Reinforcement Learning for Antenna Selection and Optimization of Irregular Reconfigurable Intelligent Surfaces

Emmanuel Obeng Frimpong*, Zhi Tian*, Yue Wang†
*Electrical and Computer Engineering Department, George Mason University, Fairfax, VA USA
†Department of Computer Science, Georgia State University, Atlanta, GA USA

***Abstract*—This paper presents a new reinforcement learning approach to the design and optimization of irregular reconfigurable intelligent surface (IRIS) for downlink communications in 6G multiuser wireless systems. Under the total power constraint of the IRIS device, we formulate a sum rate maximization problem that jointly optimizes the elements selection, the phase shift and the precoding design. For this challenging problem, we develop a deep reinforcement learning technique that can approach the optimal solution at affordable complexity. Physical constraints of the design parameters are properly incorporated into the developed DRL approach. Simulation results show that our proposed algorithm is able to learn from its environment and gradually improve its performance, and also converge to better performance compared to the state-of-the-art benchmarks when implemented in large-scale antenna systems.**

***Index Terms*—Reinforcement learning, irregular reconfigurable intelligent surfaces, element selection, joint optimization.**

I. Introduction

With recent developments in programmable meta-materials, low-cost reconfigurable intelligent surfaces (RIS) have been widely considered for adoption in wireless systems to enhance system capacity and throughput [1]. An RIS is typically a uniform array that consists of a large number of reflecting elements with high-resolution phase shifters [2], and it serves as a relay in a fully passive mode. However, power consumption in adjusting the phases of all elements is non-negligible [3], which limits the size of practical RIS devices.

To collect the diversity benefits of large-size RIS while saving the power consumption, the concept of irregular RIS (IRIS) was introduced [4], which only selects a limited number of RIS elements from a large-size regular RIS structure to maximize system capacity. Joint optimization of the antenna selection and reflection beamforming design for IRIS was formulated and implemented in [4], [5]. IRIS significantly enhances the sum rate by activating elements distributed over an enlarged surface, in contrast with conventional RIS structure by packing the same number of active antennas. This line of work is optimization-based, given known channel conditions. As a result, IRIS parameters need to be re-designed whenever the channel changes, which does not adapt to dynamic environments and not meet the real-time implementation needs given high computational costs.

Artificial intelligent (AI) has been introduced in wireless communications [6]–[8], such as beamformer designs and channel estimation for large-scale MIMO systems using supervised deep learning (DL) [9], [10], and wideband spectrum sensing via attention-based and distributed DL [11], [12]. These DL approaches significantly reduce the complexity and computation time during online prediction, after offline training. However, this requires large labeled training dataset. To overcome this issue, deep reinforcement learning (DRL) provides an alternative paradigm of training deep neural network on an agent. DRL allows this agent to take actions and observe the environment so as to maximize a cumulative reward [13]–[15]. DRL-based solutions have been developed for RIS design involving all elements. In [14], the signal-to-interference-plus-noise ratio (SINR) is maximized by jointly designing the beamforming, power control and interference coordination using DRL. An RIS-aided MISO NOMA system is developed in [16], where the DRL agent selects phase shift of the RIS element to maximize the sum rate. In quest to maximize energy efficiency, DRL is adopted to jointly select the base station beamforming vector and RIS configuration [17], [18]. In [19], a DRL-RIS empowered multihop terahertz communication is proposed to jointly select both BS beamforming vector and RIS phase shifts for each of the multi-RIS involved. However, these prior works involve all RIS elements during communications, which entails high power consumption and limits practical use of large-scale RIS.

There is few work that considers DRL for IRIS. Unlike RIS, IRIS requires antenna element selection in its design, leading to an integer programming problem. Therefore direct extension of DRL-RIS methods is not applicable for IRIS. In [20], IRIS antenna element problem is solved using DRL in a separate-approach manner. A single DRL structure first selects the RIS elements, and then a signal processing (SP) based phase estimation algorithm is applied to further find the phases of these selected elements. Because an SP module is blended into the data-driven DRL model, the DRL training has to be customized to involve a predefined threshold, and the training converges only when the accumulated rewards over several episodes exceed the threshold. Setting the reward threshold low may prevent the DRL agent from fully exploring the environment to reach the optimal solution. On the other hand, setting it too high may prevent the DRL agent from converging. Hence determining an judicious threshold is challenging, and it renders this algorithm impractical. This is an inherent drawback of such a DRL structure with a hybrid SP module, which prevents to fully leverage the strength and benefit of DRL. Also, the SP-based phase shift estimation is

This work was supported in part by the National Science Foundation grants #2128596, #2231209 and #2413622.

a nonconvex optimization problem without closed-form solution. It is usually solved via greedy search or approximation, e.g. iterative majorization-minimization (MM) methods [20], which is not only computationally involved but also subject to suboptimal performance. Therefore, it is still a challenging problem to design efficient antenna selection schemes for IRIS to achieve high sum rate in real time, as the focus of this work.

This paper investigates IRIS element selection to maximize the sum rate for a typical IRIS-aided communication system utilizing DRL. Assuming full channel state information (CSI), we focus on solving the non-convex mixed integer programming problem. We propose an IRIS optimization scheme based on deep deterministic policy gradient (IRIS-DDPG) to jointly optimize the transmit beamforming, the IRIS element selection and the phase shift. Our contributions are summarized below:

- In our IRIS-DDPG, we define the reward function using the sum rate of multiple users, and design the DDPG procedure to find the optimal action (i.e., transmit beamforming, element selection and phase shift) policy in a given wireless environment. In lieu of exhaustive search, the computational bottleneck of integer programming for element selection is resolved by properly designing the deep neural network (DNN)-based action network using differentiable activation functions in the output neurons for antenna section.
- While traditional DRL applies to unconstrained optimization problems, we design the DDPG structure for DRL to tailor for the physical constraints of IRIS systems, given a total power budget. To the best of our knowledge, this is the first work to introduce a fully learning-based DRL framework to IRIS optimization, without invoking ad hoc design components such as thresholding. Such a learning-based IRIS systems can interact with the complex wireless environment and improve the performance by constantly adjusting the DDPG model parameters.
- Numerical results demonstrate that the proposed IRIS-DDPG is able to achieve the desired sum rate performance which comes closer to the optimal results by exhaustive search. Compared with existing algorithms with polynomial complexity, IRIS-DDPG offers better performance at lower computational complexity. Such advantages are particularly attractive for real-time operations of large-scale IRIS systems.

II. System model and problem formulation

Consider a downlink RIS-aided communication system where a BS equipped with M antennas communicates to K single-antenna users. An N-element RIS plays as a relay in Fig. 1, where only N_s elements are activated during communications to save power. The BS transmitter employs a precoding vector for each user and superposes the precoded symbols from all K users to form the transmitted signal:

$$\mathbf{x} = \mathbf{W}\ \mathbf{s}, \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2 \cdots \mathbf{w}_k]$ is the precoding matrix and $\mathbf{w}_k \in \mathbb{C}^{M\times 1}$ denotes the precoding vector for user k, and

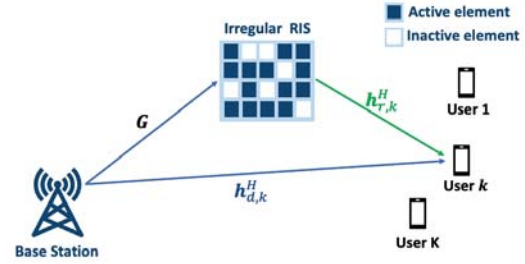

Fig. 1: A wireless communication system aided by IRIS.

$\mathbf{s} = [s_1, s_2, \cdots s_K]^T \in \mathbb{C}^{K\times 1}$ denotes the transmitted symbol vector for K users satisfying $E[\mathbf{s}\mathbf{s}^H] = \mathbf{I}_K$. The signal is transmitted to each user k through two channel paths: one direct path between the BS and user k and the other reflected channel from BS to IRIS and from IRIS to user k. To reflect antenna selection, we introduce a selection matrix $\mathbf{Z} = \text{diag}(\mathbf{z})$, where $\mathbf{z} = [z_1, \cdots, z_N]^T$ is a binary-valued indicator vector representing the activation state of the N RIS reflecting elements, that is, $z_n = 1$ if the n-th element is selected, and $z_n = 0$ otherwise, $n = 1, \ldots, N$.

The reflection coefficient matrix for IRIS N elements has

$$\mathbf{\Theta} = \text{diag}\left(\left[e^{j\theta_1}, e^{j\theta_2}, \cdots, e^{j\theta_N}\right]\right), \tag{2}$$

where $\forall n = 1 \cdots N$, and $\theta_n \in [0, 2\pi]$ represent continuous phase shift of the n-th RIS element in the IRIS. We define $\mathbf{G} \in \mathbb{C}^{N\times M}$ as the BS-RIS channel. We let $\mathbf{H}_r^H \in [\mathbf{h}_{r,1}, \mathbf{h}_{r,2} \cdots, \mathbf{h}_{r,K}]^H \in \mathbb{C}^{K\times N}$ and $\mathbf{H}_d^H \in [\mathbf{h}_{d,1}, \mathbf{h}_{d,2} \cdots, \mathbf{h}_{d,K}]^H \in \mathbb{C}^{K\times M}$ where $\mathbf{h}_{r,k}^H$ and $\mathbf{h}_{d,k}^H$ represents the channel between the RIS and user k and direct channel from BS to user k respectively. The received signal $\mathbf{y} \in \mathbb{C}^{K\times 1}$ for all K users can be expressed as

$$\mathbf{y} = (\mathbf{H}_r^H \mathbf{Z}\mathbf{\Theta}\mathbf{G} + \mathbf{H}_d^H)\mathbf{x} + \mathbf{u}, \tag{3}$$

where $\mathbf{x} \in \mathbb{C}^{M\times 1}$ is the transmitted signal at the BS in (1), $\mathbf{u} \in \mathbb{C}^{K\times 1}$ denotes the additive white Gaussian noise (AWGN) with zero mean and variance σ^2, Based on the signal model in Fig. 1, the SINR of user k is given by

$$\text{SINR}_k = \frac{|(\mathbf{h}_{r,k}^H \mathbf{Z}\mathbf{\Theta}\mathbf{G} + \mathbf{h}_{d,k}^H)\mathbf{w}_k|^2}{\sum_{i\neq k}^{K} |(\mathbf{h}_{r,k}^H \mathbf{Z}\mathbf{\Theta}\mathbf{G} + \mathbf{h}_{d,k}^H)\mathbf{w}_i|^2 + \sigma^2}, k = 1, \ldots, K. \tag{4}$$

A. Problem Formulation

In this paper, our aim is to maximize the sum rate of all users by jointly optimizing the element selection $\mathbf{Z}$, the corresponding phases $\mathbf{\Theta}$ and the precoding matrix $\mathbf{W}$ of the IRIS-aided system. The transmit power at the BS is given by $P_t = \sum_{k=1}^{K} \|\mathbf{w}_k\|_2^2$, which follows the allowable power budget,

that is, it cannot be larger than the maximum transmit power $P_{\max}$. We formulate the sum rate maximization problem as

$$(\mathbf{P1}): \max_{\mathbf{Z},\mathbf{W},\mathbf{\Theta}} \; R_s = \sum_{k=1}^{K} \log_2(1+\text{SINR}_k), \tag{5}$$

$$\text{s.t., } (\mathbf{C1}): P_t \leq P_{\max}, \tag{6}$$

$$(\mathbf{C2}): \theta_n \in [0, 2\pi], \quad \forall n = 1, 2, \cdots, N, \tag{7}$$

$$(\mathbf{C3}): z_n \in \{1, 0\}, \quad \forall n = 1, 2, \cdots, N, \tag{8}$$

$$(\mathbf{C4}): \mathbf{1}^T \mathbf{z} = N_s. \tag{9}$$

Here (**C1**) depicts the transmission power constraint, (**C2**) reflects the continuous-valued phase-shift range, and (**C3**) and (**C4**) denote the antenna topology constraints that there are N_s ones (i.e., activated elements) and $N - N_s$ zeros (i.e., deactivated elements) in the topology matrix $\mathbf{Z}$.

The problem (**P1**) is a mixed integer programming problem due to the binary-valued vector $\mathbf{z}$. Finding the optimal solution entails exhaustive search over all 2^N possible values for $\mathbf{z}$, which is inefficient especially for large-scale cases with large N. In this paper, we opt to solving this challenging optimization problem by reformulating it in the context of advanced DRL method to obtain computationally feasible solutions to $\mathbf{Z}, \mathbf{W}, \mathbf{\Theta}$.

III. Deep Reinforcement Learning Empowered IRIS

This section starts from DRL and DDPG which are the foundation and the enabling techniques to our IRIS algorithm.

A. Fundamentals of DRL

RL is a learning framework where an agent gradually makes the best decision by interacting with the environment – performing actions in the environment, observing the instant rewards and the transitions of the state in the environment.

State: Let $\mathcal{S}$ denote the set of all possible states describing the environment. The state $s^{(t)} \in \mathcal{S}$ is the observation at time t.

Action: We use $\mathcal{A}$ to represent the set of actions. Action is a set of options that an agent takes to transition between states of the environment. At time t, once the agent performs action $a^{(t)} \in \mathcal{A}$ following a policy π, the current state $s^{(t)}$ transits to next state $s^{(t+1)}$ and the agent gets rewards $r^{(t)}$.

State transition probability: Transitioning between states is usually random and the environment is the source of randomness. The transition probability from state s to s' after taking action a is $P^a_{ss'} = Pr(s^{(t+1)} = s' | s^{(t)} = s, a^{(t)} = a)$.

Reward: A value rewarded to the agent after an action is taken. At a given time t, reward $r^{(t)}$ shows how good action $a^{(t)}$ is given state $s^{(t)}$.

Experience buffer: Over episodes of the agent's interaction with the environment, its experience is stored in a buffer as a collection of the quadruplets $(s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)})$, $\forall t$, which are used for training.

The agent aims an optimal policy to maximize the cumulative reward

$$R^{(t)} = \sum_{\tau=0}^{\infty} \gamma^\tau r^{(t+\tau+1)}, \tag{10}$$

where $\gamma \in [0, 1]$ is the discount rate. To this end, Q-learning, a model-free RL algorithm, can be used to find the optimal action-selection policy [21]. To assess an action under the current state, the Q function defines the expected reward as

$$Q_\pi(s^{(t)}, a^{(t)}) = E_\pi[R^{(t)} | s^{(t)} = s, a^{(t)} = a]. \tag{11}$$

For a huge state-action space, a function approximator is used to obtain optimal $Q^*(s^{(t)}, a^{(t)})$.

B. Deep Deterministic Policy Gradients - DDPG

For continuous action space, DDPG as an actor-critic DRL [22], is adopted in this work due to its ability to stabilize the learning process and provides a more efficient approach for learning in complex environments. DDPG has both actor and critic architectures [13]. The actor network learns the optimal policy to choose actions, while the critic network evaluates the state-action pairs using Q function. Due to the huge state-action space, DNN has been introduced to approximate both the Q function and the action. With DRL, the Q function is:

$$Q(s(t), a(t)) = Q_\theta(s(t), a(t)), \tag{12}$$

where θ is the weight parameters of DNN and will be updated by gradient descent:

$$\theta^{(t+1)} \triangleq \theta^{(t)} - \mu \Delta_\theta \mathbb{L}(\theta), \tag{13}$$

where μ is the learning rate for the update on θ and Δ_θ is the gradient of the loss $\mathbb{L}(\theta)$ with respect to θ. The loss function is the difference between the NN's predicted value and the actual target value. In RL, the actual target value is unknown. To address this problem, DDPG introduces two NNs with identical architectures. The training NN and the target NN with value functions $Q(\theta^{(train)} | s^{(t)}, a^{(t)})$ and $Q(\theta^{(target)} | s^{(t)}, a^{(t)})$ respectively. The actual target value is estimated as

$$y = r^{(t)} + \gamma \max_{a'} Q(\theta^{(target)} | s^{(t+1)}, a'). \tag{14}$$

The loss function is given as

$$\mathbb{L}(\theta) = (y - Q(\theta^{(train)} | s^{(t)}, a^{(t)}))^2. \tag{15}$$

In DDPG, the actor takes state as input and output an action, which together with the state is fed as input to the critic. The critic then calculates the Q value which is used to evaluate the performance of the current action. The training critic network (c, train) is updated by

$$\theta^{(t+1)}_{c,train} = \theta^{(t)}_{c,train} - \mu_{c,train} \Delta \mathbb{L}(\theta^{(t)}_{c,train}), \tag{16}$$

$$\begin{aligned} \mathbb{L}(\theta^{(t)}_{c,train}) = (r^{(t)} + \gamma q(\theta^{(t)}_{c,target} | s^{(t+1)}, a') \\ - q(\theta^{(t)}_{c,train} | s^{(t)}, a^{(t)}))^2, \end{aligned} \tag{17}$$

where $\mu_{c,train}$ is the learning rate for the update on training critic network. a' is the action output from the target actor network and $\Delta \mathbb{L}(\theta^{(t)}_{c,train})$ denotes the gradient with respect to the training critic network $\theta_{c,train}$. The training and target critic network, $\theta_{c,train}$ and $\theta_{c,target}$ respectively. The update on the training actor network (a, train) is given as

$$\begin{aligned} \theta^{(t+1)}_{a,train} =& \theta^{(t)}_{a,train} \\ & - \mu_{a,train} \Delta q(\theta^{(t)}_{c,target} | s^{(t)}, a^{(t)}) \Delta \pi(\theta^{(t)}_{a,train} | s^{(t)}), \end{aligned} \tag{18}$$

where $\mu_{a,train}$ denotes the learning rate for the training actor network. $\Delta\pi(\theta^{(t)}_{a,train}|s^{(t)})$ is the gradient of the training actor network with respect to its parameters $\theta^{(t)}_{a,train}$. The gradient of the target critic network with respect to the action is given by $\Delta q(\theta^{(t)}_{c,target}|s^{(t)}, a^{(t)})$. The target network are updated after a specified time interval O by synchronizing it with the training network, which is actively trained in each iteration.

$$\theta_{c,target} \leftarrow \tau_c\theta_{c,train} + (1-\tau_c)\theta_{c,target}, \tag{19}$$

$$\theta_{a,target} \leftarrow \tau_a\theta_{a,train} + (1-\tau_a)\theta_{a,target}, \tag{20}$$

where τ_c, τ_a are the soft update rate of the target critic network and the target actor network respectively. This soft update ensures stability and convergence during training.

C. IRIS-DDPG

In this section, we discuss our proposed IRIS-DDPG algorithm. The key steps are to properly define the state, actions and rewards for the IRIS systems at hand, and design the double DNNs that can effectively address that physical constraints in our optimization formulation (P1). Note that the standard DDPG is designed for unconstrained problems only.

State: In this work, we define $s^{(t)}$ to be the transmit power and the received power of all users at the t^{th} time step as:

$$s^{(t)} = \left(\left\{P^{(t)}_{\text{Tx},k}\right\}_{k=1}^{K}, \left\{P^{(t)}_{\text{Rx},k}\right\}_{k=1}^{K}\right), \tag{21}$$

where $P^{(t)}_{\text{Tx},k}$ is the transmit power for user k at time t given by $P^{(t)}_{\text{Tx},k} = \|\mathbf{w}_k^H\mathbf{w}_k\|^2$ and $P^{(t)}_{\text{Rx},k}$ is the received power for user k at time t given by $P^{(t)}_{\text{Rx},k} = |(\mathbf{h}^H_{r,k}\mathbf{Z}\boldsymbol{\Theta}\mathbf{G} + \mathbf{h}^H_{d,k})|^2$.

Action: We define the action to include the IRIS struture, corresponding phases and the precoding design:

$$a^{(t)} = \left\{\mathbf{Z}^{(t)}, \boldsymbol{\Theta}^{(t)}, \mathbf{W}^{(t)}\right\}. \tag{22}$$

It is important to note that the action space of (22) should be defined to obey constraints (6) – (9).

Reward: Given the instantaneous channels $\mathbf{G}, \mathbf{h}_{r,k}, \mathbf{h}_{d,k}\ \forall k$ and the action $\mathbf{W}^{(t)}$, $\mathbf{Z}^{(t)}$ and $\boldsymbol{\Theta}^{(t)}$, we compute the sum rate R_s (5) as the reward. For the output of the critic networks, we define the reward function as:

$$r = \begin{cases} R_s, & P_t = \text{Tr}(WW^H) \le P_{\max} \quad \text{(23a)} \\ R_s - C, & P_t = \text{Tr}(WW^H) > P_{\max} \quad \text{(23b)} \end{cases}$$

where C is a large value to penalize any violation of the power constraint, say $C = 100$.

D. IRIS-DDPG DNN Architecture

The DNN structures of the actor and critic network are fully connected DNNs, consisting of one input layer, one output and 3 hidden layers as shown in Fig. 2. The input and output dimension of the actor network is the cardinality of the state and action respectively. Specifically, the numbers of tunable elements in $\mathbf{Z}, \boldsymbol{\Theta}, \mathbf{W}$ are N, N and MK, respectively, which correspond to $(2N + MK)$ total neurons at the output layer of the action network. For the critic networks, the output layer has one neuron to yield the Q value, which is based on the reward function defined in (23) and the Q value function (11), an indicator to evaluate the performance of current action.

A key step in our DRL design is to holistically confine the agent within its constrained action space. Tailoring to the specific constraints on design parameters of IRIS, we construct the action and critic multi-layer neural networks as in Fig. 2. First, to satisfy the power constraint (6) on $\mathbf{W}$, we employ linear activation functions for the MK output neurons corresponding to $\mathbf{W}$. Violation of (6) will be panelized in the reward function (23). Then, the N RIS elements along the diagonal of Z are chosen based on softmax algorithm. Specifically, we select the highest N_s probabilities and the rest of $N - N_s$ are the RIS element not selected, to satisfy constraints (8) and (9). Lastly, the corresponding N phases θ of RIS element are chosen according to sigmoid algorithm and then multiplied by 2π to satisfy constraint (7).

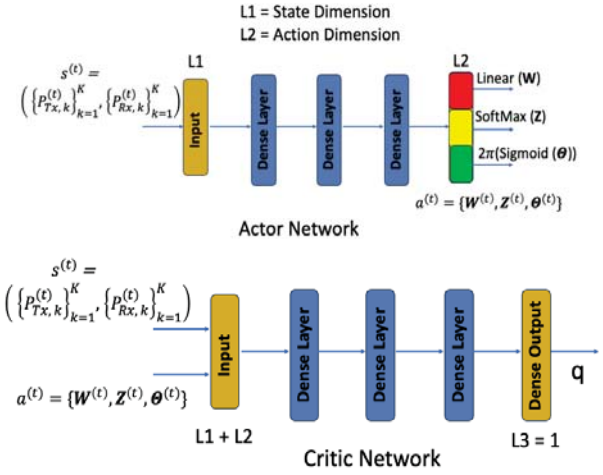

Fig. 2: The actor and critic networks in IRIS-DDPG.

The optimizer used for both the training critic network and training actor network is Adam optimizer with adaptive learning rate $\mu_c^{(t)} = \lambda_c\mu_c^{(t-1)}$ and $\mu_a^{(t)} = \lambda_a\mu_a^{(t-1)}$, where λ_c and λ_a are the decaying rate for the training critic and training actor network. The complete training process of proposed IRIS-DDPG is summarized in Algorithm 1.

IV. Numerical Results

A. Simulation settings and benchmarks

In the IRIS systems of interest, K single-antenna users are served by a BS equipped with M antennas and an irregular RIS equipped with N elements of which N_s elements are selected. The uncorrelated Rayleigh fading channel model is adopted. The hyperparamters in our algorithm are described in table I. We consider three state-of-the-art algorithms as benchmarks: the ATS-NECE (NECE) algorithm [4], successive refinement (SR) [23] and the optimal solution via exhaustive search.

Fig. 3 depicts the achieved sum rate performances of various algorithms as a function of the total transmit power constraint,

Algorithm 1 IRIS-DDPG Algorithm

Require: $\mathbf{G}, \mathbf{h}_{r,k}, \mathbf{h}_{d,k}, \forall k$
Ensure: Action: $\mathbf{W}, \mathbf{Z}, \mathbf{\Theta}$, Reward: R, Q-value function

Initialize the experience buffer B with size D, training actor network parameter $\theta_{a,train}$, target actor network parameter $\theta_{a,target} = \theta_{a,train}$, training critic network with parameter $\theta_{c,train}$, target critic network with parameter $\theta_{c,target} = \theta_{c,train}$, transmit beamforming matrix $\mathbf{W}$, RIS element selection $\mathbf{Z}$ and phase shift matrix $\mathbf{\Theta}$
for episode $= 0, 1, \cdots, N-1$ **do**
  Collect $\mathbf{G}, \mathbf{h}_{r,k}, \mathbf{h}_{d,k}, \forall k$ to obtain first state $s^{(0)}$
  for $t = 0, 1, 2 \cdots, T-1$ **do**
    Obtain output from output layer $\theta_a^{(train)}$
    Compute Tr $(\mathbf{W}\mathbf{W}^H) = P_t$ from section of layer with linear activation
    Choose highest N_s probabilities as selected elements in $\mathbf{Z}^{(t)}$ from section of layer with softmax activation
    Choose respective phases $\mathbf{\Theta}^{(t)}$ from section of layer with sigmoid activation and multiply by 2π
    Obtain action $a^{(t)} = \mathbf{W}^{(t)}, \mathbf{Z}^{(t)}, \mathbf{\Theta}^{(t)}$
    if $P_t \leq P_{\max}$ **then**
      Compute instant reward as (23a) with $a^{(t)}$
    else
      Compute instant reward (23b) with $a^{(t)}$
    end if
    Obtain new state $s^{(t+1)}$ given action $a^{(t)}$
    Store in experience buffer $(s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)})$
    Update our network parameters by sampling random batch size U from experience buffer
    Calculate target value by (14)
    Update the training critic network $\theta_{c,train}$ by (16)
    Update the training actor network $\theta_{a,train}$ by (18)
    Update target critic network $\theta_{c,target}$ after every O steps by (19)
    Update target actor network $\theta_{a,target}$ after every O steps by (20)
  end for
end for

for M=4, N=20, N_s=10 and K=4. It shows that IRIS-DDPG outperforms the state-of-the-art and comes closest to that of the optimal exhaustive search method. In general, the sum rate increases with the transmit power.

Consider a large-scale system with M=4, N=100, N_s=50, and K=4. The large value of N makes it infeasible to simulate the exhaustive search method, which is thus dropped from the comparison. From Fig. 4, we observe that IRIS-DDPG method outperforms the NECE and SR methods, which confirms the effectiveness of IRIS-DDPG for large-scale RIS systems.

B. Computational Complexity

We analyze the complexity order of the proposed IRIS-DDPG algorithm, along with that of other benchmarks.

In the exhaustive search method, $\binom{N}{N_s}$ possible IRIS structures are searched. For each structure, we quantize to

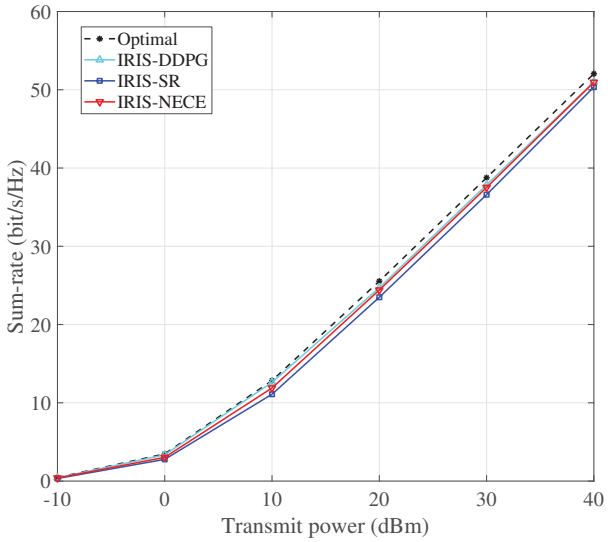

Fig. 3: Sum rate versus transmit power for small scale system with M=4, N=20, N_s=10 and K=4.

TABLE I: Hyperparamter Descriptions

Simulation Parameters	Value
Discounted rate γ	0.95
Learning rate $\mu_c, \mu_a, \tau_c, \tau_a$	0.001
Decaying rate λ_c, λ_a	10^{-5}
Experience replay buffer size D	100000
Training episode N	100
Training steps T	10000
Mini-batch size U	16
Synchronization interval O	20
Noise power σ^2	-80dBm

have up to L phase shift combinations for each N_s active elements. Hence, the complexity is on the order of $\mathcal{O}(\binom{N}{N_s} L^{N_s} K \log N^3)$, which is exceedingly high for large values of N and N_s.

The SR algorithm in [23] is an approximate search algorithm that sequentially select the antenna elements one by one in a greedy manner. The complexity is on the order of $\mathcal{O}(\epsilon N(LN_s)K \log N^3)$, $1 \leq \epsilon \leq 10$ [24].

NECE is a population-based optimization algorithm that also resorts to an approximate search strategy. Its complexity order turns out to be $\mathcal{O}(\epsilon N^2 K \log N^3)$, $1 \leq \epsilon \leq 10$ [24].

For complexity of IRIS-DDPG, L denotes the layers of the model, U_0 denotes the size of the input layer, U_l represents the size of the l-th layer. There are N_{epi} episodes and T steps per episode. Then, the whole training computation is defined as $\mathcal{O}\left(N_{epi} T \left(\sum_{l=1}^{L} U_{l-1} U_l\right)\right)$. However, the size of the hidden layers $\left(\sum_{l=2}^{L-1} U_{l-1} U_l\right)$ are constant C. The size of the input layer U_0 is $2K$ and that of the output, U_L, is $2N + MK$. Hence the complexity order turns out to be $\mathcal{O}(2K + 2N + MK + C)$. Once the actor and critic networks is trained, IRIS-DDPG can be used to adapt to different channel environments directly, without retraining. This is a huge computational advantage over signal processing

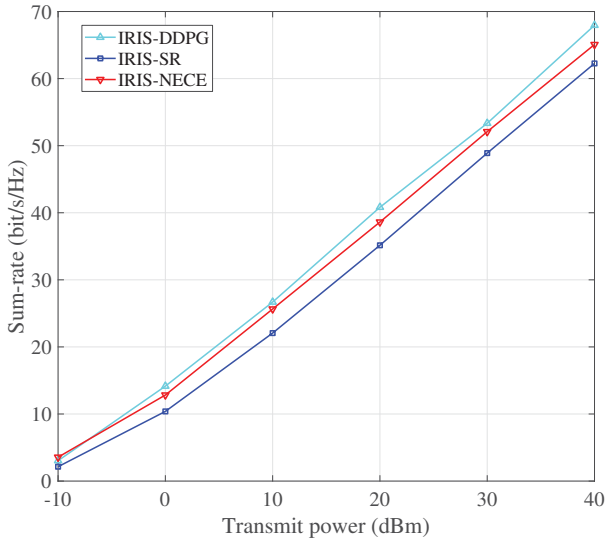

Fig. 4: Sum rate versus transmit power for large scale system with M=4, N=100, N_s=50 and K=4.

based approaches. By comparison, the complexity order of IRIS-DDPG is linear in the RIS size N, which has evident complexity advantages over the benchmarking methods.

V. Conclusion

This work develops a new joint design of transmit beamforming, RIS element selection and phase shifts based on the DRL technique. The proposed IRIS-DDPG method, by virtue of its judicious design of the embedded double DNN structures, efficiently overcomes the bottleneck of the mixed integer programming problem imposed by antenna selection. In addition, the DDPG structure is enhanced to accommodate the total power constraint of the RIS systems. Simulation results verify that the proposed IRIS-DDPG outperforms the start-of-the-art methods in terms of both sum-rate performance and computational complexity, making it attractive for high-data-rate wireless systems with large-scale RIS.

References

[1] S. Hu, F. Rusek, and O. Edfors, "Beyond massive mimo: The potential of positioning with large intelligent surfaces," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1761–1774, 2018.

[2] R. Méndez-Rial, C. Rusu, N. González-Prelcic, A. Alkhateeb, and R. W. Heath, "Hybrid mimo architectures for millimeter wave communications: Phase shifters or switches?," *IEEE access*, vol. 4, pp. 247–267, 2016.

[3] C. Huang, G. C. Alexandropoulos, A. Zappone, M. Debbah, and C. Yuen, "Energy efficient multi-user miso communication using low resolution large intelligent surfaces," in *2018 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, IEEE, 2018.

[4] R. Su, L. Dai, J. Tan, M. Hao, and R. MacKenzie, "Capacity enhancement for irregular reconfigurable intelligent surface-aided wireless communications," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6, IEEE, 2020.

[5] J.-C. Chen, "Capacity improvement for intelligent reflecting surface-assisted wireless systems with a limited number of passive elements," *IEEE Wireless Communications Letters*, vol. 11, no. 4, pp. 801–805, 2022.

[6] Y. Wang, Z. Tian, X. Fan, Z. Cai, C. Nozari, and K. Zeng, "Distributed swarm learning for edge internet of things," *IEEE Communications Magazine*, pp. 1–7, Early Access, 2024.

[7] X. Fan, Y. Wang, Y. Huo, and Z. Tian, "Joint optimization of communications and federated learning over the air," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4434–4449, 2022.

[8] L. Huang, Y. Wang, Q. Zhang, J. Han, W. Tan, and Z. Tian, "Machine learning for underwater acoustic communications," *IEEE Wireless Communications*, vol. 29, no. 3, pp. 102–108, 2022.

[9] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive mimo for hybrid precoding," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3027–3032, 2019.

[10] H. Huang, W. Xia, J. Xiong, J. Yang, G. Zheng, and X. Zhu, "Unsupervised learning-based fast beamforming design for downlink mimo," *IEEE Access*, vol. 7, pp. 7599–7605, 2018.

[11] W. Zhang, Y. Wang, X. Chen, and Z. Tian, "Spectrum transformer: Wideband spectrum sensing using multi-head self-attention," in *2023 IEEE 24th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 101–105, 2023.

[12] W. Zhang, Y. Wang, F. Yu, Z. Qin, X. Chen, and Z. Tian, "Wideband spectrum sensing based on collaborative multi-task learning," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 01–06, 2022.

[13] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3133–3174, 2019.

[14] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5g networks: Joint beamforming, power control, and interference coordination," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1581–1592, 2019.

[15] W. Xu, Z. Yang, D. W. K. Ng, M. Levorato, Y. C. Eldar, and M. Debbah, "Edge learning for b5g networks with distributed signal processing: Semantic communication, edge computing, and wireless sensing," *IEEE journal of selected topics in signal processing*, vol. 17, no. 1, pp. 9–39, 2023.

[16] Z. Yang, Y. Liu, Y. Chen, and J. T. Zhou, "Deep reinforcement learning for ris-aided non-orthogonal multiple access downlink networks," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6, IEEE, 2020.

[17] G. Lee, M. Jung, A. T. Z. Kasgari, W. Saad, and M. Bennis, "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces," in *ICC 2020-2020 IEEE international conference on communications (ICC)*, pp. 1–6, IEEE, 2020.

[18] J. Lin, Y. Zout, X. Dong, S. Gong, D. T. Hoang, and D. Niyato, "Deep reinforcement learning for robust beamforming in irs-assisted wireless communications," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pp. 1–6, IEEE, 2020.

[19] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, and Z. Zhang, "Hybrid beamforming for ris-empowered multi-hop terahertz communications: A drl-based method," in *2020 IEEE Globecom Workshops (GC Wkshps*, pp. 1–6, IEEE, 2020.

[20] S. Zan, Y. Pang, R. Gravina, E. Cao, Y. Li, and W. Zang, "A deep reinforcement learning based approach for intelligent reconfigurable surface elements selection," in *2022 IEEE Intl Conf on DASC/PiCom/CBDCom/CyberSciTech*, pp. 1–7, IEEE, 2022.

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[22] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[23] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1838–1851, 2019.

[24] R. Y. Rubinstein and D. P. Kroese, *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, vol. 133. Springer, 2004.