Deep Reinforcement Learning Based Intelligent Reflecting Surface Optimization for MISO Communication Systems

Keming Feng, Qisheng Wang, Xiao Li, *Member, IEEE,* and Chao-Kai Wen, *Member, IEEE*

***Abstract*—This letter investigates the intelligent reflecting surface (IRS)-aided multiple-input single-output wireless transmission system. Particularly, the optimization of the passive phase shift of each element at IRS to maximize the downlink received signal-to-noise ratio is considered. Inspired by the huge success of deep reinforcement learning (DRL) on resolving complicated control problems, we develop a DRL based framework to solve this non-convex optimization problem. Numerical results reveal that the proposed DRL based framework can achieve almost the upper bound of the received SNR with relatively low time consumption.**

***Index Terms*—Intelligent reflecting surface, non-convex optimization, deep reinforcement learning, phase shift design**

I. Introduction

Recently, the intelligent reflecting surface (IRS) technology has drawn great amount of attention due to its capability of providing remarkable massive MIMO-like gains with low cost [1]–[4]. These surfaces are usually made of almost passive reconfigurable units, each of which can reflect the incident signal independently with different phase shifts. Through adjusting these phase shifts dynamically, a more preferable propagation condition can be obtained. Additionally, these surfaces can be easily coated on facades of outdoor buildings or indoor walls, thus can be implemented with low complexity.

To utilize the IRS effectively and efficiently, some work has been done on the configuration of phase shifts [5]–[8]. A semidefinite relaxation (SDR) method was introduced to optimize the phase shifts of each unit so as to maximize the received signal-to-noise ratio (SNR). Since SDR method is of high computational complexity, a relatively low complexity fix point iteration (FPI) algorithm was proposed in [6]. However, when the user is located far away from the BS, the performance loss is relatively high. In [7] and [8], the phase shifts of each unit is optimized one by one iteratively in a greedy manner. Thus, it is less efficient for large-scale systems.

Due to the recent advances of artificial intelligence, especially deep learning (DL), in wireless communication, [9] and [10] utilized DL methods to the phase shift design. However, this supervised learning requires enormous training labels being calculated in advance. In many cases, these training labels themselves are difficult to obtain, if not impossible. On the contrary, deep reinforcement learning (DRL) based methods do not need training labels and possess the property of online learning and sample generation, which is more storage-efficient.

In this letter, we investigate the phase shift design of the IRS utilizing deep reinforcement learning (DRL). A DRL-based framework is proposed to tackle the non-convexity induced by the unit modulus constraints. We introduce the deep deterministic policy gradient (DDPG) algorithm into the DRL framework. Simulation results indicate that the performance of the proposed algorithm surpasses the state-of-the-art algorithms in terms of received SNR and running time.

II. System Model And Problem Formulation

Consider a single-user multiple-input single-output (MISO) downlink system, as illustrated in Fig. 1. The BS employs a uniform linear array (ULA) with M antenna elements, the IRS is deployed with $N = N_x \times N_y$ passive phase shifters, where N_x and N_y are the number of passive units in each row and column. All phase shifters on the IRS are configurable via a smart controller. All channels are assumed to be quasi-static frequency flat-fading and available at both the BS and IRS. The channels of the BS-user, IRS-user, and BS-IRS links are denoted as $\mathbf{h}_d \in \mathbb{C}^{M\times 1}$, $\mathbf{h}_r \in \mathbb{C}^{N\times 1}$, and $\mathbf{G} \in \mathbb{C}^{N\times M}$, respectively.

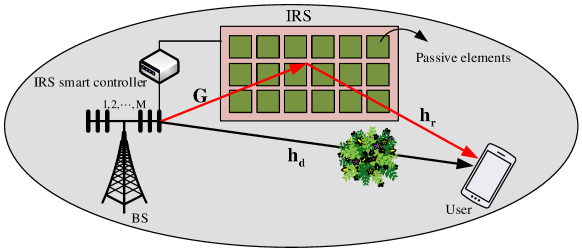

Fig. 1. IRS-aided single-user MISO system

For the considered system, the received signal at the user is

$$y = (\mathbf{h}_r^H \mathbf{\Phi} \mathbf{G} + \mathbf{h}_d^H)\mathbf{b}s + n, \tag{1}$$

where $\mathbf{\Phi} = \mathrm{diag}\left(e^{j\theta_1}, e^{j\theta_2}, \cdots e^{j\theta_N}\right)$ is the phase shift matrix at the IRS, $\mathrm{diag}(a_1, \cdots, a_N)$ denotes a diagonal matrix with $a_1, \cdots, a_N$ as its diagonal entries, $\theta_i \in [0, 2\pi]$ represents the phase shift of the i-th element on the IRS, $\mathbf{b} \in \mathbb{C}^{M\times 1}$ is the

Manuscript received December 13, 2019; accepted January 18, 2020. The work of X. Li was supported in part by the National Natural Science Foundation of China under Grants 61971126 and 61831013. The work of C.-K. Wen was supported in part by the Ministry of Science and Technology of Taiwan under grants MOST 108-2628-E-110-001-MY3. The associate editor coordinating the review of this paper and approving it for publication was Prof. Jianhua Zhang. *(Corresponding author: Xiao Li.)*

K. Feng, Q. Wang, and X. Li are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, P. R. China (e-mail: keming_feng@seu.edu.cn; qishengw@seu.edu.cn; li_xiao@seu.edu.cn).

C.-K. Wen is with the Institute of Communications Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan (e-mail: chaokai.wen@mail.nsysu.edu.tw).

beamforming vector at the BS with the constraint $\|\mathbf{b}\|^2 \leq P_{\max}$, $P_{\max}$ is the maximum transmit power of the BS, s is the transmitted signal satisfying $\mathbb{E}[s^2]=1$, $n \sim \mathcal{CN}(0,\sigma^2)$ is the noise. Then, the received SNR can be obtained as

$$\gamma = \left|(\mathbf{h}_r^H \mathbf{\Phi} \mathbf{G} + \mathbf{h}_d^H)\mathbf{b}\right|^2/\sigma^2. \tag{2}$$

Note that, for a fixed phase shift matrix $\mathbf{\Phi}$, the optimal beamforming method that maximizes the received SNR is the maximum-ratio transmission (MRT) [8], i.e.,

$$\mathbf{b}^* = \sqrt{P_{\max}}\frac{\left(\mathbf{h}_r^H \mathbf{\Phi} \mathbf{G} + \mathbf{h}_d^H\right)^H}{\left\|\mathbf{h}_r^H \mathbf{\Phi} \mathbf{G} + \mathbf{h}_d^H\right\|}. \tag{3}$$

The optimization problem for the phase shift matrix $\mathbf{\Phi}$ to maximize γ can be formulated as

$$\begin{aligned} \text{(P1):} \quad & \max_{\mathbf{\Phi}} \quad \|\mathbf{h}_r^H \mathbf{\Phi} \mathbf{G} + \mathbf{h}_d^H\|^2, \\ & \text{s.t.} \quad |\mathbf{\Phi}_{i,i}| = 1, \quad \forall i = 1,2,\cdots,N, \end{aligned} \tag{4}$$

where $\mathbf{\Phi}_{i,i}$ is the i-th diagonal element of $\mathbf{\Phi}$. Note that (P1) is a NP-hard problem owing to the non-convexity of the objective function and the unit modulus constraints. A SDR method was proposed in [5] to solve this problem. However, it is computational expensive with complexity of $O((N+1)^6)$ [6]. In this letter, we focus on the design of the phase shift matrix $\mathbf{\Phi}$, we propose a robust DRL based framework to deal with (P1) efficiently, which will be described in the next section.

III. DRL Based Framework

In this section, we first briefly introduce the DRL techniques involved. Then, the proposed DRL based framework will be described in detail.

A. Deep Reinforcement Learning Basics

A reinforcement learning (RL) system consists of two major parts, i.e., the agent and the environment. Interactions between them can be described as a Markov Decision Process (MDP) [11]. During time step t in each episode, the agent obtains the state s_t from the environment, and chooses an action a_t from the action space based on a policy π. Once the action is done, the environment updates the current state to s_{t+1}, and emits a reward r_t which measures the performance of a_t under current state. Learning of the agent is to determine the optimal policy that maximizes the long-term reward. Two kinds of algorithms, i.e., the value based and policy based algorithms, are usually applied to determine the optimal policy.

Deep Q network (DQN) [12] is a value based algorithm for discrete action space. Under a policy π, an action-state-Q function of the agent for an action a under state s, which evaluates the current action-state pair, is defined as

$$Q_\pi(s,a;\boldsymbol{\theta}) = \mathbb{E}_\pi\left[G_t | s_t = s, a_t = a\right], \tag{5}$$

where $\mathbb{E}[\cdot]$ represents the expectation, $G_t = \sum_{t=0}^{\infty}\lambda^t r_t$ is the expected cumulative reward, $\lambda \in (0,1]$ is a discounting factor, $\boldsymbol{\theta}$ represents the parameters of the deep neural network (DNN) used in DQN. This algorithm aims at maximizing the Q value (5) of a certain action-state pair by training the DNN [11]. The training batch is randomly sampled from a relay buffer with $\{s_t, a_t, r_t, s_{t+1}\}$ as one piece of previous data.

Policy gradient (PG) is a policy based algorithm aiming at maximizing the expectation of the discounted cumulative reward of each episode when the action space is continuous. At each time step t, the agent chooses the action according to a policy $\pi_{\boldsymbol{\theta}}$. Therefore, training of the policy can be represented as a gradient ascent procedure [13]

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \beta \mathbb{E}_{\pi_{\boldsymbol{\theta}_t}}[\nabla_{\boldsymbol{\theta}_t} \log \pi_\theta(s,a) Q_{\pi_{\boldsymbol{\theta}_t}}(s,a)], \tag{6}$$

where β is the learning rate, $Q_{\pi_{\boldsymbol{\theta}_t}}(s,a)$ is the action-state-Q function under current policy $\pi_{\boldsymbol{\theta}_t}$. The drawback of this algorithm is that the policy network can be updated only after an episode is done, which slows down the convergence rate.

B. Phase Shift Design Framework Using DDPG

According to the above description, the DQN algorithm is not suitable to solve the problem (P1), since it can only deals with discrete action spaces. As for PG algorithm, its convergence performance is unsatisfactory under wireless communication context. In this letter, a DDPG based algorithm is developed to solve problem (P1), it can overcome the limitations of the DQN and PG algorithm. The proposed framework is illustrated in Fig. 2.

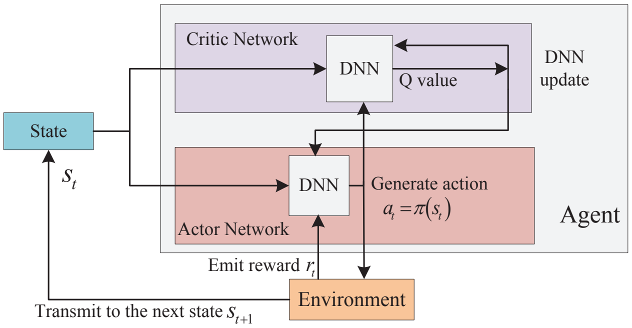

Fig. 2. The DRL Based Phase Shift Design Framework Using DDPG

1) Deep Deterministic Policy Gradient: DDPG is a model-free, off-policy actor-critic (AC) algorithm combining the advantages of DQN and PG [14]. It can learn deterministic policy optimally under high-dimensional continuous action space. In DDPG, a deterministic policy network (DPN) is used as an actor to chose actions from a continuous action space $\mathcal{A}$, i.e., $a = \mu(s;\boldsymbol{\theta}_\mu)$, where $\boldsymbol{\theta}_\mu$ is the parameters of the DPN. A Q network $Q(s,a;\boldsymbol{\theta}_q)$ is modeled as a critic to measure the performance of the chosen action, where $\boldsymbol{\theta}_q$ is the parameter of the critic network. The goal of DDPG is to maximize the output Q value. To achieve this goal, as in DQN, an experience replay is reserved to reduce the correlation of different training samples. Moreover, to solve the problem that the Q value update is incline to divergence under a single Q network [12], a copy is created for the actor and critic networks, i.e., $\mu'(s;\boldsymbol{\theta}_{\mu'})$ and $Q'(s,a;\boldsymbol{\theta}_{q'})$, which are referred to as target networks and used to calculate the corresponding target values. The networks being copied are referred to as evaluation networks. Note that, the target network shares the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.2969167, IEEE Wireless Communications Letters

same structure with its corresponding evaluation network, but with different parameters, i.e. $\boldsymbol{\theta}_q \neq \boldsymbol{\theta}_{q'}$, $\boldsymbol{\theta}_\mu \neq \boldsymbol{\theta}_{\mu'}$. The target networks are then updated through soft update, which can be written as

$$\boldsymbol{\theta}_{i'} = \tau\boldsymbol{\theta}_i + (1-\tau)\boldsymbol{\theta}_{i'}, i = \mu \text{ or } q, \tag{7}$$

where $\tau \ll 1$. Soft update moves the unstable problem in learning the action-state-Q function and accelerates the convergence of AC method [14]. Since the DPN learns a deterministic policy, for exploration, DDPG treats it independently from the learning process. Moreover, an exploration policy $\widetilde{\mu}$ is created through the adding of a noise sampled from a stochastic process $\mathcal{N}$

$$\widetilde{\mu}(s_t) = \mu(s_t;\boldsymbol{\theta}_\mu) + \mathcal{N}, \tag{8}$$

where $\mathcal{N}$ can be chosen to suit the environment.

2) The DRL Formulation: In this letter, the communication system is regarded as the environment and the IRS is treated as an agent. We define the corresponding elements as follows.

- *State space:* The state s_t is defined as
$$s_t = \left[\gamma^{(t-1)}, \theta_1^{(t-1)}, \cdots, \theta_N^{(t-1)}\right], \tag{9}$$
where $\gamma^{(t-1)}$ is the received SNR at time step $t-1$.
- *Action space:* During time step t, the agent uses state s_t as input to update the phase shifts induced by the IRS under current channel states. When the update is done, new phase shifts are obtained. Therefore, the action vector $a_t \in \mathbb{R}^N$ is defined as
$$a_t = \left[\theta_1^{(t)}, \cdots, \theta_N^{(t)}\right]. \tag{10}$$
- *Reward function:* In this letter, the objective is to maximize the received SNR. Thus, the received SNR defined in (9) is used as the reward, i.e., $r_t = \gamma^{(t)}$.

3) Working Procedure: At the initialization stage, four networks are generated, i.e., the actor target net $\boldsymbol{\theta}_{\mu'}$, the actor evaluation net $\boldsymbol{\theta}_\mu$, the critic target net $\boldsymbol{\theta}_{q'}$ and the critic evaluation net $\boldsymbol{\theta}_q$, whose parameters are all uniformly distributed. Besides, an experience replay $\mathcal{D}$ with capacity $\mathcal{C}$ is built as well. Without loss of generality, the phase shifts of all elements are chosen randomly from 0 to 2π at the beginning of each episode. During each episode, we first calculate all channels involved. Then, taking the state s_t as excitation, the actor evaluation net gives out a corresponding action a_t. Reforming a_t into a phase shift matrix $\boldsymbol{\Phi}^{(t)} = \text{diag}(e^{j\theta_1^{(t)}}, \cdots, e^{j\theta_N^{(t)}})$ to calculate the current reward r_t by (2), the next state s_{t+1} can then be obtained by (9). Storing $\{s_t, a_t, r_t, s_{t+1}\}$ as one transition into $\mathcal{D}$. The critic evaluation net then samples a N_B-size minibatch $\{s_j, a_j, r_j, s_{j+1}\}$ $(j = 1, \cdots, N_B)$ from the experience replay $\mathcal{D}$ to calculate the target Q value y_j, i.e.,

$$y_j = \begin{cases} r_j, & j = N_B, \\ r_j + \lambda Q'(s_{j+1}, \mu'(s_{j+1};\boldsymbol{\theta}_{\mu'});\boldsymbol{\theta}_{q'}), & j < N_B. \end{cases} \tag{11}$$

The loss function of the critic evaluation net is given as

$$L(\boldsymbol{\theta}_q) = \frac{1}{N_B}\sum\nolimits_{j=1}^{N_B}(y_j - Q(s_j, a_j;\boldsymbol{\theta}_q))^2. \tag{12}$$

Then, the critic evaluation net can be updated by SGD. Afterwards, using policy gradient to update the actor evaluation net with the ascent factor

$$\Delta_{\boldsymbol{\theta}_\mu} = \frac{1}{N_B}\sum_{j=1}^{N_B}(\nabla_a Q(s_j, \mu(s_j;\boldsymbol{\theta}_\mu);\boldsymbol{\theta}_q)|\nabla_{\boldsymbol{\theta}_{\boldsymbol{\mu}}}\mu(s_j;\boldsymbol{\theta}_\mu)). \tag{13}$$

Finally, the actor target net and the critic target net is updated using soft update (7). The detail of the DRL based framework is shown in the following algorithm.

Algorithm 1 The DRL based Framework

```
Input: The discount factor $\lambda$, the soft update coefficient $\tau$,
  the learning rate $\alpha$, the experience replay capacity $\mathcal{C}$, and
  the batchsize $N_B$.
  Randomly initialize the critic evaluation network
  $Q(s, a;\boldsymbol{\theta}_q)$ and the actor evaluation network $\mu(s;\boldsymbol{\theta}_\mu)$.
  Initialize the critic target network $Q'(s, a;\boldsymbol{\theta}_{q'})$ and the
  actor target network $\mu'(s;\boldsymbol{\theta}_{\mu'})$ with the parameters of the
  corresponding evaluation networks.
  Empty the experience replay $\mathcal{D}$.
Output: The optimal phase shift matrix $\boldsymbol{\Phi}^*$ and the maximized received SNR $\gamma^*$ under current channel state.
for episode $j = 1, \cdots, K$ do
    Obtain the current CSI $(\mathbf{h}_r^{(j)}, \mathbf{G}^{(j)}, \mathbf{h}_d^{(j)})$;
    Randomly chose phase shifts to obtain $\boldsymbol{\Phi}^{(0)}$ and $\gamma^{(0)}$
    as initial state $s_1$;
    Initialize a random process $\mathcal{N}$;
    for $t = 1, \cdots, T$ do
        Action $a_t = \mu(s_t;\boldsymbol{\theta}_\mu) + \mathcal{N}$;
        Reform $a_t$ into phase shift matrix $\boldsymbol{\Phi}^{(t)} =$
        $\text{diag}(e^{j\theta_1^{(t)}}, \cdots, e^{j\theta_N^{(t)}})$ to calculate $\gamma^{(t)}$. Obtain
        the next state $s_{t+1}$. Then, store the transition
        $\{s_t, a_t, r_t, s_{t+1}\}$ into $\mathcal{D}$.
        Sample a $N_B$ minibatch transitions $\{s_j, a_j, r_j, s_{j+1}\}$
        from $\mathcal{D}$.
        Set target Q value according to (11).
        Update $Q(s, a;\boldsymbol{\theta}_q)$ by minimizing the loss in (12).
        Update the policy $\mu(s;\boldsymbol{\theta}_\mu)$ using the sampled policy
        gradient in (13).
        Soft update the target networks according to (7).
        Update the sate $s_t = s_{t+1}$.
    end for
end for
```

IV. Numerical Results

This section demonstrates the performance of the proposed framework. The channel between the BS and the user is assumed to be Rayleigh fading which suggests that the line-of-sight signal between them is blocked (note that it could be other fading as well), i.e.,

$$\mathbf{h}_d = \sqrt{PL_d}\widetilde{\mathbf{h}}_d, \tag{14}$$

where $\widetilde{\mathbf{h}}_d \in \mathbb{C}^{M\times 1}$ contains independent and identical (i.i.d) distributed $\mathcal{CN}(0,1)$ elements. The channel between the BS and IRS as well as that between the IRS and the user, are

Rician fading, i.e.,

$$\mathbf{G} = \sqrt{PL_G}\left(\sqrt{\frac{K_1}{K_1+1}}\overline{\mathbf{G}} + \sqrt{\frac{1}{K_1+1}}\widetilde{\mathbf{G}}\right), \tag{15}$$

$$\mathbf{h}_r = \sqrt{PL_r}\left(\sqrt{\frac{K_2}{K_2+1}}\overline{\mathbf{h}}_r + \sqrt{\frac{1}{K_2+1}}\widetilde{\mathbf{h}}_r\right), \tag{16}$$

where K_1 and K_2 are the Rician-K factors, $\widetilde{\mathbf{G}} \in \mathbb{C}^{N\times M}$ and $\widetilde{\mathbf{h}}_r \in \mathbb{C}^{N\times 1}$ are the random components with i.i.d and $\mathcal{CN}(0,1)$ distributed elements. The distances between two adjacent antenna elements at BS and IRS are both half of the carrier frequency. Thus, the deterministic components $\overline{\mathbf{G}}$ and $\overline{\mathbf{h}}_r$ can be expressed as [15, Eq. (3)][16, Eq. (6)]

$$\overline{\mathbf{G}} = \left[\mathbf{a}^H_{N_x}(\theta_{\text{AoA,h}}) \otimes \mathbf{a}^H_{N_y}(\theta_{\text{AoA,v}})\right]\mathbf{a}_M(\theta_{\text{AoD,b}}), \tag{17}$$

$$\overline{\mathbf{h}}_r = \mathbf{a}^H_{N_y}(\theta_{\text{AoD,v}}) \otimes \widetilde{\mathbf{a}}^H_{N_x}(\theta_{\text{AoD,v}}, \theta_{\text{AoD,h}}), \tag{18}$$

with

$$\mathbf{a}_i(\theta) = [1, e^{-j2\pi\frac{d}{\lambda}\sin(\theta)}, \cdots, e^{-j2\pi(i-1)\frac{d}{\lambda}\sin(\theta)}], \tag{19}$$

$$\widetilde{\mathbf{a}}_{N_x}(\theta_{\text{AoD,v}}, \theta_{\text{AoD,h}}) = [1, e^{-j2\pi\frac{d}{\lambda}\phi}, \cdots, e^{-j2\pi(N_x-1)\frac{d}{\lambda}\phi}], \tag{20}$$

where $\phi = \cos(\theta_{\text{AoD,v}})\sin(\theta_{\text{AoD,h}})$, $\theta_{\text{AoA/D,h/v}}$ represent the angles of arrival/departure in horizontal/vertical directions at the IRS, and $\theta_{\text{AoD,b}}$ is the angle of departure at the BS. The distance between the BS and IRS is 51 m, $M = 10$, $N = 50$ $(N_x = 10, N_y = 5)$ (if not specified otherwise), $P_{\text{max}} = 5$ dBm, and $\sigma^2 = -80$ dBm. The user moves on a line in parallel to that connects the BS and IRS, and the vertical distance between these two lines is 1.5 m. The path loss is modeled as $PL = PL_0 - 10\xi\log_{10}(\frac{d}{D_0})$ dB, where $PL_0 = -30$ dB, $D_0 = 1$ m, ξ is the path loss exponent, and d is the BS-user horizontal distance. The penetration loss of 5 dB is assumed in both BS-user link and IRS-user link. The antenna gain of 0 dBi is assumed at both the BS and user, and that of the IRS is 5 dBi. The path loss exponents of the BS-IRS, BS-user, and IRS-user links are set to $\xi_{bi} = 2$, $\xi_{bu} = \xi_{iu} = 2.8$, respectively. Over 500 realizations of the channels' random components are averaged to obtain the simulation results.

In the proposed DRL framework, all neural networks are considered to be a four layered DNN. The actor evaluation net and the critic evaluation net both use Adam optimizer for parameters update. The input layer of the actor network contains $N+1$ neurons while the output layer contains N neurons (these two numbers change to $2N+1$ and 1, respectively in the critic network). The two hidden layers contain 300 and 200 neurons, respectively. The first three layers are all followed by a ReLU function while the output layer uses tanh$(\cdot)$ function to provide enough gradient. Furthermore, we set the batchsize $N_B = 16$, the number of step in each episode $T = 1000$, the learning rate $\alpha = 10^{-3}$, the discount factor $\lambda = 0.95$, the soft update coefficient $\tau = 0.005$, and the experience replay capacity $\mathcal{C} = 50000$. The additional noise $\mathcal{N}$ is selected as complex Gaussian noise with zero mean and variance 0.1.

Fig. 3 demonstrates the received SNR of the proposed algorithm vs. the horizontal distance between the BS and the user, denoted as d. In this figure, We consider a scenario similar with [5], where the IRS is coated on the facade of a tall building and is aware of the BS's location, so that $K_1 \to \infty$ and $K_2 = 0$. The performance of the SDR algorithm [5] serving as an upper bound, the fix point iteration algorithm [6] with random initialization, as well as the system without IRS, are also shown. It can be easily observed that the proposed DRL based framework almost achieves the upper bound of the received SNR, which testifies its optimality. Note that it brings less gain when the user is close to the BS, this is because in this setting, the user is far from the IRS and thus get less signal power from the IRS. In the absence of IRS, the received SNR decreases rapidly as the user moves apart from the BS. This performance degradation can be substantially improved by placing an IRS between the BS and the user. It is also noted that the performance of the proposed algorithm is obviously superior to the fix point iteration when $d \geq 40$ m.

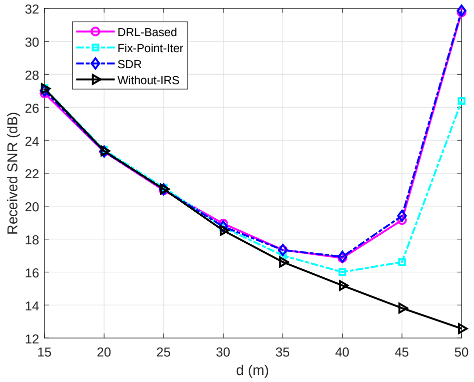

Fig. 3. Received SNR vs. BS-user horizontal distance

In Fig. 4, the received SNRs under different number of passive element on IRS are compared. In this figure, we consider a scenario similar with [7], where the IRS is placed without the knowledge of the BS's location, the Rician K-factors are set to $K_1 = K_2 = 10$, and the horizontal BS-user distance $d = 48$ m. As can be observed, the performance of all algorithms get better as the number of passive unit on the IRS increases. This is because the power reflected by the IRS increases. Particularly, the difference between the received SNRs with $N = 50$ and $N = 100$ is approximately 6 dB, which suggests that $O(N^2)$ gain can be attained by doubling the number of passive unit on the IRS.

The running time of the three algorithms under different number of passive element on IRS is given in Table I. The other simulation parameters are the same with Fig. 4. We can see that the SDR algorithm is extremely time-consuming, and its running time increases enormously as N increases. The fix point iteration algorithm has the lowest running time [6], while the time consumed increases promptly as N increases, which is as expected since more optimization variables are involved. In contrast, the time consumed by the proposed DRL based framework remains around 34 ms for all N values, which is explicable since the number of hidden layer neurons remain unchanged as N grows. This property verifies that the proposed framework is efficient and robust. More importantly,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LWC.2020.2969167, IEEE Wireless Communications Letters

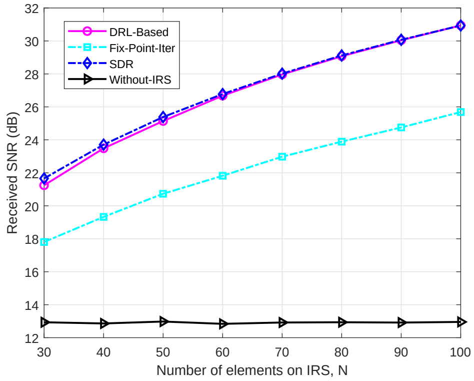

Fig. 4. Received SNR vs. number of elements on IRS

TABLE I
RUNNING TIME COMPARISON

N	Running Time (ms)		
	SDR	FPI	DRL
30	416.13	4.68	34.23
50	619.26	11.34	34.47
70	919.73	13.99	34.72
90	1175.31	18.11	34.94

it can achieve almost optimal received SNR with relatively low time consumption.

Fig. 5 compares the performance of the proposed framework under different numbers of antenna element at BS and different number of passive unit at IRS. The curve “DRL-based-M” is obtained by fixing the number of passive unit on IRS to 30 ($N_x = 10$ and $N_y = 3$) and varying the number of antenna at the BS from 30 to 100. The curve “DRL-based-N” is obtained by fixing the number of antenna at BS to 30 and varying the number of passive unit on the IRS from 30 to 100 (N_y changes from 3 to 10). In the curve “DRL-based-N”, each point needs to be trained again. The other simulation parameters are the same with Fig. 4. It can be seen that the increase in the number of passive unit on the IRS leads to higher performance gain, which indicates that increasing the number of low-cost passive elements on the IRS is more energy-efficient than enlarging the scale of costly RF chains on the BS. It is worth noting that the performance gain becomes more pronounced as the number of element grows.

V. CONCLUSION

In this letter, we investigated the phase shifts design for the IRS-aided downlink MISO wireless communication system to maximize the received SNR. An efficient DRL based framework were proposed to tackle the non-convex unit modulus constraints, which are major concerns for optimizing phase shifts introduced by the IRS. Numerical results reveal that the proposed framework can obtain significant performance gain compared to the fix point iteration algorithm and achieve almost the upper bound calculated by the SDR algorithm with much less time consumption.

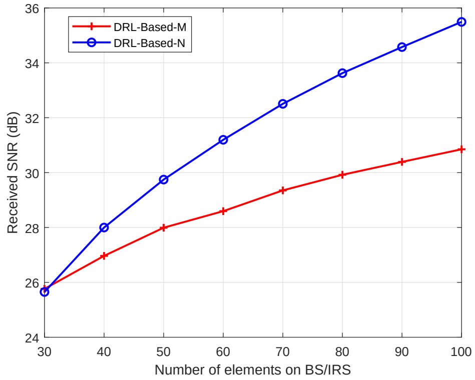

Fig. 5. Received SNR vs. number of elements at BS/IRS

REFERENCES

[1] Q. Wu and R. Zhang, “Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network,” *arXiv preprint arXiv:1905.00152*, 2019.

[2] W. Tang, M. Z. Chen, J. Y. Dai, Y. Zeng, X. Zhao, S. Jin, Q. Cheng, and T. J. Cui, “Wireless communications with programmable metasurface: New paradigms, opportunities, and challenges on transceiver design,” *arXiv preprint arXiv:1907.01956*, 2019.

[3] W. Tang, J. Y. Dai, M. Chen, X. Li, Q. Cheng, S. Jin, K.-K. Wong, and T. J. Cui, “Programmable metasurface-based RF chain-free 8PSK wireless transmitter,” *Electron. Lett.*, vol. 55, no. 7, pp. 417–420, Apr. 2019.

[4] S. Abeywickrama, R. Zhang, and C. Yuen, “Intelligent reflecting surface: Practical phase shift model and beamforming optimization,” *arXiv preprint arXiv:1907.06002*, 2019.

[5] Q. Wu and R. Zhang, “Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design,” in *Proc. IEEE GLOBECOM*, 2018, pp. 1–6.

[6] X. Yu, D. Xu, and R. Schober, “MISO wireless communication systems via intelligent reflecting surfaces,” *arXiv preprint arXiv:1904.12199*, 2019.

[7] Y. Han, W. Tang, S. Jin, C.-K. Wen, and X. Ma, “Large intelligent surface-assisted wireless communication exploiting statistical CSI,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8238–8242, 2019.

[8] Q. Wu and R. Zhang, “Beamforming optimization for intelligent reflecting surface with discrete phase shifts,” in *Proc. IEEE ICASSP*, 2019, pp. 7830–7833.

[9] C. Huang, G. C. Alexandropoulos, C. Yuen, and M. Debbah, “Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces,” *arXiv preprint arXiv:1905.07726*, 2019.

[10] A. Taha, M. Alrabeiah, and A. Alkhateeb, “Enabling large intelligent surfaces with compressive sensing and deep learning,” *arXiv preprint arXiv:1904.10136*, 2019.

[11] R. W. Picard, S. Papert *et al.*, “Affective learning–a manifesto,” *BT technology journal*, vol. 22, no. 4, pp. 253–269, 2004.

[12] V. Mnih, K. Kavukcuoglu, D. Silver *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, “Policy gradient methods for reinforcement learning with function approximation,” in *Advances in NIPS*, 2000, pp. 1057–1063.

[14] T. P. Lillicrap, J. J. Hunt, A. Pritzel *et al.*, “Continuous control with deep reinforcement learning,” *arXiv preprint arXiv:1509.02971*, 2015.

[15] Y. Cao and T. Lv, “Intelligent reflecting surface aided multi-user millimeter-wave communications for coverage enhancement,” *arXiv preprint arXiv:1910.02398*, 2019.

[16] X. Li, S. Jin, H. A. Suraweera, J. Hou, and X. Gao, “Statistical 3-D beamforming for large-scale MIMO downlink systems over rician fading channels,” *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1529–1543, 2016.