Intelligent Reflecting Surfaces (IRSs): A Promising Technology for 6G Networks

A Project Report

of

IE 692 (MSc. Ph.D Phase-II Project)

Submitted by

by

Chesta Pahuja

(Roll No. - 19i190008)

Under Guidance of

Prof. Manjesh Hanawal

Indian Institute of Technology Bombay

Powai – 400 076, Maharashtra, India

June 2021

DECLARATION

I, Chesta Pahuja, hereby declare that the project report **Intelligent Reflecting Surfaces (IRSs): A Promising Technology for 6G Networks**, submitted for partial fulfillment of Course IE692: MSc. Ph.D Phase-II Project, IIT Bombay is a bonafide work done by me under the supervision of Prof. Manjesh Hanawal

This submission represents my ideas in my own words and where ideas or words of others have been included, we have adequately and accurately cited and referenced the original sources.

I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in our submission.

Chesta Pahuja

17-06-2021

Abstract

In this work, we examine a downlink MISO scenario with an intelligent reflecting surface (IRS) to maximise the SNR for the user. The IRS optimization problem is complex and non-convex because it necessitates the tuning of the phase shift reflection matrix with unit modulus constraints. We use deep reinforcement learning (DRL) to forecast and optimally adjust the IRS phase shift matrices, owing to the increasing use of DRL approaches capable of tackling non-convex optimization problems. The Deep deterministic policy gradient (DDPG) algorithm in [1] has been studied and implemented in Python from the scratch. The IRS-assisted MISO system based on the DRL scheme produces a high SNR, according to implementation and simulation results. Furthermore, we modify the IRS-DRL framework to account for unknown channel gains between IRS and user, which would occur in practice because IRS is uninformed of the user's position and cannot forecast the angles of departures, as required by the assumed Rician Channel fading model.

Contents

Chapter 1

Introduction

In order to achieve a sustainable wireless network development with scalable costs in the future, innovative research on finding both spectrum and energy efficient approaches with low hardware costs is required. Future communication systems will be able to handle larger data speeds, improved spectral efficiency, lower latency, greater coverage regions, and a large number of connections, among other things, as wireless communications improve.
Wireless networks promise to meet the demand for the vast number of connections as wireless technologies have increased tremendously over the previous few decades. To enable a fully linked and sustainable community, next-generation networks will be an end-to-end ecosystem.The major goal of these networks is to give users with seamless and ubiquitous wireless communications with faster throughput, low latency, and low energy consumption, as well as to facilitate escalation.

For billions of connections, there has been a significant increase in mobile data use. While 5G networks are being deployed, 6G network technologies are being developed and evaluated in order to achieve more dependable and quicker communication. Among these technologies is the Intelligent reflecting surface (IRS) proposed as a viable new approach to achieve the aforesaid goal. IRS is a planar array made up of a large number of re-configurable passive elements (e.g., low-cost printed dipoles), each of which can independently induce a phase shift on the incident signal (controlled by an attached smart controller), thereby collaboratively modifying the reflected signal propagation. Additionally, these surfaces can be simply coated

on the facades of outdoor structures or internal walls, resulting in a low-complexity implementation.

Most of the available literature focuses on passive IRS, which just applies a phase shift to the incident signal. As a result, the IRS will use **no transmit power**. As a result, the phase-shift matrix is the focus of the IRS optimization problem.

The IRS concept is similar to massive multiple-input multiple-output (MIMO) technology, which uses vast arrays of antennas to increase spectrum and energy efficiency. As a result, IRS, like large MIMO in 5G networks, is envisioned to play a critical role in 6G communication networks. As a result, IRS can be utilised to achieve massive MIMO 2.0. IRSs differ from massive MIMO in that they adjust the wireless propagation environment for communication. Other existing technologies that have been compared to IRSs in recent research include backscatter communication, millimetre (mm)-wave communication, and network densification, in addition to massive MIMO. In the next section, we review some of the key features of IRS device and the edge it gives to the 6G networks

Our contributions to this project are mentioned in section 2.4. We provide a survey of existing literature in Section 2.3. We give details on the framework and problem formulation in section 3, and the method and approach in Section 3.2. Section 4 consists of a summary of the experiments and results obtained with our experiments. We conclude with a short summary and pointers to forthcoming work in Section 5.

Chapter 2

Fundamental research

2.1 IRS: Faster, Smarter, greener

We look at the IRS's distinguishing characteristics, to devise an optimal phase-shift design:

- **Nearly passive:** IRS comprises of a large number of low-cost passive reflecting elements that only reflect signals and do not transmit them. As a result, IRS is nearly inactive and, in theory, does not require a dedicated energy source.
- **Programmable control:** IRS is facilitated by controller which uses programming to regulate the scattering, reflection, and refraction characteristics of radio waves, overcoming the disadvantages of natural wireless propagation. As a result, IRS-assisted wireless communications may intelligently modify the wavefront of impinging signals, such as phase, amplitude, frequency, and even polarisation, without requiring sophisticated decoding or encoding and radio frequency processing operations.
- **Easy deployment:** IRS is easy to install and remove because of its compact size, low weight, conformal geometry, and smaller thickness than wavelengths. As a result, IRS may be simply installed on the outer walls of buildings, billboards, factory and indoor space ceilings, human apparel, and so on.
- **Good compatibility:** IRS devices can be integrated into existing communication networks by just altering the network protocol rather than modifying their

hardware or software. Meanwhile, IRS has a full-band response that, in theory, can operate on any operating frequency.

2.2 6G Networks: Vision

In the next decade, the intelligent information society, which is highly digitised, intelligence-inspired, and data-driven internationally, will be deployed. The key to achieving this great vision is 6G networks, which are projected to provide connectivity for everything, full wireless coverage, and the integration of all functions to support full vertical applications. 6G networks, which are expected to provide connectivity for everything, full wireless coverage, and the integration of all functions to support full vertical applications, are the key to realising this magnificent ambition. As a result, 6G networks must process a large amount of data in a short amount of time, with exceptionally high throughput and low latency. The following is a summary of the vision 6G networks holds:

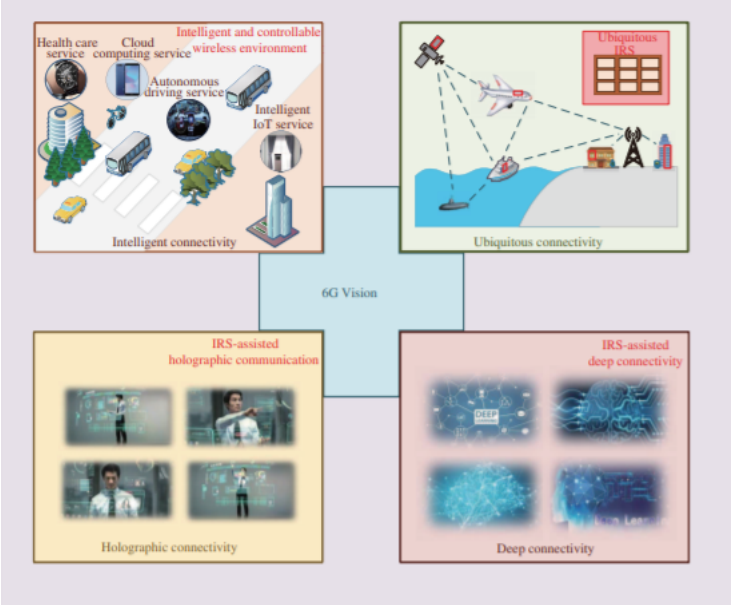

Figure 2.1: The vision of 6G networks with IRS technology

- **Ubiquitous connectivity**: With the rapid advancement of aircraft and deep-sea exploration technologies, human and intelligent device active space will be substantially enlarged, putting more demands on communication network coverage. As a result, one of the goals of 6G networks is to create ubiquitous

connection by combining satellite communications, aerial communications, terrestrial communications, and underwater communications to provide global coverage

- **Deep connectivity**: The types and scenarios of information interaction are becoming increasingly complicated as intelligent information services, such as the intelligent Internet of things (IoT), grow. There is reason to anticipate that in 6G networks, information interaction will be substantially enlarged in terms of both space and types of information. As a result, "deep connectivity" will necessitate that the active space of each connected intelligent device is expanded in depth, and the network itself has deep sensing, deep learning, and deep mind capabilities, all of which are expected to enable mind-to-mind interaction with intelligent devices.

- **Intelligent connectivity:** 6G networks will be an autonomous system with human-like intelligence, with many methods to interface with intelligent terminals, including voice, sight, and brain waves. As a result, "intelligent connectivity" will necessitate that the complicated network should have intelligent administration all connected devices should be intelligent and the related information services need to be intelligent.

- **Holographic connectivity:** With the rapid advancement of augmented reality (AR) and virtual reality (VR) technologies, information interaction in the next decade will most certainly evolve into high-fidelity VR/AR interaction, and possibly even holographic information interaction. As a result, 6G networks will be necessary to deliver ubiquitous high-fidelity VR/AR services as well as holographic communications.

With its flexibility in deployment and reconfiguration, low implementation cost, and low power consumption, IRS is expected to improve the transmission performance in the next 6G wireless networks.IRS is a promising technology for 6G wireless networks. The overall concept of IRS-assisted 6G wireless networks includes an intelligent and regulated wireless environment, IRS-assisted ubiquitous connection, IRS-assisted deep connectivity, and IRS-assisted holographic communications. IRS-

assisted intelligent connectivity can help in a variety of scenarios in 6G networks. In the metropolis, for example, a major portion of the outside walls of high-rise buildings are composed of glass. To successfully regulate the propagation of radio waves, smart glasses with specific IRSs can be used.

2.3 Recent studies

IRS is a new paradigm and many new studies are being conducted to improve the phase-shift design.The design of tunable elements is the first challenge in implementing IRS-assisted 6G networks. Continuously modifying the reflection coefficients of each element is undoubtedly beneficial to network performance, yet, due to the complicated design and expensive technology of enormous high-precision elements, it is too expensive to implement. The main optimization formulations and solutions proposed for IRS-assisted wireless systems are:

- **SNR or Capacity Maximization:** In [2] The authors focus on an IRS-assisted multiple-input single-output(MISO) system, where one IRS with N passive scattering elements is deployed to assist the downlink information transmission. A joint beamforming problem is formulated to maximize the received signal power at the user, by jointly optimizing the AP's transmit beamforming and the continuous phase shift of each scattering element.

  Solution: ***Sem-definite Relaxation(SDR)*** technique is proposed to obtain an approximate solution as a performance upper bound.

  Alternating Optimization:isthen employed to update the active and passive beamforming strategies iteratively. Given the fixed passive beamforming, the BS's optimal beamforming is easily obtained by the maximum-ratio transmission strategy. In [3], the authors consider a similar IRS-assisted MISO downlink system. The joint optimization of the AP's transmit beamforming and the IRS's phase shifts is solved by **fixed point iteration** and **manifold optimization** techniques, respectively, which are shown to be effective in tackling the IRS's unit modulus constraints. These two algorithms not only achieve a higher data rate but also have a reduced computational complexity.

  Deep Learning based solutions: Deep learning is particularly powerful in

extracting the features from the raw data and providing a "meaning" to the input by constructing a model-free data mapping with huge number of learnable parameters. In [1] a Deep Reinforcement Learning based solution is proposed for channel estimation using SNR maximization.

- **Power Minimization or EE/SE Maximization:** Besides SNR and rate maximization, the IRS-assisted wireless networks can also help minimize the BS's transmit power or maximize the overall EE/SE performance. The IRS can configure wireless channels in favor of information transmission between transceivers. This results in a more energy-efficient communication paradigm. Focusing on an IRS-assisted MISO downlink scenario as that in [2], the authors minimize the BS's transmit power under individual users' SINR constraints by jointly optimizing the AP's transmit beamforming and the IRS's passive beamforming strategies. Following SDR procedure and alternating optimization, the AP's transmit beamforming can be efficiently optimized by solving SOCP, and the optimization of the IRS's passive beamforming is degenerated to a conventional relay beamforming optimization problem.

- **IRS-assisted Physical Layer Security:** The IRS's wave manipulation has the flexibility of simultaneously creating enhanced beams to an intended receiver and suppressed beams to unintended receivers. This can be used to enhance physical layer security in wireless communications. The authors in [5] use IRS to defend against eavesdroppers.. In particular, IRS can be used as a very effective tool to prevent wireless eavesdropping attacks by simultaneously controlling the transmissions at the LT and the reflection at the IRS. As a result, the achievable secrecy rates obtained by the IRS-assisted systems can be significantly improved compared with the conventional methods only relying on the LT's transmission control. Manysimulation results verify the improvement of secrecy performance of IRS-assisted systems.

2.4 Our Contributions

We focus on the DRL based solution for channel estimation proposed in [1]. The study above assumes that the IRS and user communication channels are well-known. Such an assumption, however, contradicts the practical case, in which IRSs are passive elements incapable of channel estimation. The following contributions are made by this project:

- We implement the DDPG based solution for IRS optimization problem in Colab Environment as proposed in the original paper [1] and conduct numerical simulations and experiments, clear details of which lacked in the original paper.
- Apart from this, we modify the DDPG framework to take into consideration that channel gains for the IRS-user link is unknown, and return best phase shift matrix in which the IRS learns the best way for reflecting the incident signals by modifying the phase.

Chapter 3

Detailed analysis of problem

3.1 System Model

We consider the downlink of an IRS assisted MISO system with one user as shown in Fig. 3.1

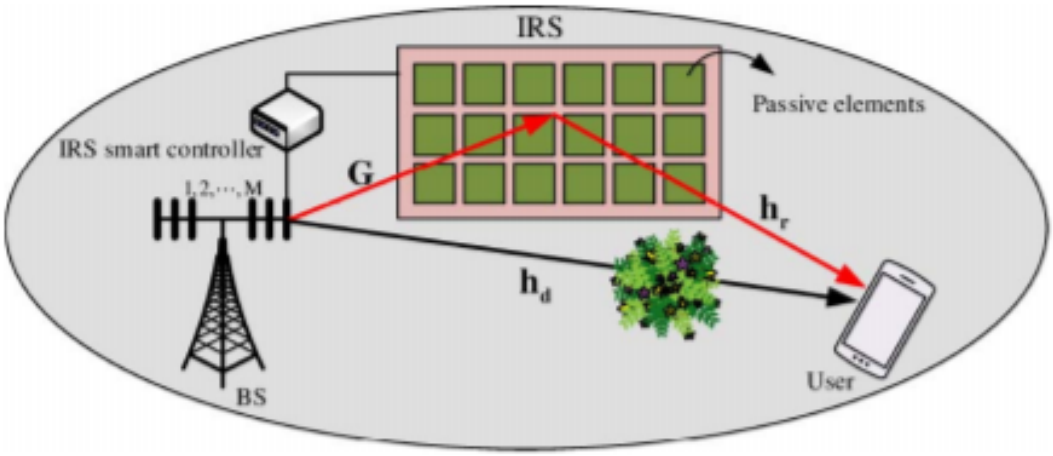

Figure 3.1: IRS-assisted single-user MISO system

This is a three-node communication system with M-antenna transmitter BS, a single-antenna reciever R and IRS deployed with N = $N_x * N_y$ elements. 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 transmitted signal at the base station is

$$x = bs \tag{3.1}$$

where b$\in C^{Mx1}$ is the beamforming vector at the BS with the constraint $b^2 \leq$ Pmax where Pmax is the maximum transmit power of the BS and s is the transmitted signal satisfying E[s^2]=1.

For the considered system, received signal for the user can be written as:

$$y = (h_r^H \phi G + h_d^H) bs + n \tag{3.2}$$

Here $\phi = diag(e^{j\theta_1} e^{j\theta_2} \dots e^{j\theta_n})$ is the phase-shift matrix at the IRS, $\theta_i \in [0, 2\pi]$ is the phase-shift angle which needs to be optimized. The phase-shift matrix satisfies constant-modulus constraints $|\phi_i|^2 = 1 \; \forall$ i = 1...N, and n CN(0, σ^2) is the Additive White Gaussian Noise (AWGN)

$h_r \in C^{NxM}$ and $h_r \in C^{Nx1}$ are the channels between BS-IRS and IRS-user. In practice, the IRS is usually deployed in a position with line-of-sight (LOS) to both transmitter and receiver, it is desirable to adopt the Rician fading model. Both channels follow the Rician fading model:

$$G = \sqrt{PL_G} \left(\sqrt{\frac{K_1}{K_1 + 1}} \bar{G} + \sqrt{\frac{1}{K_1 + 1}} \tilde{G} \right) \tag{3.3}$$

$$h_r = \sqrt{PL_h} \left(\sqrt{\frac{K_2}{K_2 + 1}} \bar{h_r} + \sqrt{\frac{1}{K_2 + 1}} \tilde{h_r} \right) \tag{3.4}$$

where K_1 and K_2 are Rician-K factors, $\tilde{G} \in C^{NXM}$ $\tilde{h_r} \in C^{NX1}$ are the non-LOS (nLOS) random components with i.i.d distributed $CN(0, 1)$ elements. The LoS components $\bar{G}$ and $\bar{h_r}$ are:

$$\bar{G} = \left[a_{N_x}^H (sin(\theta_{AOA,h})) \otimes a_{N_y}^H (sin(\theta_{AOA,v})) \right] a_M (sin(\theta_{AOD,b})) \tag{3.5}$$

$$\bar{h_r} = a_{N_y}^H (sin(\theta_{AOD,v})) \otimes a_{N_x}^H (cos(\theta_{AoD,v}) * sin(\theta_{AoD,h})) \tag{3.6}$$

where $a_i(\phi) = \left[1 \;\; e^{-j2\pi \frac{d}{\lambda} \phi} \;\; e^{-j2\pi * 2 * \frac{d}{\lambda} \phi} \;\; e^{-j2\pi * (i-1) * \frac{d}{\lambda} \phi} \right]$

where, $\theta_{AOA/AOD}$ represents angles of departure/arrival in horizontal/vertical directions at the IRS, $\theta_{AOD,b}$ is the angle of departure at the BS. Here, path loss model followed is :

$$PL = PL_0 - 10\epsilon log_{10}(d/D_0) dB \tag{3.7}$$

ϵ is the path-loss exponent, d is the distance between the BS-user.

The channel between the BS-user follows Rayleigh fading model where:

$$h_d = \sqrt{(PL_d)}\tilde{h_d} \tag{3.8}$$

where, $\tilde{h_d} \in C^{Mx1}$ is the nLOS component of the channel.

Thus, the recieved SNR for the user can be written as:

$$\gamma = \left|(h_r^H \phi G + h_d^H)b\right|^2 / \sigma^2 \tag{3.9}$$

The problem is a joint optimization problem to maximize the SNR by finding the optimum beam-forming vector at the BS and the optimum phase-shift design.For a fixed phase-shift matrix ϕ, the optimum beamforming vector that maximizes the SNR is the Maximum Transmission Rate(MRT) i.e.:

$$b^* = \sqrt{P_{max}} \frac{(h_r^H \phi G + h_d^H)^H}{||h_r^H \phi G + h_d^H||} \tag{3.10}$$

Substituting this in 3.9, the optimization problem for the phase-shift matrix ϕ to maximize γ reduces to:

$$\max_{\phi} ||h_r^H \phi G + h_d^H||^2 \tag{3.11}$$

$$s.t. |\phi_{i,i}| = 1 \forall i = 1, ..., N \tag{3.12}$$

This is a NP-hard problem owing to the non-convexity of the objective function and the unit modulus constraints. Many methods discussed in the previous section have been proposed but are not efficient/computationally complex. An effective solution proposed to is the Deep Reinforcement Learning (DRL) based framework which we investigate further.

3.2 Proposed DRL-Based Phase Control for IRS

Overview of Deep Reinforcement Learning

Reinforcement Learning(RL) is a type of machine learning technique that enables an agent to learn in an interactive environment by trial and error using feedback from

its own actions and experiences. The agent, environment, state, reward, policy, and value are the key variables that describe the RL problem. The environment is the physical environment in which the agent functions, the state is the current state of the agent, the reward is the feedback the agent receives from the environment, the policy is the process of mapping the agent's current state to actions and value is the agent's future reward that it receives when taking action in a specific state.

The agent at time t gets state s_t from the environment and selects action a_t based

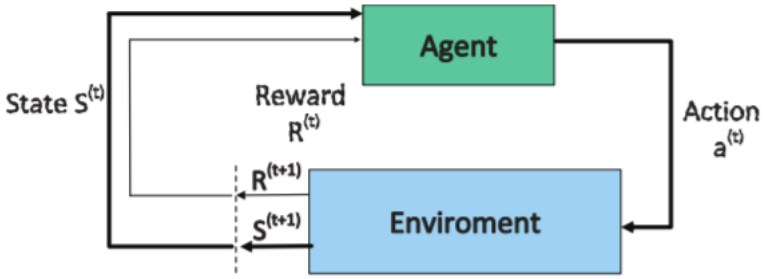

Figure 3.2: DRL framework

on policy π. After selecting the action the state changes from s_t to s_{t+1} and generate reward r_t. The objective of RL is to maximize the total reward, the reward is the performance measure of a specific action during the current state. RL models the interaction between the agent and the environment as Markov's decision process (MDP). Many RL algorithms use dynamic programming methods, but the prime difference between RL algorithms and dynamic programming is that RL algorithms don't assume the awareness of a specific MDP's mathematical model. Two kinds of algorithms, i.e., the value based and policy based algorithms, are usually applied to determine the optimal policy RL algorithms are also designed for big MDPs where specific processes are impractical.

Q-Learning

Q-learning (QL) is a model-free RL method since it does not require a model, it obtains the best action under the current state, where "Q" denotes the quality. Further, it is an off-policy algorithm since QL learns actions that are outside the current policy so it may take actions that are random and therefore the policy is not needed here. When the agent interacts with the environment and needs to update the action-state pairs in

the Q-table, it has two choices, either to explore or exploit. Exploration is to explore the environment in order to obtain information and knowledge about it, in this case the agent acts randomly. Whereas exploitation is to exploit the available "already known" information about the environment so as to maximize the reward. Precisely, QL selects the optimal policy in order to maximize the total reward. The trade-off between exploration and exploitation is balanced by using a specific parameter which can be set based on how often we want to exploit or explore. Moreover, QL is based on the concept of a Q-function, which is a function of states and actions. The Q-function $Q_\pi(s, a)$ measures the expected discounted sum of rewards or the return achieved from selecting action a in a state s based on policy π. It evaluates how good a particular action is, for a given state. The optimal Q-value Q(s, a) is defined as the maximum return achieved from a given state s and action a while obeying the optimal policy. In order to update the Q-value for any action Bellman equation is used. It provides us the best reward and the optimal policy to achieve this reward. It means that the maximum value of the return from the action and state is equal to the expectation of the current maximum possible reward plus the maximum possible long-term reward achieved from the next state s' discounted by $\lambda \in [0, 1]$ which is the discount factor, and the equation is expressed as:

$$Q*(s, a) = \mathcal{E}[r(s, a) + \lambda \max_{a'} Q*(s', a')] \tag{3.13}$$

where E[.] denotes the expectation. The updates will occur after each action and it will end when the episode finishes. In order to converge and learn optimal values, the agent needs to learn and explore many episodes. Furthermore, there are three important steps in the update, the first step is that the agent will take action for each state and receive a reward. The second step is that the agent will either select the action by checking the highest value in the Q-table or by random . The third step is to update the Q-values employing the following formula:

$$Q^{new}(s, a) = (1 - \alpha)Q(s, a) + \alpha[r(s, a) + \lambda \max_{a'} Q*(s', a')] \tag{3.14}$$

The step size is adjusted using the learning rate α which measures the acceptance of

the new value compared to the old one.

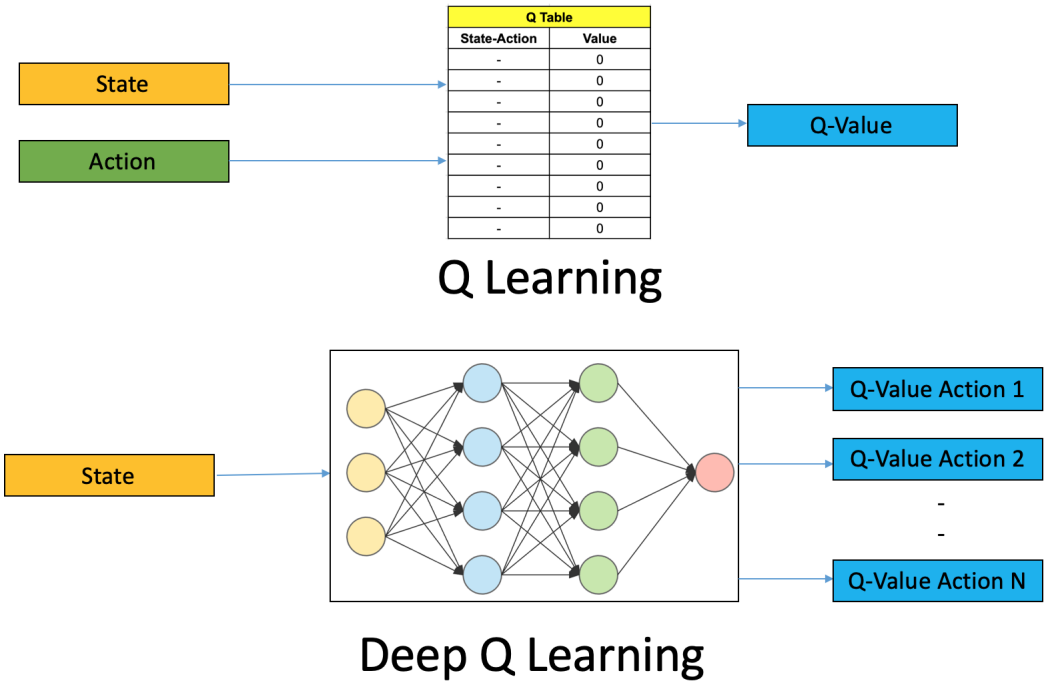

Figure 3.3: Q-Learning vs. Deep Q-Learning(DQN

Deep Q-Learning

Deep Q-Learning (DQL) uses neural network that approximates the Q-function. Here, instead of calculating Q-values immediately for each action and state via iterations, we will utilize a function approximator for estimating Q-values for each possible action. To perform this neural networks are employed. The Neural network accepts states as input from the environment and generates the estimated Q-values for every action the agent can select. Furthermore, in many problems in DQL it is not practical to represent the Q-function based on s and a only, because the neural network is trained with parameters θ to evaluate the Q-values such that Q(s, a, θ) ≡ Q* (s, a)

$$Q_\pi(s, a, \theta) = E_\pi[R^{(t)}_{s(t)=s,a(t)=a}] \tag{3.15}$$

where $R^{(t)} = \sum \gamma^{(t)} r^{(t)}$ is the cumulative expected reward. The aim of the DQN is to maximize the value of the Q-function in 3.15 by DNN training. The optimal Q-value is obtained from the Bellman equation. Then, we can calculate the loss by comparing

the Q-value for a given state-action pair to the target value which is the right hand side of the Bellman equation expression yielding the following expressions:

$$Loss = Q*(s,a,\theta) - Q(s,a,\theta) \tag{3.16}$$

$$E[r(s,a,\theta) + \lambda \max_{a'} Q*(s',a',\theta)] - E[\sum_{t=0}^{\infty} \lambda^t r^t] \tag{3.17}$$

Next, the NN will update its weight values in the policy network based on the gradient descent and back propagation algorithms. This process is repeated periodically for many episodes until we sufficiently minimize the loss.

Policy Gradient:

Policy gradient (PG) is on-policy, model-free RL method, used for **continuous action space**. The PG agent is a RL agent which is policy-based, it aims directly to maximize the expected reward by obtaining a parametrized policy that generates a trajectory τ , where the trajectory represents the states, actions and rewards i.e. $s_0, a_0, r_1, s_1, a_1, r_2, ...$ In other words, we need to find out the parameters θ that maximizes J, where θ represents the weights of the neural networks.

$$J(\theta) = E_\pi[r(\tau)] \tag{3.18}$$

where r(τ)denotes the total reward for a specified trajectory τ. A well known approach in machine learning to solve the maximization problem is the gradient ascent to step through the parameters. Therefore, the policy training is demonstrated as a gradient ascent process:

$$\theta_{t+1} = \theta_t + \alpha E_{\pi_{\theta_t}}[r(\tau)\delta_{\theta t} log(\pi_\theta(\tau)] \tag{3.19}$$

where α is the learning rate. where $Q_{\pi_{\theta_t}}(\tau)$ is the Q-value for the trajectory τ , and policy π_{θ_t}. Throughout the training, the PG agent estimates the probability of selecting each action and chooses actions by random depending on the probability distribution. Before it learns from experience and updates the policy parameters, the PG agent performs a full training episode utilizing the existing policy. The disadvantage of PG

method is that the network policy is updated after the completion of the episode only. This leads to slow convergence of the policy gradient algorithm.

3.3 DDPG based framework for IRS Phase-control

A DDPG based algorithm is developed in [1] to solve the IRS optimization problem. Since they only deal with discrete time spaces, Deep Q-Networks are ineffective. Furthermore, in the setting of wireless communication, the convergence of the policy gradient (PG) method is insufficient. DDPG merges the Q-networks and the PG scheme and overcomes the disadvantages of both algorithms.

DDPG is a model-free reinforcement learning technique which combines the advantages of policy gradients and Q-learning. DDPG uses the Bellman equation and the off-policy data to learn the Q-function, and then uses the Q-function to learn the policy. DDPG consists of four neural networks ; one for actor network, one for critic network, one for target actor network, and one for target critic network, which ensures the stability. The optimization problem in can be solved using DDPG by learning the policy.

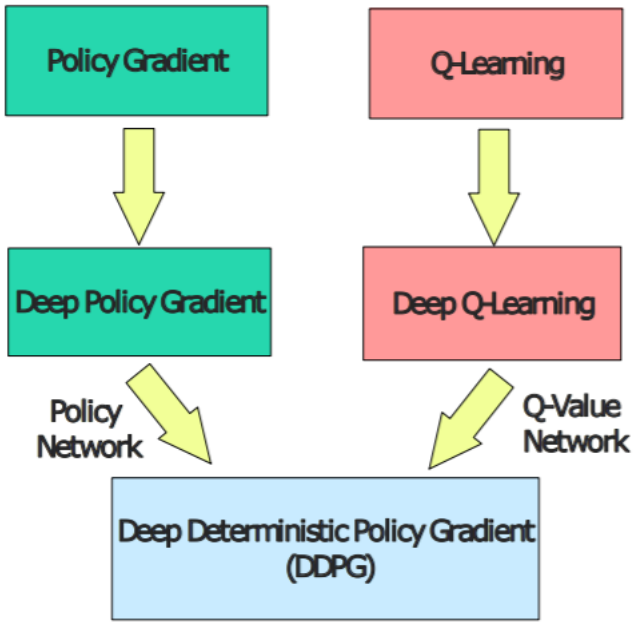

Figure 3.4: DQN+ DPG = DDPG

The actor network is a policy network that accepts state as input and generates the **precise** action continuously. In deep Q-networks the optimal action is obtained by calculating the *argmax* over all the Q-values for a finite number of discrete actions. In DDPG the actor network performs the same but for continuous action spaces, it generates the actions **directly** by taking the argmax, and selects actions a = $\mu(s|\theta^{\mu})$ from a continuous action settings $\boldsymbol{A}$, where μ is the policy, s denotes the states, and θ^{μ} denotes the parameters of the **deterministic policy** network (DPN). Thus, the actor is a DPN that calculates the action directly, rather than generating the probability distributions over all actions. The critic network accepts states and actions as input and generates the Q-value, so it is considered as the Q-value network Q(s, a—θ^{Q}), where θ^{Q} represents the parameters of the Q-network. The critic network evaluates the performance of the selected action. Therefore, DDPG is an enhancement for the actor-critic vanilla network and its aim is to maximize the Q-value which is the output, and it can only be utilized for environments having continuous action settings. The optimal action is expressed as:

$$\mu * (s|\theta^{\mu}) = arg \max_{a} Q * (s, a|\theta^{Q}) \tag{3.20}$$

where Q*(s, a|θ^{Q}) is the optimal Q-value function. Further, to maximize the Q-value a replay memory $\mathcal{D}$ is utilized to minimize the correlation of various training samples. This is significant for the algorithm behavior in order to be stable. The replay memory needs to be adequately large to include a broad range of previous experiences. Moreover, DDPG makes use of target networks to increase the stability during the training.

A copy from the actor and critic network are formed to find out the Q-value for the next state i.e.a = $\mu'(s|\theta^{\mu'})$ and Q'(s, a—$\theta^{Q'}$). Based on the main networks, the weights of these target networks are updated periodically. In deep Q-networks, the weight of the main network is copied periodically to the target network and this called ”hard update”, whereas in DDPG ”soft update” is performed where a only fraction of the weights of the main network are transferred to the target network:

$$\theta^{Q'} = \tau\theta^{Q} + (1 - \tau)\theta^{Q'} \tag{3.21}$$

$$\theta^{\mu'} = \tau\theta^{\mu} + (1-\tau)\theta^{\mu'} \tag{3.22}$$

where $\tau \leq 1$. Soft updates are significant in order to accelerate the convergence of the Actor-Critic process since it stabilizes learning. The main networks which are copied are called evaluation networks. The target networks and the evaluation networks have the same structure but the difference is in parameters.

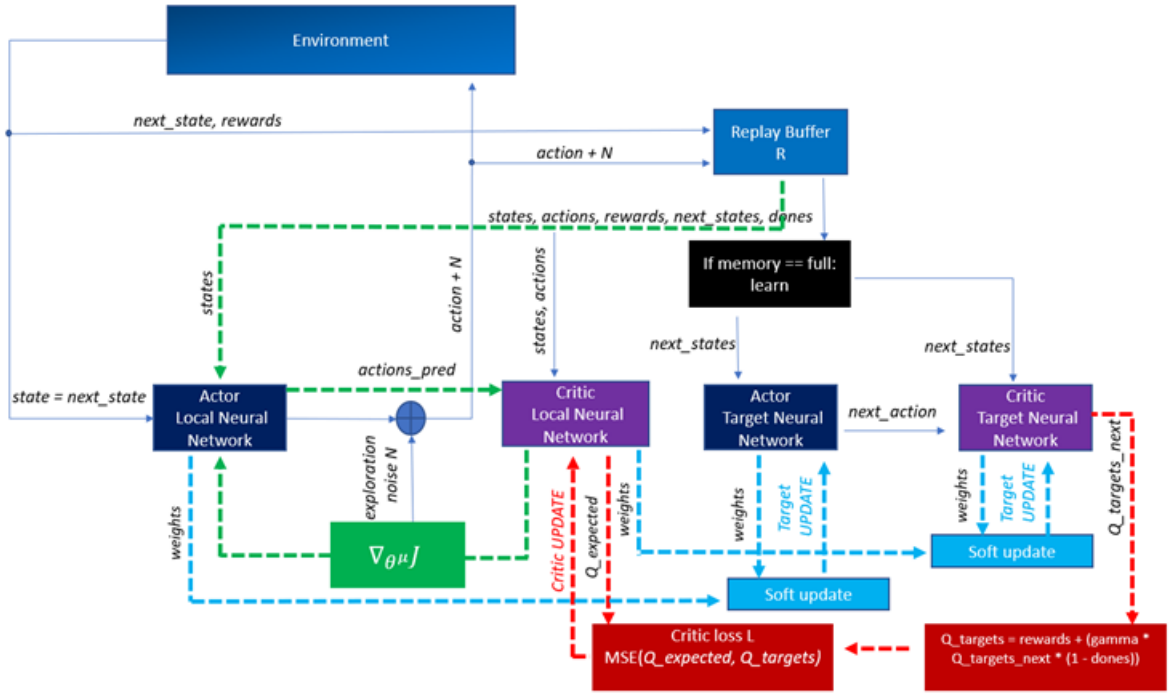

Figure 3.5: DDPG block diagram

Mapping IRS problem to DDPG framework

Next, we discuss the mapping of IRS problem to DDPG components namely action-space, state-space, reward function and also the working procedure of the DDPG algorithm.

- State Space: The state space of the DDPG agent at timestep (t) can be defined as follows:

$$s^t = \left[\theta_1^{t-1}, \theta_2^{t-1}\theta_n^{t-1}, \gamma^{t-1}\right] \tag{3.23}$$

  where $\theta_i i = 1...N$ are the phase-shift angles and γ is the SNR.

- Action Space: The action space definition is defined by the policy function as

follows:

$$a^t = \mu(s^t|\theta^\mu) + n_t \tag{3.24}$$

where, μ is the policy function and θ^μ its parameters. n(t) is the random process-Ornstein-Uhlenbeck (OU) process-based action noise. Final output is an array that defines the phase of each element in the IRS.

- Reward function: The objective is to maximize the received SNR. Thus, the received SNR defined in used as the reward, i.e.

$$r_t = \gamma^{(t)} \tag{3.25}$$

- Exploration vs Exploitation: Since the action space of the DDPG is continuous, the exploration of action space is handled with noise generated by the OU process. OU process samples noise from a correlated normal distribution.

DDPG Algorithm design and working procedure

- As shown in Algorithm 1, we begin by initializing the actor and critic evaluation networks, along with actor and critic target networks. The target networks are initialized by copying the same weights as evaluation networks. Then we initialize the experience replace D with maximum capacity C

- Step 1 represents looping through the episodes. In step 2, we obtain the current state information(CSI) for the episode. In step 3, we randomly initialize the phase-shift angles to obtain the initial state s^0. In step 4, OU process is initialized.

- From step 5 to 13, represents each iteration (i.e., timestep t). In each time-step, we observe the state s for the agent (IRS), determine an action (i.e., phase value) with exploration noise based on OU process in step 6. After the agent determined and executed the action, a reward r(t) is received and new state s(t+1) is observed, and transactions are stored in respective replay memories in step 7. A random mini-batch of transitions are sampled in step 8. Using the Bellman equation, the actors and critic networks' targets are computed in

Algorithm 1 The DRL Based Framework

Input: The discount factor λ, the soft update coefficient τ, the learning rate α, 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 γ^* under current channel state.

1: **for** episode $j = 1, \cdots, K$ **do**
2: Obtain the current CSI $(\mathbf{h}_r^{(j)}, \mathbf{G}^{(j)}, \mathbf{h}_d^{(j)})$;
3: Randomly chose phase shifts to obtain $\boldsymbol{\Phi}^{(0)}$ and $\gamma^{(0)}$ as initial state s_1;
4: Initialize a random process $\mathcal{N}$;
5: **for** $t = 1, \cdots, T$ **do**
6: Action $a_t = \mu(s_t; \boldsymbol{\theta}_\mu) + \mathcal{N}$;
7: 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 $\mathcal{D}$.
8: Sample a N_B minibatch transitions $\{s_j, a_j, r_j, s_{j+1}\}$ from $\mathcal{D}$.
9: Set target Q value according to (11).
10: Update $Q(s, a; \boldsymbol{\theta}_q)$ by minimizing the loss in (12).
11: Update the policy $\mu(s; \boldsymbol{\theta}_\mu)$ using the sampled policy gradient in (13).
12: Soft update the target networks according to (7).
13: Update the sate $s_t = s_{t+1}$.
14: **end for**
15: **end for**

step 9. The critic network weights are updated by minimizing the loss using computed targets in step 10. The actor network weights are updated for the sampled policy gradient in step 11.

- Finally, the agent's target networks are updated using the update rate (τ) for stability in step 12.

Chapter 4

Numerical Results

The algorithm and the procedure developed in the paper along with numerical simulations has been coded in Python using Colab environment.

4.1 Simulation experiments

First and foremost, we simulate the channel gains for different channels involved in the IRS problem. The details are discussed below: The BS-IRS and IRS-user channels follow Rician fading model and are given by:

$$G = \sqrt{PL_G}\left(\sqrt{\frac{K_1}{K_1+1}}\bar{G} + \sqrt{\frac{1}{K_1+1}}\tilde{G}\right) \tag{4.1}$$

$$h_r = \sqrt{PL_h}\left(\sqrt{\frac{K_2}{K_2+1}}\bar{h_r} + \sqrt{\frac{1}{K_2+1}}\tilde{h_r}\right) \tag{4.2}$$

$$PL = PL_0 - 10\epsilon log_{10}(d/D_0)dB \tag{4.3}$$

For the experiments, the reference path-loss PL_0 = -30dB corresponding to reference distance D_0 = 1m. The path loss exponent ϵ for BS-User and IRS-User links is 2.8 whereas for BS-IRS link is 2. For all the experiments, we consider a scenario where the IRS is coated on the facade of a tall building and is aware of the BS's location, so that K1 $-> \infty$ and K2$-> 0$. Over 500 realizations of the channels' random components are averaged to obtain the simulation results

Simulation parameter	Value
Number of reflecting elements(M)	10
Number of IRS elements (N)	50 (Nx = 10, Ny = 5)
Distance between BS and IRS	51m
BS transmit power(Pmax)	5dBM
σ^2	80 dBM
T, No. of time-steps in each episode	1000
No. of episodes	50
Learning rate, α	10^{-3}
Discount factor, λ	0.95
Soft-update coefficient, τ	0.005
Buffer capacity,C	50000
Batch-size, N_B	16
IRS: $\theta_{AOA,h}$, $\theta_{AOA,v}$	$\pi/3$, $5\pi/3$
BS: $\theta_{AOD,b}$	$\pi/6$
IRS: $\theta_{AOD,h}$, $\theta_{AOD,v}$	$\pi/6$, $5\pi/6$

Actor and Critic network architecture

In the proposed DDPG algorithm, the actor and critic networks are both dense neural networks (DNN) with 4 layers. The input of the actor network is the number of states that contains N+1 neurons while the output is the number of actions which contains N neurons. The hidden layers in the actor and critic network contain 300 and 400 neurons each, followed by ReLU activation function. The output layer of the actor and critic network uses the tanh(·) function in order to provide enough gradient. For the critic network, the input layer is the number of states and the number of actions. The state input with 2N+1 neuronsand gives Qvalue output with 1 neuron. Both actor and critic main networks use Adam optimizer to update parameters.

4.1.1 Results

Fig 4.1 demonstrates the recieved SNR for each episode for a single run of the algorithm. The average reward over all the episodes converges to 20dB with BS-

user distance, d = 48m. This demonstrates the convergence of the DDPG algorithm.

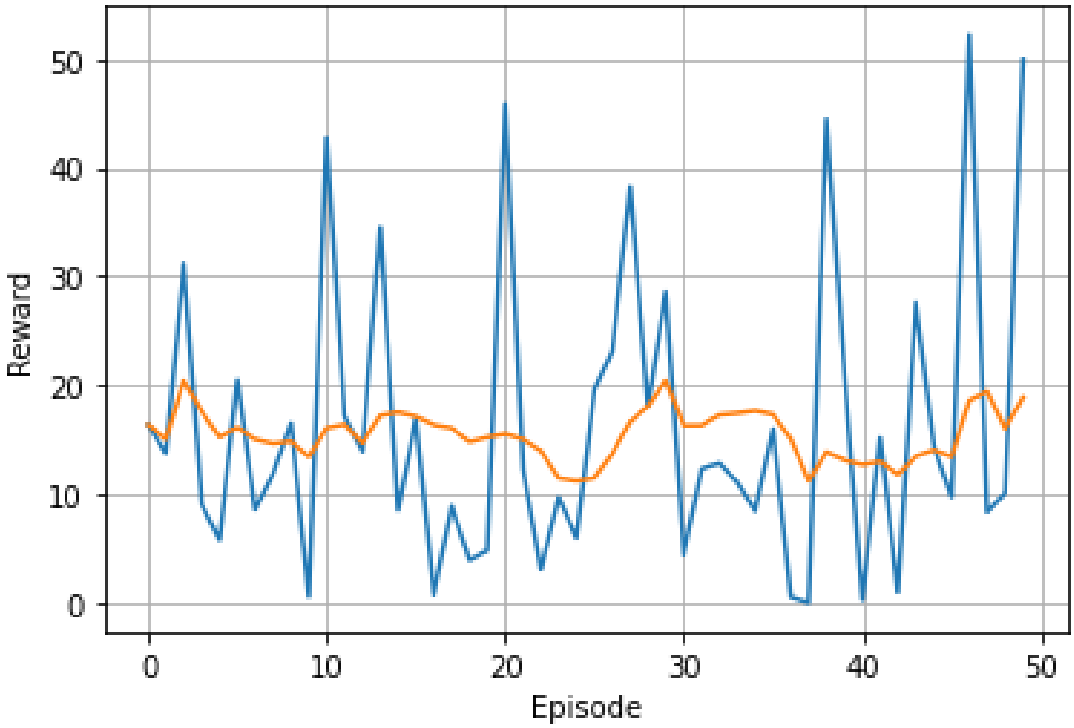

Figure 4.1: Epoch vs. SNR for the proposed algorithm

Next, Fig 4.2 gives the analysis of recieved SNR with varying BS-user distance d. The signal strength, SNR improves as the user moves away from the BS towards the IRS.The best BS-user distance seems to be 90m.

Fig 4.3 clearly demonstrates the effectiveness of using IRS protocol in the MISO system. As we increase the size of (number of elements in) the IRS, the SNR significantly improves.

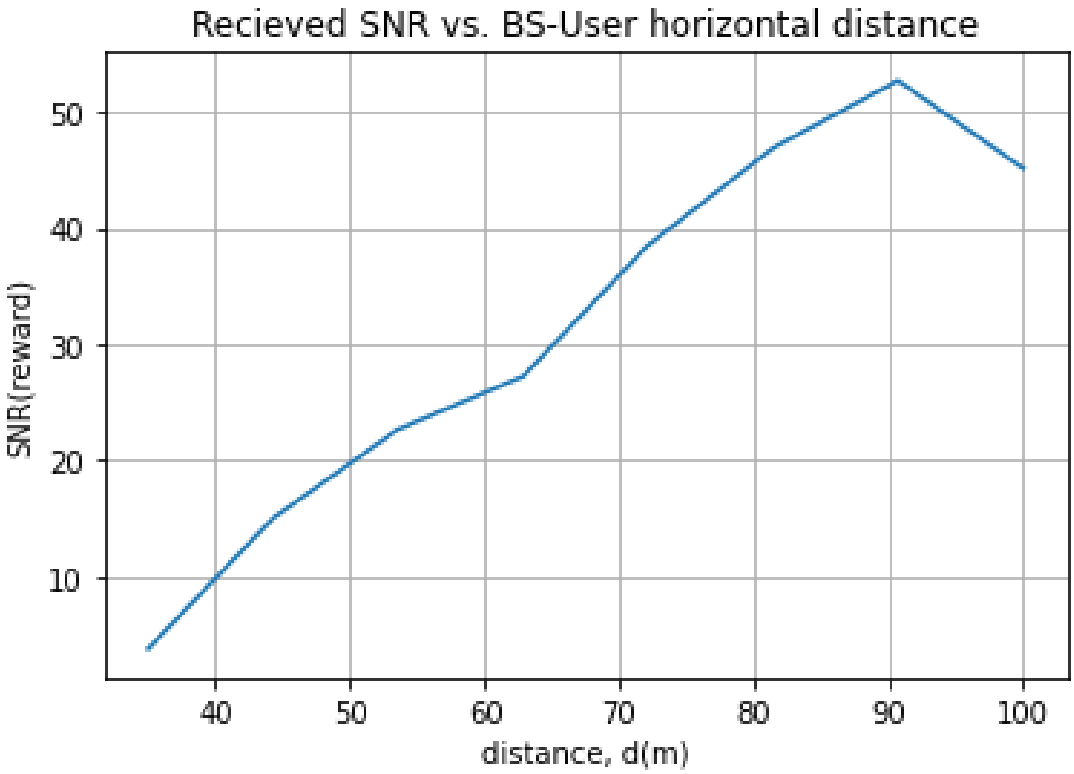

Figure 4.2: SNR vs BS-User horizontal distance

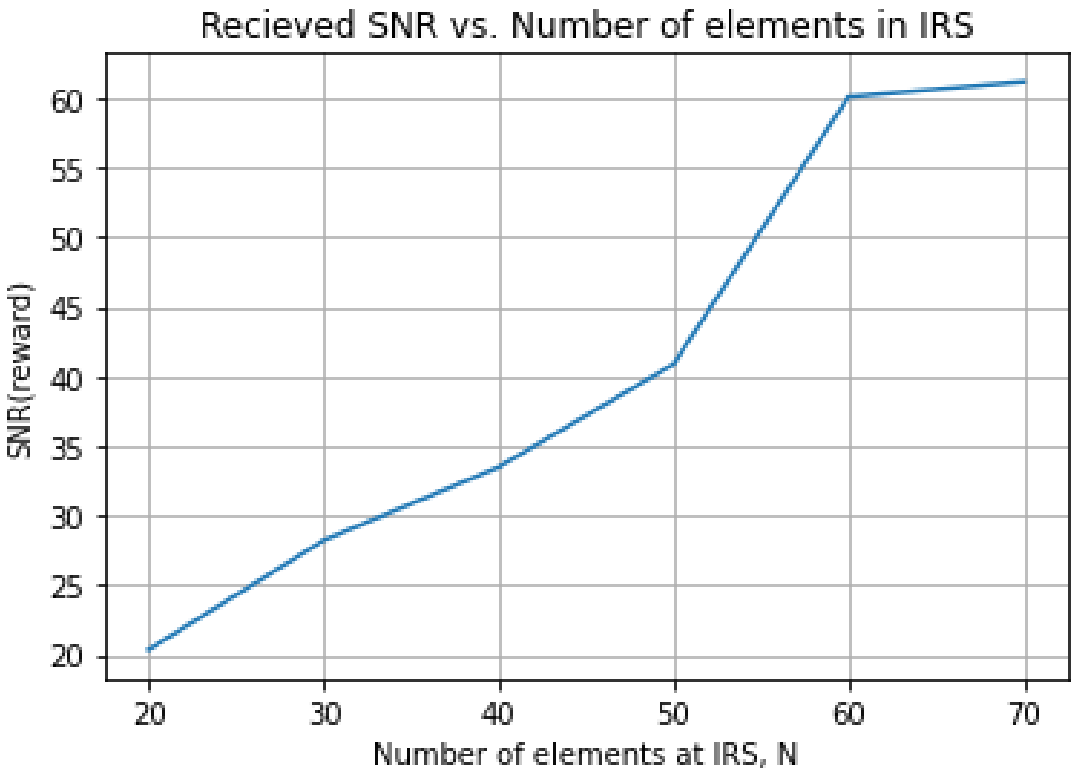

Figure 4.3: SNR vs No. of elements in IRS(N)

4.2 Modifications

The DDPG algorithm is modified to take into consideration that the channels between IRS-User are unknown. The RL mapping in this case changes to:

- State Space: The state space of the DDPG agent at timestep (t) now can be defined as follows:

$$s^t = \left[h_{r1}^t h_{r2}^t .. h_{rN}^t \theta_1^{t-1}, \theta_2^{t-1} \theta_n^{t-1}, \gamma^{t-1}\right] \tag{4.4}$$

  where h_r^t is the IRS-user unknown channel, $\theta_i i = 1...N$ are the phase-shift angles and γ is the SNR.

- Action Space: The action space definition is defined by the policy function as follows:

$$a^t = \mu(s^t|\theta^\mu) + n_t \tag{4.5}$$

  where, μ is the policy function and θ^μ its parameters. n(t) is the random process-Ornstein-Uhlenbeck (OU) process-based action noise. Final output is an array that defines the channel gain vector for IRS-USer link and the phase of each element in the IRS.

- Reward function: The objective is to maximize the received SNR. Thus, the received SNR defined in used as the reward, i.e.

$$r_t = \gamma^{(t)} \tag{4.6}$$

- Exploration vs Exploitation: Since the action space of the DDPG is continuous, the exploration of action space is handled with noise generated by the OU process. OU process samples noise from a correlated normal distribution.

With this modification, the actor and critic neural network architecture is modified to- The input of the actor network is the number of states that contains 2N+1 neurons while the output is the number of actions which contains 2N neurons. For the critic network, the input layer is the number of states and the number of actions. The state input with 3N+1 neurons and gives Qvalue output with 1 neuron. Rest all

specifications are same.

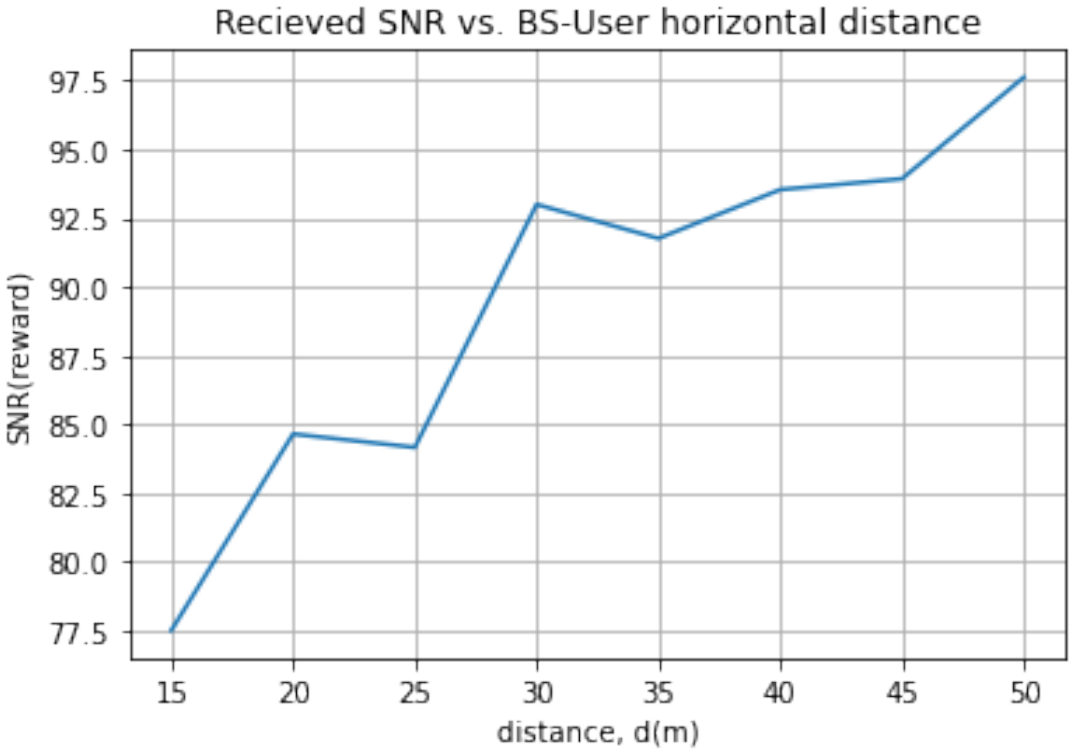

Figure 4.4: SNR vs BS-USer horizontal distance,d

Fig 4.4 demonstrates the increasing trend of SNR as BS-User horizontal distance increases, when the channel gains are also learned by the network.

Chapter 5

Conclusion and Future Work

We researched and examined the most promising technology for upcoming 6G networks in this study. We focused on the MISO downlink scenario, which was aided by Intelligent Reflecting Surface. The formulated problem is non-convex because it has a constant modulus constraint and a non-convex goal function. As a result, DRL learning approaches are appropriate for this situation. We've employed the DDPG, a DRL method, to solve the SNR maximisation problem, which results in the optimal phase shift matrix by tuning the IRS elements. To give a more realistic approach to the given problem, we modified the IRS-DDPG framework to consider the channel gains between the IRS and the user to be unknown. In both situations, the observed results are similar, with an increasing trend in SNR.

Future directions suggested by the study and evaluation are as follows: To begin with, this study, like the majority of others on IRSs and their applicability to wireless communications, is based on theoretical analysis with simulations serving as validation. As a result, confirming theoretical results with data from real-world system implementations and tests is an essential study field. Second, if data is received in an online fashion, we can reformulate the aforementioned optimization problem—this is an interesting study direction (contextual bandits). Next, existing models for how IRSs alter incident signals are straightforward. The behaviour of an IRS, on the other hand, is determined by its physical materials and manufacturing techniques. Models taking these issues into account can more accurately guide the optimization of IRSs for aiding wireless communications.

Bibliography

[1] Keming Feng, Qisheng Wang, Xiao Li,Chao-Kai Wen, P., 2020. Deep Reinforcement Learning Based Intelligent Reflecting Surface Optimization for MISO Communication Systems. arXiv preprint arXiv:2162-2345.

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

[3] X. Yu, D. Xu, and R. Schober, "MISO wireless communication systems via intelligent reflecting surfaces," in Proc. IEEE/CIC ICCC, Aug. 2019, pp. 735–740.

[4] Yuhang Jia, Graduate Student Member, IEEE, Chencheng Ye, and Ying Cui, Member, IEEE, Analysis and Optimization of an Intelligent Reflecting Surface-assisted System with Interference.

[5] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," IEEE Wireless Commun. Lett., vol. 9, no. 6, pp. 760–764, Jun. 2020.

[6] Q. Wu and R. Zhang, "Beamforming optimization for intelligent reflecting surface with discrete phase shifts," in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), May 2019, pp. 7830–7833

[7] Z. Chu, W. Hao, P. Xiao, and J. Shi, "Intelligent reflecting surface aided multi-antenna secure transmission," IEEE Wireless Commun. Lett., vol. 9, no. 1, pp. 108–112, Jan. 2020..