_t, a_t, r_t, s_{t+1})$ as reference tuple
    end if
    compare current reward $r_t$ with the data from memory pool $\mathcal{R}$, and update the current tuple $(s, a_t, r_t, s)$ into $\mathcal{R}$ if $r_t$ is higher then the lowest reward in $\mathcal{R}$
    randomly sample a batch of data from $\mathcal{R}$ with the size of $N$, which is denoted as $(s_i, a_i, r_i, s_{i+1})$
    $y_i = r_i$
    update Q network with minimum loss:
        $$L = \frac{1}{N} \sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$$
    update policy network with sampled policy gradient
        $$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$
    update target networks:
        $$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
        $$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$
end for
```

Regarding the parameter settings for DDPG, since $t = 1$ always holds, there is no potential reward for the system in the future. Therefore, the reward attenuation coefficient is set to 0, which means that the DDPG method only makes decisions with intention of maximizing the current reward.

The constraints of the optimization problem in this paper focus on the AP active beamforming $\mathbf{B}$ and IRS passive beamforming $\mathbf{\Theta}$. To implement the constraints into DDPG, we first limit the range of output that the decision network produces to $[-1, 1]$ by utilizing the activation function tanh. Then for the part of the output action that represents $\mathbf{B}$, we use the normalization method so that the active beamforming modulus does not exceed p_0; for the part that represents $\mathbf{\Theta}$, we extend its range to $[-\pi, \pi)$ so that all the phase shifts satisfy the constraint that $|\theta_n|^2 = 1$.

To improve the stability of DDPG training, we alter the replay mechanism by updating the memory pool in a different way. The original replay mechanism updates the experience data in the memory pool sequentially, once the memory pool reaches its predetermined capacity. Ideally, as the number of iterations increases, the optimization results gradually improve, and the rewards obtained from earlier experience data are not as good as the newer data. With the sequential update of the memory pool, the experience data gradually adjusts towards higher rewards, allowing the network to learn from experience data with higher rewards and improve its effect naturally. However, in actual training, due to the instability of network approximation, the training can converge to sub-optimal or even poor results, which cannot be avoided completely. To address this issue, we maintain a memory pool that preserves better experiences. This memory pool continuously saves a batch of data with the best results in the training history, and judges whether the result of the current iteration can enter the memory pool by comparing its reward with the lowest reward in the pool. If the training result is better than the worst data in the existing memory pool, the current result will overwrite the worst data; otherwise, it will not enter the memory pool. The update mechanism of this memory pool ensures that the network can always learn from better experiences, so as to achieve better optimization results.

To accelerate the convergence of DDPG and achieve higher performance, we introduce TAO into the training process. Before the training, we use TAO to obtain an initial optimization result, along with its corresponding state and action, which are recorded as a reference state transition tuple $(s_{ref}, a_{ref}, r_{ref}, s_{ref})$. This tuple can be used as a reference for training data and is compared with the current training result in each iteration. If the current reward is not better than the reward of the reference data, the reference tuple will be updated into the memory pool, so as to improve the quality of the experiences. In the pre-optimized stage, TAO does not need to traverse a lot of random seeds as Section III, as only a better reference result is needed for DDPG, therefore the optimizing time is still a lot shorter than the TAO.

By improving the replay mechanism and introducing TAO, the stability and converging speed of the proposed TAO-DDPG is better than its original version, making it more suitable to deal with the optimization problem in this paper. The whole algorithm is shown in Algorithm 2.

5. Numerical Results

In this section, we demonstrate numerical results to evaluate the performance of the proposed algorithms in this paper. The simulated system settings are illustrated in Fig. 3, where the AP is assumed to be located at (0m, 0m), and the IRS is at (50m, 0m) in a space rectangular coordinate system. The users are uniformly distributed on a circle vertical to the x-axis with a radius of 2m. The channels follow the log-distance propagation model. The number of AP antennas M is set to 4, the power efficiency of IRS η is 0.8, and the noise power for each user is uniformly set to -80 dBm.

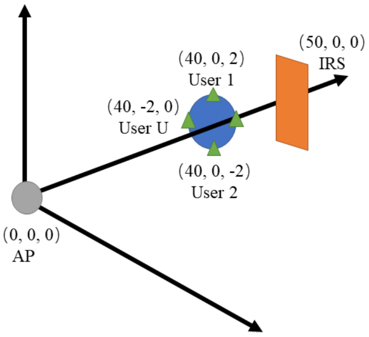
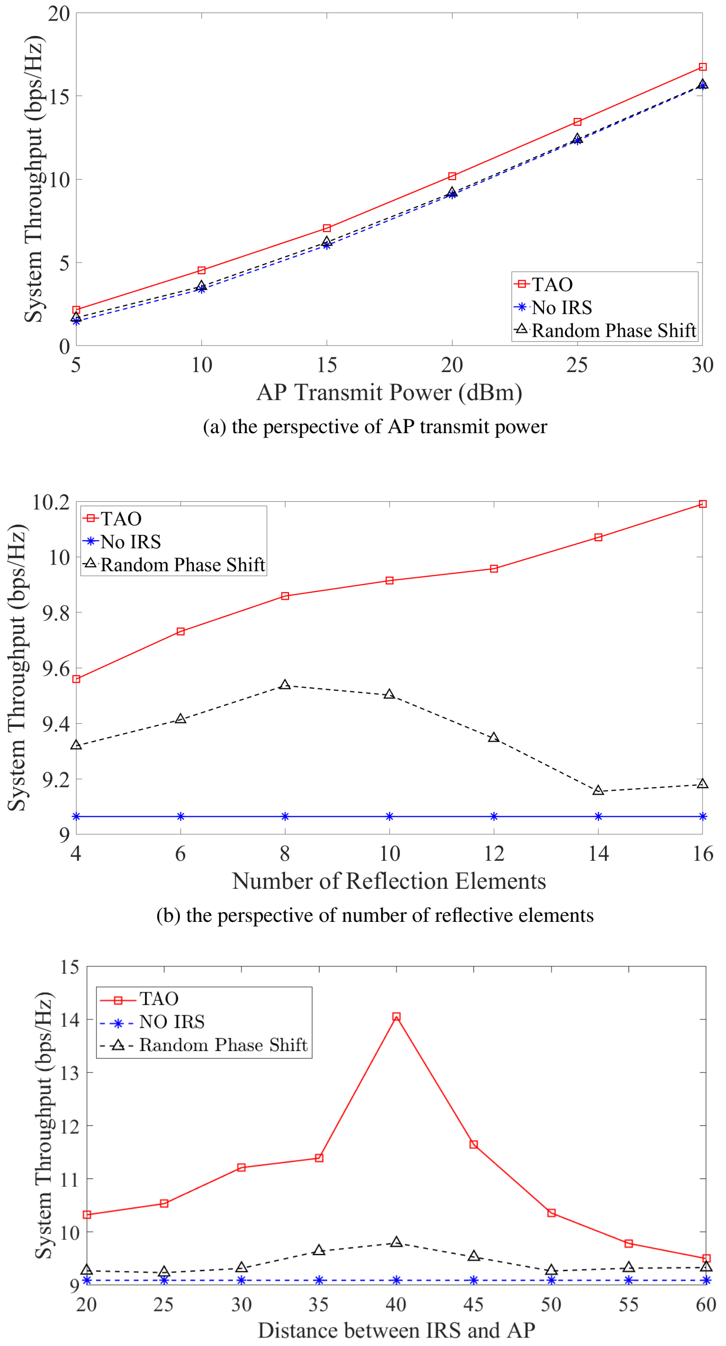


Figure 3: Settings of Simulated System

5.1. *Comparison between TAO and Common IRS Schemes*

We demonstrate the feasibility of TAO by comparing its performance with two other common schemes in wireless communication systems when confronted with IRS: random phase shift and no IRS. The random phase shift scheme means the phase shifts of reflective elements on IRS are randomly set without further adjustments. No IRS means the system excludes IRS, and the AP directly transmits signals to users. We compare the performance of these schemes from two perspectives:AP transmit power and the number of reflective elements.

Fig. 4a compares the performance of three schemes in terms of AP transmit power. The number of reflective elements N is set as 16, the number of users U is 4, and the energy cost of each reflective element is $4\mu W$. It is observed that the system throughput escalates with the increase of AP transmit power under all schemes. This is reasonable since higher AP transmit power enables the system to transmit more information faster. As for the comparison between the schemes, it is clear that TAO brings better system throughput than no IRS and random phase shift regardless of the change of AP transmit power. This demonstrates the superiority of TAO and its feasibility. In Fig. 4b, the performance of three schemes is compared in terms of the number of reflective elements. The AP transmit



(a) the perspective of AP transmit power

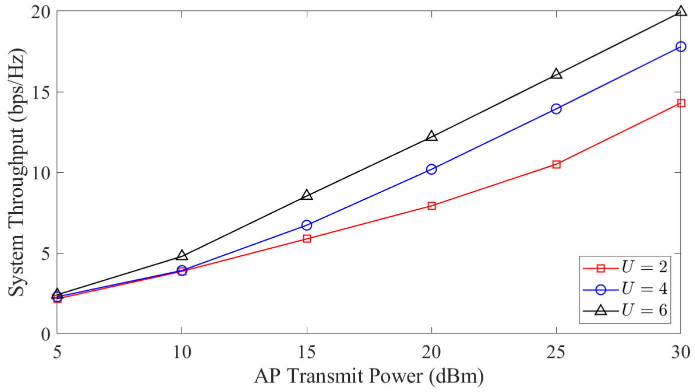
(b) the perspective of number of reflective elements

(c) the perspective of distance between IRS and AP

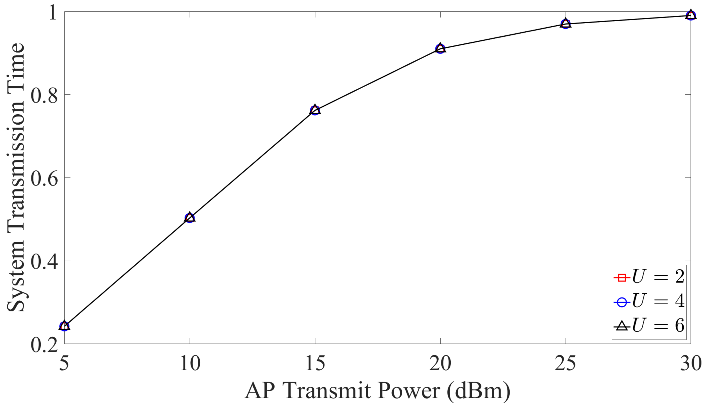
Figure 4: Comparison of Performance between Different Schemes

power p_0 is fixed at 20 dBm, the number of users U is still 4, and the energy cost of each reflective element is $4\mu W$ as well. It is observed that with the increase of the number of reflective elements, the system throughput under TAO rises in a stable way, while random phase shift shows large fluctuations, and no IRS remains constant. This means that TAO can effectively utilize the change in the number of reflective elements and maximize the benefits of IRS in assisting the system transmission. As for the comparison between the schemes, it is clear that TAO brings better system throughput than no IRS and random phase shift, regardless of the change in the number of reflective elements. This further proves the feasibility of TAO in optimizing wireless powered IRS-assisted multi-user MISO system. In Fig. 4c, the effect of IRS deployment on the system throughput for the three schemes is illustrated. The positions of the AP and users are the same as in Fig. 3. The IRS is positioned in the same line with the AP and users, with 16 reflective elements, and the energy cost of each reflective element is $4\mu W$. The AP transmit power p_0 is set to 20 dBm as well. It can be observed that when the IRS is deployed at different positions, the system throughput under the TAO scheme is better than that of both the no IRS and random phase shift schemes. As the IRS moves away from the AP and closer to the users, its improvement on the system throughput becomes more significant, indicating that the system performance can be optimized by phase adjustment. However, as the IRS moves further away from the AP, the improvement of the IRS on the system throughput becomes less effective, and there is no significant improvement at a distance of 60 between the IRS and the AP. The overall system throughput variation trend of the random phase shift scheme is similar to that of the TAO scheme, but the optimization effect of the TAO scheme is more pronounced.

To investigate the impact of the number of users on the system, we compare the performance after TAO's optimization with different numbers of users from the perspective of AP transmit power. Both system throughput and system transmission time are observed in the simulation. The number of users U varies in the set {2, 4, 6}. AP transmit power p_0 is set as 20 dBm, the number of reflective elements N is 16, and the energy cost of each reflective element is $4\mu W$.



(a) system throughput



(b) system transmission time

Figure 5: Impact of the Number of Users on System Performance

Fig. 5a depicts the system throughput with different numbers of users. It shows that with given AP transmit power, the system throughput escalates as the number of users rises. With the increase of AP transmit power, the gaps between different U enlarges. This implies that the proposed operating model with TAO's optimization enables the system to support the increase of the number of users. On the other hand, the gap between $U = 4$ and $U = 6$ is smaller compared to the gap between $U = 2$ and $U = 4$. This shows that when U is large, the performance improvement caused by the increase of U is limited due to the small size of IRS. According to the results shown in Fig. 5b, the system transmission time is compared with different numbers of users, and it is observed that with a given AP transmit power, the system transmission time remains constant as the number of users increases. This indicates that the number of users has no influence on the system transmission time, which is consistent with the previous deduction on system transmission time t_{IAT}. Furthermore, with the increase in the AP transmit power, the system transmission time increases and approaches a value of 1. This phenomenon can be attributed to the fact that as the AP transmit power increases, the time required for the IRS to harvest enough energy for its operation decreases, and hence, the IRS can complete charging faster, which increases the time the system can transmit signals accordingly. At high AP transmit power levels, the proportion of the system transmission time approaches its theoretical upper bound, which is 100%. However, in practical applications, it is important to consider the charging requirements of the IRS itself, and therefore, this upper bound may not be achievable.

5.2. *Comparison between TAO-DDPG and TAO*

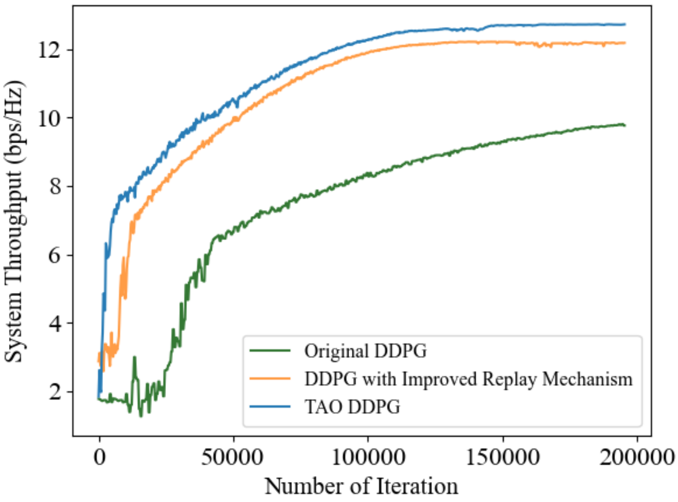
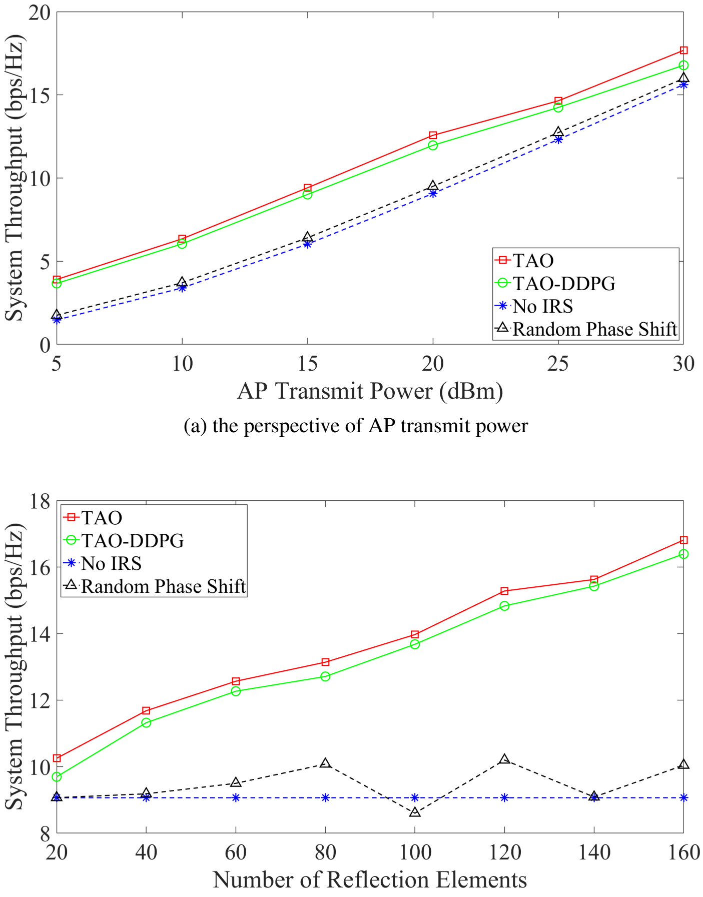


Figure 6: Feasibility of TAO-DDPG

To demonstrate the feasibility of the TAO-DDPG, we compare the performance of the original DDPG, the DDPG with an improved replay mechanism, and the DDPG with the introduction of TAO. We set the AP transmit power p_0 to 20 dBm, the number of reflective elements N to 60, the number of users U to 4, and the energy cost of each reflective element to $4\mu W$.

Fig. 6 depicts the performance comparison of three aforementioned algorithms. The original DDPG exhibited oscillations in the early stages of iterations and failed to obtain stable training results. The reward rose steadily after about 25,000 iterations, lagging compared to the other two algorithms, and eventually converged to a slightly lower result. This finding indicates that the original replay mechanism cannot ensure stable DDPG training and converged more slowly when applied to the problem in this paper. For the DDPG with the improved replay mechanism, after exhibiting short-term oscillations in the early stage, the training curve started to rise relatively steadily after about 10,000 iterations, and the final convergence result was slightly better than that of the original DDPG. The oscillation in the early stage is considered inevitable due to the accumulation of experiences during the initial iterations of DDPG with the introduction of randomness, while the stable rise of the curve in the later stage and the marked improvement in the final result confirm the effectiveness of the improved replay mechanism. The finding that TAO-DDPG started to steadily improve in fewer iterations after the introduction of TAO implies that the introduction of TAO enabled DDPG to update the memory pool with better training records earlier, allowing the Q network and the action value network to learn from better experiences and make better decisions, and the reward obtained can increase stably earlier. On the other hand, throughout the entire training process, TAO-DDPG consistently outperforms DDPG without TAO, which confirms that the introduction of TAO significantly improves the performance of DDPG, making it more suitable for solving the optimization problem in this paper.



(a) the perspective of AP transmit power

(b) the perspective of the number of reflective elements

Figure 7: Comparison between TAO and TAO-DDPG

We further compare the performance between the TAO-DDPG and TAO to demonstrate that the TAO-DDPG could achieve similar performance with TAO. Same as the previous simulation, we carry out the comparison of performance between TAO-DDPG and TAO based on AP transmit power and the number of reflective elements.

Fig. 7a presents the comparison of system throughput in terms of AP transmit power. The number of reflective elements N is set as 60, the number of users U is 4, and the energy cost of each reflective element is $4\mu W$. The optimization results of TAO-DDPG show a steady increase in system throughput with the increase of AP transmit power and outperform the random phase shift and no IRS cases. Although the performance of TAO-DDPG is slightly lower than that of TAO, the overall result is close to TAO, achieving up to 97% of TAO's performance.

From Fig. 7b, we can clearly observe that the comparison of system throughput in terms of the number of reflective elements. AP transmit power p_0 is set as 20 dBm, the number of users U is 4, and the energy cost of each reflective element is $4\mu W$. The optimization results of TAO-DDPG steadily increase with the increase of the number of reflective elements, achieving up to 98% of TAO's performance. On the other hand, the performance of TAO-DDPG is significantly better than the random phase shift and no IRS cases, demonstrating the effectiveness of

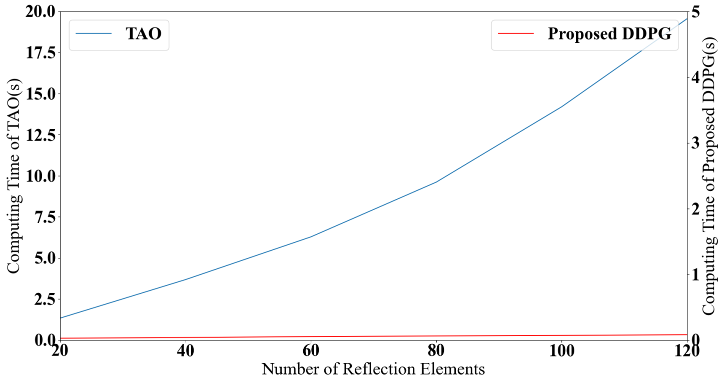


Figure 8: Computing Time of Single Iteration of TAO and TAO-DDPG

TAO-DDPG for the optimization problem in this paper. Additionally, the performance of random phase shift under a large size IRS is no longer stably higher than no IRS, indicating that a large size IRS does not necessarily improve system performance, and IRS passive beamforming must be optimized to assist transmission effectively.

In order to demonstrate that TAO-DDPG can significantly reduce the optimizing time while providing similar performance to TAO, especially in large size IRS scenarios, we compare the computing time spent on a single iteration between the two methods. Tab. 1 displays the time for a single iteration in terms of the number of reflective elements. We set AP transmit power p_0 as 20 dBm, the number of users U is 4, and the energy cost of each reflective element is $4\mu W$. The results show that as the number of reflective elements increases, the computing time of both TAO and TAO-DDPG increases, but with a significant difference in magnitude. For the same number of reflective elements, the computing time of TAO for a single iteration is dozens or even hundreds of times that of TAO-DDPG, which means TAO-DDPG can reduce the optimizing time up to 80% compared to TAO.

Table 1: Comparison of Computing Time Spent on Single Iteration between TAO and TAO-DDPG

number of reflective elements	TAO	TAO-DDPG
20	1.333s	0.028s
40	3.676s	0.038s
60	6.276s	0.053s
80	9.607s	0.062s
100	14.190s	0.070s
120	19.547s	0.081s

To provide a more intuitive representation of the difference in computing time spent on a single iteration between the two methods, Fig. 8 is presented. The left ordinate represents the computing time of TAO, and the right ordinate represents TAO-DDPG. As shown in the figure, with the increase of the number of reflective elements, the computing time for a single iteration of TAO increases significantly, while TAO-DDPG maintains a rather low level of computing time, and the gap becomes more obvious with the increase of the number of reflective elements. These results confirm that TAO-DDPG is an effective method for reducing the optimizing time in large size IRS scenarios.

6. Conclusion

In this paper, we focused on the optimization of a multi-user MISO system assisted by IRS, while integrating WPT to provide energy for the IRS. To address the system throughput maximization problem, we designed an operating model based on TS and proposed TAO. Additionally, we proposed TAO-DDPG to reduce the optimizing time in large size IRS scenarios, which improved the replay mechanism of DDPG while combining TAO to boost its convergence.

The numerical results demonstrate that TAO significantly improves the system performance compared to common IRS schemes. The system throughput escalates with the increase of AP transmit power and the number of reflective elements when TAO is applied. Meanwhile, the system performance improves as the number of users increases. Furthermore, TAO-DDPG achieves similar performance to TAO while significantly shortening the optimizing time, makeing it more suitable to optimize system throughput when dynamic changes occur at intervals in the system with large size IRS.

Acknowledgements

This work was funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 824019 and No 101022280, Horizon Europe MSCA programme under grant agreement number 101086228, and EPSRC with RC Grant reference EP/Y027787/1. For the purpose of open access, the author has applied a CC BY licence to any Author Accepted Manuscript arising from this submission.

References

[1] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, Y.-C. Liang, Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey, IEEE Communications Surveys & Tutorials 22 (4) (2020) 2283–2314. doi:10.1109/COMST.2020.3004197.

[2] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. D. Renzo, M. Debbah, Holographic mimo surfaces for 6g wireless networks: Opportunities, challenges, and trends, IEEE Wireless Communications 27 (5) (2020) 118–125. doi:10.1109/MWC.001.1900534.

[3] E. Bjarnson, A. Azdogan, E. G. Larsson, Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?, IEEE Wireless Communications Letters 9 (2) (2020) 244–248. doi:10.1109/LWC.2019.2950624.

[4] S. Zhou, W. Xu, K. Wang, M. Di Renzo, M.-S. Alouini, Spectral and energy efficiency of irs-assisted miso communication with hardware impairments, IEEE Wireless Communications Letters 9 (9) (2020) 1366–1369. doi:10.1109/LWC.2020.2990431.

[5] H. Han, J. Zhao, W. Zhai, Z. Xiong, D. Niyato, M. Di Renzo, Q.-V. Pham, W. Lu, K.-Y. Lam, Reconfigurable intelligent surface aided power control for physical-layer broadcasting, IEEE Transactions on Communications 69 (11) (2021) 7821–7836. doi:10.1109/TCOMM.2021.3104871.

[6] G. Yang, X. Xu, Y.-C. Liang, M. D. Renzo, Reconfigurable intelligent surface-assisted non-orthogonal multiple access, IEEE Transactions on Wireless Communications 20 (5) (2021) 3137–3151. doi:10.1109/TWC.2020.3047632.

[7] M. Fu, Y. Zhou, Y. Shi, K. B. Letaief, Reconfigurable intelligent surface empowered downlink non-orthogonal multiple access, IEEE Transactions on Communications 69 (6) (2021) 3802–3817. doi:10.1109/TCOMM.2021.3066587.

[8] J. Zhu, Y. Huang, J. Wang, K. Navaie, Z. Ding, Power efficient irs-assisted noma, IEEE Transactions on Communications 69 (2) (2021) 900–913. doi:10.1109/TCOMM.2020.3029617.

[9] B. Zheng, Q. Wu, R. Zhang, Intelligent reflecting surface-assisted multiple access with user pairing: Noma or oma?, IEEE Communications Letters 24 (4) (2020) 753–757. doi:10.1109/LCOMM.2020.2969870.

[10] C. Huang, A. Zappone, M. Debbah, C. Yuen, Achievable rate maximization by passive intelligent mirrors, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 3714–3718. doi:10.1109/ICASSP.2018.8461496.

[11] H. Guo, Y.-C. Liang, J. Chen, E. G. Larsson, Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks, IEEE Transactions on Wireless Communications 19 (5) (2020) 3064–3076. doi:10.1109/TWC.2020.2970061.

[12] M. Jung, W. Saad, M. Debbah, C. S. Hong, On the optimality of reconfigurable intelligent surfaces (riss): Passive beamforming, modulation, and resource allocation, IEEE Transactions on Wireless Communications 20 (7) (2021) 4347–4363. doi:10.1109/TWC.2021.3058366.

[13] M.-M. Zhao, Q. Wu, M.-J. Zhao, R. Zhang, Intelligent reflecting surface enhanced wireless networks: Two-timescale beamforming optimization, IEEE Transactions on Wireless Communications 20 (1) (2021) 2–17. doi:10.1109/TWC.2020.3022297.

[14] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, M.-S. Alouini, Asymptotic max-min sinr analysis of reconfigurable intelligent surface assisted miso systems, IEEE Transactions on Wireless Communications 19 (12) (2020) 7748–7764. doi:10.1109/TWC.2020.2986438.

[15] M. Hua, Q. Wu, Joint dynamic passive beamforming and resource allocation for irs-aided full-duplex wpcn, IEEE Transactions on Wireless Communications 21 (7) (2022) 4829–4843. doi:10.1109/TWC.2021.3133491.

[16] L. R. Varshney, Transporting information and energy simultaneously, in: 2008 IEEE International Symposium on Information Theory, 2008, pp. 1612–1616. doi:10.1109/ISIT.2008.4595260.

[17] H. Xing, L. Liu, R. Zhang, Secrecy wireless information and power transfer in fading wiretap channel, IEEE Transactions on Vehicular Technology 65 (1) (2016) 180–190. doi:10.1109/TVT.2015.2395725.

[18] W. Lu, P. Si, X. Liu, B. Li, Z. Liu, N. Zhao, Y. Wu, Ofdm based bidirectional multi-relay swipt strategy for 6g iot networks, China Communications 17 (12) (2020) 80–91. doi:10.23919/JCC.2020.12.006.

[19] R. Zhang, C. K. Ho, Mimo broadcasting for simultaneous wireless information and power transfer, IEEE Transactions on Wireless Communications 12 (5) (2013) 1989–2001. doi:10.1109/TWC.2013.031813.120224.

[20] Q. Wu, R. Zhang, Weighted sum power maximization for intelligent reflecting surface aided swipt, IEEE Wireless Communications Letters 9 (5) (2020) 586–590. doi:10.1109/LWC.2019.2961656.

[21] Y. Tang, G. Ma, H. Xie, J. Xu, X. Han, Joint transmit and reflective beamforming design for irs-assisted multiuser miso swipt systems, in: ICC 2020 - 2020 IEEE International Conference on Communications (ICC), 2020, pp. 1–6. doi:10.1109/ICC40277.2020.9148892.

[22] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, L. Hanzo, Intelligent reflecting surface aided mimo broadcasting for simultaneous wireless information and power transfer, IEEE Journal on Selected Areas in Communications 38 (8) (2020) 1719–1734. doi:10.1109/JSAC.2020.3000802.

[23] S. Gong, J. Lin, B. Ding, D. Niyato, D. I. Kim, M. Guizani, When optimization meets machine learning: The case of irs-assisted wireless networks, IEEE Network 36 (2) (2022) 190–198. doi:10.1109/MNET.211.2100386.

[24] Y. Zou, Y. Long, S. Gong, D. T. Hoang, W. Liu, W. Cheng, D. Niyato, Robust beamforming optimization for self-sustainable intelligent reflecting surface assisted wireless networks, IEEE Transactions on Cognitive Communications and Networking 8 (2) (2022) 856–870. doi:10.1109/TCCN.2021.3133839.

[25] K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, Deep reinforcement learning: A brief survey, IEEE Signal Processing Magazine 34 (6) (2017) 26–38. doi:10.1109/MSP.2017.2743240.

[26] M. Shehab, B. S. Ciftler, T. Khattab, M. M. Abdallah, D. Trinchero, Deep reinforcement learning powered irs-assisted downlink noma, IEEE Open Journal of the Communications Society 3 (2022) 729–739. doi:10.1109/OJCOMS.2022.3165590.

[27] J. Zhang, H. Zhang, Z. Zhang, H. Dai, W. Wu, B. Wang, Deep reinforcement learning-empowered beamforming design for irs-assisted miso interference channels, in: 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), 2021, pp. 1–5. doi:10.1109/WCSP52459.2021.9613575.

[28] R. Huang, V. W. S. Wong, Joint user scheduling, phase shift control, and beamforming optimization in intelligent reflecting surface-aided systems, IEEE Transactions on Wireless Communications 21 (9) (2022) 7521–7535. doi:10.1109/TWC.2022.3159187.

[29] B. Lyu, P. Ramezani, D. T. Hoang, S. Gong, Z. Yang, A. Jamalipour, Optimized energy and information relaying in self-sustainable irs-empowered wpcn, IEEE Transactions on Communications 69 (1) (2021) 619–633. doi:10.1109/TCOMM.2020.3028875.

[30] Y. Cao, T. Lv, W. Ni, Intelligent reflecting surface aided multi-user mmwave communications for coverage enhancement, in: 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, 2020, pp. 1–6. doi:10.1109/PIMRC48278.2020.9217160.