February 2023
Master’s Degree Thesis

Deep Reinforcement Learning-Based RIS-Aided Wireless Communication Systems

Graduate School of Chosun University

Department of Computer Engineering

K M Faisal

Deep Reinforcement Learning-Based RIS-Aided Wireless Communication Systems

심층강화학습 기반 RIS 지원 무선 통신 시스템 연구

February 24, 2023

Graduate School of Chosun University

Department of Computer Engineering

K M Faisal

Deep Reinforcement Learning-Based RIS-Aided Wireless Communication Systems

Advisor: Prof. Wooyeol Choi, Ph.D.

A thesis submitted in partial fulfillment of the
requirements for a Master’s degree

October 2022

Graduate School of Chosun University

Department of Computer Engineering

K M Faisal

This is to certify that the master’s thesis of

K M Faisal

has been approved by examining committee for the thesis requirement for the master’s degree.

파이살 케이엠의
석사학위논문을 인준함

위원장	조선대학교	교수	신석주	
위원	조선대학교	교수	강문수	(인)
위원	조선대학교	교수	최우열	(인)

2022년 12월

조선대학교 대학원

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Deep Reinforcement Learning-Based RIS-Aided Wireless Communication Systems

K M Faisal
Advisor: Prof. Choi, Wooyeol, Ph.D.
Department of Computer Engineering
Graduate School of Chosun University

Reconfigurable Intelligent Surface (RIS) can offer a customizable wireless transmission and is regarded as an incredibly crucial enabling technology when used as reflectors for current wireless base stations (BSs) to overcome the blockage challenges of millimeter wave (mmWave) wireless communications systems. It is a promising sixth-generation (6G) and beyond wireless services strategy to offer Gbps data throughput for networks operating at frequencies beyond 28 GHz. Furthermore, RIS has a tremendous opportunity to mitigate the blockage impact and significantly lower the needless switching because of its capacity to enhance the scattering environment and produce reflecting signal multipath. However, because there are so many RIS elements, optimising the BS and reflector RIS configuration is complex and will result in performance loss. Due to the growing popularity of deep reinforcement learning (DRL) in this thesis, we employ a twin-delayed deep deterministic policy gradient (TD3) approach to solve non-convex optimization problems where the BS gets state data from the RIS, composed of feedback from the channel states of users. Consequently, for real-world systems with continuous phase-shift and beamforming matrix control, the BS ensures optimal action constituted of the transmission power allotment of BS and phase-shift configuration in the Nakagami-m fading environment. The experimental findings demonstrate that the suggested solutions outperform other existing benchmarks.

한 글 요 약

심층강화학습 기반 RIS 지원 무선 통신 시스템 연구

파이살 케이엠
지도교수: 최우열
컴퓨터공학과
조선대학교 대학원

재구성 가능한 지능형 표면(RIS)은 맞춤형 무선 전송을 제공할 수 있으며 밀리미터파(mmWave) 무선 통신 시스템의 차단 문제를 극복하기 위해 현재 무선 기지국(BS)의 반사경으로 사용될 때 매우 중요한 활성화 기술로 간주됩니다. 28GHz 이상의 주파수에서 작동하는 네트워크에 Gbps 데이터 처리량을 제공하는 것은 유망한 6세대(6G) 및 무선 서비스를 넘어서는 전략입니다. 또한 RIS는 산란 환경을 개선하고 반사 신호 다중 경로를 생성할 수 있는 능력 때문에 막힘 영향을 완화하고 불필요한 전환을 크게 낮출 수 있는 엄청난 기회를 가지고 있습니다. 그러나 RIS 요소가 너무 많기 때문에 BS 및 반사경 RIS 구성을 최적화하는 것이 복잡하고 성능 손실이 발생합니다. 이 백서에서 심층 강화 학습(DRL)의 인기가 높아짐에 따라 우리는 BS가 피드백으로 구성된 RIS에서 상태 데이터를 가져오는 비볼록 최적화 문제를 해결하기 위해 쌍 지연 심층 결정적 정책 기울기(TD3) 접근 방식을 사용합니다. 사용자의 채널 상태에서. 결과적으로 연속 위상 편이 및 빔포밍 매트릭스 제어가 있는 실제 시스템의 경우 BS는 Nakagami-m 페이딩 환경에서 BS의 전송 전력 할당 및 위상 편이 구성으로 구성된 최적의 동작을 보장합니다. 실험 결과는 제안된 솔루션이 기존의 다른 벤치마크를 능가함을 보여줍니다.

I. INTRODUCTION

Web-enabled gadgets, such as smartphones, have emerged as vital tools for global communication, information transfer, and entertainment. The academia and industry are now focusing on sixth generation wireless technology as the wireless sector is in a highly exciting moment where the fifth generation (5G) technology has been largely standardized and commercialized. According to the Cisco Annual Internet Report (2018-2023), mobile connectivity is expected to be available to more than 70% of the global population by 2023 and the number of overall mobile subscribers is expected to increase from 5.1 billion in 2018 to 5.7 billion in 2023 [1]. Inter-cell synchronization approaches have been built to solve the interference as cellular networks have become denser owing to more aggressive frequency reuse. However, the bandwidth of a network is still constrained owing to the irregularity of wireless transmission and accessible spectrum [2].

The limited availability of spectrum for communication systems is encouraging a gradual migration towards the higher frequency bands with abundant unoccupied spectra. However, as the radio frequency increases, the electromagnetic (EM) waves become more susceptible to obstruction from objects such as buildings in metropolitan regions. Adding more relays and base stations (BSs) to minimize communication distances and provide better network coverage consumes more energy. As a result, employing traditional cellular methods to assure wireless service coverage is challenging. To deal with the spectrum scarcity of communication systems,reconfigurable intelligent surfaces (RISs) have evolved as an important wireless network resolution for attaining high spectrum and energy efficiency [3]. For upcoming wireless communication networks such as beyond 5G, the RIS is projected as viable technology with the potential to significantly increase link quality and minimize the possibility of blockages. Small, low-cost passive components are piled together in the RIS to reflect incoming signals with a controllable phase shift toward the receiver. The comparatively simple deployment of RIS-assisted communications with affordable passive parts makes them valuable in smart radio contexts.

However, some certain challenges must be addressed before obtaining the advantages of RISs. To enhance the phase-shift configuration, several investigations are being undertaken. The majority of currently used methods are built on the convex optimization concept, which can lead to poor performance and requires a lot of time because the method needs to go through several rounds before it converges. The performance of the network would surely benefit from adjusting the reflection coefficients of every component on a continuous basis. However, doing so would be cost-prohibitive given the intricate design and high-tech nature of the massive high-precision components and with more reflecting components, their complexity would rise as well. Thus the first difficulty in establishing 6G networks with RIS support is the layout of adjustable components.

Owing to their ability to learn and the requirement of operating over wider search areas, machine learning (ML) techniques have attracted attention in wireless communications [4]–[8], especially in the field of RISs. Over the last few years, several researchers have attempted to overcome these obstacles. They have been working with various ML algorithms for the communication sector so that the infrastructure can independently solve all challenges. Most ML methods work by learning the parameters and constructing an optimization model from the input information for the goal function. In the present arena, as a massive amount of data must be handled, the efficiency and effectiveness of mathematical optimization procedures significantly impact the popularity and application of ML models [9]. Theoretically challenging nonlinear and non-convex problems can be solved using artificial intelligence (AI) approaches. Machine learning strategies have been used to accomplish phase control in wireless systems with RIS assistance and are more adaptable to stochastic system models than optimization techniques [10]. Specifically, deep reinforcement learning (DRL), which is regarded as an incredible potential contender in the future communication network confronting a variety of requirements, pursues the ideal approach through an agent-environment interaction learning process[11]. DRL occupies an important place for optimizing the RIS phase shifts with no need for offline training and dataset with labeling.

A. Related Works

The phase shift optimizations in the communication systems with RIS support have been researched in various early works of literature. The received signal will considerably affect the phase alterations on the RIS.The study in [12] focused on a single-user multi-user multiple-input, single-output (SU-MISO) system supported by RIS and determined the values of the phases to optimize the overall signal level obtained at the user end. The adaptive transmission situation of a RIS-aided uplink orthogonal frequency division multiplexing (OFDM) system was explored by the authors of [13], which relied on semidefinite relaxation technology to increase the mean achievable rate. By enhancing the RIS with discrete phase shift and transmit beamforming which is continuous of the BS, the study in [14] examines a RIS-aided downlink system architecture that reduces the transmit power of BS. Adjusting the incoming wave characteristics while taking into account realistic reflection coefficients and constrained RIS operating connection, a passive beamformer is suggested in [15] to attain an an asymptotic optimum result. Using statistical channel state information, the authors of [16] address the issue of optimising downlink capacity to build the ideal RIS phase shift. The authors of [17] take into account a comparable MISO downlink system with RIS assistance. Manifold optimization and fixed point iteration approach, which have been demonstrated to be beneficial in overcoming the unit modulus limitations of RIS-aided system, are used to resolve the collaborative optimization of the phase shifts and the access point transmit beamforming, respectively.

To increase the attainable rate of OFDM, it was suggested by [18] to optimize reflection coefficients of the RIS and allocation of transmit power. The performance of Spectral Efficiency in MISO systems with zero-forcing in RIS-assisted systems was evaluated by the authors in [19]. In order to maximize SE, nonlinear proportional rate constraints are applied to both the phase shift at the RIS and the transmission power allotment at the BS. A mathematical structure for understanding the error performance of communication networks relying on RIS is offered in [20].

By simultaneous phase shifts optimization at the RIS and precoding matrices at the BS under the power and unit modulus limitations imposed on each BS, [21] focused on enhancing the weighted sum rate. The authors of [22] investigated the through the optimal linear precoder (OLP), which enhances the minimal level SINR related to a specified power restriction for specified phase matrix of RIS and developed deterministic approximations for asymptotically OLP parameters, to have better the RIS phase matrix. To optimize the WSR while adhering to the BS transmit power restriction, a combined passive and active beamforming issue is focused in a multiuser downlink MISO system with RIS assistance using a fractional programming technique in [23]. In [24], convex radio resource allocation was dealt with by integrating alternating optimization and majorization-minimization to produce a low-complexity and convergent method in a sum-rate maximization approach for a RIS-based, multiuser MIMO system. The RIS phase components and the BS transmit beamforming vector were optimized for the purpose of maximization of the secrecy rate using the alternative optimization (AO) algorithm by the authors in [25] and [26]. However, their concepts were not extended to secure multi-user RIS-assisted communication systems. The reflecting beamforming at RIS and the beamforming at BS in RIS-aided communication networks, which are less suitable for extensive systems, were optimized primarily using conventional optimization approaches in the research above. The processing demands of large-scale heterogeneous communication networks are challenging for complicated numerical optimization and mathematical calculation techniques. The model-free machine learning approach has emerged recently as an outstanding tool for addressing theoretically unsolvable non-linear difficulties and high-dimension, complex EM environments of communication systems [27]–[30]. ML Methods can be utilized in upcoming 6G wireless communication systems to cope with non-trivial difficulties caused by extraordinarily high dimensions in big-scale MIMO systems, according to an enormous body of research concerns and findings [31]. A minimal variance unbiased estimator-based RIS channel estimation approach is utilized in [32]. By assigning the arriving pilots to the cascaded and direct channels, a supervised learning architecture was employed in [33] for the estimation of channel. Utilizing methods from both supervised and unsupervised deep learning, the authors of [34] suggested strategies to the overhead problem of beam training. However, the solutions in [32]–[34] considered that RIS does not operate independently and is managed by a different base station. Moreover, supervised deep learning necessitates a significant dataset-gathering stage prior to training. Owing to its effectiveness in actual environments,reinforcement learning has, most recently, gained much interest in the subject of study [35]. The combined non-convex optimization issue was addressed in [11] by using the DRL while taking into account beamforming interference management restrictions and intricate resource allocation. The optimization of network coverage [36] and hybrid beamforming matrices [37] for mmWave systems was developed using DRL.In the RIS-assisted communication system, an optimization-driven DRL approach was investigated for the optimization issue of joint beamforming [38] and optimizing the user SNR [39]. Subject to the worst budget power restriction of RIS and the least desirable data rate demand of the receiver, a power reduction problem in a MISO system with RIS assistance was solved using the DRL technique in [40].

B. Contributions

The Rayleigh fading model is well acknowledged to be a sound concept for the fading that occurs in several wireless systems for communication [41], [42]. However, it might not be a viable option in case of a realistic RIS-aided communicative context as the RISs are deliberately located for taking the advantage of line-of-sight (LoS) connections between the endpoints. Numerous assessment projects [43], [44] demonstrate that considering the Nakagami-m distribution provides a much better match for the fading channel distribution. There is greater freedom with the Nakagami-m distribution due to the extra variable. Earlier attempts to enhance phase shift and transmit beamforming in RIS-aided communications relied on DRL algorithms such as DDPG, which are unsteady and strongly rely on determining the appropriate hyperparameters. These conventional DRL algorithms consistently overestimate the Q values of the critic value network, and when estimation errors accumulate over time, the agent may eventually find itself in a local optimum. With the help of the TD3 approach, we propose a collaborative design of phase shifts and transmit beamforming to optimize the sum rate in RIS-aided systems, which would overcome the shortcomings of the current DRL algorithm. The employment of the TD3 approach in a Nakagami-m fading environment has not yet been discussed in the literature, despite the fact that there have only been a few research on the application of DRL for the optimization of phase shift and transmit beamforming in RIS-aided communication systems. The contributions of this thesis are described as follows:

- To optimize the average sum rate of the users, we formulate an optimization problem that aims to appropriately allocate the downlink transmit beamforming and effectively determine the phase shifts of the RIS components in RIS-aided communication systems. With the assistance of RIS, the suggested framework is able to serve users who experienced blocked conditions, hence enhancing the overall data rate.

- The sum rate maximization problem is a non-convex problem owing to multiuser interaction. We elicit an effective algorithm-based TD3 approach that tackles the drawbacks of traditional DRL by concentrating on lowering the overestimation bias and eradicating the requirement for gathering massive training datasets in order to resolve this NP-hard problem of continuous action space.

- We investigate a real-world RIS-aided communication system assuming Nakagami-m fading between the RIS and the user as well as between the BS and the RIS. To analyze the performance of the system under consideration, we select helpful metrics. The simulation results show that our suggested DRL approach effectively solves the joint optimization issue and outperforms the other benchmarks.

C. Thesis Layout

The thesis is organized as follows. In Chapter II, an overview of the structural design of the RIS is discussed. Section III introduces RL designs that have been applied in the literature. Then in chapter IV, we describe the system model and problem formulation of joint beamforming and phase shift design. Next in Chapter V, we describe the proposed solution and simulation analysis of RIS-TD3 framework. And finally, we conclude the thesis in Chapter VI.

II. Overview of Reconfigurable Intelligent Surface

RIS models are primarily created using metamaterials, which are periodically aligned subwavelength elements capable of providing complete control over EM actions of the metasurface and consist of unit cells [45]–[47]. This man-made EM material surface can be controlled electrically via integrated electronics and has unique wireless communication characteristics [48]. More precisely, an RIS functions by the placement of a large number of low-cost antenna components with the goal of controlling re-radiation and capturing energy. In the literature, varactor-centered and positive-intrinsic-negative diode based control methods were the standard techniques used [49]–[51]. To enhance the user communication quality and improve the properties of incident waves, control signals are transmitted by a BS to an RIS controller in an RIS-supported wireless network. The RIS does not perform digitizing because it operates as a reflector. Consequently, if properly implemented, the energy consumption of the RIS will be significantly lower than that of standard relays such as amplify-and-forward relay [52]–[54]. As illustrated in Figure 1, the practical EM wave-based tasks that RISs can employ in wireless communications are as follows:

- **Reflection:** An impacting radio wave is reflected in a particular direction, which may not be in the same direction as the incidence wave direction.

- **Refraction:** An impacting radio wave is refracted which may not be in the same direction as the incidence wave.

- **Absorption:** This entails creating a smart surface that cancels the refracted and reflected radio waves corresponding to a certain incident radio wave.

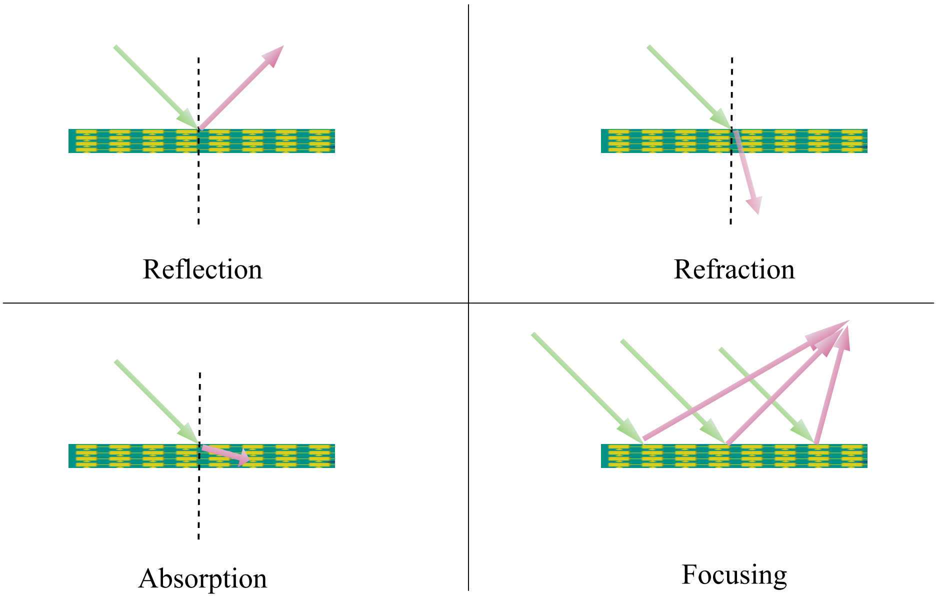


Figure 1: Electromagnetic wave-based activities of a reconfigurable intelligent surface.

- **Focusing:** It entails directing an impinging radio beam to a certain point.

A. Perspective of physics

EM waves encounter dispersed particles while traveling across space, which attenuates the signal. The physics-based bedrock of surface electromagnetism is the surface equivalence theorem. The Huygens principle asserts that each point across a wavefront is a generator of spherical wavelets, and additional wavelets emerging from various sites overlap. The wavefront is formed by the addition of several spherical wavelets. The EM field radiated by an RIS can be computed and analyzed based on the Huygens principle.

Figure 2 (a) [55], [56] illustrates a volume V occupied by several EM radiation sources consisting of charges q_i and currents J_i. Just outside the volume V, these sources generate a magnetic induction field B and an electric field E. The arrangement of scatterers can be substituted by an arbitrarily thin layer of particular magnetic currents J_m and electric currents J_e that completely covers the volume V, as per the Huygens principle. Magnetic currents can only be created by cycles of electric currents with a limited depth. Hence, the layer thickness can be electrically negligible but not zero. EM fields are scattered exclusively outside the volume V by the corresponding surface currents J_m and J_e, and all these EM fields are identical to those formed by the original sources. Huygens' surfaces that are related to currents that disperse EM fields solely with one side may be extended to metamaterials.

The boundary conditions are based on the fact that when an average tangential field is applied to a thin sheet of polarizable objects, it induces magnetic J_ms and electric J_es surface currents, which may be linked to the applied fields using magnetic surface admittance Y_m and electric surface impedance Z_e. Figure 2(b) [57]–[59] shows a magnetic surface admittance Y_m(x,y) and an electric surface impedance Z_e(x,y), which define the physical configuration of a generic sheet of the metasurface. The mean applied field induces magnetic and electric currents on the metasurface, creating a discontinuity between the fields above and below the surface, thereby allowing wavefront modification.

B. Interaction between the cells

The RIS modulation is dependent on the intercell connection of tunable chips, which regulate the scattering components of the metasurface to provide the desired tunable functions. Wireless or cable communication is possible among the underlying chip controllers. Because wired communication is easier

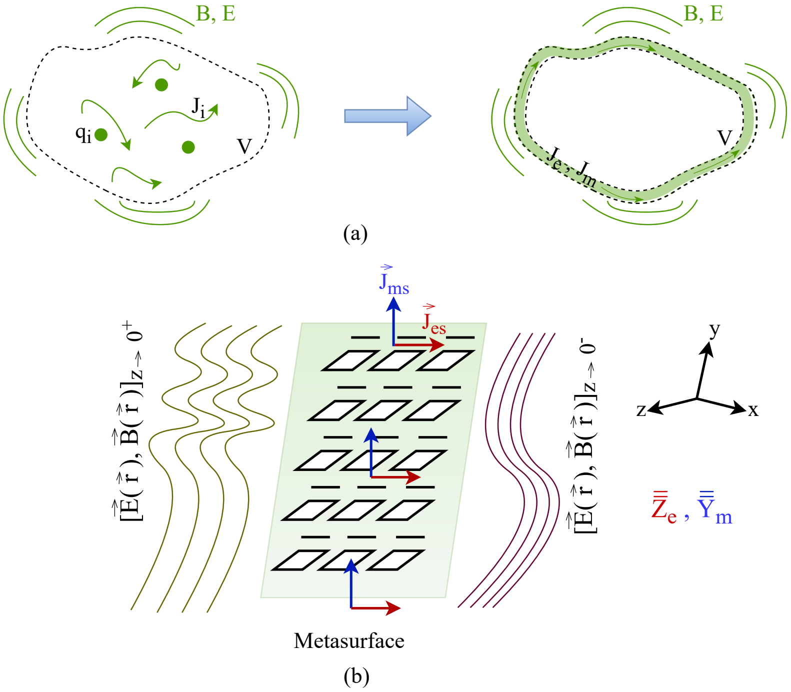
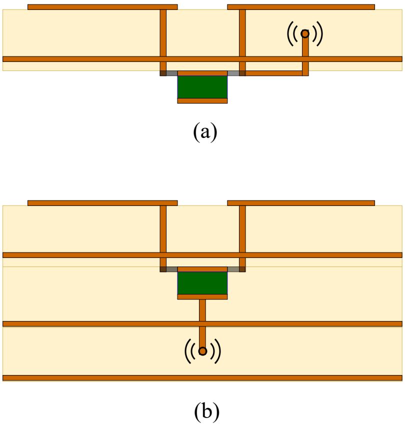


Figure 2: Representation of (a) surface equivalent theorem scattered EM source and (b) physical layout of the metasurface.

to combine with the controllers on the same chip, it is a better option; however, in a significantly compact or large sized metasurface, wireless intercell communication is an effective solution. With strict robustness requirements and energy latency, the design guidelines for inter-communication procedures must be practiced [18]. The exact application is determined by either the size of the tile or the desired wavelength. Two separate connection pathways are shown in the Figures 3 (a), (b) [24]. In case (a), the metasurface layer, which is the gap between the plane at the back and the metasurface patches, is the first channel. The antenna is a part of the chip, whereas the role of the waveguide is performed by the plane at the back and metasurface patches. In case (b), a separate control



(a)

(b)

Figure 3: Communication channels: (a) metasurface substrate, (b) waveguide with specific parallel plates.

plane is constructed by inserting additional metal slabs beneath the chip for the second channel. As in the aligned-plate waveguide, monopoles supplied from the chip could generate waves that travel in this barrier condition.

C. Relationship between the metasurface and the RIS

Metasurface is a two-dimensional planar metamaterial with EM properties. Metamaterials have not been discovered in natural supplies, and they are composed several tightly placed subwavelength resonating structures known as meta-atoms or pixels [20]. The distinguishing characteristics are their ability to shape EM waves in a variety of ways. Owing to their petite size, a significant number of these closely packed atoms provide large degrees of freedom in altering the incident EM waves. For instance, a metasurface can impose arbitrary quasi-continuous [60] amplitude or phase profiles on the incident wavefronts and exert fine-grained control over the dispersed electric field by carefully incorporating its meta-atoms.

In general, software and meta-atom oriented controllers are essential elements of the RIS that influence the metasurface reconfiguration rate. The related power consumption of static and reconfigurable metasurfaces is significantly different because no active electrical circuits are required for static metasurfaces; they can be completely passive. As energy is required to control the received signals and switches for reconfiguration, metasurfaces with reconfigurable properties can only be virtually passive. However, a specialized power supply is not required for signal transmission after the metasurface has been appropriately calibrated.

D. Passive beamforming and RIS

When multiple antennas produce identical signal copies of the postponed signal, beamforming occurs. Constructive interference occurs in geographic places where the signal copies are collected simultaneously, whereas at other places, destructive interference occurs. When multiple antenna send signals, the receiver will collect better signals than when a single antenna transmits signals while consuming the same total power. The time delays at the transmitting antennas are set to create constructive interference at the receiver. This traditional array gain demonstrates that the beamformed signal becomes more spatially concentrated if there is an increase in array size. The received signal strength and surface area are proportional, and depend on the number of elements of the transmitter. With the delay in time, when the RIS re-radiates the chosen signal, an array gain is produced to beamform the signal at the receiver, similar to the traditional manner. The process of passive beamforming by the RIS between the BS and the user by reflecting the signals to aid in communication is shown in Figure 4. The RIS reflection coefficients can be modified by the BS using an RIS controller.

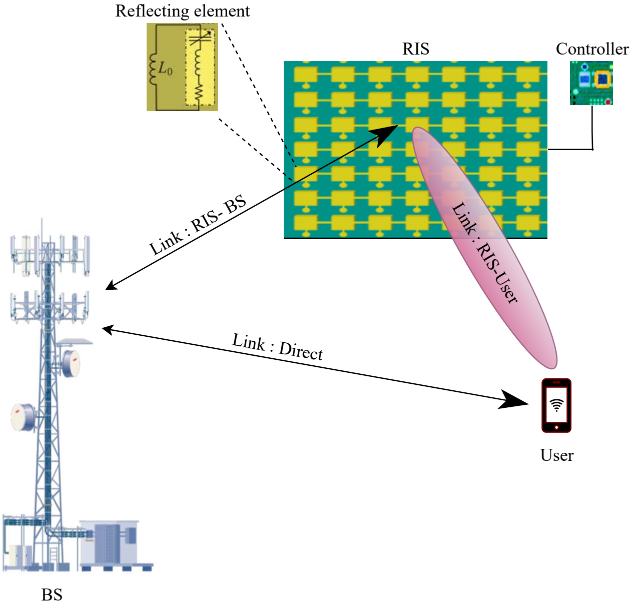


Figure 4: Passive beamforming using an RIS.

Furthermore, passive beamforming at the RIS and transmit beamforming at the BS must be developed together to increase communication performance [61].

III. Reinforcement learning

The section of science that studies the theory and characteristics of learning algorithms, their performance, and associated systems is known as ML. ML is a wide multidisciplinary area that draws concepts from a variety of domains, including information theory, AI, statistics, optimal control, optimization theory, and a variety of other scientific, mathematical, and engineering disciplines [27]–[30]. ML has touched nearly every scientific subject owing to its deployment in diverse applications, which has a significant influence on research and society [36]. Currently, ML is predominantly applied in autonomous systems, suggestion engines, informatics, data mining, and recognition systems [31]. The ML technique typically comprises two major phases: training and decision making. In the training phase, a dataset is used to train and understand the model of the system. During the decision-making process, the trained model is employed to derive the projected output for every new input given to the system. This is a commonly applied and effective ML method that learns about the environment by performing various actions and determining the best operation strategy. The two fundamental factors of RL are the environment and the agent. By applying the Markov decision process (MDP) [62], the agent investigates the surroundings and determines the action that must be implemented for the optimum result.

Q-learning (QL) is a straightforward and effective RL method in which a model of the environment is not required; the goal is achieved based on the reward.

When action a is chosen, $Q(s,a)$ is the current value of the state s; $0 < \alpha < 1$ is the learning constant and $0 < \phi < 1$ is the discounting factor. The algorithm operates as follows: the agent chooses an action at some state s. Given that the action a is implemented, it discovers the highest feasible Q-value in the following state $(s+1)$ and changes the current Q-value. The discounting factor provides the choice of either rewarding in the future $(if >> 0)$ or presenting immediate rewards $(if\phi << 1)$. To improve the convergence and stability of the algorithm, a constant is used to adjust the learning rate. QL has been previously used in various wireless situations, such as in wireless sensor network routing [63]–[65]. It is simple to set up and exhibits an acceptable balance between memory and energy needs.

A. Deep reinforcement learning

A subset of ML is deep learning (DL), which allows an algorithm to generate projections and classifications without being explicitly programmed based on the decisions of input data. Some cases of DL include QL, k-nearest neighbor classifiers, and linear regression. DL algorithms may extract information from raw data in a hierarchical manner by utilizing nonlinear processing components of multiple layers for forecasting outcomes based on the desired objective [66]. Recently, DL has attracted more interest from the academic community because of its superior performance in areas such as computer vision, information retrieval, speech recognition, and language processing [67]–[70]. As computing power and graphic processors are improving daily [71], it is increasingly becoming important for areas involving big data sets to deliver projected analytic solutions.

1. Twin Delayed Deep Deterministic Policy Gradient

Reinforcement learning is an AI technology that tries to understand about the environment by choosing the best operating policy depending on various actions. The RL is made up of two parts: an environment and an agent. Utilizing the Markov decision process (MDP), the agent analyzes the environment and chooses the appropriate action. The learning goal of the agent is to find out the optimal policy for maximizing future rewards. To establish the best policy, two types of techniques are commonly used: policy-based and value-based approaches. The value function, the state, the instant reward, the action, and the policy are a few essential aspects necessary to define the RL learning process [72].

The tuple $\left(\mathscr{A}, \mathscr{S}, \mathscr{R}_a, \mathscr{P}_a(_{s \to s_{t+1}})\right)$ represents various parameters in MDP: $a_t \in \mathscr{A}$ is a finite action space at time t, $s_t \in \mathscr{S}$ is a finite state space at time t, $r_t \in \mathscr{R}_a$ is the instantaneous reward which is delivered by the environment for action a_t at time t, $\mathscr{P}_a(_{s \to s_{t+1}})$ is the transition probability towards the next state $s \to s_{t+1}$ after performing action a_t from the current state s_t [73]. At time t, the state belonging to the environment will migrate from the present state s_t to another state s_{t+1} when the agent executes action a_t. The possibility of performing an action a_t depending on the state s_t is represented by the policy π. The policy function meets the condition $\sum_{a_t \in \mathscr{A}} \pi = 1$. The reward function evaluates the instant result from the action a in a particular state s, but the value function evaluates the forthcoming rewards obtained through the action a in that state s taken by the agent. The total amount of forthcoming discounted rewards can be written as:

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

where, the discount factor is denoted by γ that is set to be $0 \leq \gamma \leq 1$. Upcoming rewards are discounted in order to focus on the present reward. The significance of long-term gains in the present state is determined by the respect of γ. The Q value function is stated as follows in the perspective of a specific policy π:

$$Q_{\pi}\left(s_t, a_t\right) = E_{\pi}\left[R \mid s = s_t, a = a_t\right] \tag{2}$$

Adjusting the Q-table is the main goal of Q-Learning by employing Bellman's equation as

$$\begin{aligned} Q_{\pi}\left(s_t, a_t\right) =& E_{\pi}\left[r_{t+1} \mid s = s_t, a = a_t\right] \\ & + \gamma \sum_{s_{t+1} \in \mathscr{S}} \mathscr{P}_a\left(s \rightarrow s_{t+1}\right) \\ & \left(\sum_{a_{t+1} \in \mathscr{A}} \pi\left(s_{t+1}, a_{t+1}\right) Q_{\pi}\left(s_{t+1}, a_{t+1}\right) \right) \end{aligned} \tag{3}$$

To find the most optimum policies π^*, the Q-learning method is employed. For the best policy, the optimum Q function from (21) may be written as

$$\begin{aligned} Q\pi^*\left(s_t, a_t\right) = r_{t+1} + \gamma \sum_{s_{t+1} \in \mathscr{S}} \mathscr{P}_a\left(s \rightarrow s_{t+1}\right) \\ \max_{a_{t+1} \in \mathscr{A}} Q\pi^*\left(s_{t+1}, a_{t+1}\right) \end{aligned} \tag{4}$$

It is possible to solve the Bellman equation cyclically, and recapitulating (22) produces the best Q function. As a result, the recursive solution upon this update technique of the Q function may be written as

$$\begin{aligned} Q\pi^*(s_t, a_t) \leftarrow &(1-\alpha)Q\pi^*(s_t, a_t) + \alpha \Big(r_{t+1} \\ &+\gamma \max_{a_{t+1}} Q_\pi(s_{t+1}, a_{t+1})\Big), \end{aligned} \tag{5}$$

where, for updating the Q function, α is used as the learning rate.

Throughout a relatively limited state and action space, the Q-learning method is effective. Nevertheless, when the action and state space are huge, the algorithm gets more difficult. Due to the huge Q-table, the Q-learning technique cannot provide an efficient strategy under this circumstance. As a result, function estimation is developed for solving issues with large action and state spaces. The function approximator contains deep Q-learning (DQL), which is designed for substituting the Q-table with a deep neural network (DNN). DQL is a deep reinforcement learning (DRL) technique that utilises Q-values in a similar manner to Q-learning, but without the Q-table [74]. The DNN estimates the system model, policy function, and the action and state value function as a combination of multiple non-linear functions, in which both the action and Q function are represented by DNN instead of using exact mathematical modeling. The input recieved by the DNN is the state obtained from the surrounding, which provides approximated Q-values for each action the agent can take. Additionally, since the NN is trained with parameters θ to assess the Q-values, it can not be desirable to describe the Q-function solely on action and state alone in many DQL aspects. The Q function of the agent, which assesses the current state-action combination according to a policy, can be given by

$$Q_\pi(s_t, a_t; \omega) = E_\pi \left[R \mid s = s_t, a = a_t \right] \tag{6}$$

where, θ denotes the weighting parameters of the DNN employed in DQN. Instead of actively adjusting the Q function just like in (21), the ideal Q value function can be addressed utilizing stochastic optimization techniques employing DRL as

$$\omega_{(t+1)} = \omega_t - \eta \nabla_\omega L(\omega) \tag{7}$$

where, ∇_ω denotes the gradient of the loss function $L(\omega)$ and η indicates the learning rate for updating ω. DNN is used as an estimator for the Q-value function in the DQN method, using the parameter of weight ω. The present state s_t is given as the input of the DNN, and the anticipated action a_t is retrieved from the output of the DNN. After every epoch, the DNN parameter ω is modified to provide a more refined Q-value estimate. Through training, the DNN can decrease the loss function. The distinctions between the anticipated value of NN and actual target values is expressed by the loss function as

$$L(\omega) = \left(y - Q\left(s_t, a_t \mid \omega\right)\right)^2 \tag{8}$$

The target y is determined by

$$y = r_{t+1}(s_t, a_t) + \gamma \max_{a_{t+1}} Q\left(s_{t+1}, a_{t+1} \mid \omega_{trg}\right) \tag{9}$$

where, ω_{trg} and ω are the target and training networks, respectively, having similar configurations.

Built on the actor-critic algorithm, DDPG is just an RL framework capable of supporting continuous action sets. DDPG originates and decides actions based on a parametric technique rather than the usual way of producing actions based on chances.This algorithm combines an actor-based on value functions and a critic-based on policy searches. The critic network uses NNs for training to imitate an actual Q-table while avoiding the dimensional constraint. Rather than the policy gradient that picks an unexpected action out of a predetermined distribution, the actor-network is taught to generate deterministic policies. To boost the convergence speed and reduce superfluous computation, the DDPG method employs experience target network and replay buffer approaches. In comparison to DQN [75], DDPG may adapt explicitly from raw inputs and needs minimal training stages. The goal of DNN training is to enhance value function prediction by changing the parameterization strategy v explicitly in the way of the gradient as

$$J(\Upsilon) = \sum_{s\in\mathscr{S}} s_d \sum_{a\in\mathscr{A}} \pi_\Upsilon(a \mid s) Q(s,a \mid \omega) \tag{10}$$

where the Q value estimated by the DNN using parameter ω is $Q(s,a \mid \omega)$, and according to the policy π_Υ, static distribution of state is indicated by s_d. The gradient of (28) following the theorem of deterministic policy can be expressed as

$$\nabla_\Upsilon J(\Upsilon) = E_{s\sim s_d}\left[\nabla_a Q(s,a \mid \omega) \nabla_\Upsilon \pi_\Upsilon(s)|_{a=\pi_\Upsilon(s)}\right] \tag{11}$$

The actor-critic architecture is driven by the policy gradient in (29) and modifies the DNN parameters ω, Υ independently. The Q-network is updated by the critic network as

$$\omega_{t+1} = T_d s_\omega \nabla_\omega Q\left(s_t, a_t \mid \omega_t\right) + \omega_t \tag{12}$$

The policy parameter Υ is updated by the actor-network in a gradient direction as

$$\Upsilon_{t+1} = \Upsilon_t + s_\Upsilon \nabla_a Q\left(s_t, a_t \mid \omega_t\right) \nabla_\Upsilon \pi_\Upsilon(s)|_{a_t=\pi_\Upsilon(s)} \tag{13}$$

where, s_ω and s_Υ indicates the sizes of the steps. The temporal difference between y and $Q\left(s_t, a_t \mid \omega_t\right)$ is given by T_d.

The actor-network can estimate the action that ensures the optimum Q value function while upcoming state is provided. The actor network is updated as follows:

$$\omega^a_{\text{up}} \omega^a_t - \alpha^a \nabla^a Q\left(\omega^c_{trg} \mid s_t, a_t\right) \nabla_{\omega^a_{trn}} \pi\left(\omega^a \mid s_t\right) \tag{14}$$

where, the gradient of the actor-network and target critic-network are $\nabla_{\omega^a_{trn}} \pi\left(\omega^a \mid s_t\right)$ and $\nabla^a Q\left(\omega^c_{trg} \mid s_t, a\right)$ respectively. Supplied with the input s_t and parameter ω^a, $\pi\left(\omega^a \mid s_t\right)$ signifies the actor network.

The critic network updates can be expressed by

$$\begin{aligned} & \omega_{\mathrm{up}}^{c} \omega_{t}^{c}-\alpha^{c} \nabla_{\omega_{trn}^{c}} L\left(\omega^{c}\right), && (15) \\ & L\left(\omega^{c}\right)=\left(r_{t}+\gamma Q\left(\omega_{trg}^{c} \mid s_{t+1}, a_{t+1}\right)\right. \\ & \left.-Q\left(\omega^{c} \mid s_{t}, a_{t}\right)\right)^{2}, && (16) \end{aligned}$$

where, $\nabla_{\omega_{trn}^{c}} L\left(\omega^{c}\right)$ is the gradient, concerning the critic network ω^{c}. a_{t+1} is the actor target network action output, and critic network is updated by learning rate α^{c}. The training network updates significantly faster than the target network. As shown in (32), with regard to the action the target critic network influences the update of the actor network. With τ as their soft update coefficient, target actor and critic networks update may be represented as

$$\begin{aligned} & \omega_{trg}^{a} \leftarrow(1-\tau) \omega_{trg}^{a}+\tau \omega^{a} \\ & \omega_{trg}^{c} \leftarrow(1-\tau) \omega_{trg}^{c}+\tau \omega^{c} && (17) \end{aligned}$$

IV. System model and problem formulation

In this chapter, we describe the proposed scheme to for the collaborative design of beamforming matrix and Phase shift matirix in RIS-aided system under multiple performance metrics. To describe the proposed solution, we derive the system model and formulate the constraints of RIS-assisted system into a deep reinforcement learning problem.

A. System model

We take into account a downlink system that includes a MISO configuration, as shown in Figure 5. The reflecting RIS is used to address signal obstruction between the user and the BS. Acting as a reflecting array, RIS causes phase shifting of impinging signals, and might be smartly configured depending on the wireless communication system by using the meta-surfaces embedded with electronic circuits. In the proposed system, potential obstacles block the straight path between the user and the BS. While the RIS uses $N = N_a \times N_b$ passive phase shifters, with N_a and N_b denoting the number of passive components for every row and column respectively, the BS uses a uniform linear array (ULA) with M antenna array. Considering a crowded metropolitan condition, it is anticipated that the line-of-sight (LoS) between the BS and users is absent. The channels are $H_{BR} \in C^{N\times M} and$ and $\mathbf{h}_{RU} \in C^{N\times 1}$ for the BS-RIS and RIS-user respectively. The Nakagami-m distribution models both of them as independent random variables. An intelligent controller is used to configure all phase shifters on the RIS. At the user, the received signal with the constraint of total transmit power can be expressed as:

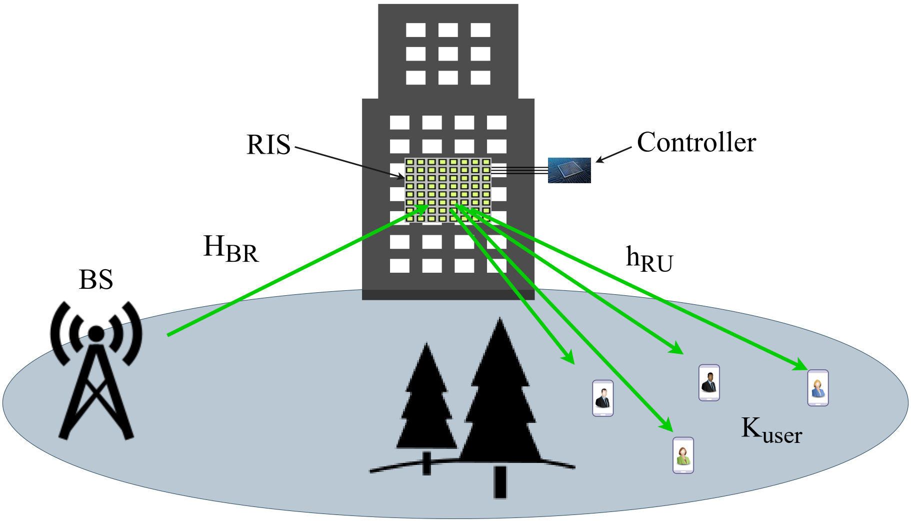


Figure 5: RIS-aided multiuser communication system

$$
\begin{aligned}
y = H_{\mathrm{BR}}\Psi\mathbf{h}^{T}_{\mathrm{RU}_k})Bu + w_k& \\
B \in C^{M\times K},& \qquad (18a)\\
E\left\{\Lambda\right\} \leq \mathscr{P},& \qquad (18b)
\end{aligned}
$$

where, $\mathscr{P}$ represents the total BS transmit power and B denotes beamforming matrix for the system. The transmitted signal is indicated by u fulfilling $E\left[u^2\right] = 1$. The additive white Gaussian noise (AWGN) is states by w_k with variance of σ^2 i.e $w_k \sim \mathscr{CN}\left(0, \sigma^2\right)$. $\Lambda = Bu(Bu)^H$ is the conjugate transpose of beamforming matrix and transmitted signal. The phase shift matrix of the RIS is defined as:

$$\Psi = \mathrm{diag}\left(e^{j\varphi_1}, e^{j\varphi_2}, e^{j\varphi_3}, \ldots, e^{j\varphi_N}\right), \tag{19}$$

The n-th element phase shift angle of the RIS is represented by $\varphi_i \in [0, 2\pi]$ and $\mathrm{diag}\,(x_1, x_2, \ldots, x_n)$ signifies a diagonal matrix with diagonal entries $x_1, ..., x_n$. The probability density function (PDF) for the Nakagami-m distribution [76] of $|H_{\mathrm{BR}}|$ and $\left|\mathbf{h}_{\mathrm{RU}_k}^T\right|$ are as follows

$$f_{|H_{\mathrm{BR}}||\mathbf{h}_{\mathrm{RU}_k}|}(x) = \frac{2m^m}{\Gamma(m)\Omega^m} x^{2m-1} e^{-\frac{m}{\Omega}x^2}, \tag{20}$$

where $|.|$ stands for the absolute value, Ω denotes the spread and m is the shape of the parameters. The interference portion of the received signal with the beamforming vector b_l can be regarded as

$$I = \sum_{l, l \neq k}^{K} (H_{\mathrm{BR}} \Psi \mathbf{h}_{\mathrm{RU}_k}^T) b_l u_l + w_k \tag{21}$$

At the k^{th} user, the SINR obtained from the signal received can be represented as

$$\gamma_k = |(H_{\mathrm{BR}} \Psi \mathbf{h}_{\mathrm{RU}_k}^T) b_k|^2 / \sum_{l, l \neq k}^{K} |(H_{\mathrm{BR}} \Psi \mathbf{h}_{\mathrm{RU}_k}^T) b_l|^2 + \sigma^2 \tag{22}$$

1. Problem formulation

With the goal of improving the performance of the RIS-aided communication system, the phase-shift matrix Ψ of RIS and the transmit beamforming matrix B are collaboratively addressed using a framework for optimizing it. The objective of this framework is to maximize the sum rate $\mathscr{R}$, subject to the constraints of transmit power, by formulating it as an optimization problem. Accordingly, the problem is formulated as:

$$(P1)\max_{\Psi,B}\mathscr{R}(\Psi,B)$$

$$\mathscr{R}=\sum_{k=1}^{K}\log_2(1+\gamma_k) \tag{23a}$$

$$\text{s.t. } |\Psi_n|=1, \quad \forall n=1,\cdots,N \tag{23b}$$

$$B(B)^H \leq \mathscr{P} \tag{23c}$$

where the $n-th$ diagonal component of Ψ is represented by Ψ_n. The unit modulus restrictions and non-convexity of the objective function make (6a) an NP-hard problem. Conventional phase shift optimization strategies for RIS-assisted systems need fully updated channel information. Using the traditional mathematical methods to find the best solutions becomes unfeasible to do, especially for massive networks. Instead of tackling the problematic optimization issue theoretically, the purpose of the thesis is to develop optimal phase shift and transmit beamforming matrix that can be modified continuously in the setting of a sophisticated TD3 framework to comply with $P1$ effectively.

V. Solution based on DRL

In this part, we explain the RIS-TD3 structure in-depth and its rationale. Eventually, for the suggested RIS-TD3 framework, we present the learning technique.

A. RIS-TD3 framework

Applying DQN, policy gradient, and Q-Learning techniques, the RIS beamforming policy may be quantitatively accomplished. The policy gradient approach can handle continuous state-action spaces, although this may settle to an inferior outcome. Moreover, Q-Learning has a poor learning time and cannot handle the continuous state space, so it is not a productive learning method. Furthermore, solving the optimization issue in a high-dimensional input state space is difficult for Q-learning and policy gradient techniques. While DQN excels at policy learning in state spaces with high-dimension, the non-linear Q-function approximator can cause the learning operation to become unsteady [77]. With the succeeding adjustments, the TD3 algorithm enhances the DDPG technique and covers some of the gaps left by DDPG. Target networks are employed to limit error propagation by postponing the update of the policy network till the Q-value converges, hence reducing the high variance and noisy gradients and minimizing the value error for each update. Since the policy network updates are less frequent, more reliable policy changes are made possible. Target policy smoothing is carried out using the regularisation approach, where clipped noise is introduced to the target action deduced from the policy in order to decrease the variation in the target action values. The notion of clipped double Q-learning is utilized to tackle the problem of overestimation bias. Two

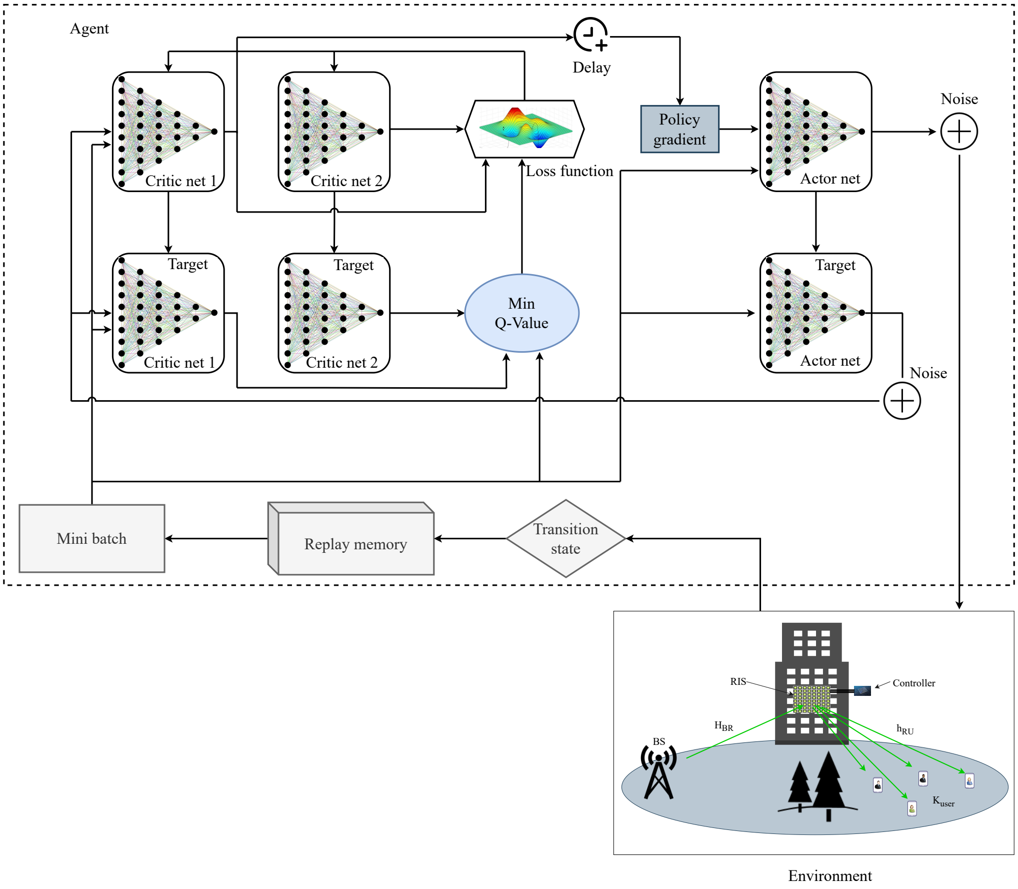


Figure 6: TD3 framework for RIS

different critic networks are used by TD3, forming the targets using the smallest of the two values as shown in Figure 6.

State space: The agent exclusively understands local data since it engages with the environment to improve the sum rate. At the time step t, the channel matrix H_{BR} and $\mathbf{h}_{RU}$, previous step action, transmit and received power makes up the state space. With the unit variance of symbols, the received power R_p and transmit power T_p for the k^{th} user are stated as

$$R_p = |(H_{\text{BR}}\Psi\mathbf{h}_{\text{RU}_k}^T)B(l)|^2 \tag{24}$$

$$T_p = \|B_k\|^2 = \left|B_k^H B_k\right|^2 \tag{25}$$

The following definition applies to the situation at t^{th} time step:

$$s_t = \left[H_{\text{BR}}, \mathbf{h}_{\text{RU}_k}, T_p, R_p, a^{(t-1)}\right] \tag{26}$$

Action space: The phase shifts and transmit beamforming prompted by the RIS within the present channel conditions are updated by the agent using the input of state s_t at time interval t. The action vector a_t is stated as follows

$$a_t = \left[\Psi^t, B^t\right] \tag{27}$$

Reward: Whenever an action is chosen by the agent in real-time, the reward serves as a marker to assess how effective is the policy. When every learning step of the reward function aligns with the intended outcome, the performance of the system will be improved. Therefore, it is critical to provide an effective incentive mechanism to raise overall satisfaction. In this study, the optimization target is represented by the reward function, and our goal is to increase the system sum rate $\mathscr{R}$. The reward function may be represented as follows using the goal mentioned above:

$$r_t = \mathscr{R}^t \tag{28}$$

B. Operational steps

In the beginning, six networks are created, consisting of actor-network ω^a, target actor network ω^a_{trg}, critic networks ω^{c1}, ω^{c2} and target critic networks ω^{c1}_{trg}, ω^{c2}_{trg} with uniformly distributed parameters. The target networks' parameters are built by copying the coefficients of the actor and critic networks. In addition, a memory $|B|$ for experience replay with a specified cardinality is constructed. The phase shifts of all components are, without losing generality, determined at random at the start of each episode, ranging from 0 to 2π. Every episode starts with information on all the channels at work. Following the initial state s_1 observation, action a_t is chosen from the actor-network and assigned with noise.

$$a_t = \omega^a(s_t) + \rho \tag{29a}$$

$$\rho \sim \mathcal{N}(0, \sigma) \tag{29b}$$

where ρ denotes the exploration noise. The next state s_{t+1} and instant reward r_t can be determined by modifying the action into a transmit beamforming matrix Λ and a phase shift matrix Ψ. Preserving (s_t, a_t, r_t, s_{t+1}) as a transition towards experience replay memory $|B|$. Transitions (s_t, a_t, r_t, s_{t+1}) from the experience replay $|B|$ are sampled into a minibatch M_B. The deterministic action a^n_t for each state s_{t+1} is output by each target actor network, and this action is then given a clipped noise.

$$a^n_t = \omega^a_{trg}(s_{t+1}) + \rho \tag{30a}$$

$$\rho \sim \text{clip}(\mathcal{N}(0, \sigma'), -c, c) \tag{30b}$$

where σ and c are the policy noise variance and the noise clip, respectively. The target Q-network receives the state s_t and the target action a_t^n as inputs and estimates the target Q value. Then the target value y is determined by choosing the smaller value of the two Q values.

$$y \leftarrow r_t + \gamma \min_{i=1,2} Q_{\omega^c_{trg}}\left(s_{t+1}, a_t^n\right) \tag{31}$$

Using the stochastic gradient descent (SGD) optimizer, the critical loss is backpropagated before updating the two critical models' parameters. The critic networks are updated by

$$\begin{aligned} \omega^c_{\mathrm{up}_i} = \mathrm{argmin}_{\omega^c_i} {M_B}^{-1} \sum \left(y - Q_{\omega^c_i}(s_t, a_t)\right)^2 \\ i = 1,2 \end{aligned} \tag{32}$$

The deterministic policy gradient updates the actor network if it is time to update (t *mod* p) the policy network.

$$\begin{aligned} \omega^a_{\mathrm{up}} = {M_B}^{-1} \sum \nabla_{a_t} Q_{\omega^{c1}}(s_t, a_t)\Big|_{a_t} \nabla_{\omega^a} \pi_{\omega^a}(s_t) \\ a_t = \pi_{\omega^a}(s_t) \end{aligned} \tag{33}$$

Then the target networks get updates from $\omega^c_{trg_i}$, ω^a_{trg}

$$\omega^c_{trg_i} \leftarrow (1-\tau)\,\omega^c_{trg_i} + \tau \omega_i^c \qquad i = 1,2 \tag{34}$$

$$\omega^a_{trg} \leftarrow (1-\tau)\,\omega^a_{trg} + \tau \omega^a \tag{35}$$

The algorithm 1 outlines the steps of the RIS-TD3 framework.

Algorithm 1 RIS-TD3 framework

```
Input: Channel state ($H_{\mathrm{BR.}}$, $\mathbf{h}_{\mathrm{RU}_k}$) of all the users.
Initialize: learning rate $\alpha$, soft update coefficient $\tau$, batch size $|B|$, phase shift matrix $\Psi$, empty experience replay memory $\mathscr{R}$, discount factor $\gamma$, beamforming matrix $\Lambda$, actor $\omega^a$ and target actor $\omega^a_{trg}$ networks, critic $\omega^{c1}$, $\omega^{c2}$ and target critic $\omega^{c1}_{trg}$, $\omega^{c2}_{trg}$ networks.
Output: Sum rate $\mathscr{R}$ and Optimal action $a^*_t$
for each episode do
    Get instant channel information $H_{\mathrm{BR}}$, $\mathbf{h}_{\mathrm{RU}_k}$.
    Randomly reset the environment.
    Obtain the initial state $s_1$.
    for each time step do
        Observe the initial state $s_1$.
        Select an action $a_t = \omega^a(s_t)$
        Add noise $\rho \sim N(0,\sigma)$ to $a_t$.
        Execute the action $a_t = \omega^a(s_t) + \rho$.
        Observe the reward $r^t$ and next state $s_{t+1}$.
        Store the transition($s_t$, $a_t$, $r_t$, $s_{t+1}$)
         into $|B|$.
        Sample a minibatch of $M_B$ randomly
        from $|B|$.
        Compute target action $a^n_t = \omega^a_{trg}(s_{t+1}) + \rho$
                                  $\rho \sim \mathrm{clip}(N(0,\sigma'), -c, c)$.
        Compute the target value
            $y \leftarrow r_t + \gamma \min_{i=1,2} Q_{\omega^c_{trg}}(s_{t+1}, a^n_t)$.
        Update critics $\omega^c_{\mathrm{up}_i}$.
        if t mod p then
            Update actor network by
                  DPG $\nabla_{\omega^a} J(\nabla_{\omega^a})$.
            Update target networks $\omega^c_{trg_i}$,$\omega^a_{trg}$.
        end if
    end for
end for
```

C. Numerical evaluation and discussion

This section discusses the experiment findings for different traditional DRL techniques used in wireless communication networks supported by RIS, whose top view layout is shown in Figure 7. To assess the suggested method, we ran several simulations. We present the design of the networks and list the TD3 parameters in Table 1 to begin this section. The simulation treats a scenario with one RIS, one BS, and multiple users. The channel matrices $H_{\mathrm{BR.}}$ and $\mathbf{h}_{\mathrm{RU}_k}$ are constructed at random using the Nakagami-m distribution. The critic and actor networks in the suggested method are dense neural networks. The input of the actor network is the number of states, and the output is the number of actions. Using two fully linked hidden layers with [512, 512] neurons, we implement a TD3-based algorithm, and to deal with the negative inputs, the activation function employed in this architecture is *tanh*. Moreover, Adam optimizer is employed to update parameters in both actor and critic networks. In analyzing the numerical findings, we take into account the following methods.

- **The suggested method:** To solve the combined optimization problem of the phase shift matrix and transmit power of the RIS, we implement the TD3 algorithm.

- **Soft actor critic (SAC) and DDPG:** To optimize the combined design of the phase shift matrix and transmit beamforming of the RIS panel, we employ the SAC and DDPG algorithms.

- **Random:** With random phase shift matrix generation, we optimize the combined phase shift and transmit beamforming matrix of the RIS-assisted system.

Table 1: Simulation parameters.

Parameter	**Value**
Experience replay buffer size R_n	50000
Batch size R_b	16
Soft update rate τ	0.005
Policy noise variance σ	0.2
Noise clip c	0.3
Delay update parameter	2
Episode	3000
Training samples	60×10^4
Actor learning rate LR_a	0.001
Critics learning rate LR_c	0.001
Target actor learning rate LR_{a_t}	0.001
Target critics learning rate LR_{c_t}	0.001
Discount factor γ	0.99

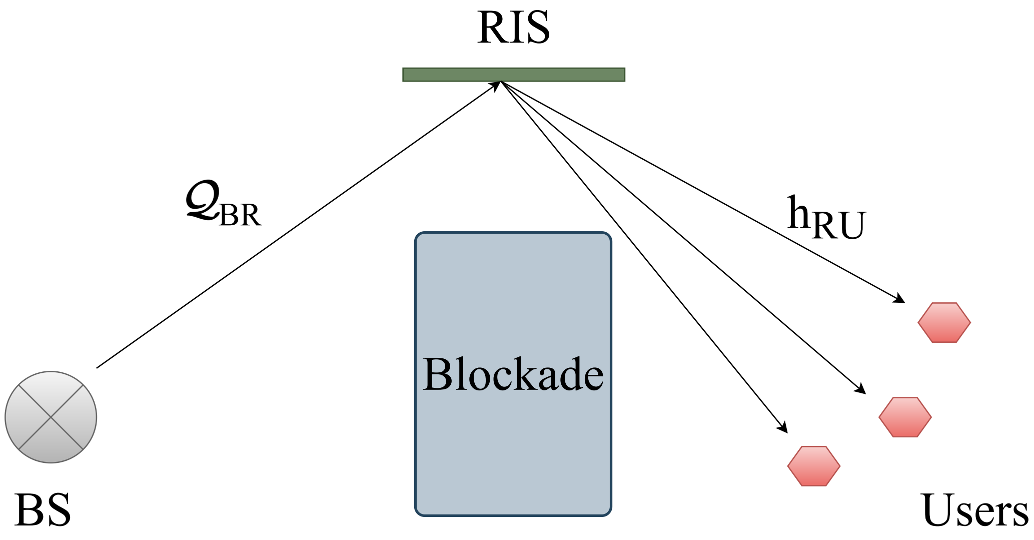


Figure 7: (Top view) RIS-assisted communication system

First, we evaluate the average network sum rate obtained by our proposed methodology with that of other schemes, namely the DDPG, SAC, and random. The environmental parameters are set as Pwr_t= $25dBm$, $M = 16$, N = 16 (N_a = 4, N_b = 4) and $K = 16$ for the number of BS antennas, the reflecting components of RIS, and the number of users, respectively. The average sum rate vs. the number of training samples is shown in Figure 8. It is noted that the SAC and DDPG algorithms are followed by the suggested TD3-based DRL method, which produces the best reward and surpasses other techniques. The random approach performs poorly considering the average of the average network sum rate.

The Nakagami-m fading shape parameter m determines the degree of fading. In Figure 9, the shape parameter m is compared. The mean of the distribution grows as the value of m increases, as can be observed, suggesting that the channel is in better condition the higher the value of m.

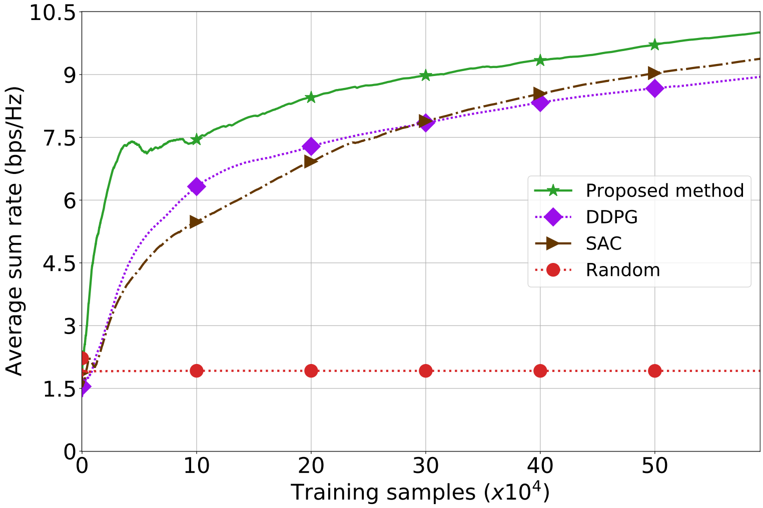


Figure 8: Comparison with the benchmarks

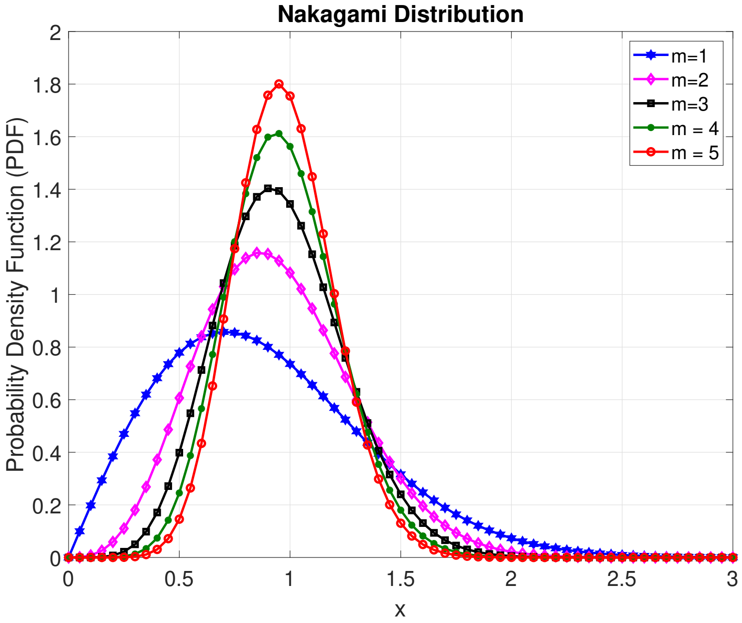


Figure 9: PDF of Nakagami-m distribution

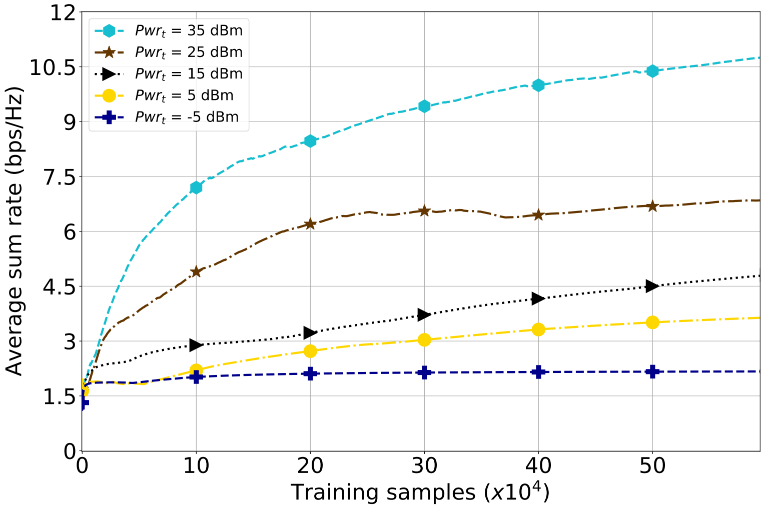


Figure 10: Different transmit power effect

The average sum rate for different transmit power Pwr_t= $-5dBm$, $5dBm$, $15dBm$, $25dBm$, and $35dBm$ is illustrated in Figure 10. Setting the environmental parameters to $M = 8$, $N = 8$ ($N_a = 4$, $N_b = 4$) and $K = 8$ allows us to compare the scenarios. When the transmission power of the base station is increased, a rise in the average sum rate of the system can be seen. Convergence can be attained faster using the suggested DRL approach. Under Pwr_t= $15dBm$, however, performance degrades and affects convergence as well.

With varying numbers of RIS components $N_r = 64, 32, 16, 8$ undergoing continuous phase shifts, the average sum rate is depicted in Figure 11, with the setting of Pwr_t= $25dBm$, $M = 8$ and $K = 8$ being the environment. The sum rate of users rises as the number of RIS elements increases because the data rate of each user rises as N_r increases.

The impact of Pwr_t for BS transmission power on the system average sum rate is demonstrated in Figure 12. To compare, we selected three distinct $M = 8$, $N =$

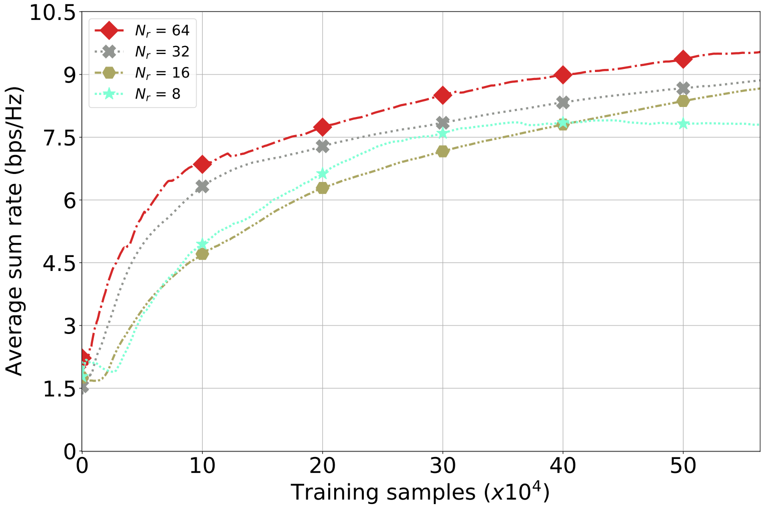


Figure 11: Influence of the number of components in RIS

8 (N_a = 4, N_b = 2) and $M = 4$, N = 4 (N_a = 2, N_b = 2) environmental parameters. When there is a variation in the transmission power of BS, a rise in the system sum rate is seen. In order to increase performance using Pwr_t, it is possible to effectively eliminate channel interference by the simultaneous design of phase shifts and transmit beamforming.

The average sum rate versus training samples for various learning rates (10^{-5}, $0.0001, 0.001, 0.01$ and 0.1) is shown in Figure 13. It is apparent that different learning rates affect the deep reinforcement learning performance of the algorithm in distinct ways. We recognize there is no benefit to training when the learning rate is 0.1. With the growth of the training samples, there is a rise in the average sum rate with a learning rate of 0.01. If the learning rate reduces, the reward often rises with minor variations. With a learning rate of 0.001, the model can clearly learn the situation. As a result, we choose an appropriate learning rate that is neither too high nor too low for our test, the value of 0.001. The average

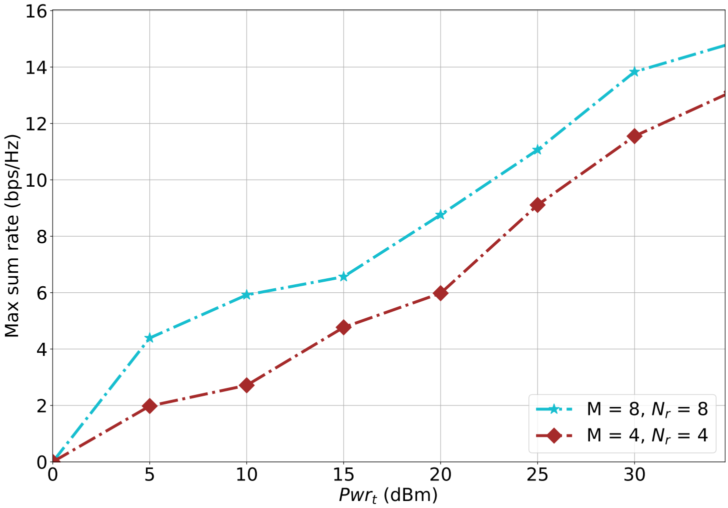


Figure 12: Effect of power in max sum rate

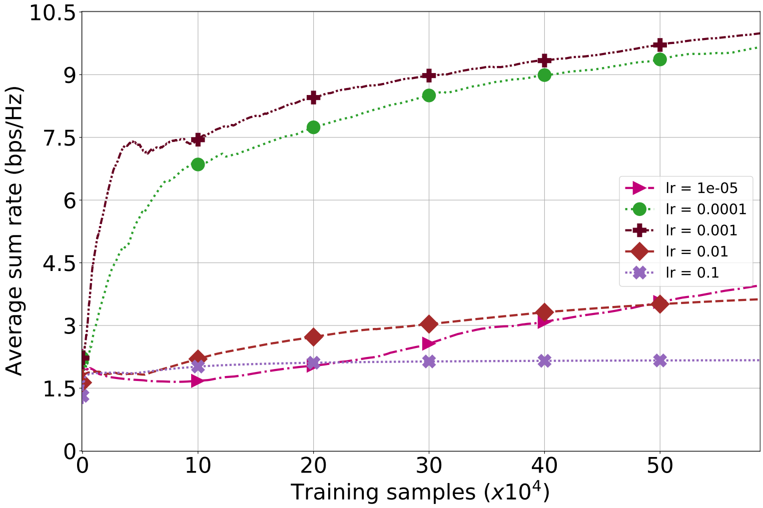


Figure 13: Hyperparameter effect on performance

sum rate will rapidly saturate at an unsatisfactory value if the learning rate is too big. The outcome can develop more quickly within a suitable range if the learning rate is larger. On the other hand, oscillations will be lessened at the expense of

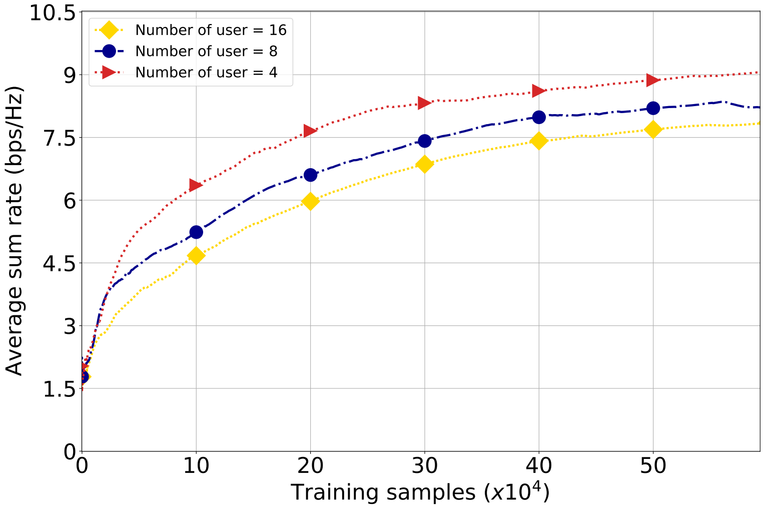


Figure 14: Convergence for different number of users

the pace of performance if the learning rate is lower. Therefore, it is important to choose the right learning rate that is neither too high nor too low.

Finally, Figure 14 shows the convergence of the suggested method for multiple users counts at a learning rate of 0.01 and $Pwr_t = 25dBm$. The convergent graphs represent the excellent adaptability and reliability of an expanding user base under the RIS-aided system. Since each UE has a rate demand that must be met, the sum rate falls as the count of users increases. The findings confirm that the suggested method can deliver a performance with maximum effectiveness.

VI. CONCLUSION

The principal focus of this thesis is on developing a collaborative design for phase shifts and transmit beamforming that optimizes the sum rate while complying with BS transmit power constraints in RIS-assisted communication systems. To provide a more practical solution to the problem, we thought of a multi-user MISO communication system while considering the effects of Nakagami-m fading channels. The critical issue is how to appropriately manage beamforming and phase shifts, which can not be resolved using conventional optimization techniques since the system is so complicated and dynamic. This non-convex optimization problem and the massive continuous action space can be addressed by leveraging breakthroughs in machine learning technology. We provide a practical DRL-based framework for optimizing the phase shifts induced by the RIS MU-MISO system to address the non-trivial optimization problem. By employing the TD3 technique, the limitations with the Q values of the critic value network of traditional DRL traditional algorithms are also eliminated. Without previous knowledge of the wireless network, the agent appropriately determines the network parameters. Simulation findings reveal that our suggested technique can perform better than traditional DRL benchmarks because after getting the feedback from the reward the agent can correctly adapt the action converging to the optimum value. Additionally, in our future works, we intend to expand the scope of the solutions we have put forth while taking into account more complicated scenarios in the MU-MIMO system for UAV and NOMA.

PUBLICATIONS

A. Journals

1. K. M. Faisal **and** W. Choi, “Machine learning approaches for reconfigurable intelligent surfaces: A survey,” *IEEE Access*, **jourvol** 10, **pages** 27 343–27 367, 2022. DOI: `10.1109/ACCESS.2022.3157651`.

B. Conferences

1. K. M. Faisal **and** W. Choi, “A study on machine learning-based approaches for reconfigurable intelligent surface,” **in***2021 International Conference on Information and Communication Technology Convergence (ICTC)* 2021, **pages** 227–232. DOI: `10.1109/ICTC52510.2021.9620993`.

REFERENCES

[1] E. Perspectives **and** C. Report, *Cisco annual internet report - cisco annual internet report (2018–2023) white paper*, 2021. Accessed: Dec. 28, 2021. [Online]. Available: `https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html`.

[2] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han **and** G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Transactions on Cognitive Communications and Networking*, **jourvol** 6, **number** 3, **pages** 990–1002, 2020. DOI: `10.1109/TCCN.2020.2992604`.

[3] B. Sheen, J. Yang, X. Feng **and** M. M. U. Chowdhury, "A deep learning based modeling of reconfigurable intelligent surface assisted wireless communications for phase shift configuration," *IEEE Open Journal of the Communications Society*, **jourvol** 2, **pages** 262–272, 2021. DOI: `10.1109/OJCOMS.2021.3050119`.

[4] C.-X. Wang, M. D. Renzo, S. Stanczak, S. Wang **and** E. G. Larsson, "Artificial intelligence enabled wireless networking for 5g and beyond: Recent advances and future challenges," *IEEE Wireless Communications*, **jourvol** 27, **number** 1, **pages** 16–23, 2020. DOI: `10.1109/MWC.001.1900292`.

[5] X. Liu, M. Chen, Y. Liu, Y. Chen, S. Cui **and** L. Hanzo, "Artificial intelligence aided next-generation networks relying on uavs," *IEEE*

Wireless Communications, **jourvol** 28, **number** 1, **pages** 120–127, 2021. DOI: 10.1109/MWC.001.2000174.

[6] Q.-V. Pham, N. Thanh Nguyen, L. B. Le, K. Lee, W.-J. Hwang *et al.*, “Intelligent radio signal processing: A contemporary survey,” *arXiv e-prints*, arXiv–2008, 2020.

[7] N. T. Nguyen, K. Lee **and** H. DaiIEEE, “Application of deep learning to sphere decoding for large mimo systems,” *IEEE Transactions on Wireless Communications*, **jourvol** 20, **number** 10, **pages** 6787–6803, 2021. DOI: 10.1109/TWC.2021.3076527.

[8] L. V. Nguyen, D. H. N. Nguyen **and** A. L. Swindlehurst, “Dnn-based detectors for massive mimo systems with low-resolution adcs,” **in***ICC 2021 - IEEE International Conference on Communications* 2021, **pages** 1–6. DOI: 10.1109/ICC42927.2021.9501054.

[9] S. Sun, Z. Cao, H. Zhu **and** J. Zhao, “A survey of optimization methods from a machine learning perspective,” *IEEE Transactions on Cybernetics*, **jourvol** 50, **number** 8, **pages** 3668–3681, 2020. DOI: 10.1109/TCYB.2019.2950779.

[10] K. M. Faisal **and** W. Choi, “Machine learning approaches for reconfigurable intelligent surfaces: A survey,” *IEEE Access*, **jourvol** 10, **pages** 27 343–27 367, 2022. DOI: 10.1109/ACCESS.2022.3157651.

[11] 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*, **jourvol** 68, **number** 3, **pages** 1581–1592, 2020. DOI: 10.1109/TCOMM.2019.2961332.

[12] Q. Wu **and** R. Zhang, “Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design,” **in***2018 IEEE Global Communications Conference (GLOBECOM)* 2018, **pages** 1–6. DOI: 10.1109/GLOCOM.2018.8647620.

[13] S. Lin, B. Zheng, G. C. Alexandropoulos, M. Wen, F. Chen **and** S. sMumtaz, “Adaptive transmission for reconfigurable intelligent surface-assisted ofdm wireless communications,” *IEEE Journal on Selected Areas in Communications*, **jourvol** 38, **number** 11, **pages** 2653–2665, 2020. DOI: 10.1109/JSAC.2020.3007038.

[14] Q. Wu **and** R. Zhang, “Beamforming optimization for intelligent reflecting surface with discrete phase shifts,” **in***ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2019, **pages** 7830–7833. DOI: 10 . 1109 / ICASSP . 2019 . 8683145.

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

[16] Y. Han, W. Tang, S. Jin, C.-K. Wen **and** X. Ma, “Large intelligent surface-assisted wireless communication exploiting statistical csi,” *IEEE Transactions on Vehicular Technology*, **jourvol** 68, **number** 8, **pages** 8238–8242, 2019. DOI: 10.1109/TVT.2019.2923997.

[17] X. Yu, D. Xu **and** R. Schober, “Miso wireless communication systems via intelligent reflecting surfaces : (invited paper),” **in***2019 IEEE/CIC*

International Conference on Communications in China (ICCC) 2019, **pages** 735–740. DOI: 10.1109/ICCChina.2019.8855810.

[18] Y. Yang, S. Zhang **and** R. Zhang, "Irs-enhanced ofdm: Power allocation and passive array optimization," **in***2019 IEEE Global Communications Conference (GLOBECOM)* 2019, **pages** 1–6. DOI: 10.1109/GLOBECOM38437.2019.9014204.

[19] Y. Gao, C. Yong, Z. Xiong, D. Niyato, Y. Xiao **and** J. Zhao, "Reconfigurable intelligent surface for miso systems with proportional rate constraints," **in***ICC 2020 - 2020 IEEE International Conference on Communications (ICC)* 2020, **pages** 1–7. DOI: 10.1109/ICC40277.2020.9148766.

[20] E. Basar, "Transmission through large intelligent surfaces: A new frontier in wireless communications," **in***2019 European Conference on Networks and Communications (EuCNC)* 2019, **pages** 112–117. DOI: 10.1109/EuCNC.2019.8801961.

[21] C. Pan, H. Ren, K. Wang *et al.*, "Multicell mimo communications relying on intelligent reflecting surfaces," *IEEE Transactions on Wireless Communications*, **jourvol** 19, **number** 8, **pages** 5218–5233, 2020. DOI: 10.1109/TWC.2020.2990766.

[22] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah **and** M.-S. Alouini, "Asymptotic max-min sinr analysis of reconfigurable intelligent surface assisted miso systems," *IEEE Transactions on Wireless Communications*, **jourvol** 19, **number** 12, **pages** 7748–7764, 2020. DOI: 10.1109/TWC.2020.2986438.

[23] H. Guo, Y.-C. Liang, J. Chen **and** E. G. Larsson, "Weighted sum-rate optimization for intelligent reflecting surface enhanced wireless networks," *arXiv preprint arXiv:1905.07920*, 2019.

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

[25] H. Shen, W. Xu, S. Gong, Z. He **and** C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Communications Letters*, **jourvol** 23, **number** 9, **pages** 1488–1492, 2019. DOI: 10.1109/LCOMM.2019.2924214.

[26] M. Cui, G. Zhang **and** R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Communications Letters*, **jourvol** 8, **number** 5, **pages** 1410–1414, 2019. DOI: 10.1109/LWC.2019.2919685.

[27] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor **and** S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, **jourvol** 20, **number** 1, **pages** 269–283, 2021. DOI: 10.1109/TWC.2020.3024629.

[28] J. Gao, C. Zhong, X. Chen, H. Lin **and** Z. Zhang, "Unsupervised learning for passive beamforming," *IEEE Communications Letters*, **jourvol** 24, **number** 5, **pages** 1052–1056, 2020. DOI: 10.1109/LCOMM.2020.2965532.

[29] R. S. Sutton **and** A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

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

[31] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen **and** L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Communications*, **jourvol** 24, **number** 2, **pages** 98–105, 2017. DOI: 10.1109/MWC.2016.1500356WC.

[32] T. L. Jensen **and** E. De Carvalho, "An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator," **in***ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2020, **pages** 5000–5004. DOI: 10.1109/ICASSP40776.2020.9053695.

[33] A. M. Elbir, A. Papazafeiropoulos, P. Kourtessis **and** S. Chatzinotas, "Deep channel learning for large intelligent surfaces aided mm-wave massive mimo systems," *IEEE Wireless Communications Letters*, **jourvol** 9, **number** 9, **pages** 1447–1451, 2020. DOI: 10.1109/LWC.2020.2993699.

[34] A. Taha, M. Alrabeiah **and** A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," *IEEE Access*, **jourvol** 9, **pages** 44 304–44 321, 2021. DOI: 10.1109/ACCESS.2021.3064073.

[35] F. Zhou, G. Lu, M. Wen, Y.-C. Liang, Z. Chu **and** Y. Wang, "Dynamic spectrum management via machine learning: State of the art, taxonomy,

challenges, and open research issues," *IEEE Network*, **jourvol** 33, **number** 4, **pages** 54–62, 2019. DOI: `10.1109/MNET.2019.1800439`.

[36] R. Shafin, H. Chen, Y.-H. Nam *et al.*, "Self-tuning sectorization: Deep reinforcement learning meets broadcast beam optimization," *IEEE Transactions on Wireless Communications*, **jourvol** 19, **number** 6, **pages** 4038–4053, 2020. DOI: `10.1109/TWC.2020.2979446`.

[37] Q. Wang, K. Feng, X. Li **and** S. Jin, "Precodernet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," *IEEE Wireless Communications Letters*, **jourvol** 9, **number** 10, **pages** 1677–1681, 2020. DOI: `10.1109/LWC.2020.3001121`.

[38] S. Gong, J. Lin, B. Ding, D. Niyato, D. I. Kim **and** M. Guizani, "When optimization meets machine learning: The case of irs-assisted wireless networks," *IEEE Network*, **jourvol** 36, **number** 2, **pages** 190–198, 2022. DOI: `10.1109/MNET.211.2100386`.

[39] A. Feriani, A. Mezghani **and** E. Hossain, "On the robustness of deep reinforcement learning in irs-aided wireless communications systems," *arXiv preprint arXiv:2107.08293*, 2021.

[40] 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* 2020, **pages** 1–6. DOI: `10.1109/GLOBECOM42002.2020.9322372`.

[41] X. Qian, M. Di Renzo, J. Liu, A. Kammoun **and** M.-S. Alouini, "Beamforming through reconfigurable intelligent surfaces in single-user mimo systems: Snr distribution and scaling laws in the presence of

channel fading and phase noise," *IEEE Wireless Communications Letters*, **jourvol** 10, **number** 1, **pages** 77–81, 2021. DOI: 10.1109/LWC.2020.3021058.

[42] Z. Zhang, Y. Cui, F. Yang **and** L. Ding, "Analysis and optimization of outage probability in multi-intelligent reflecting surface-assisted systems," *arXiv preprint arXiv:1909.02193*, 2019.

[43] T. Aulin, "Characteristics of a digital mobile radio channel," *IEEE Transactions on Vehicular Technology*, **jourvol** 30, **number** 2, **pages** 45–53, 1981. DOI: 10.1109/T-VT.1981.23882.

[44] H. Suzuki, "A statistical model for urban radio propogation," *IEEE Transactions on Communications*, **jourvol** 25, **number** 7, **pages** 673–680, 1977. DOI: 10.1109/TCOM.1977.1093888.

[45] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis **and** I. Akyildiz, "A new wireless communication paradigm through software-controlled metasurfaces," *IEEE Communications Magazine*, **jourvol** 56, **number** 9, **pages** 162–169, 2018. DOI: 10.1109/MCOM.2018.1700659.

[46] M. Di Renzo, M. Debbah, D.-T. Phan-Huy *et al.*, "Smart radio environments empowered by reconfigurable ai meta-surfaces: An idea whose time has come," *EURASIP Journal on Wireless Communications and Networking*, **jourvol** 2019, **number** 1, **pages** 1–20, 2019. DOI: 10.1186/s13638-019-1438-9.

[47] Y.-C. Liang, R. Long, Q. Zhang, J. Chen, H. V. Cheng **and** H. Guo, "Large intelligent surface/antennas (lisa): Making reflective radios smart," *Journal of Communications and Information Networks*, **jourvol** 4, **number** 2, **pages** 40–50, 2019. DOI: 10.23919/JCIN.2019.8917871.

[48] E. Basar **and** I. Yildirim, “Reconfigurable intelligent surfaces for future wireless networks: A channel modeling perspective,” *IEEE Wireless Communications*, **jourvol** 28, **number** 3, **pages** 108–114, 2021. DOI: 10.1109/MWC.001.2000338.

[49] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini **and** R. Zhang, “Wireless communications through reconfigurable intelligent surfaces,” *IEEE Access*, **jourvol** 7, **pages** 116 753–116 773, 2019. DOI: 10.1109/ACCESS.2019.2935192.

[50] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis **and** I. Akyildiz, “Realizing wireless communication through software-defined hypersurface environments,” **in***2018 IEEE 19th International Symposium on ”A World of Wireless, Mobile and Multimedia Networks” (WoWMoM)* 2018, **pages** 14–15. DOI: 10.1109/WoWMoM.2018.8449754.

[51] G. Lavigne, K. Achouri, V. S. Asadchy, S. A. Tretyakov **and** C. Caloz, “Susceptibility derivation and experimental demonstration of refracting metasurfaces without spurious diffraction,” *IEEE Transactions on Antennas and Propagation*, **jourvol** 66, **number** 3, **pages** 1321–1330, 2018. DOI: 10.1109/TAP.2018.2793958.

[52] S. V. Hum **and** J. Perruisseau-Carrier, “Reconfigurable reflectarrays and array lenses for dynamic antenna beam control: A review,” *IEEE Transactions on Antennas and Propagation*, **jourvol** 62, **number** 1, **pages** 183–198, 2014. DOI: 10.1109/TAP.2013.2287296.

[53] J. Huang, Q. Li, Q. Zhang, G. Zhang **and** J. Qin, “Relay beamforming for amplify-and-forward multi-antenna relay networks with energy harvesting

constraint," *IEEE Signal Processing Letters*, **jourvol** 21, **number** 4, **pages** 454–458, 2014. DOI: 10.1109/LSP.2014.2305737.

[54] M.
Di Renzo, K. Ntontin, J. Song *et al.*, "Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison," *IEEE Open Journal of the Communications Society*, **jourvol** 1, **pages** 798–807, 2020. DOI: 10.1109/OJCOMS.2020.3002955.

[55] F. Yang **and** Y. Rahmat-Samii, *Surface electromagnetics: with applications in antenna, microwave, and optical engineering*. Cambridge University Press, 2019.

[56] M. Di Renzo, A. Zappone, M. Debbah *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE Journal on Selected Areas in Communications*, **jourvol** 38, **number** 11, **pages** 2450–2525, 2020. DOI: 10.1109/JSAC.2020.3007211.

[57] S. Tretyakov, *Analytical modeling in applied electromagnetics*. Artech House, 2003.

[58] E. Kuester, M. Mohamed, M. Piket-May **and** C. Holloway, "Averaged transition conditions for electromagnetic fields at a metafilm," *IEEE Transactions on Antennas and Propagation*, **jourvol** 51, **number** 10, **pages** 2641–2651, 2003. DOI: 10.1109/TAP.2003.817560.

[59] A. Epstein **and** G. V. Eleftheriades, "Huygens' metasurfaces via the equivalence principle: Design and applications," *JOSA B*, **jourvol** 33, **number** 2, A31–A50, 2016. DOI: 10.1364/JOSAB.33.000A31.

[60] C. Huang, S. Hu, G. C. Alexandropoulos *et al.*, "Holographic mimo surfaces for 6g wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Communications*, **jourvol** 27, **number** 5, **pages** 118–125, 2020. DOI: 10.1109/MWC.001.1900534.

[61] E. Björnson, Ö. Özdogan **and** E. G. Larsson, "Reconfigurable intelligent surfaces: Three myths and two critical questions," *IEEE Communications Magazine*, **jourvol** 58, **number** 12, **pages** 90–96, 2020. DOI: 10.1109/MCOM.001.2000407.

[62] M. Ponsen, M. E. Taylor **and** K. Tuyls, "Abstraction and generalization in reinforcement learning: A summary and framework," **in***International Workshop on Adaptive and Learning Agents* Springer, 2009, **pages** 1–32. DOI: 10.1007/978-3-642-11814-2_1.

[63] P Beyens, M Peeters, K Steenhaut **and** A Nowe, "Routing with compression in wsns: A q-learning approach," *Proc. of the 5th Eur. Wksp on Adaptive Agents and Multi-Agent Systems (AAMAS)*, 2005.

[64] J. A. Boyan **and** M. L. Littman, "Packet routing in dynamically changing networks: A reinforcement learning approach," **in***Advances in neural information processing systems* 1994, **pages** 671–678.

[65] R. Sun, S. Tatsumi **and** G. Zhao, "Q-map: A novel multicast routing method in wireless ad hoc networks with multiagent reinforcement learning," **in***2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. TENCOM '02. Proceedings.* **volume** 1, 2002, 667–670 vol.1. DOI: 10.1109/TENCON.2002.1181362.

[66] A. Mukherjee, V. Keshary, K. Pandya, N. Dey **and** S. C. Satapathy, “Flying ad hoc networks: A comprehensive survey,” *Information and decision sciences*, **pages** 569–580, 2018. DOI: 10.1007/978-981-10-7563-6_59.

[67] D. Yu **and** L. Deng, “Deep learning and its applications to signal and information processing [exploratory dsp],” *IEEE Signal Processing Magazine*, **jourvol** 28, **number** 1, **pages** 145–154, 2011. DOI: 10.1109/MSP.2010.939038.

[68] D. C. Cireşan, U. Meier, L. M. Gambardella **and** J. Schmidhuber, “Deep, big, simple neural nets for handwritten digit recognition,” *Neural computation*, **jourvol** 22, **number** 12, **pages** 3207–3220, 2010. DOI: 10.1162/neco_a_00052.

[69] G. Hinton, L. Deng, D. Yu *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, **jourvol** 29, **number** 6, **pages** 82–97, 2012. DOI: 10.1109/MSP.2012.2205597.

[70] G. E. Dahl, D. Yu, L. Deng **and** A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, **jourvol** 20, **number** 1, **pages** 30–42, 2012. DOI: 10.1109/TASL.2011.2134090.

[71] X.-W. Chen **and** X. Lin, “Big data deep learning: Challenges and perspectives,” *IEEE Access*, **jourvol** 2, **pages** 514–525, 2014. DOI: 10.1109/ACCESS.2014.2325029.

[72] M. Ponsen, M. E. Taylor **and** K. Tuyls, "Abstraction and generalization in reinforcement learning: A summary and framework," **in***Adaptive and Learning Agents* M. E. Taylor **and** K. Tuyls, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, **pages** 1–32. DOI: 10.1007/978-3-642-11814-2_1.

[73] K. TUYLS **and** A. NOWÉ, "Evolutionary game theory and multi-agent reinforcement learning," *The Knowledge Engineering Review*, **jourvol** 20, **number** 1, 63–90, 2005. DOI: 10.1017/S026988890500041X.

[74] N. C. Luong, D. T. Hoang, S. Gong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Communications Surveys Tutorials*, **jourvol** 21, **number** 4, **pages** 3133–3174, 2019. DOI: 10.1109/COMST.2019.2916583.

[75] K. Cho **and** D. Yoon, "On the general ber expression of one- and two-dimensional amplitude modulations," *IEEE Transactions on Communications*, **jourvol** 50, **number** 7, **pages** 1074–1080, 2002. DOI: 10.1109/TCOMM.2002.800818.

[76] J. D. Parsons **and** P. J. D. Parsons, *The mobile radio propagation channel*. wiley New York, 2000, **volume** 2.

[77] A. Gosavi, "Reinforcement learning: A tutorial survey and recent advances," *INFORMS Journal on Computing*, **jourvol** 21, **number** 2, **pages** 178–192, 2009. DOI: 10.1287/ijoc.1080.0305.

ACKNOWLEDGEMENTS

I would like to express my gratitude to the Almighty Allah and all the individuals who assisted me in completing my master's degree and research. First of all, I sincerely want to thank my supervisor, Prof. Wooyeol Choi, for letting me pursue my master's degree at Chosun University. His constant inspiration, support, and insightful recommendations have led and pushed me throughout my studies and research. His ongoing oversight and direction have helped me produce high-quality research. I will be eternally grateful to him for instilling in me the values of professionalism, organizational skills, and concentration. I also would like to convey my heartfelt gratitude to Prof. Seok Joo Shin and Prof. Moon Soo Kang, members of the thesis committee, for their constructive remarks and helpful ideas. Furthermore, I am glad for the opportunity to work in the Department of Computer Engineering at Chosun University with such a diversified batch of students, teachers, and staff. I want to thank Smart Networking Lab for giving me such an excellent opportunity and an environment to develop academically. My lab mates have provided me with both moral and academic assistance. I also want to thank all my Bangladeshi friends and seniors at Chosun University for their kindness and support in making my time in South Korea enjoyable and simple. Finally, I want to thank my parents, relatives, and friends for their continuous support throughout my challenging times. I couldn't have accomplished anything without their inspiration and guidance.