Deep Reinforcement Learning for Practical Phase Shift Optimization in RIS-assisted Networks over Short Packet Communications

Ramin Hashemi, Samad Ali, Ehsan Moeen Taghavi, Nurul Huda Mahmood, and Matti Latva-aho

Centre for Wireless Communications (CWC), University of Oulu, Oulu, Finland,
Emails: {ramin.hashemi, samad.ali, seyed.moeentaghavi, nurulhuda.mahmood, matti.latva-aho}@oulu.fi

***Abstract*—We study the practical phase shift design in a non-ideal reconfigurable intelligent surface (RIS)-aided ultra-reliable and low-latency communication (URLLC) system under finite blocklength (FBL) regime by leveraging a novel deep reinforcement learning (DRL) algorithm named as twin-delayed deep deterministic policy gradient (TD3). First, assuming industrial automation system with multiple actuators, the signal-to-interference-plus-noise ratio (SINR) and achievable rate in FBL regime are identified for each actuator in terms of the phase shift configuration matrix at the RIS. The channel state information (CSI) variations due to feedback delay are also considered that result in channel coefficients' obsolescence. Then, the problem framework is proposed where the objective is to maximize the total achievable FBL rate in all ACs, subject to the practical phase shift constraint at the RIS elements. Since the problem is intractable to solve using conventional optimization methods, we resort to employing an actor-critic policy gradient DRL algorithm based on TD3, which relies on interacting RIS with FA environment by taking actions which are the phase shifts at the RIS elements, to maximize the expected observed reward, which is defined as the total FBL rate. The numerical results show that optimizing the practical phase shifts in the RIS via the proposed TD3 method is highly beneficial to improve the network total FBL rate in comparison with typical DRL methods.**

***Index Terms*—Block error probability, deep reinforcement learning (DRL), finite blocklength (FBL), factory automation, reconfigurable intelligent surface (RIS), twin delayed DDPG (TD3), ultra-reliable low-latency communications (URLLC).**

I. INTRODUCTION

Industrial wireless systems involving devices, actuators (AC) and robots that require ultra-reliable and low-latency communications (URLLC) is anticipated to grow in the future 6th generation of wireless communications (6G) [1]. Industrial Internet of things (IIoT) is the industrial application of IoT connectivity along with networking and cloud computing based on data analytic collected from IoT devices. The industrial environments are very diverse and heterogeneous as they are characterized by a large number of use-cases and applications. An underlying commonality among these diverse applications is that the wireless industrial automation connectivity solutions envisioned in Industry 4.0 (initialized in 5G) [2] will leverage cloud computing and machine learning throughout the manufacturing process. The expected URLLC key performance indicators (KPIs) are *reliability* in the order of $1 - 10^{-9}$, *latency* around $0.1 \sim 1$ ms round-trip time and *jitter* in the order of 1 μs for industrial control networks [1]. There is also high data rate demand due to increased number of sensors and their resolution, e.g., for robots. In URLLC both the data and meta data sizes are small while both parts need to be very robust and have minimal error [3]. Thus, joint encoding of data and meta data is beneficial in terms of coding gain [4]. In addition, as the packets in URLLC are usually short-length, the finite blocklength (FBL) theory is leveraged to investigate the penalty term in the achievable rate due to coding in FBL regime [5].

Recently, re-configurable intelligent surface (RIS) has been introduced as a promising technology to enhance the energy efficiency, and spectral efficiency of wireless communications [6]. An RIS is composed of meta-materials where the phase and amplitude of each element can be adjusted. This allows the reflected signal to have a desired effect, e.g., enhance the received signal-to-interference-plus-noise ratio (SINR) at a given location. Because of this feature, the distribution of the received signal, when only the reflected channel through the RIS is available due to blockage in the presence of obstacles, will be as deterministic as possible depending on the quantization levels at each phase shift element or circuitry impairments [7]. Thus, the application of the RIS technology in factory automation (FA) environments in order to ensure high reliable and low-latency links due to no processing overhead is very promising.

Several existing works such as [8]–[12] have studied deep reinforcement learning (DRL) applications in phase shift design at the RIS. In [8] the secrecy rate of a wireless channel with RIS technology was maximized with quality of service (QoS) constraints on the secrecy rate and data rate requirements of the users. The resulting problem is solved by a novel DRL algorithm based on post-decision state and prioritized experience replay methods. The authors in [9] considered a downlink multiple-input single-output (MISO) system to adjust the RIS phase shifts as well as the coordinate of the flying UAV and transmit power via a decaying deep Q network (DQN) algorithm. A novel actor-critic DRL algorithm named as deep deterministic policy gradient (DDPG) that is a model-free and off-policy method for learning continuous

---

This research has been supported by the Academy of Finland, 6G Flagship program under Grant 346208.

actions is employed in [10] to maximize the secrecy rate in a downlink MISO system via adjusting the phase shifts at the RIS. The work in [11] studied maximizing the total achievable rate in infinite blocklength regime over a multi-hop multi-user RIS-aided wireless terahertz (THz) communication system. The maximization of the mmWave secrecy rate by jointly optimizing the UAV trajectory, beamforming vectors and RIS phase shift is conducted in [12] where two independent DDPG networks, i.e., twin DDPG were leveraged to allocate the action strategies.

The resource allocation problems in RIS-assisted URLLC systems over short packet communications is a relatively new topic and have only been investigated in a few papers [13]–[16]. In [13] the authors studied an optimization problem for beamforming and phase shift control in a RIS-enabled orthogonal frequency division multiple access (OFDMA) URLLC systems where the cooperation of a set of base stations (BSs) to serve the URLLC traffic was discussed. In [14] the UAV trajectory and channel blocklength in FBL regime as well as phase shift optimization in a RIS-aided network to minimize the total error probability was investigated. In [15] the user grouping, channel blocklength and the reflective beamforming optimization in a URLLC system was studied where a dedicated RIS assists the BS in transmitting short packets in FBL scenario. The proposed optimization problem was tackled by semi-definite relaxation (SDR) method and the user grouping problem is solved by a greedy algorithm. The authors in [16] studied the applicability of the RIS in joint multiplexing of enhanced mobile broadband (eMBB) and URLLC traffic to optimize the admitted URLLC packets while minimizing the eMBB rate loss to ensure the quality of service of the two traffic types by designing RIS phase shift matrices.

Despite the interesting results in the aforementioned works on the phase shift control in RIS-aided communications, the performance of a URLLC system over finite blocklength regime as well as assessing the impact of quantization at the RIS has not been investigated before. In other words, most of the prior studies, assumed that the RIS is ideal and the scenario is infinite blocklength regime. Moreover, the exploited DRL algorithm in the aforementioned papers is typically DDPG which has issues that are addressed in recent studies on machine learning literature as twin delayed DDPG (TD3) method [17]. Motivated by the compelling works on resource allocation via DRL methods in RIS communications, we aim to elaborate the phase shift control problem where the objective is to maximize the total FBL rate in a factory automation scenario with multiple actuators subject to practical phase shift constraints. Moreover, the proposed DRL algorithm is robust to channel variations over different time-slots due to being outdated because of feedback delay which deteriorates the performance.

In this paper, $\mathbf{h} \sim \mathcal{CN}(\mathbf{0}_{N\times 1}, \mathbf{C}_{N\times N})$ denotes circularly-symmetric (central) complex normal distribution vector with zero mean $\mathbf{0}_{N\times 1}$ and covariance matrix $\mathbf{C}$. The operations $[\cdot]^{\mathrm{H}}$, $[\cdot]^{\mathrm{T}}$ denote the transpose and conjugate transpose of a matrix or vector, respectively. In Section II, the system model and the FBL rate is proposed, then the optimization framework of phase shift design at the RIS is presented. In Section III the DRL solution approach is reviewed and the considered method is analyzed. The numerical results are presented in Section IV. Finally, Section V concludes the paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

Consider the downlink (DL) of an RIS-assisted wireless network in a factory setting which consists of a BS with $M = M_x \times M_y$ uniform planar array (UPA) antennas and K ACs. The RIS which has $N = N_x \times N_y$ phase shift elements constructs a communication channel between the ACs and multi-antenna BS. The total channel response between the BS and an AC consists of a reflected path from the RIS as illustrated in Fig. 1 since the direct paths between BS and the ACs are blocked due to obstacles. The channel matrix between BS and the RIS is denoted as

$$\mathbf{H} = \overline{\mathbf{H}}_{\text{LoS}} + \mathbf{H}_{\text{NLoS}} = [\mathbf{h}_1^{\text{inc}}, ..., \mathbf{h}_M^{\text{inc}}] \in \mathbb{C}^{N\times M}, \tag{1}$$

where the column vector $\mathbf{h}_m^{\text{inc}} = \sqrt{\frac{\zeta}{\zeta+1}}\overline{\mathbf{h}}_m^{\text{inc}} + \sqrt{\frac{1}{\zeta+1}}\tilde{\mathbf{h}}_m^{\text{inc}}$ for $\forall m \in \{1, ..., M\}$ which $\tilde{\mathbf{h}}_m^{\text{inc}} \sim \mathcal{CN}(\mathbf{0}_{M\times 1}, \boldsymbol{\beta}^{\text{inc}})$, $\boldsymbol{\beta}^{\text{inc}} = \text{diag}(\beta_1^{\text{inc}}, ..., \beta_M^{\text{inc}})$ denotes the covariance matrix of non-line-of-sight (NLoS) path containing the pathloss coefficients from BS to the RIS. Additionally, the line-of-sight (LoS) channel $\overline{\mathbf{H}}_{\text{LoS}}$ is defined as $\overline{\mathbf{H}}_{\text{LoS}} = \sqrt{\beta^{\text{inc}}}\mathbf{a}^{\text{H}}(\phi_1^a, \phi_1^e, N_x, N_y) \times \mathbf{a}(\phi_2^a, \phi_2^e, M_x, M_y)$ where $\phi_1^{a/e}$ denote the azimuth (elevation) angle of a row (column) of the UPA at the RIS and the projection of the transmit signal from BS to the RIS on the plane of the UPA at the RIS. Similarly, $\phi_2^{a/e}$ shows the azimuth (elevation) angle between the direction of a row (column) of the UPA at BS and the projection of the signal from BS to the RIS on the plane of the UPA at BS. In addition [18]

$$\mathbf{a}(x, y, N_1, N_2) \coloneqq \text{rvec}\left(\left(e^{\text{j}\mathcal{G}(x,y,n_1,n_2)}\right)_{\substack{n_1=1,2,...,N_1,\\ n_2=1,2,...,N_2}}\right), \tag{2}$$

where rvec$(\cdot)$ denotes the row vectorization of a matrix and

$$\mathcal{G}(x, y, n_1, n_2) = 2\pi\frac{d}{\lambda}\left[(n_1 - 1)\cos x + (n_2 - 1)\sin x\right]\sin y,$$

in which λ is the operating wavelength, and d is the antenna/element spacing. Similarly, the channel between RIS and AC k as $\mathbf{h}_k^{\text{RIS}} = \sqrt{\frac{\zeta_k^{\text{RIS}}}{\zeta_k^{\text{RIS}}+1}}\overline{\mathbf{h}}_k^{\text{RIS}} + \sqrt{\frac{1}{\zeta_k^{\text{RIS}}+1}}\tilde{\mathbf{h}}_k^{\text{RIS}}$ where the Rician parameter ζ_k^{RIS} controls the proportion of LoS to the none-LoS power in AC k. The NLoS channel is distributed as $\tilde{\mathbf{h}}_k^{\text{RIS}} \sim \mathcal{CN}(\mathbf{0}_{N\times 1}, \boldsymbol{\beta}_k^{\text{RIS}})$ and $\boldsymbol{\beta}_k^{\text{RIS}} = \text{diag}(\beta_1^{\text{RIS}_k}, ..., \beta_N^{\text{RIS}_k})$ is the covariance matrix containing the pathloss coefficients from RIS to AC k. The LoS channel $\overline{\mathbf{h}}_k^{\text{RIS}} \in \mathbb{C}^{N\times 1}$ is modeled by $\overline{\mathbf{h}}_k^{\text{RIS}} = \sqrt{\beta_k^{\text{RIS}}}\mathbf{a}(\phi_3^{a,k}, \phi_3^{e,k}, N_x, N_y)$ for $\forall k$ in which $\phi_3^{a,k}, \phi_3^{e,k}$ are the azimuth/elevation angles between RIS and the AC k.

The complex reconfiguration matrix $\boldsymbol{\Theta}_{N\times N}$ is denoted by

$$\boldsymbol{\Theta}_{N\times N} = \text{diag}(\beta_1 e^{j\theta_1}, \beta_2 e^{j\theta_2}, ..., \beta_N e^{j\theta_N}), \ \ \forall n \in \mathcal{N} \tag{3}$$

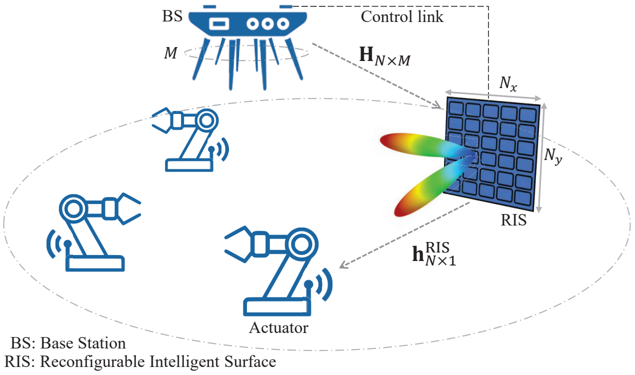
BS
Control link
M
$\mathbf{H}_{N\times M}$
N_x
N_y
RIS
$\mathbf{h}^{\mathrm{RIS}}_{N\times 1}$
Actuator
BS: Base Station
RIS: Reconfigurable Intelligent Surface

Fig. 1: The considered system model.

where $\theta_n \in [-\pi, \pi)$, $\beta_n \in [0,1]$, and $\mathcal{N} = \{1,2,...,N\}$. Note that in our model we have assumed that the RIS elements have no coupling or there is no joint processing among elements [6]. Furthermore, the phase shifts of the RIS are picked up from

$$\hat{\theta}_n = \mathcal{Q}(\theta_n) \in \Phi = \{-\pi, -\pi+\Delta, ..., -\pi+(Q-1)\Delta\}, \quad \forall n \in \mathcal{N} \tag{4}$$

where $\mathcal{Q}(\cdot)$ denotes the quantizer system response and $Q = 2^b$ is the number of quantization levels, b denotes the number of bits assigned to a discrete and quantized phase and $\Delta = \frac{\pi}{2^{b-1}}$ is the quantization step. Through this process, a phase error e distributed uniformly over $\frac{-\Delta}{2} \le e \le \frac{\Delta}{2}$ appears which deteriorates the system performance. In practical RIS, phase shifters have phase-dependent amplitude response expressed as [19]

$$\beta_n(\theta_n) = (1-\beta_{\min})\left(\frac{\sin(\theta_n - \phi)+1}{2}\right)^{\alpha} + \beta_{\min}, \tag{5}$$

where $\beta_{\min} \ge 0$ (minimum amplitude), $\alpha \ge 0$ (the steepness) and $\phi \ge 0$ (the horizontal distance between $-\frac{\pi}{2}$ and $\beta_{\min}$) are circuit implementation parameters. Note that, $\beta_{\min} = 1$ results in an ideal phase shifter.

For the considered system model, the received signal at the AC k is

$$\begin{aligned} y_k[t] =& \overbrace{\left(\mathbf{h}_k^{\mathrm{RIS}^{\mathrm{H}}} \mathbf{\Theta}\mathbf{H}\right)\mathbf{s}_k x_k[t]}^{\text{Actuator } k \text{ signal}} \\ &\underbrace{+\left(\mathbf{h}_k^{\mathrm{RIS}^{\mathrm{H}}} \mathbf{\Theta}\mathbf{H}\right)\sum_{i=1, i\neq k}^{K} \mathbf{s}_i x_i[t] + n_k[t],}_{\text{Interference plus noise}} \end{aligned} \tag{6}$$

where $\mathbf{s}_k$ is the beamforming vector applied at the transmitter to the symbol $x_k \sim \mathcal{CN}(0,1)$ of AC k with $||\mathbf{s}_k||_2^2 = p_k$ in which p_k is the transmit power allocated for AC k, and $n[t]$ is the additive white Gaussian noise with $\mathbb{E}[|n[t]|^2] = N_0 W = \sigma^2$ where N_0, W are the noise spectral density and the system bandwidth, respectively.

Based on the received signal at actuator k, linear minimum mean square error (MMSE) signal detection is performed to maximize the output SINR. Thus, the linear receiver is represented by vector $\mathbf{s}_k^{\mathrm{MMSE}} = \mathbf{R}_k^{-1}\mathbf{h}_k(t)$ where $\mathbf{R}_k = \sigma^2\mathbf{I}_M + \sum_{i=1, i\neq k}^{K} p_i \mathbf{h}_i(t)\mathbf{h}_i^{\mathrm{H}}(t)$ is the covariance matrix of the interference plus noise signal. By substituting $\mathbf{s}_k^{\mathrm{MMSE}}$ to the received signal in (6), the corresponding SINR achieved at time instance t is simplified as

$$\mathrm{SINR}_k = p_k \mathbf{h}_k^{\mathrm{H}}(t)\left(\sigma^2\mathbf{I}_M + \sum_{i=1, i\neq k}^{K} p_i \mathbf{h}_i(t)\mathbf{h}_i^{\mathrm{H}}(t)\right)^{-1}\mathbf{h}_k(t), \tag{7}$$

where $\mathbf{h}_k(t) = \mathbf{H}^{\mathrm{H}}(t)\mathbf{\Theta}^{\mathrm{H}}(t)\mathbf{h}_k^{\mathrm{RIS}}(t)$ and $\mathbf{\Theta}(t) \in \mathbb{C}^{N\times N}$ denotes the reconfiguration phase matrix and t is the time index.

In practice, the channel estimates become outdated after a delay time T_d which results in imperfect CSI. To model the channel variations, delayed feedback is expressed as $\mathbf{h}_k(t+T_d) = \rho_d \mathbf{h}_k(t) + \mathbf{e}_k(t)$ where T_d is the feedback delay, $\mathbf{e}_k(t) \sim \mathcal{CN}(\mathbf{0}, \sqrt{1-\rho_d^2}\mathbf{I}_M)$ and $\rho_d = \frac{\mathbb{E}[\mathbf{h}_k^{\mathrm{H}}(t)\mathbf{h}_k(t+T_d)]}{\mathbb{E}[\mathbf{h}_k^{\mathrm{H}}(t)\mathbf{h}_k(t)]}$ is the normalized correlation coefficient between the current and the outdated channel response [12]. According to Clarke's fading spectrum model for band-limited channels $\rho_d = \mathrm{J}_0(2\pi T_s f_d T_d)$[1] in which f_d is Doppler frequency and T_s is the symbol block duration.

Based on the received SINR at AC k the number of information bits that can be transmitted through m_k channel uses over a quasi-static additive white Gaussian channel (AWGN) is given by [5]

$$L_k = \mathcal{V}_k(\mathbf{\Theta}) - Q^{-1}(\varepsilon_k)\mathcal{W}_k(\mathbf{\Theta}) + \mathcal{O}\left(\log_2(m_k)\right), \tag{8}$$

where $\mathcal{V}_k(\mathbf{\Theta}) = m_k \mathrm{C}(\mathrm{SINR}_k)$ in which $\mathrm{C}(x) = \log_2(1+x)$ is the Shannon capacity term defined in infinite blocklength regime, $\mathcal{W}_k(\mathbf{\Theta}) = \sqrt{m_k \mathrm{V}(\mathrm{SINR}_k)}$ where the channel dispersion is defined as $\mathrm{V}(x) = \frac{1}{(\ln 2)^2}\left(1 - \frac{1}{(1+x)^2}\right)$, and ε_k is the target error probability for AC k. Also, $Q^{-1}(.)$ is the inverse of Q-function[2]. Also, note that from (8) when the blocklength m_k asymptotically goes infinity the achievable rate will be simplified to the conventional Shannon capacity formula. By solving (8) in order to find the decoding error probability ε_k at the AC k it yields $\varepsilon_k = Q\left(f(\mathrm{SINR}_k)\right)$ such that

$$f(\mathrm{SINR}_k) = \sqrt{\frac{m_k}{\mathrm{V}(\mathrm{SINR}_k)}}\left(\log_2(1+\mathrm{SINR}_k) - \frac{L}{m_k}\right). \tag{9}$$

B. Problem Formulation

The optimization problem for each transmission time to optimize total FBL rate of the ACs by configuring the phase matrix of the RIS is formulated as:

$$\begin{aligned} \textbf{P1} \quad & \max_{\mathbf{\Theta}} \; L_{\mathrm{tot}}(\mathbf{\Theta}) = \sum_{k=1}^{K}\left[\mathcal{V}_k(\mathbf{\Theta}) - Q^{-1}(\varepsilon_k^{\mathrm{th}})\mathcal{W}_k(\mathbf{\Theta})\right] \\ \textbf{s.t.} \quad & \mathrm{C}_1\colon \theta_n \in \Phi, \quad \forall n, \\ & \mathrm{C}_2\colon \beta_n(\theta_n) = (1-\beta_{\min})\left(\frac{\sin(\theta_n - \phi)+1}{2}\right)^{\alpha} + \beta_{\min}, \forall n, \end{aligned}$$

---

[1] $\mathrm{J}_0(\cdot)$ is the zeroth order of the Bessel function of the first kind.

[2] As usual $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-\nu^2/2} d\nu$.

where the objective is to maximize the total number of information bits across all ACs in FBL regime while ensuring the block error probability at a target value $\varepsilon_k^{\text{th}}\ \forall k \in \{1,2,...,K\}$, and the variables are the reflective phase shift matrix $\boldsymbol{\Theta}$ at the RIS. In **P1** the constraint C_1 denotes that the phase adjustment variable is selected from the discrete set Φ given in (4), and C_2 implies the practical phase shift model which affects the amplitude response at the RIS. **P1** belongs to a class of mixed-integer nonlinear combinatorial problems that are difficult to solve with optimization methods. It is rational to use DRL for such problems since in DRL, the solution to the problem is the output of the forward pass to the neural network, which is a computationally simple process since it is often a set of simple operations. Further, the training of the neural networks that is done in different steps is performed in the background. Once the training is completed, the neural networks are updated. Therefore, the process to find the optimized phase shifts in our problems is the inference of the neural networks that can be done in real-time [10]. Such a real-time solution cannot be obtained using optimization methods. Consequently, we employ a model-free DRL algorithm based on the TD3 algorithm described in the following section.

III. DRL-BASED FORMULATION

The goal of the agent in reinforcement learning (RL) is to *learn* to find an optimal policy that maps states to actions based on its interaction with the environment so that the accumulated discounted reward function over a long time is maximized. A state contains all useful information from the sequence of the observations, actions and rewards. These kind of problems are tackled by representing them as Markov decision processes (MDP). An MDP is characterized by $(\mathcal{S},\mathcal{A},\mathcal{R},\mathcal{P}_{s\to s'})$ in which $\mathcal{S}$ is the set of environment states, $\mathcal{A}$ denotes the set of possible actions that is defined in terms of the RIS phase shift values, $\mathcal{R}$ is the reward function, and $\mathcal{P}_{s\to s'}$ is the transition probabilities from state s to s', $\forall s, s' \in \mathcal{S}$. Mathematically, a Markov property means that the probability of next state (future state) is independent of the past given the present state. In RL algorithms, the environment can be fully observable where the agent directly observes the environment or partially observable [20]. The aim of the agent is to find an optimal policy to maximize the accumulated and discounted reward function over time-steps, i.e., to find π^* in which the set of states $\mathcal{S}$ is mapped into the set of actions $\mathcal{A}$ as $\pi^*: \mathcal{S} \to \mathcal{A}$. The optimal policy π^* maximizes the action-value function defined as $Q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_{t+k+1}|S_t = s, A_t = a\right]$ where the variable $0 \le \gamma \le 1$ is the discount factor to uncertainty of future rewards and r_i is the acquired reward in step i.

One of the efficient model-free and off-policy actor-critic methods that deals with the continuous action-space is DDPG [20]. In this algorithm, four deep neural networks (DNNs) are employed, two of them are for actor-critic networks and the other two are called target networks. The actor network directly gives the action by importing the states through a DNN with parameter set $\boldsymbol{\xi}^{\text{act}}$, i.e., $a = \mu(s;\boldsymbol{\xi}^{\text{act}})$ where $\mu(\cdot)$ denotes the deterministic policy meaning that the output is a value instead of a distribution. The critic network that has a DNN with $\boldsymbol{\xi}^{\text{crit}}$ weights evaluates the action-value function based on the action given by the policy network and the current state. The other two networks which are named as target networks give the target action-values in order to minimize the mean-squared Bellman error (MSBE) which is defined as

$$\mathcal{L}(\boldsymbol{\xi}^{\text{crit}},\mathcal{B}) = \mathbb{E}\left[\left(Q(s,a;\boldsymbol{\xi}^{\text{crit}}) - \overbrace{(r+\gamma\max_{a'}Q(s',a';\boldsymbol{\xi}^{\text{crit}}))}^{\text{target value}}\right)^2\right], \tag{10}$$

where $\mathcal{B}$ is the experience replay memory which has stored the set of states, actions, rewards and the next states as a tuple (s,a,r,s') over previous steps. From (10) the next optimal action a' is calculated by the target actor network with parameter set $\boldsymbol{\xi}^{\text{targ-act}}$ where $a' = \mu(s';\boldsymbol{\xi}^{\text{targ-act}})$ and the corresponding action-value $Q(s',a';\boldsymbol{\xi}^{\text{targ-crit}})$ is then evaluated using the target critic network with weights $\boldsymbol{\xi}^{\text{targ-crit}}$. The two networks weights are usually just copied over from the main network by polyak averaging which is

$$\boldsymbol{\xi}^{\text{targ-act}} \leftarrow \tau\boldsymbol{\xi}^{\text{act}} + (1-\tau)\boldsymbol{\xi}^{\text{targ-act}}, \tag{11}$$

$$\boldsymbol{\xi}^{\text{targ-crit}} \leftarrow \tau\boldsymbol{\xi}^{\text{crit}} + (1-\tau)\boldsymbol{\xi}^{\text{targ-crit}}, \tag{12}$$

where $\tau << 1$ is the soft update hyperparameter used to control the updating procedure.

Before proceeding with TD3 method, we restate the following Lemma from [17]:

Lemma 1. *For the true underlying action-value function which is not known during the learning process, i.e.,* $Q_\pi(s,a)$ *and the estimated* $Q(s,a;\boldsymbol{\xi}^{crit})$ *the following inequality holds*

$$\mathbb{E}\left[Q\left(s, a=\mu(s;\boldsymbol{\xi}^{act});\boldsymbol{\xi}^{crit}\right)\right] \ge \mathbb{E}\left[Q_\pi\left(s, a=\mu(s;\boldsymbol{\xi}^{act})\right)\right],$$

Based on Lemma 1, as the DDPG algorithm leverages the typical Q-learning methods, it overestimates the Q-values during the training which propagates throughout the next states and episodes. This resultant deteriorates the policy network which utilizes the Q-values to update its weights and hyperparameters and results in poor policy updates. As seen in Algorithm 1, the TD3 introduces the followings to address the challenges [17]:

- TD3 recruits two DNNs for estimating the action-value function in the Bellman equation, then the minimum value of the output of Q-values is used in the (10).
- In this method, the target and policy networks are being updated less frequently than critic networks.
- A regularization of the actions that can incur high peaks and failure to the Q-value in DDPG method is leveraged so that the policy network will not try these actions in the next states. Therefore, the action will be chosen based on adding a small amount of clipped random noise to the selected action as given by

$$a' = \text{clip}(\mu(s';\boldsymbol{\xi}^{\text{targ-act}}) + \text{clip}(\kappa, -c, +c), a_{\text{Low}}, a_{\text{High}}), \tag{13}$$

Algorithm 1: Twin Delayed DDPG (TD3) Algorithm

```
Initialization: Initial values for $\boldsymbol{\xi}^{\text{act}}$, $\boldsymbol{\xi}_1^{\text{crit}}$ and $\boldsymbol{\xi}_2^{\text{crit}}$, $\mathtt{E}, \mathtt{T}$, $d$,
  empty replay memory $\mathcal{B}$. Let $\boldsymbol{\xi}^{\text{targ-act}} \leftarrow \boldsymbol{\xi}^{\text{act}}$, $\boldsymbol{\xi}_i^{\text{targ-crit}} \leftarrow \boldsymbol{\xi}_i^{\text{crit}}$
  $i \in \{1,2\}$, and policy update iteration $d$;
for $e = 1, 2, ..., \mathtt{E}$ do
    Randomly initialize the phase matrix at the RIS;
    for $t = 1, 2, ..., \mathtt{T}$ do
        Collect current CSI $\{\mathbf{h}_1(t), \mathbf{h}_2(t), ..., \mathbf{h}_K(t)\}$;
        Select action $a = \text{clip}(\mu(s; \boldsymbol{\xi}^{\text{act}}) + \kappa, a_{\text{Low}}, a_{\text{High}})$,
          where $\kappa \sim \mathcal{N}(0, \sigma^2)$;
        Observe next state $s'$ and the reward value $r$;
        Store the tuple $(s, a, s', r)$ in $\mathcal{B}$;
        Sample mini-batch from replay buffer $\mathbb{B} \subset \mathcal{B}$;
        Compute target actions from (13);
        Compute $\mathtt{tt} = r + \gamma \min_{i \in \{1,2\}} Q(s', a'; \boldsymbol{\xi}_i^{\text{targ-crit}})$ and
          update the critic networks by performing gradient
          descent for $i \in \{1,2\}$ by computing
          $\frac{1}{|\mathbb{B}|} \nabla_{\boldsymbol{\xi}_i^{\text{crit}}} \sum_{(s,a,r) \in \mathbb{B}} \left( Q(s, a; \boldsymbol{\xi}_i^{\text{crit}}) - \mathtt{tt} \right)^2$;
        if $t$ mod $d$ then
            Compute $\frac{1}{|\mathbb{B}|} \nabla_{\boldsymbol{\xi}^{\text{act}}} \sum_{s \in \mathbb{B}} Q(s, \mu(s; \boldsymbol{\xi}^{\text{act}}); \boldsymbol{\xi}_i^{\text{crit}})$ and
              update the policy network;
            Update the target networks via (11), (12);
        end
    end
end
Output: Trained agent with DNNs' weights.
```

where $\kappa \sim \mathcal{N}(0, \tilde{\sigma}^2)$ is the added normal Gaussian noise and $a_{\text{Low}} = -\pi$, $a_{\text{High}} = +\pi$ are the lower and upper limit value for the selected action at the RIS elements that is clipped to ensure a feasible action due to added noise. Also, the constant c truncates the added noise at first stage to keep the target close to the original action.

A preliminary step to solve the problem **P1** with TD3 is to map the components and properly define the algorithm states, actions and the reward function as follows:

1) States: The agent interacts with the environment to optimize the FBL rate performance while ensuring a target block error probability. Hence, the agent only has knowledge about the local information, e.g., the channels' response. Consequently, the DRL agent state space is defined as the aggregation of the real and imaginary parts of the total channel coefficients ($\mathbf{h}_k$, $\forall k$), thus, we have $s_t = \bigcup_{k=1}^{K} \left\{ \mathbf{h}_k^{\text{real}}(t), \mathbf{h}_k^{\text{imag}}(t) \right\}$.

2) Actions: The action is determined as the value of phase shift of each element ($\theta_n(t)$, $\forall n$) and the action space set is given by $a_t = \bigcup_{n=1}^{N} \{\theta_n(t)\}$ such that each element value is chosen in the interval $\theta_n(t) \in [-\pi, \pi]$, $\forall n$.

3) Reward Function: The objective function in **P1** is considered as the step-reward function which is to be maximized over time-steps t, i.e., $r_t = L_{\text{tot}}(\boldsymbol{\Theta}(t))$.

IV. NUMERICAL RESULTS

In this section, we numerically evaluate the proposed phase shift optimization problem by using the proposed DRL method. Table I shows the considered parameters selected for the network. Since, the components and robots in an FA environment are in almost fixed position, we considered three ACs in a factory environment located in 2D-plane coordinates as at $[60, 30]$, $[30, 60]$ and $[15, 45]$ where a BS is positioned at $[0, 0]$ and the RIS is located in the edge side of the factory at $[75, 75]$. The large scale path loss for the reflected channels is modeled as $\text{PL(dB)}_{\text{ref}} = -30 - 22\log_{10}(D[\text{m}])$ where D is the distance between the transmitter and the receiver.

TABLE I: Simulation parameters.

Parameter	Default value
BS transmit power (p)	20 mW
Number of ACs (K)	3
Number of BS antennas (M)	4
Number of RIS elements (N)	25
Rician factors (ζ and ζ_k^{RIS} $\forall k$)	10
Target error probability ($\varepsilon_k^{\text{th}}, \forall k$)	10^{-7}
$f_d T_d T_s$	0.1
Outdated CSI coefficient (ρ_d)	0.9
Noise power density (N_0)	-174 dBm/Hz
Channel blocklength (M)	60
Bandwidth (W)	1 MHz
Carrier frequency	1900 MHz
	$\beta_{\min} = 0.6$
RIS phase shifter parameters	$\alpha = 1.6$
	$\phi = 0.43\pi$

The learning rate in actor networks of TD3 agent is set to 10^{-4} and for the critic networks is 10^{-3}. The activation functions in all hidden layers are considered as $ReLU(\cdot)$ except for the last layer in which for the actor network is assumed to be $tanh(\cdot)$ to provide better gradient. The experience replay buffer capacity is 5000 with batch size 32 such that the samples are uniformly selected from the buffer data. The number of steps in an episode is set to 200. Furthermore, the exploration noise in TD3 actor networks which is a zero-mean normal random variable has $0.1 \times \pi$ variance. The target actor/critic networks' soft update coefficient is $\tau = 0.005$. During the updating procedure, the policy network is being updated every two steps.

In Fig. 2 the average FBL rate for continuous phase shift control is illustrated for TD3 and DDPG methods. The curves plot the mean and standard deviation of the FBL rate throughout episodes while filling the space between the positive and negative mean error using a semi-transparent background. As observed, the TD3 has fewer fluctuations in terms of the FBL rate compared to the DDPG algorithm which is shown as shaded region around the curves. In addition, TD3 outperforms DDPG method in both final performance and learning speed in phase control. This observation highlights the applicability of proposed TD3 algorithm in phase matrix optimization of RIS-aided networks toward realizing reliable and robust wireless links over short packet communications.

In Fig. 3 the network sum rate is assessed in terms of increasing the total number of reflective elements at the RIS. A gap is also observed between the upperbound performance and lowerbound case which demonstrates that the system actual performance will lie between these two curves. Another result from these curves, is that the total achievable rate in all cases increases with the number of RIS elements, i.e., with/without ideal/non-ideal RIS or with quantized phase shifts. The similar

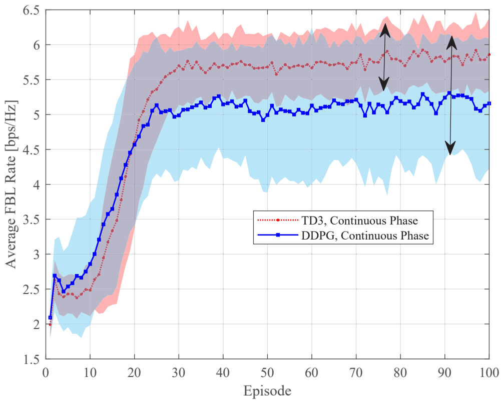

Fig. 2: Comparison of TD3 and DDPG convergence for similar DNN configurations. The fluctuations in DDPG method occurred due to frequent policy network updates.

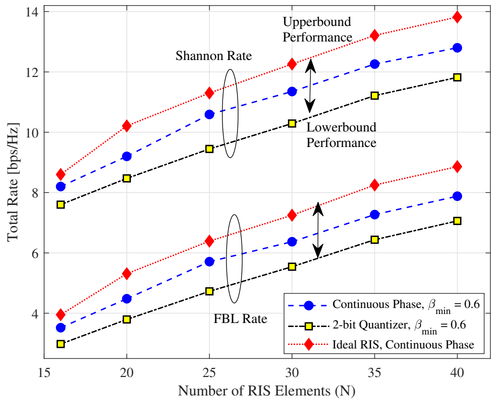

Fig. 3: The effect of increasing number of the RIS elements on the total achievable rate of the system.

performance is also shown in FBL and Shannon rates. On the other hand, the slopes of the curves are quite similar when the number of RIS elements start to increase which additionally shows the practicality of the proposed TD3 algorithm in ideal/non-ideal reflective phase shift design problems.

V. CONCLUSION

In this paper, we have studied the reflective phase shift design problem by a novel and efficient DRL algorithm in RIS-aided URLLC systems over short packet communications. First, the problem framework with the objective of maximizing total FBL rate of ACs in a factory environment has been proposed where the constraints are the discrete selection of each phase value due to quantization process. Moreover, the channel coefficients' uncertainty due to transmission and processing delay which leads to feedback delay has been taken into account in the proposed formulations. Since the proposed problem is challenging to solve via optimization-based algorithms that are usually computationally inefficient, we have introduced a policy gradient DRL framework based on unsupervised actor-critic methods to optimize the phase shifts which concurrently learns a Q-function and a policy. The utilized DRL method, i.e., TD3 has addressed the issues in the conventional DDPG method that dramatically overestimate action-value function, which then leads to the policy breaking. The numerical results demonstrate the applicability of the proposed DRL method in RIS phase shift design problems in short packet communications underlying URLLC systems.

REFERENCES

[1] N. H. Mahmood, S. Böcker *et al.*, *White paper on critical and massive machine type communication towards 6G*, ser. 6G Research Visions, nr. 11, N. H. Mahmood, O. Lopez *et al.*, Eds. Oulu, Finland: University of Oulu, Jun. 2020.

[2] G. Aceto, V. Persico *et al.*, "A survey on information and communication technologies for Industry 4.0: State-of-the-art, taxonomies, perspectives, and challenges," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 4, pp. 3467–3501, Oct. 2019.

[3] N. H. Mahmood, O. A. Lopez *et al.*, "A predictive interference management algorithm for URLLC in beyond 5G networks," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 995–999, Mar. 2021.

[4] P. Popovski, C. Stefanovic *et al.*, "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, May 2019.

[5] Y. Polyanskiy, H. V. Poor *et al.*, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[6] M. Di Renzo, A. Zappone *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Apr. 2020.

[7] R. Hashemi, S. Ali *et al.*, "Average rate and error probability analysis in short packet communications over RIS-aided URLLC systems," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 10 320–10 334, Oct. 2021.

[8] H. Yang, Z. Xiong *et al.*, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[9] X. Liu, Y. Liu *et al.*, "Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2042–2055, 2021.

[10] K. Feng, Q. Wang *et al.*, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, May 2020.

[11] C. Huang, Z. Yang *et al.*, "Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1663–1677, Jun. 2021.

[12] X. Guo, Y. Chen *et al.*, "Learning-based robust and secure transmission for reconfigurable intelligent surface aided millimeter wave UAV communications," *IEEE Wireless Commun. Lett.*, vol. 10, no. 8, pp. 1795–1799, 2021.

[13] W. R. Ghanem, V. Jamali *et al.*, "Joint Beamforming and Phase Shift Optimization for Multicell IRS-aided OFDMA-URLLC Systems," in *Proc. IEEE Wireless Commun. and Networking Conf.*, 2021, pp. 1–7.

[14] A. Ranjha and G. Kaddoum, "URLLC Facilitated by Mobile UAV Relay and RIS: A Joint Design of Passive Beamforming, Blocklength and UAV Positioning," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4618–4627, 2021.

[15] H. Xie, J. Xu *et al.*, "User grouping and reflective beamforming for IRS-aided URLLC," *IEEE Wireless Commun. Lett.*, vol. 10, no. 11, pp. 2533–2537, 2021.

[16] M. Almekhlafi, M. A. Arfaoui *et al.*, "Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1304–1319, 2022.

[17] S. Fujimoto, H. Hoof *et al.*, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1587–1596.

[18] Y. Jia, C. Ye *et al.*, "Analysis and optimization of an intelligent reflecting surface-assisted system with interference," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8068–8082, 2020.

[19] S. Abeywickrama, R. Zhang *et al.*, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, Sep. 2020.

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