A Deep Reinforcement Learning Approach for Autonomous Reconfigurable Intelligent Surfaces

Hyuckjin Choi*, Ly V. Nguyen†, Junil Choi*, and A. Lee Swindlehurst†
*School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, South Korea
†Center for Pervasive Communications and Computing, University of California, Irvine, California, USA

2024 IEEE International Conference on Communications Workshops (ICC Workshops) | 979-8-3503-0405-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICCWORKSHOPS59551.2024.10615955

***Abstract*—A reconfigurable intelligent surface (RIS) is a prospective wireless technology that enhances wireless channel quality. An RIS is often equipped with passive array of elements and provides cost and power-efficient solutions for coverage extension of wireless communication systems. Without any radio frequency (RF) chains or computing resources, however, the RIS requires control information to be sent to it from an external unit, e.g., a base station (BS). The control information can be delivered by wired or wireless channels, and the BS must be aware of the RIS and the RIS-related channel conditions in order to effectively configure its behavior. Recent works have introduced hybrid RIS structures possessing a few active elements that can sense and digitally process received data. Here, we propose the operation of an entirely autonomous RIS that operates without a control link between the RIS and BS. Using a few sensing elements, the autonomous RIS employs a deep Q network (DQN) based on reinforcement learning in order to enhance the sum rate of the network. Our results illustrate the potential of deploying autonomous RISs in wireless networks with essentially no network overhead.**

***Index Terms*—Autonomous RIS, DQN, deep learning, MU-MISO, rate maximization, wireless communication.**

I. Introduction

A reconfigurable intelligent surface (RIS) is an innovative technology that has the ability to shape a wireless channel in beneficial ways thanks to the use of adjustable reflecting elements [1], [2]. They can be used for various purposes, such as improving network throughput, coverage, or energy efficiency. It is commonly assumed that RISs are essentially passive arrays without radio-frequency (RF) chains and computing resources, and are fully controlled by an external entity such as a base station (BS). To reap the benefits of RISs, channel state information (CSI) is also generally required. However, CSI estimation in passive RIS systems is challenging, often requiring a high pilot overhead that can significantly reduce the spectral efficiency [3], [4]. In addition, the requirement of a control link through a wired cable or wireless channel limits the deployment flexibility of RISs and increases the complexity of the system installation, configuration, and maintenance costs. In some circumstances, establishing such a control link may be infeasible, for example, when an RIS is only temporarily deployed to an area.

Recently, hybrid RIS structures have been introduced in which the RIS elements are able to simultaneously reflect and sense the incoming signal [5]–[7]. Such structures pave the way for a new methodology where an RIS can "sense-then-shape" the environment itself. The benefit of hybrid RISs in terms of pilot overhead reduction has been recently reported in [7], [8]. Motivated by these recent advances, we study the concept of an autonomous RIS as depicted in Fig. 1, which is *self-configured* instead of being fully controlled by a remote BS, thus maximizing the deployment flexibility as well as simplifying the system configuration. Since an autonomous RIS is self-configured, the BS can operate using conventional protocols, e.g., estimating the effective instantaneous CSI and then performing combining/precoding. This approach is thus more practical than the one that the BS estimates the instantaneous cascaded CSI, jointly optimizes the combining/precoding and the phase shifts of the RIS, and then forwards the solution to the RIS via a control link [9], [10].

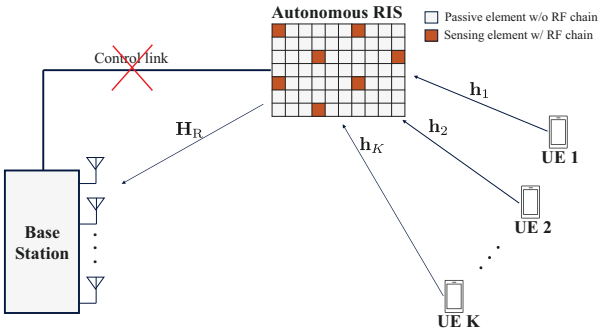

Fig. 1: Illustration of an autonomous RIS-assisted system.

The proposed autonomous RIS employs a hybrid RIS with some sensing elements and processing capability to self-configure its phase shifts using a deep Q network (DQN) [11]. DQN is a reinforcement learning method that is useful when the state, e.g., a received signal, is strongly related to the channel environment, and the action, e.g., the RIS configuration, changes the channel environment. Since it is impractical to define the received signals and wireless channels in discrete sets, DQN is widely used for communication system developments, where a deep neural network is used to create a continuous-valued Q table [9], [10], [12]. Without any control link, the partial observations provided by the RIS sensing elements are the only information available to the autonomous RIS. The key contribution of the paper is a method to convert the partial observations of the hybrid RIS into an estimate of the sum rate, which serves as the reward or Q value of the reinforcement learning-based DQN.

The paper is organized as follows. In Section II, we present a cluster-based channel model suitable for describing time variations due to changing positions of the clusters and mobile user equipment (UE). Section III provides the RIS system model and defines the observations from a few RIS sensing elements. Using the RIS observations, we develop a method for evaluating the sum-rate in Section IV. Section V details the proposed DQN approach, and Section VI illustrates its performance using simulation results. Concluding remarks follow in Section VII.

Notation: Upper-case and lower-case boldface letters are used to indicate matrices and column vectors, respectively. The element-wise absolute value of vector $\mathbf{a}$ is represented as $|\mathbf{a}|$. The p-norm of vector $\mathbf{a}$ is $\|\mathbf{a}\|_p$. The indicator operator $[\![\mathbf{a}]\!]_\infty$ is defined as $[\![\mathbf{a}]\!]_p = [a_1^p \ a_2^p \ \cdots]^{\mathrm{T}}/\|\mathbf{a}\|_p^p$ for $p \to \infty$. The i-th element of vector $\mathbf{a}$ and the i-th row of matrix $\mathbf{A}$ are given as $[\mathbf{a}]_i$ and $[\mathbf{A}]_{a,:}$, respectively. A partial vector formed from elements of vector $\mathbf{a}$ using the set of indices $\mathcal{I}$ is defined as $[\mathbf{a}]_{\mathcal{I}}$. The matrices $\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{H}}$ denote the transpose and the conjugate transpose of $\mathbf{A}$, respectively. The discrete Fourier transform (DFT) of matrix $\mathbf{A}$ is given by DFT$\{\mathbf{A}\}$, where the DFT is performed column-wise. The set of complex numbers is denoted as $\mathbb{C}$.

II. Channel Model

We consider a system with an M-antenna BS serving K single-antenna UEs with the help of an N-element RIS. We assume the cluster-based channel model, where there is no direct channel between the UEs and the BS. Our model considers signals arriving at the autonomous RIS from both the BS and UEs. The channel between the k-th UE and the RIS is modeled as

$$\mathbf{h}_k = \sum_{\ell=0}^{L_k-1} \frac{\sqrt{P_0}}{\prod_{p=0}^{P_{k,\ell}-1}(d_{k,\ell,p})^{\alpha}} e^{-j2\pi f_c \tau_{k,\ell}} \mathbf{a}_{\mathrm{R}}(\theta_{k,\ell}^{\mathrm{hor}}, \theta_{k,\ell}^{\mathrm{ver}}), \quad (1)$$

where f_c is the carrier frequency, α is the pathloss exponent, P_0 is the reference channel power, and L_k is the number of distinct channel paths, each of which can arrive via multiple reflections. The delay $\tau_{k,l}$ of channel path ℓ is given by $\tau_{k,\ell} = \sum_{p=0}^{P_{k,\ell}-1} d_{k,\ell,p}/c$ where c is the speed of light, and $d_{k,\ell,p}$ represents the length of the p-th segment of the ℓ-th channel path for the k-th UE. The uniform planar array (UPA) response at the RIS is defined as

$$\begin{aligned}\mathbf{a}_{\mathrm{R}}(\theta_{k,\ell}^{\mathrm{hor}}, \theta_{k,\ell}^{\mathrm{ver}}) = \\ [1 \ e^{j\pi\cos(\theta_{k,\ell}^{\mathrm{hor}})\sin(\theta_{k,\ell}^{\mathrm{ver}})} \ \cdots \ e^{j(N_{\mathrm{hor}}-1)\pi\cos(\theta_{k,\ell}^{\mathrm{hor}})\sin(\theta_{k,\ell}^{\mathrm{ver}})}]^{\mathrm{T}} \\ \otimes [1 \ e^{j\pi\cos(\theta_{k,\ell}^{\mathrm{ver}})} \ \cdots \ e^{j(N_{\mathrm{ver}}-1)\pi\cos(\theta_{k,\ell}^{\mathrm{ver}})}]^{\mathrm{T}}, \end{aligned} \quad (2)$$

where $\theta_{k,\ell}^{\mathrm{hor}}$ and $\theta_{k,\ell}^{\mathrm{ver}}$ are the zenith and azimuth angles of arrival (ZoA/AoA) for the ℓ-th channel path of the k-th UE.

The BS-to-RIS channel is modeled as

$$\begin{aligned}\mathbf{H}_{\mathrm{R}} = \sum_{\ell=0}^{L_{\mathrm{R}}-1} \frac{\sqrt{P_0}}{\prod_{p=0}^{P_{\mathrm{R},\ell}-1}(d_{\mathrm{R},\ell,p})^{\alpha}} e^{-j2\pi f_c \tau_{\mathrm{R},\ell}} \\ \times \mathbf{a}_{\mathrm{R}}(\theta_{\mathrm{R},\ell}^{\mathrm{hor}}, \theta_{\mathrm{R},\ell}^{\mathrm{ver}}) \mathbf{a}_{\mathrm{B}}^{\mathrm{H}}(\phi_{\mathrm{R},\ell}^{\mathrm{ver}}) \end{aligned} \quad (3)$$

where the segment lengths $d_{\mathrm{R},\ell,p}$ and the delay $\tau_{\mathrm{R},\ell}$ are defined similarly to (1). We assume a uniform linear array (ULA) at the BS whose response $\mathbf{a}_{\mathrm{B}}(\cdot) \in \mathbb{C}^{M\times 1}$ can be found as in (2) with $N_{\mathrm{hor}} = 1$, and the ZoA/AoA and zenith of departure (ZoD) are $\theta_{\mathrm{R},\ell}^{\mathrm{hor}}, \theta_{\mathrm{R},\ell}^{\mathrm{ver}}$ and $\phi_{\mathrm{R},\ell}^{\mathrm{ver}}$, respectively. The positions of the clusters, BS, RIS, and UE nodes define the parameters of the channels, including the lengths of the channel segments and the ZoDs, AoAs, and ZoDs of the channel paths. Time-variations in the positions of the clusters and UEs produce channel variations. In Section VI, we describe the model assumed for cluster and UE motion in the numerical studies.

III. System Model

The effective uplink (UL) channel for the k-th UE is given as

$$\mathfrak{h}_k = \mathbf{H}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}(\mathbf{v}) \mathbf{h}_k, \quad (4)$$

where $\mathbf{v}$ is the RIS phase shift vector defined as $\mathbf{v} = [e^{j\psi_1} \ \cdots \ e^{j\psi_N}]^{\mathrm{T}}$. Let $\mathbf{H} = [\mathfrak{h}_1, \ldots, \mathfrak{h}_K] \in \mathbb{C}^{M\times K}$ be the effective uplink channel matrix. We assume reciprocity, so the downlink channel matrix is $\mathbf{H}^{\mathrm{H}} = [\mathfrak{h}_1, \ldots, \mathfrak{h}_K]^{\mathrm{H}} \in \mathbb{C}^{K\times M}$.

In a time division duplex (TDD) system, the downlink (DL) signal received at the k-th UE is written as

$$r_k = \mathfrak{h}_k^{\mathrm{H}} \mathbf{F} \mathbf{s} + n_k, \quad (5)$$

where the DL precoding matrix $\mathbf{F}$ consists of K precoding vectors $\mathbf{F} = [\mathbf{f}_1 \ \cdots \ \mathbf{f}_K]$, and the DL symbol vector is given as $\mathbf{s} = [s_1 \ \cdots \ s_K]^{\mathrm{T}}$. The noise at the k-th UE is zero-mean Gaussian with variance σ_n^2, which we denote as $n_k \sim \mathcal{CN}(0, \sigma_n^2)$. We assume that the DL symbols are randomly generated and satisfy $\mathbb{E}[\mathbf{s}\mathbf{s}^{\mathrm{H}}] = P_{\mathrm{BS}}\mathbf{I}_K$, where P_{BS} is the BS transmit power, and $\mathbf{I}_K$ is a $K \times K$ identity matrix.

The limited number of RIS sensing elements enables the RIS to obtain partial information about the UL and DL channels, The signal from the BS to the RIS is given as

$$[\mathbf{y}_{\mathrm{R}}]_{\mathcal{I}_s} = [\mathbf{H}_{\mathrm{R}}]_{\mathcal{I}_s,:} \mathbf{F}\mathbf{s} + \mathbf{n}_{\mathrm{R},\mathcal{I}_s}, \quad (6)$$

where $\mathcal{I}_s$ is the set of indices corresponding to the sensing elements, the noise term $\mathbf{n}_{\mathrm{R},\mathcal{I}_s}$ is distributed as $\mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{|\mathcal{I}_s|})$, and $|\mathcal{I}_s|$ is the cardinality of $\mathcal{I}_s$. The received signal at the RIS from the k-th UE is given by

$$[\mathbf{y}_k]_{\mathcal{I}_s} = [\mathbf{h}_k]_{\mathcal{I}_s} x_k + \mathbf{n}_{k,\mathcal{I}_s}, \quad (7)$$

where x_k is the UL symbol from the k-th UE with transmit power $\mathbb{E}[|x_k|^2] = P_{\mathrm{UE}}$. The noise $\mathbf{n}_{k,\mathcal{I}_s}$ is also assumed to be distributed as $\mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{|\mathcal{I}_s|})$.

In the following section, we propose a method for evaluating the sum-rate using only the partial observations from the RIS sensing elements. The derived sum-rate will then be used to form the DQN reward.

IV. Proposed Sum-Rate Evaluation

A. Observation Recovery

The full-dimensional RIS received signals $\mathbf{y}_{\mathrm{R}}$ from the BS and $\mathbf{y}_k$ from the k-th UE are necessary for the sum-rate evaluation. From (3), $\mathbf{y}_{\mathrm{R}}$ can be represented as

$$\begin{aligned}\mathbf{y}_{\mathrm{R}} &= \mathbf{H}_{\mathrm{R}}\mathbf{F}\mathbf{s} + \mathbf{n}_{\mathrm{R}} \\ &= \sum_{\ell=0}^{L_{\mathrm{R}}-1} \frac{\sqrt{P_0}}{\prod_{p=0}^{P_{\mathrm{R},\ell}-1}(d_{\mathrm{R},\ell,p})^{\alpha}} e^{-j2\pi f_c \tau_{\mathrm{R},\ell}} \\ &\qquad \times \mathbf{a}_{\mathrm{R}}(\theta^{\mathrm{hor}}_{\mathrm{R},\ell}, \theta^{\mathrm{ver}}_{\mathrm{R},\ell}) \mathbf{a}^{\mathrm{H}}_{\mathrm{B}}(\phi^{\mathrm{ver}}_{\mathrm{R},\ell}) \mathbf{F}\mathbf{s} + \mathbf{n}_{\mathrm{R}} \\ &= \sum_{\ell=0}^{L_{\mathrm{R}}-1} \beta_{\mathrm{R},\ell} \mathbf{a}_{\mathrm{R}}(\theta^{\mathrm{hor}}_{\mathrm{R},\ell}, \theta^{\mathrm{ver}}_{\mathrm{R},\ell}) + \mathbf{n}_{\mathrm{R}}, \end{aligned} \tag{8}$$

by introducing $\beta_{\mathrm{R},\ell} = \frac{\sqrt{P_0}}{\prod_{p=0}^{P_{\mathrm{R},\ell}-1}(d_{\mathrm{R},\ell,p})^{\alpha}} e^{-j2\pi f_c \tau_{\mathrm{R},\ell}} \mathbf{a}^{\mathrm{H}}_{\mathrm{B}}(\phi^{\mathrm{ver}}_{\mathrm{R},\ell}) \mathbf{F}\mathbf{s}$. In (8), the parameters $\beta_{\mathrm{R},\ell}, \theta^{\mathrm{hor}}_{\mathrm{R},\ell}$, and $\theta^{\mathrm{ver}}_{\mathrm{R},\ell}$ fully define the channel. The partial observation in (6) can be written as

$$[\mathbf{y}_{\mathrm{R}}]_{\mathcal{I}_s} = \sum_{\ell=0}^{L_{\mathrm{R}}-1} \beta_{\mathrm{R},\ell} [\mathbf{a}_{\mathrm{R}}(\theta^{\mathrm{hor}}_{\mathrm{R},\ell}, \theta^{\mathrm{ver}}_{\mathrm{R},\ell})]_{\mathcal{I}_s} + \mathbf{n}_{\mathrm{R},\mathcal{I}_s}, \tag{9}$$

which still contains information about $\beta_{\mathrm{R},\ell}, \theta^{\mathrm{hor}}_{\mathrm{R},\ell}$, and $\theta^{\mathrm{ver}}_{\mathrm{R},\ell}$. If the channel parameters $\beta_{\mathrm{R},\ell}, \theta^{\mathrm{hor}}_{\mathrm{R},\ell}$, and $\theta^{\mathrm{ver}}_{\mathrm{R},\ell}$ can be extracted from (9), it is possible to reconstruct (8) assuming the noise is not excessive.

The ZoA/AoA $\theta^{\mathrm{hor}}_{\mathrm{R},\ell}$ and $\theta^{\mathrm{ver}}_{\mathrm{R},\ell}$ can be inferred using the orthogonal matching pursuit (OMP) algorithm [13]. The autocorrelation matrix for OMP can be expressed as

$$\mathbf{R}_{\mathbf{y}_{\mathrm{R}}} = \mathbb{E}\left[[\mathbf{y}_{\mathrm{R}}]_{\mathcal{I}_s} [\mathbf{y}_{\mathrm{R}}]^{\mathrm{H}}_{\mathcal{I}_s} \right], \tag{10}$$

which can be obtained by a sample average. The OMP algorithm gives the estimated ZoA/AoA $\hat{\theta}^{\mathrm{hor}}_{\mathrm{R},\ell}$ and $\hat{\theta}^{\mathrm{ver}}_{\mathrm{R},\ell}$ from $\mathbf{R}_{\mathbf{y}_{\mathrm{R}}}$. With AoAs obtained by OMP, the remaining channel parameter $\beta_{\mathrm{R},\ell}$ can be computed as

$$\begin{aligned}\hat{\beta}_{\mathrm{R},\ell} &= \frac{1}{|\mathcal{I}_s|} [\mathbf{a}(\hat{\theta}^{\mathrm{hor}}_{\mathrm{R},\ell}, \hat{\theta}^{\mathrm{ver}}_{\mathrm{R},\ell})]^{\mathrm{H}}_{\mathcal{I}_s} [\mathbf{y}_{\mathrm{R}}]_{\mathcal{I}_s} \\ &\approx \beta_{\mathrm{R},\ell} + \frac{1}{|\mathcal{I}_s|} [\mathbf{a}(\hat{\theta}^{\mathrm{hor}}_{\mathrm{R},\ell}, \hat{\theta}^{\mathrm{ver}}_{\mathrm{R},\ell})]^{\mathrm{H}}_{\mathcal{I}_s} \mathbf{n}_{\mathrm{R},\mathcal{I}_s}. \end{aligned} \tag{11}$$

With the estimated parameters $\hat{\theta}^{\mathrm{hor}}_{\mathrm{R},\ell}, \hat{\theta}^{\mathrm{ver}}_{\mathrm{R},\ell}$ and $\hat{\beta}_{\mathrm{R},\ell}$, the full-dimensional RIS received signal from the BS is recovered as

$$\hat{\mathbf{y}}_{\mathrm{R}} = \sum_{\ell=0}^{L_{\mathrm{R}}-1} \hat{\beta}_{\mathrm{R},\ell} \mathbf{a}_{\mathrm{R}}(\hat{\theta}^{\mathrm{hor}}_{\mathrm{R},\ell}, \hat{\theta}^{\mathrm{ver}}_{\mathrm{R},\ell}). \tag{12}$$

The full-dimensional received signal from the k-th UE can be obtained using the same procedure.

B. Sum-Rate Evaluation Method

The DL sum-rate defined as

$$\mathcal{R} = \sum_{k=1}^{K} \log_2 \left(1 + \frac{P_{\mathrm{BS}} |\mathfrak{h}_k^{\mathrm{H}} \mathbf{f}_k|^2}{\sum_{k' \neq k} P_{\mathrm{BS}} |\mathfrak{h}_k^{\mathrm{H}} \mathbf{f}_{k'}|^2 + \sigma_n^2} \right) \tag{13}$$

can be used as the DQN reward. However, the autonomous RIS cannot directly calculate the sum-rate in (13) since it requires precise CSI. The autonomous RIS therefore has to evaluate the sum-rate relying on the sensed observations from the RIS.

We first define the observation

$$z_k = \hat{\mathbf{y}}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \hat{\mathbf{y}}_k \tag{14}$$

with the RIS received signals $\hat{\mathbf{y}}_{\mathrm{R}}$ and $\hat{\mathbf{y}}_k$ that are recovered in Section IV-A. Assuming perfect recovery of $\mathbf{y}_{\mathrm{R}}$ and $\mathbf{y}_k$, the observation is expanded as

$$\begin{aligned} z_k =& (\mathbf{H}_{\mathrm{R}}\mathbf{F}\mathbf{s} + \mathbf{n}_{\mathrm{R}})^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} (\mathbf{h}_k x_k + \mathbf{n}_k) \\ =& \mathbf{s}^{\mathrm{H}}\mathbf{F}^{\mathrm{H}}\mathbf{H}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{h}_k x_k + \mathbf{s}^{\mathrm{H}}\mathbf{F}^{\mathrm{H}}\mathbf{H}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{n}_k \\ &+ \mathbf{n}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{h}_k x_k + \mathbf{n}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{n}_k. \end{aligned} \tag{15}$$

If we assume that the BS precoder is designed sufficiently well, e.g., using zero-forcing (ZF) or minimum mean square error (MMSE) precoders, the inter-user interference (IUI) can be assumed to be negligible, i.e., $\sum_{k' \neq k} \mathbf{f}_{k'}^{\mathrm{H}} \mathbf{H}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{h}_k \approx 0$. With this assumption, (15) can be further formulated as

$$\begin{aligned} z_k \approx& s_k^* \mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{h}_k x_k + s_k^* \mathbf{F}^{\mathrm{H}} \mathbf{H}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{n}_k \\ &+ \mathbf{n}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{h}_k x_k + \mathbf{n}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{n}_k, \end{aligned} \tag{16}$$

from which the following can be evaluated:

$$\begin{aligned} \mathbb{E}[|z_k|^2] &\approx P_{\mathrm{BS}} P_{\mathrm{UE}} |\mathbf{f}_k^{\mathrm{H}} \mathbf{H}_{\mathrm{R}}^{\mathrm{H}} \operatorname{diag}\{\mathbf{v}\} \mathbf{h}_k|^2 + N\sigma_n^4 \\ &= P_{\mathrm{BS}} P_{\mathrm{UE}} |\mathbf{f}_k^{\mathrm{H}} \mathfrak{h}_k|^2 + N\sigma_n^4, \end{aligned} \tag{17}$$

where the expectation is taken over the DL symbol s_k, UL symbol x_k, and noise signals $\mathbf{n}_{\mathrm{R}}$ and $\mathbf{n}_k$ in the DL and UL transmissions, respectively. The expectation in (17) can be evaluated using a sample average. Finally, the sum-rate can be evaluated as

$$\begin{aligned} \hat{\mathcal{R}} &= \sum_{k=1}^{K} \log_2 \left(1 + \frac{\mathbb{E}[|z_k|^2]/P_{\mathrm{UE}}}{\sigma_n^2} \right) \\ &= \sum_{k=1}^{K} \log_2 \left(1 + \frac{P_{\mathrm{BS}} |\mathfrak{h}_k^{\mathrm{H}} \mathbf{f}_k|^2 + N\sigma_n^4 / P_{\mathrm{UE}}}{\sigma_n^2} \right). \end{aligned} \tag{18}$$

The sum-rate in (18) will be close to the actual sum-rate in (13) assuming a proper beamformer $\mathbf{F}$ satisfying $\sum_{k' \neq k} |\mathfrak{h}_k^{\mathrm{H}} \mathbf{f}_{k'}|^2 \approx 0$ and sufficient transmit power $P_{\mathrm{BS}} \gg \sigma_n^2$. Section V investigates more details of the proposed DQN, where the sum-rate evaluated in (18) is used to define a reward.

V. Proposed DQN Design

DQN is a reinforcement learning method with a deep neural network. A flowchart for our proposed DQN is given in Fig. 2. The RIS recovers the received signal from its sensing elements and evaluates the sum-rate. The recovered observation is transformed into the DQN state, the input of the DQN neural network, and the sum-rate gives the target Q value. The DQN neural network outputs the Q value, after which the target Q value is updated and accumulated for DQN training. The RIS chooses the action according to a given policy, and then the RIS

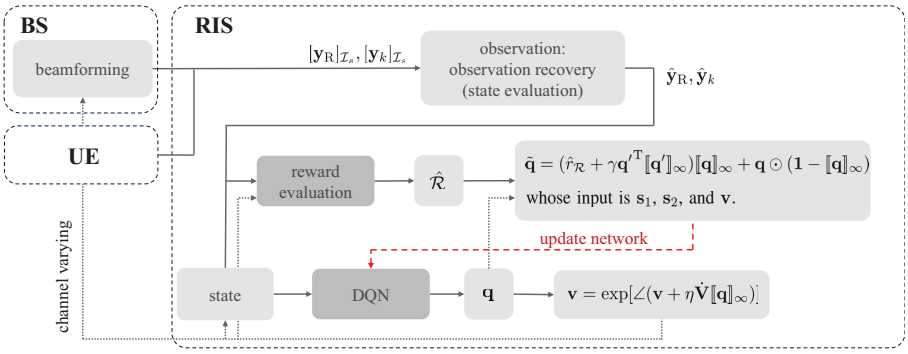

Fig. 2: Flowchart of the proposed DQN interacting with the environment.

phase shifts are updated. As a result, the RIS channel changes, and the BS modifies the beamformer for the given RIS channel. We present further details about the arguments and structure of the DQN in Sections V-A and V-B, respectively. The DQN training process is then explained in Section V-C.

A. DQN Arguments

1) Environment: The environment refers to the medium with which the agent interacts, including the BS, the wireless channels, and the UEs.

2) Agent: The RIS acts as the agent.

3) State: The DQN has two states including the combined RIS observation

$$\mathbf{s}_{1,k} = \left|\text{DFT}\left\{\text{diag}\left\{\mathbf{y}_{\text{R}}^{\text{H}}\right\}\text{diag}\left\{\mathbf{v}\right\}\mathbf{y}_k\right\}\right|^{\text{T}}, \tag{19}$$

from which we obtain $\mathbf{s}_1 = [\mathbf{s}_{1,1} \ \cdots \ \mathbf{s}_{1,K}]$, and the RIS phase shift vector

$$\mathbf{s}_2 = \left|\text{DFT}\left\{\mathbf{v}\right\}\right|^{\text{T}}, \tag{20}$$

which are the DQN inputs that go into separate pipelines of the DQN structure and merge to predict the Q value. The combined observation in (19) can be represented as

$$\begin{aligned}\mathbf{s}_{1,k} &\approx \left|\text{DFT}\left\{\text{diag}\left\{s_k^*\mathbf{f}_k^{\text{H}}\mathbf{H}_{\text{R}}^{\text{H}}\right\}\text{diag}\left\{\mathbf{v}\right\}\mathbf{h}_k x_k\right\}\right|^{\text{T}} \\ &= \left|s_k^* x_k \text{DFT}\left\{\text{diag}\left\{\mathbf{f}_k^{\text{H}}\mathbf{H}_{\text{R}}^{\text{H}}\right\}\text{diag}\left\{\mathbf{v}\right\}\mathbf{h}_k\right\}\right|^{\text{T}} \\ &= \left|s_k^* x_k \text{DFT}\left\{\text{diag}\left\{\mathbf{f}_k^{\text{H}}\right\}\boldsymbol{\mathfrak{h}}_k\right\}\right|^{\text{T}} \\ &= \left|s_k^* x_k \,\text{diag}\left\{\mathbf{f}_k^{\text{H}}\right\}\text{DFT}\left\{\boldsymbol{\mathfrak{h}}_k\right\}\right|^{\text{T}}\end{aligned} \tag{21}$$

where the approximation is due to the assumptions $\sum_{k' \neq k} |\boldsymbol{\mathfrak{h}}_k^{\text{H}}\mathbf{f}_{k'}|^2 \approx 0$, and $P_{\text{BS}} \gg \sigma_n^2$. Since the representation in (21) becomes a function of the concatenated RIS channel $\boldsymbol{\mathfrak{h}}_k$, we use the combined observation as the state. Note that the DFT operations in (19) and (20) convert the combined RIS observation $\text{diag}\left\{\mathbf{y}_{\text{R}}^{\text{H}}\right\}\text{diag}\left\{\mathbf{v}\right\}\mathbf{y}_k$ and the RIS phase shift vector $\mathbf{v}$ into the spatial domain. This transformation makes the signal sparser and more informative to the DQN. Since the absolute value varies slowly in the spatial domain and is robust to overfitting, we take the absolute values for each state.

4) Action: The action chosen by the policy determines the RIS phase update. The optimal policy selects the maximum Q value as follows:

$$a^* = \arg\max_{a \in \mathcal{A}} Q^*([\mathbf{s}_1, \mathbf{s}_2], a), \tag{22}$$

where $\mathcal{A} = \{1, \ldots, N_a\}$ represents the set of N_a possible actions, and $Q^*([\mathbf{s}_1, \mathbf{s}_2], a)$ is the action-value function. We define a vector $\mathbf{q}$ whose a-th element is $Q^*([\mathbf{s}_1, \mathbf{s}_2], a)$, and with the states $[\mathbf{s}_1, \mathbf{s}_2]$ and the Q value vector $\mathbf{q}$, we can define the DQN as

$$\mathbf{q} = f_{\text{DQN}}(\mathbf{s}_1, \mathbf{s}_2). \tag{23}$$

The RIS phase shifts are updated as

$$\mathbf{v} = \exp[j\angle(\mathbf{v} + \eta\Delta\mathbf{v})], \tag{24}$$

where the update direction is defined as $\Delta\mathbf{v} = \dot{\mathbf{V}}\mathbf{e}_{a^*}$, η is the step size, the action set matrix $\dot{\mathbf{V}} = [\dot{\mathbf{v}}_1 \ \cdots \ \dot{\mathbf{v}}_{N_a}]$ is a set of possible update directions for the RIS phase shifts, and $\mathbf{e}_{a^*}$ is a one-hot vector whose only non-zero element is in position a^* and is equal to 1. In particular, we set $\mathbf{e}_{a^*} = [\![\mathbf{q}]\!]_\infty$. In our implementation, we will choose the action set matrix to be a DFT matrix. Since the columns of the DFT matrix form the basis of an N-dimensional space, consecutive action decisions can be interpreted as different linear combinations of the basis vectors. Thus, even with a discrete action set, the resulting RIS phase shift can be any vector in N-dimensional space. To tune the convergence of the RIS update, we use the following adaptive learning rate

$$\eta = 0.1|\Delta\mathcal{R}| + 0.01, \tag{25}$$

where $\Delta\mathcal{R} = \mathcal{R}' - \mathcal{R}$ is the rate increment, and serves as a proxy for the gradient.

5) Policy: We employ an ϵ-greedy policy. During the initial stages, the Q value is not reliable, so choosing the best action as in (22) is not the best strategy. Thus initially, with probability ϵ, the action is randomly chosen. The value of ϵ gradually decreases so that the best action in (22) is more likely to be selected in later stages.

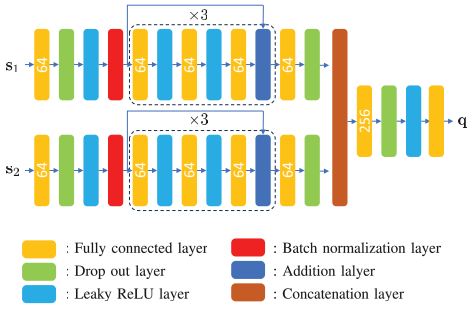

Fig. 3: Structure of the DQN neural network with two input pipelines. The number of weight parameters is written on each layer.

6) Reward: Instead of directly using the sum-rate as the reward, we propose to define the reward as the sum-rate ratio $\hat{r}_{\mathcal{R}} = \hat{\mathcal{R}}'/\hat{\mathcal{R}}$, where $\hat{\mathcal{R}}$ is the present estimated sum-rate and $\hat{\mathcal{R}}'$ is the next estimated sum-rate. Maximizing the sum-rate ratio is equivalent to maximizing the sum-rate, but it is more robust to arbitrary scaling.

The DQN neural network works as a Q table for conventional reinforcement learning. Since there are no labeled data for the reinforcement learning, the DQN training needs the target Q data. With the reward, the target Q value vector $\tilde{\mathbf{q}}$ can be evaluated as

$$[\tilde{\mathbf{q}}]_{a^*} = \hat{r}_{\mathcal{R}} + \gamma \max_{a'} Q^*([\mathbf{s}_1', \mathbf{s}_2'], a') = \hat{r}_{\mathcal{R}} + \gamma \mathbf{q}'^{\mathrm{T}} [\![\mathbf{q}']\!]_{\infty}, \tag{26}$$

where γ is the discount factor, and $\mathbf{q}'$ is the Q value vector for the next state $[\mathbf{s}_1', \mathbf{s}_2']$. The other elements in $\tilde{\mathbf{q}}$ are equal to those in $\mathbf{q}$. The target Q value vector $\tilde{\mathbf{q}}$ is the desired DQN output for the next state, which trains the DQN $f_{\text{DQN}}(\cdot)$. However, if the a-th elements of $\tilde{\mathbf{q}}$ and $f_{\text{DQN}}(\mathbf{s}_1, \mathbf{s}_2)$ are not of the same scale, which means $[\tilde{\mathbf{q}}]_a$ is too large or too small, the target Q value vector $\tilde{\mathbf{q}}$ might not be useful for training. Thus, we propose to use the following normalized target Q vector:

$$\check{\mathbf{q}} = \frac{\eta}{\sigma_{\tilde{\mathbf{q}}}} (\tilde{\mathbf{q}} - m_{\tilde{\mathbf{q}}} \mathbf{1}) + \frac{1}{1-\gamma}, \tag{27}$$

where $m_{\tilde{\mathbf{q}}}$ and $\sigma_{\tilde{\mathbf{q}}}^2$ are the sample mean and sample variance of the elements in $\tilde{\mathbf{q}}$. We still use the learning rate η here to change the variance such that $\check{\mathbf{q}}$ becomes comparable to $f_{\text{DQN}}(\mathbf{s}_1, \mathbf{s}_2)$. The loss function for training DQN is defined as

$$\text{Loss} = |[\check{\mathbf{q}} - f_{\text{DQN}}(\mathbf{s}_1, \mathbf{s}_2)]_{a^*}|^2, \tag{28}$$

where only the difference for the a^*-th element gives the DQN loss for the backpropagation.

B. DQN structure

The autonomous RIS exploits the DQN structure in Fig. 3 which is similar to [10], but the component layers are slightly different. The DQN input states are the RIS observations and the RIS phase shift vector in Eqs. (19) and (20), and the output is the Q value vector $\mathbf{q}$ whose label can be obtained as in (26). Since gradient descent is a critical issue for the DQN training, we add a batch normalization layer. The dropout layers are connected to the fully connected layers to address the overfitting issue. The activation layer here uses the leaky rectified linear unit (ReLU) function since the rate decrement should be considered in the DQN training and propagated in a negative direction [14].

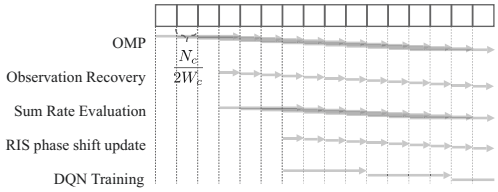

Fig. 4: The procedure associated with the proposed DQN for autonomous RIS. Each arrow means the conducting time for every operations. The training sample period is $N_c/2W_s$, where W_s is baseband bandwidth, and N_c is the sampling interval.

C. DQN Training

The proposed DQN presented above is an online machine learning approach, i.e., the DQN keeps updating its parameters using consecutive observations obtained from the RIS sensing elements. The processing sequence is shown in Fig. 4. The successively accumulated data provides the expected value of the data required for every time slot. The observation recovery and the RIS phase shift update employ instantaneous data from every training sample period. As a result, the DQN state and reward can be continuously evaluated to determine the necessary action. The DQN training is, however, not conducted with every sample. The training data is stacked for a given period and shuffled for the DQN training, which prevents overfitting. The computational complexity of the DQN is largely influenced by the training process. However, since actions can be selected concurrently during training, the proposed system is relatively unburdened by computational delay.

VI. Simulation Results

In this section, we present simulation results to show the effectiveness of our proposed DQN-based approach for the considered autonomous RIS system. The BS and the RIS are respectively located at $(0, 0, 35)$ m and $(-50, 0, 10)$ m, and the UEs are randomly placed in an area of size 100×50 m^2 centered at $(0, 0)$ m. There are two UEs, and the UE height is fixed to 1 m. The size of the BS ULA is 4×1, and the size of the RIS UPA is 4×8. The direct path between the BS and UEs is assumed to be blocked. The BS employs the MMSE precoder [15]. The first row and column of the RIS UPA are assumed to be the sensing elements, for a total of 13 RIS receivers. The system bandwidth is 20 MHz. The BS and UE transmit power is 30 dBm and 10 dBm, respectively.

The BS-to-RIS and UE-to-RIS channels consist of both line-of-sight (LoS) and non-LoS (NLoS) channels, where the NLoS channels are generated using clusters with random scattering in an area of size $200 \times 100 \times 50$ m^3, as illustrated in Fig. **??**. There are three clusters for each BS-to-RIS and UE-to-RIS

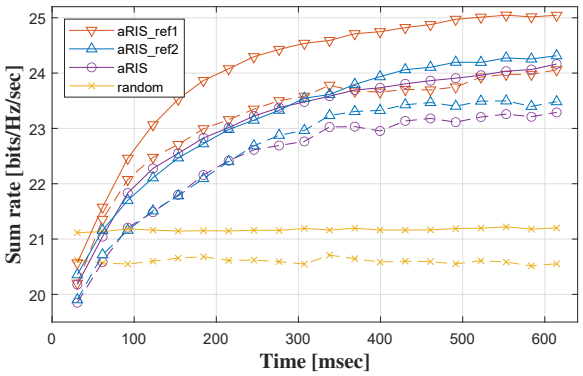

Fig. 5: Sum-rates for DQNs with different levels of information. The bold and dotted lines are for the cases of 1 m/sec and 5 m/sec UE and cluster movement, respectively.

channel. The clusters and UEs are also assumed to move at a fixed speed along a linear trajectory in random directions.

As explained in Section V, the autonomous RIS needs the full-dimensional observations and the sum-rate to obtain the DQN state and reward. To investigate the importance of each stage, we compare three DQNs: *i)* aRIS_ref1, with noise-free observations and precisely known sum-rate, *ii)* aRIS_ref2, with noise-free observations and the estimated sum-rate in (18), and *iii)* aRIS, with the proposed evaluated sum-rate after observation recovery as explained in Section IV-A. We also compare these approaches with the use of random RIS phase shifts, which does not require any channel information.

Fig. 5 shows a sum rate comparison for two scenarios where the UEs and clusters move at speeds of 1 m/sec (bold lines) and 5 m/sec (dotted lines), respectively. The DQN for the case involving UE and cluster motion at 1 m/sec converges to a higher sum rate more rapidly than the case of 5 m/sec, since more rapidly varying channels are more challenging for DQN adaptation. For a given UE speed, aRIS_ref1 achieves the best performance since it uses perfect knowledge of the channels, while the proposed a_RIS is comparable to aRIS_ref2 for both low and high mobility cases. This clearly shows the potential of autonomous RIS in practice.

VII. Conclusion

In this paper, we have proposed an autonomous RIS that does not require external control. The autonomous RIS is equipped with a few sensing elements whose measured data are used by a DQN to find a self-configured RIS phase shift solution. The proposed DQN updates the RIS phase shifts in a way that enhances the sum-rate. Due to the relatively small number of RIS sensing elements, the DQN state and reward must be established using only partial observations. Simulation results show that even with limited sensing information, the proposed DQN can still enhance the RIS channel and outperform the a random RIS configuration. Although the proposed RIS system may not be as energy-efficient as RIS systems devoid of computational resources, leveraging task-oriented hardware such as neural network processing units (NPUs) could substantially reduce power consumption.

Acknowledgement

This work was supported in part by the Korean Ministry of Science and ICT (MSIT) under the Information Technology Research Center support program (IITP-2024-2020-0-01787) supervised by the Institute of Information & Communications Technology Planning & Evaluation, in part by the U.S. National Science Foundation under grants CNS-2107182 and ECCS-2030029, and in part by a Korea Institute for Advancement of Technology (KIAT) grant funded by the Ministry of Trade, Industry and Energy (MOTIE) through the International Cooperative R&D program (P0022557).

References

[1] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116 753–116 773, Sept. 2019.

[2] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Select. Areas in Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[3] C. Pan, G. Zhou, K. Zhi, S. Hong, T. Wu, Y. Pan, H. Ren, M. D. Renzo, A. Swindlehurst, R. Zhang, and A. Y. Zhang, "An overview of signal processing techniques for RIS/IRS-aided wireless systems," *IEEE J. Select. Topics Signal Process.*, vol. 16, no. 5, pp. 883–917, Aug. 2022.

[4] S. Kim, H. Lee, J. Cha, S.-J. Kim, J. Park, and J. Choi, "Practical channel estimation and phase shift design for intelligent reflecting surface empowered mimo systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6226–6241, 2022.

[5] G. C. Alexandropoulos, N. Shlezinger, I. Alamzadeh, M. F. Imani, H. Zhang, and Y. C. Eldar, "Hybrid reconfigurable intelligent metasurfaces: Enabling simultaneous tunable reflections and sensing for 6G wireless communications," *arxiv preprint 2104.04690*, Apr. 2021.

[6] I. Alamzadeh, G. C. Alexandropoulos, N. Shlezinger, and M. F. Imani, "A reconfigurable intelligent surface with integrated sensing capability," *Sci. Rep.*, vol. 11, no. 1, Oct. 2021, Art. no. 20737.

[7] H. Zhang, N. Shlezinger, G. C. Alexandropoulos, A. Shultzman, I. Alamzadeh, M. F. Imani, and Y. C. Eldar, "Channel estimation with hybrid reconfigurable intelligent metasurfaces," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 2441–2456, Apr. 2023.

[8] L. V. Nguyen and A. Swindlehurst, "Decision-directed hybrid RIS channel estimation with minimal pilot overhead," *arxiv preprint 2309.11485*, 2023.

[9] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb, "Deep reinforcement learning for intelligent reflecting surfaces: Towards standalone operation," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1–5.

[10] W. Wang and W. Zhang, "Intelligent reflecting surface configurations for smart radio using deep reinforcement learning," *IEEE J. Select. Areas Commun.*, vol. 40, no. 8, pp. 2335–2346, Aug. 2022.

[11] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arxiv preprint 1312.5602*, 2013.

[12] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1581–1592, 2020.

[13] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Tran. on Signal Process.*, vol. 41, no. 12, pp. 3397–3415, 1993.

[14] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arxiv preprint 1505.00853*, 2015.

[15] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.