Deep Reinforcement Learning-Based Adaptive IRS Control with Limited Feedback Codebooks

Junghoon Kim*, Seyyedali Hosseinalipour*, Andrew C. Marcum†, Taejoon Kim‡,
David J. Love* and Christopher G. Brinton*
*Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
†Raytheon BBN Technologies, Cambridge, MA, USA
‡Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA
*{kim3220, hosseina, djlove, cgb}@purdue.edu, †andrew.marcum@raytheon.com, ‡taejoonkim@ku.edu

***Abstract*—Intelligent reflecting surfaces (IRS) consist of configurable meta-atoms, which can alter the wireless propagation environment through design of their reflection coefficients. We consider adaptive IRS control in the practical setting where (i) the IRS reflection coefficients are attained by adjusting *tunable elements* embedded in the meta-atoms, (ii) the IRS reflection coefficients are affected by the *incident angles* of the incoming signals, (iii) the IRS is deployed in multi-path, time-varying channels, and (iv) the feedback link from the base station (BS) to the IRS has a low data rate. Conventional optimization-based IRS control protocols, which rely on channel estimation and conveying the optimized variables to the IRS, are not practical in this setting due to the difficulty of channel estimation and the low data rate of the feedback channel. To address these challenges, we develop a novel adaptive codebook-based limited feedback protocol to control the IRS. We propose two solutions for adaptive IRS codebook design: (i) *random adjacency (RA)*, which utilizes correlations across the channel realizations, and (ii) *deep neural network policy-based IRS control (DPIC)*, which is based on a deep reinforcement learning. Numerical evaluations show that the data rate and average data rate over one coherence time are improved substantially by the proposed schemes.**

I. Introduction

The intelligent reflecting surface (IRS) is a technology for 6G-and-beyond [2], [3]. An IRS is a software-controlled meta-surface, consisting of configurable meta-atoms with flexible reflection coefficients. By fine-tuning these meta-atoms, the IRS can change the wireless propagation environment, resulting in power savings, throughput increase, etc. Compared to a traditional antenna array with radio frequency (RF) chains for active relaying/beamforming, an IRS is made of low cost meta-surfaces, which consume low energy for tuning [4]. These benefits have motivated research on utilizing IRS in communications/signal processing literature. We aim to address three shortcomings of the current art, as discussed below.

A. Shortcomings of Existing Works and Motivations

1) Dependency between Meta-Atoms' Reflection Phase Shift and Attenuation: Much of the existing work on IRS reflection coefficient design for communications has treated the reflection phase and attenuation of each meta atom independently [5]. In reality, the phase shift and attenuation are *interdependent* because the reflection behavior is determined by adjusting the *tunable elements* inside the meta-atoms, i.e., their controllable capacitance, as revealed in physics literature [6]. This inter-dependency has only been considered in a few works in the communications area [7].

2) Dependency between Meta-atoms' Reflection Coefficient and Incident Angle of Incoming Signals: Another practical consideration neglected in existing works is the dependency between the IRS reflection behavior and the *incident angles* of the electromagnetic (EM) waves [8], [9]. In [9], the authors propose an angle-dependent reflection coefficient model for each meta-atom using an equivalent circuit model. To the best of our knowledge, the angle-dependent property of the IRS reflection coefficient has not been incorporated into uplink/downlink signal transmission models for wireless communication systems.

3) Low Overhead Feedback Channel: The configuration of IRS meta-atoms is usually controlled via reception of some information from the base station (BS) through a *feedback* link. This link typically has a low data rate since its channel state information is unknown to the BS [4]. To reduce the feedback overhead for IRS control, some recent works have considered codebook structures [8], [10], where the codebook refers to a set of IRS reflection coefficients. In these works, the BS feeds back a specific codeword index to the IRS, using which the IRS recovers the desired reflection coefficients from the codebook. In doing so, these works directly design the IRS *reflection coefficients* and thus do not consider the practical IRS reflection behavior mentioned in Sec. I-A1, I-A2.

B. Overview of Methodology and Contributions

We propose a new methodology for IRS control that explicitly considers the above three practical design aspects. In doing so, we consider that the IRS is deployed in realistic multi-path, time-varying channels. These considerations render current optimization-based methods for IRS control [5], [7], which rely on channel estimation, impractical: it is difficult

A more comprehensive version of this paper is under review in IEEE Transactions on Wireless Communications [1]. C. G. Brinton and D. J. Love were supported in part by the National Spectrum Consortium (NSC) under Grant W15QKN-15-9-1004 and Office of Naval Research (ONR) under Grant N00014-21-1-2472. T. Kim was supported in part by the National Science Foundation (NSF) under Grants CNS1955561.

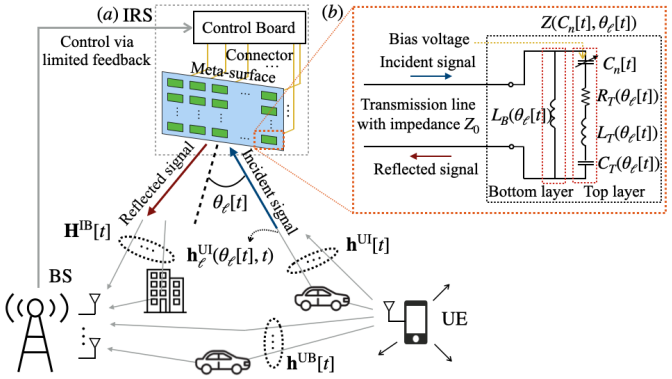


Fig. 1: An uplink point-to-point communication system consisting of a UE, IRS, and BS, where the IRS is controlled by the BS via a limited feedback link. (a) IRS as two interconnected systems: meta-surface and control board. (b) Equivalent circuit model of each meta-atom.

to measure the incident angles of incoming signals at the IRS since the IRS typically does not have active sensors.

Specifically, we propose a novel *adaptive codebook-based limited feedback protocol.* We directly design the meta-atom *capacitance values*, instead of their reflection coefficients as in existing methods. With the codebook as a set of capacitance values for the meta-atoms, we develop two *adaptive* codebook design methods: (i) *random adjacency (RA)* and (ii) *deep neural network policy-based IRS control (DPIC)*. These approaches only require the end-to-end channel from the user equipment (UE) to the BS, which can be readily measured in real-time.

II. System Model for IRS-assisted Communications

We begin by formalizing IRS meta-atom reflection behavior in Sec. II-A. Then, we describe the signal model of IRS-assisted uplink communications in Sec. II-B.

A. Reflection Behavior of IRS Meta-atoms

An IRS is composed of two interconnected systems as shown in Fig. 1(a): a meta-surface and a control board. A meta-surface is an ultra-thin sheet composed of periodic sub-wavelength metal/dielectric structures, i.e., meta-atoms. Each meta-atom generally contains a semiconductor device, i.e., the tunable element, such as a positive-intrinsic-negative (PIN) diode and a variable capacitor (varactor) [6]. A control board, e.g., a field programmable gate array (FPGA) [11], adjusts the bias voltage applied to the semiconductor in each meta-atom and changes its capacitance. Given a range of potential bias voltage values, the capacitance $C_n[t]$ at meta-atom n at time t satisfies

$$C_{\min} \leq C_n[t] \leq C_{\max}, \tag{1}$$

where $C_{\min}$ and $C_{\max}$ depend on the semiconductor used.

Through tuning the capacitance of the meta-atoms, their impedance can be adjusted. However, the impedance is also dependent on the incident angle of the incoming EM wave [8], [9]. Both of these factors should be considered in IRS reflection behavior design. As an example, we provide the impedance and reflection coefficient of a meta-atom equipped with a *varactor* using its equivalent circuit model [9] depicted in Fig. 1(b). Let $\theta_\ell[t]$ denote the incident angle of the ℓ-th channel path to the IRS.[1] Under a far-field assumption where $\theta_\ell[t]$ is the same across all the meta-atoms, the impedance of meta-atom n can be described as [9]

$$Z(C_n[t], \theta_\ell[t]) = \frac{j\omega L_B(\theta_\ell[t])\big(R_T(\theta_\ell[t]) + j\omega L_T(\theta_\ell[t]) + \frac{1}{j\omega C_T(\theta_\ell[t])} + \frac{1}{j\omega C_n[t]}\big)}{j\omega L_B(\theta_\ell[t]) + R_T(\theta_\ell[t]) + j\omega L_T(\theta_\ell[t]) + \frac{1}{j\omega C_T(\theta_\ell[t])} + \frac{1}{j\omega C_n[t]}}, \tag{2}$$

where $L_T(\cdot)$, $C_T(\cdot)$, and $R_T(\cdot)$ are the inductance, capacitance, and resistance of the top circuit layer in Fig. 1(b), respectively, $L_B(\cdot)$ is the bottom layer inductance, $C_n[t]$ is the variable capacitance, and ω is the angular frequency of the EM waves.

Considering the impedance discontinuity between the free space impedance $Z_0 \approx 376.73\,\Omega$ and the meta-atom impedance $Z(C_n[t], \theta_\ell[t])$, the reflection coefficient[2] of meta-atom n is [9]

$$\Gamma(C_n[t], \theta_\ell[t]) = \frac{Z(C_n[t], \theta_\ell[t]) - Z_0}{Z(C_n[t], \theta_\ell[t]) + Z_0}. \tag{3}$$

The expressions in (2)&(3) reveal two practical considerations for tuning the meta-atoms. First, the reflection attenuation $|\Gamma(C_n[t], \theta_\ell[t])|$ and phase shift $\angle\Gamma(C_n[t], \theta_\ell[t])$ are jointly controlled by the variable capacitance $C_n[t]$, as also reported in [7]. Thus, it is beneficial to design the variable capacitance instead of the reflection coefficient since some combinations of attenuation and phase shifts may not be feasible. Second, the reflection coefficient is a function of the incident angle $\theta_\ell[t]$, posing new challenges for applications of IRS in multi-path and time-varying channels, which will be discussed in Sec. III-A. While observed in [8], [9], this dependency has not yet been incorporated in the canonical signal model for IRS-assisted communications. In this paper, we incorporate these practical considerations into our signal model and methodology.

B. Signal Model for IRS-assisted Uplink Communications

We consider IRS-assisted uplink communications with a UE, a BS, and an IRS, shown in Fig. 1. The UE posesses a single antenna, while the BS has N_{BS} antennas. We assume a block fading channel model with time index $t = 0, 1, ...$, where channels are constant during each block. Let N_{IRS} denote the number of IRS meta-atoms. We define the *capacitance vector* across the meta-atoms at time t as

$$\mathbf{c}[t] = \big[C_1[t], ..., C_{N_{\text{IRS}}}[t]\big] \in \mathbb{R}^{N_{\text{IRS}}}. \tag{4}$$

We also formulate the *reflection coefficient matrix* $\mathbf{\Phi}(\mathbf{c}[t], \theta_\ell[t]) \in \mathbb{C}^{N_{\text{IRS}} \times N_{\text{IRS}}}$ across the IRS meta-atoms as

$$\mathbf{\Phi}(\mathbf{c}[t], \theta_\ell[t]) = \text{diag}\big(\Gamma(C_1[t], \theta_\ell[t]), ..., \Gamma(C_{N_{\text{IRS}}}[t], \theta_\ell[t])\big), \tag{5}$$

[1] We are discussing the angle-dependent reflection model provided in [9] where only azimuth coordinates of the incident angle are considered.

[2] We assume a narrowband system with a few tens of MHz in bandwidth. Then, we can approximate the reflection coefficients as constant across ω [8], [9].

where the n-th diagonal $\Gamma(C_n[t], \theta_\ell[t])$ is the reflection coefficient at meta-atom n, $n \in \{1, ..., N_{\text{IRS}}\}$, given the incident angle $\theta_\ell[t]$.

We consider multi-path single tap channels and adopt a geometric channel model representation [12]. The channel from the UE to the IRS is described as

$$\mathbf{h}^{\text{UI}}[t] = \sum_{\ell=1}^{L[t]} \mathbf{h}_\ell^{\text{UI}}(\theta_\ell[t], t) \in \mathbb{C}^{N_{\text{IRS}} \times 1}, \quad (6)$$

in which $\mathbf{h}_\ell^{\text{UI}}(\theta_\ell[t], t)$ is the ℓ-th path channel with incident angle $\theta_\ell[t]$ and $L[t]$ is the number of paths. The received signal at the BS at time t is given by

$$\mathbf{y}[t] = \mathbf{h}_{\text{eff}}(\mathbf{c}[t], t)\sqrt{P}x[t] + \mathbf{n}[t] \in \mathbb{C}^{N_{\text{BS}} \times 1}, \quad (7)$$

where $P \geq 0$ is the transmit power and $x[t] \in \mathbb{C}$ is the transmit symbol of the UE with $\mathbb{E}[|x[t]|^2] = 1$. The noise vector $\mathbf{n}[t]$ follows the complex Gaussian distribution $\mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ denotes the identity matrix and σ^2 is the variance. In (7), the *end-to-end compound channel* $\mathbf{h}_{\text{eff}}(\mathbf{c}[t], t) \in \mathbb{C}^{N_{\text{BS}} \times 1}$ is

$$\mathbf{h}_{\text{eff}}(\mathbf{c}[t], t) = \mathbf{h}^{\text{UB}}[t] + \mathbf{H}^{\text{IB}}[t] \sum_{\ell=1}^{L[t]} \mathbf{\Phi}(\mathbf{c}[t], \theta_\ell[t]) \mathbf{h}_\ell^{\text{UI}}(\theta_\ell[t], t), \quad (8)$$

where $\mathbf{h}^{\text{UB}}[t] \in \mathbb{C}^{N_{\text{BS}} \times 1}$ is the direct channel from the UE to the BS and $\mathbf{H}^{\text{IB}}[t] \in \mathbb{C}^{N_{\text{BS}} \times N_{\text{IRS}}}$ is the channel from the IRS to the BS. $\mathbf{h}_{\text{eff}}(\mathbf{c}[t], t)$ encapsulates all the channels (i.e., $\mathbf{h}^{\text{UB}}[t]$, $\mathbf{H}^{\text{IB}}[t]$, and $\mathbf{h}^{\text{UI}}[t]$) and the IRS configuration (i.e., $\mathbf{c}[t]$).

III. Problem Formulation and Limited Feedback Protocol

We first formulate the data rate maximization problem for IRS control and discuss the challenges associated with solving it in Sec. III-A. To address the challenges, we propose an adaptive codebook-based limited feedback protocol in Sec. III-B.

A. Problem Formulation and Challenges

We formulate the data rate maximization at time t as

$$\underset{\mathbf{c}[t]}{\text{maximize}} \quad R(\mathbf{c}[t], t) = W \log_2 \left(1 + \frac{P\|\mathbf{h}_{\text{eff}}(\mathbf{c}[t], t)\|_2^2}{\sigma^2}\right) \quad (9)$$

$$\text{subject to} \quad C_{\min} \leq C_n[t] \leq C_{\max}, \; n = 1, ..., N_{\text{IRS}}, \quad (10)$$

where W is the channel bandwidth (in Hz). In other words, the objective is to adapt $\mathbf{c}[t]$ based on the time-varying channels.

Operationally, we aim for (9)-(10) to be solved at the BS since the BS can obtain measurements and has abundant computing resources. The BS would then generate feedback information for the IRS. However, solving (9)-(10) presents three key challenges. First, conventional optimization-based methods, relying on channel estimations, cannot be applied: the channel estimation requires the IRS reflection coefficients, which depend on the incident angles of incoming signals that cannot readily be measured (the IRS has no active sensors). Second, adaptive control of $\mathbf{c}[t]$ is necessary for efficient IRS operation in time-varying channels, which requires periodic information acquisition from the BS. Related to this is the third challenge: the feedback link has a low data rate [4]. The main contribution of our work is developing methodology to jointly address these challenges.

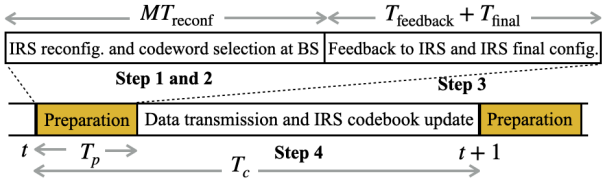


Fig. 2: Time frame structure of the proposed adaptive codebook-based limited feedback protocol for IRS-assisted communication.

B. Adaptive Codebook-based Limited Feedback Protocol

Motivated by the low overhead feedback requirement, we propose to exploit a *codebook* structure for IRS control, where the BS sends only a quantized codeword index to the IRS. Further, we consider *adaptive* design of this codebook based on channel variations. We denote the instantaneous codebook as $\mathcal{C}[t] = \{\mathbf{q}_m[t]\}_{m=1}^M$, where $\mathbf{q}_m[t] \in \mathbb{R}^{N_{\text{IRS}}}$ is the m-th codeword (capacitance vector) and M is the codebook size. The codebook is stored and updated at the IRS through its control board (see Fig. 1(a)). We propose a novel *limited feedback protocol* consisting of four steps conducted per each coherence time block t, depicted in Fig. 2:

Step 1. IRS channel sounding and reconfiguration. While the UE transmits pilot symbols, the IRS explores all of the M capacitance vectors, i.e., $\mathbf{q}_m[t] \in \mathcal{C}[t]$, $m = 1, ..., M$.

Step 2. Codeword selection at BS. The BS measures the effective channel $\mathbf{h}_{\text{eff}}(\mathbf{q}_m[t], t)$ and calculates the data-rate $R(\mathbf{q}_m[t], t)$ from (9), as IRS applies $\mathbf{q}_m[t]$, $m=1, ..., M$. The BS obtains the codeword index

$$m^\star[t] = \underset{m \in \{1,...,M\}}{\arg\max} \; R(\mathbf{q}_m[t], t). \quad (11)$$

Step 3. Feedback to IRS and IRS final configuration. The BS feeds back $m^\star[t]$ to the IRS with $\lceil \log_2 M \rceil$ feedback bits. Then, the IRS tunes its meta-atoms with $\mathbf{q}_\star[t] = \mathbf{q}_{m^\star[t]}[t] \in \mathcal{C}[t]$.

Step 4. Data transmission and IRS codebook update. The data transmission is conducted during the rest of the coherence time. During this period, the IRS obtains the next codebook $\mathcal{C}[t+1]$ either locally or with assistance from the BS.

The benefits of our protocol include its (i) simple procedure for IRS configuration, (ii) low-overhead feedback, and (iii) adaptation to dynamic channels. Careful design of the codebook $\mathcal{C}[t]$ is critical to obtaining high data rates, since the codewords $\{\mathbf{q}_m[t]\}_{m=1}^M$ are the solution candidates and the best one ($\mathbf{q}_\star[t]$ in ***Step 3***) among them is selected.[3] We next propose two adaptive codebook design approaches for selecting $\mathcal{C}[t+1]$ in ***Step 4***, which utilize the previous IRS decisions and

[3] M should be limited due to the finite coherence time and non-negligible IRS reconfiguration time. We consider that M is predetermined in the protocol.

responses, with the understanding that the channels are in practice correlated between consecutive coherence times.

IV. Adaptive Codebook Design

For adaptive codebook design, we propose a low-overhead perturbation-based approach in Sec. IV-A and a deep neural network (DNN) policy-based approach in Sec. IV-B. Then, we present a group control strategy, and quantify the time overhead and average data rate over one channel coherence block in Sec. IV-C.

A. Random Adjacency (RA) Approach

We first propose the *random adjacency* (RA) approach, which can be viewed as a random perturbation-based method [13] for codebook design. Since the optimization (9)-(10) is conducted successively in time-correlated channels, the optimal solutions in adjacent time blocks are expected to be close to one another. The RA approach exploits this intuition by generating multiple solution candidates around the previous solution. The codebook resides and is updated at the IRS in this method.

Formally, the IRS obtains the codebook $\mathcal{C}[t+1] = \{\mathbf{q}_m[t+1]\}_{m=1}^{M}$, where the m-th codeword is updated by adding a random perturbation vector $\mathbf{z}_m[t] \in \mathbb{R}^{N_{\text{IRS}}}$ to the previous solution $\mathbf{q}_\star[t]$ (obtained in ***Step 4*** in Sec. III-B) as

$$\mathbf{q}_m[t+1] = \text{clip}(\mathbf{q}_\star[t] + \mathbf{z}_m[t], [C_{\min}, C_{\max}]), \tag{12}$$

where $\text{clip}(\cdot, [C_{\min}, C_{\max}])$ is an element-wise clip function ensuring constraint (10). Each entry of $\mathbf{z}_m[t]$ is generated from the uniform distribution $\mathcal{U}(-\delta, \delta)$, where δ is the maximum step size. The RA approach follows ***Step 1–Step 4*** in Sec. III-B, while in ***Step 4*** the codewords are updated by (12).

Intuitively, the RA approach becomes more effective as the number of codewords M grows larger because more random points increase the chance of obtaining better codewords. However, M is limited due to the non-negligible IRS reconfiguration time and finite coherence time. This makes the performance of the RA approach restricted due to the nature of the randomness and motives us to develop our next codebook update algorithm.

B. DNN Policy-based IRS Control (DPIC) Approach

We next propose a DNN policy-based IRS control (DPIC) approach, aiming to learn *policies* for updating the codewords. In DPIC, the codebook resides at the IRS, as in RA. However, the IRS now updates the codebook via information reception from the BS through the feedback link. We consider that each codeword is updated *independently* based on its prior deployments. Henceforth, without loss of generality, we focus on the updates of m-th codeword.

1) Low Overhead IRS Control via Direction Codebook: To conduct the low overhead codeword update, we introduce a fixed *direction codebook* $\mathcal{D} = \{\mathbf{d}_k\}_{k=1}^{K}$ where $\mathbf{d}_k \in \mathbb{R}^{N_{\text{IRS}}}$, $k = 1, ..., K$. The BS only transmits the index of a codeword in $\mathcal{D}$ to the IRS, which enables low feedback overhead for the codeword update. We assume that $\mathcal{D}$ is generated once at the beginning of the policy learning and shared at both the BS and IRS. The BS employs a *learning architecture* (discussed in Sec. IV-B2) to first obtain a continuous direction vector $\boldsymbol{u}_m[t] \in \mathbb{R}^{N_{\text{IRS}}}$, from which it finds the index $k_m[t] \in \{1, ..., K\}$ such that $k_m[t]$-th codeword $\mathbf{d}_{k_m[t]}$ in $\mathcal{D}$ has the highest similarity to $\boldsymbol{u}_m[t]$. The BS then feeds back the index $k_m[t]$ to the IRS, which the IRS uses to recover $\mathbf{d}_{k_m[t]}$ from $\mathcal{D}$, and updates the m-th codeword as

$$\mathbf{q}_m[t+1] = \text{clip}(\mathbf{q}_m[t] + \mathbf{d}_{k_m[t]}, [C_{\min}, C_{\max}]). \tag{13}$$

2) IRS Control via Successive Decision Making: The learning architecture at the BS consists of a DNN policy that determines $\boldsymbol{u}_m[t]$ and a quantization process that determines $k_m[t]$. Our learning architecture consists of two phases: *training phase* and *utilization phase*. In the training phase, the BS aims to train the DNN policy to have an improved $\boldsymbol{u}_m[t]$ over time, while in the utilization phase the BS exploits the trained DNN policy without additional training. In both phases, the BS first determines $\boldsymbol{u}_m[t]$ with the DNN policy based on the current information (i.e., the codeword $\mathbf{q}_m[t]$ in use and the effective channel $\mathbf{h}_{\text{eff}}(\mathbf{q}_m[t], t)$). Subsequently, the BS obtains $k_m[t]$ via a quantization process applied to $\boldsymbol{u}_m[t]$ (described in Sec. IV-B4&IV-B5). The BS then feeds back $k_m[t]$ to the IRS, from which the IRS obtains the next codeword $\mathbf{q}_m[t+1]$ through (13). The next codeword affects the subsequent information at the BS (i.e., $\mathbf{q}_m[t+1]$ and $\mathbf{h}_{\text{eff}}(\mathbf{q}_m[t+1], t+1)$). The codeword update can thus be formulated as a successive decision making process (Sec. IV-B3).

3) MDP for Codeword Update: We construct a Markov decision process (MDP) for the codeword update with the following state, action, and reward.

State. The state consists of information pertinent to the environment evolution, which we define as

$$\mathbf{s}_m[t] = \{\mathbf{h}_{\text{eff}}(\mathbf{q}_m[t], t), \mathbf{q}_m[t]\} \in \mathcal{S} = \mathbb{R}^{2N_{\text{BS}}+N_{\text{IRS}}}, \tag{14}$$

where the real and imaginary parts of $\mathbf{h}_{\text{eff}}(\mathbf{q}_m[t], t)$ are stored as separate state dimensions.

Action. The action is the direction vector $\boldsymbol{u}_m[t]$:

$$\mathbf{a}_m[t] = \boldsymbol{u}_m[t] \in \mathcal{A} = [-\delta, \delta]^{N_{\text{IRS}}}, \tag{15}$$

where each entry of the action is bounded to the maximum step size, i.e., $[-\delta, \delta] \subset \mathbb{R}$. The action $\mathbf{a}_m[t]$ is used to determine the index $k_m[t]$ based on different processes in the training (Sec. IV-B4) and utilization (Sec. IV-B5) phases. The next codeword $\mathbf{q}_m[t+1]$ is then obtained from $k_m[t]$ by (13).

Reward. The reward provides a measure of efficacy for policy learning. We define the reward as

$$r_m[t] = R(\mathbf{q}_m[t+1], t+1) - N_{\text{clip},m}[t] \in \mathbb{R}, \tag{16}$$

where $R(\mathbf{q}_m[t+1], t+1)$ denotes the data rate measured at time $t+1$ using codeword $\mathbf{q}_m[t+1]$, and $N_{\text{clip},m}[t]$ denotes the number of elements/dimensions in vector $\mathbf{q}_m[t+1] \in \mathbb{R}^{N_{\text{IRS}}}$ that hit the clipping threshold in (13). $N_{\text{clip},m}[t]$ is added as a penalty to avoid actions that result in the capacitance vectors violating (10).

4) Training Phase for DNN Policy Learning: We tailor a deep reinforcement learning (DRL) methodology to train the DNN policy with the formulated MDP. We assume that the BS trains M_{A} different learning architectures, which are referred to as *agents*. Agent $m \in \{1, ..., M_{\mathrm{A}}\}$ has the DNN policy $\pi(\mathbf{s}_m[t]; \mathbf{w}_{\pi,m}) : \mathcal{S} \rightarrow \mathcal{A}$ where $\mathbf{w}_{\pi,m}$ is the respective DNN weight parameters. We consider that agent m is trained with codeword m, $m \in \{1, ..., M_{\mathrm{A}}\}$.

The *actual action* of the agent m (i.e., $\mathbf{d}_{k_m[t]}$ in (13)) is determined at the BS via the two following steps. First, the BS adds a random noise vector $\mathbf{v}_m[t]$ to the output of the policy $\pi(\mathbf{s}_m[t]; \mathbf{w}_{\pi,m})$ to have more diverse responses and avoid getting trapped in local optima [14]. The BS uses the clip function to confine the output result to the feasible action space. Second, the BS applies the *quantization process* using the direction codebook $\mathcal{D}$, through which the BS determines the codeword index $k_m[t] \in \{1, ..., K\}$ with closest Euclidean distance. In other words,

$$k_m[t] = \underset{k \in \{1,...,K\}}{\arg\min} \ \|\mathrm{clip}(\pi(\mathbf{s}_m[t]; \mathbf{w}_{\pi,m}) + \mathbf{v}_m[t], [-\delta, \delta]) - \mathbf{d}_k\|_2. \tag{17}$$

To obtain $\pi(\mathbf{s}_m[t]; \mathbf{w}_{\pi,m})$, we exploit the actor-critic network architecture of DRL. This architecture consists of an actor network $\pi(\mathbf{s}_m[t]; \mathbf{w}_{\pi,m})$ and a critic network $Q(\mathbf{s}_m[t], \mathbf{a}_m[t]; \mathbf{w}_{Q,m})$ with DNN parameters $\mathbf{w}_{Q,m}$. The actor selects an action using a policy, and the critic evaluates/criticizes the action to guide the actor network to take better actions. However, using DNNs for reinforcement learning has been known to cause learning instability [14]. To stabilize the learning, we adopt the deep deterministic policy gradient (DDPG) approach [15] for training the actor-critic networks. The overall algorithm for training M_{A} agents in the limited feedback protocol is given in Algorithm 1.

5) Utilization Phase with Trained DNN Policies: In the utilization phase, we utilize the M_{A} trained agents to conduct codebook updates without additional training of the agents. The BS partitions M codewords among M_{A} agents. We consider that codeword $m \in \{1, ..., M\}$ is allocated to agent $j[m] = \mathrm{mod}\ (m-1, M_{\mathrm{A}}) + 1 \in \{1, ..., M_{\mathrm{A}}\}$. If $M_{\mathrm{A}} = 1$, a single agent handles all M codeword updates, which we refer to as a single-agent DPIC (SDPIC). If $M_{\mathrm{A}} > 1$, multiple agents handle the M codeword updates, which we refer to as multi-agent DPIC (MDPIC). Utilizing more agents often improves performance due to the ensemble learning principle [16].

The DPIC approach follows ***Step 1***–***Step 4*** of the protocol in Sec. III-B. Additionally, in ***Step 2***, the BS constructs $\mathbf{s}_m[t] = \{\mathbf{h}_{\mathrm{eff}}(\mathbf{q}_m[t], t), \mathbf{q}_m[t]\}$ and determines $k_m[t]$ in (17) by using $\mathbf{v}_m[t] = 0$ (no random noise addition) and $\mathbf{w}_{\pi,j[m]}$ (instead of $\mathbf{w}_{\pi,m}$). In ***Step 3***, the BS additionally feeds back $\{k_m[t]\}_{m=1}^{M}$ to the IRS for codebook updates, which incurs a total of $\lceil \log_2 M \rceil + M \lceil \log_2 K \rceil$ feedback bits. In ***Step 4***, the BS also updates the codebook.

Algorithm 1 Training M_{A} agents with actor-critic architecture

Input. N_{episode} (the number of learning episodes), N_{timestep} (the duration of each episode), $C_{\min}$, $C_{\max}$, and M_{A}.
Initialize $\mathbf{w}_{Q,m}$ and $\mathbf{w}_{\pi,m}$. Empty the replay buffer $\mathcal{B}_m$, $m \in \{1, ..., M_{\mathrm{A}}\}$. Set $\epsilon_0 = (C_{\max} - C_{\min})/5$ and $\epsilon_{\min} = \epsilon_0/300$.
for $e = 0, ..., N_{\mathrm{episode}} - 1$ **do**
  Randomly generate $\mathcal{C}[0] = \{\mathbf{q}_m[0]\}_{m=1}^{M_{\mathrm{A}}}$. Update $\epsilon_e = \max\{\epsilon_{\min}, 0.99\epsilon_{e-1}\}$, if $e \geq 1$.
  for $t = 0, ..., N_{\mathrm{timestep}} - 1$ **do**
    Step 1. IRS channel sounding and reconfiguration. The IRS meta-atoms are tuned following $\{\mathbf{q}_m[t]\}_{m=1}^{M_{\mathrm{A}}}$.
    Step 2. Inference at BS. Each agent $m \in \{1, ..., M_{\mathrm{A}}\}$ computes $r_m[t-1]$ using (16), forms $\mathbf{s}_m[t]$ in (14), and determines $k_m[t]$ using (17) where $\mathbf{v}_m[t] \sim \mathcal{CN}(\mathbf{0}, \epsilon_e \mathbf{I})$.
    Step 3. Feedback to IRS. The BS feeds back $\{k_m[t]\}_{m=1}^{M_{\mathrm{A}}}$ to the IRS.
    Step 4. IRS codebook update and BS training. The IRS obtains $\mathcal{C}[t+1] = \{\mathbf{q}_m[t+1]\}_{m=1}^{M_{\mathrm{A}}}$ by (13). Each agent m stores $(\mathbf{s}_m[t-1], \mathbf{a}_m[t-1], r_m[t-1], \mathbf{s}_m[t])$ in $\mathcal{B}_m$, samples $\{(\mathbf{s}_i, \mathbf{a}_i, r_i, \mathbf{s}'_i)\}_{i=1}^{N_{\mathrm{batch}}}$ from $\mathcal{B}_m$, and updates $\mathbf{w}_{Q,m}$ and $\mathbf{w}_{\pi,m}$ via the DDPG algorithm in [15].
  end for
end for

C. Group Control, Time Overhead, and Effective Data Rate

We consider a *group control* [5], where IRS meta-atoms are partitioned into multiple groups and the same capacitance is applied for the meta-atoms belonging to the same group. This reduces the dimension of the design variables, and thus the BS can conduct the training/inference in a timely manner. We thus neglect the computation time overhead of our methods, and define the *time overhead* T_p shown in Fig. 2 as

$$T_p = M T_{\mathrm{reconf}} + T_{\mathrm{feedback}} + T_{\mathrm{final}}. \tag{18}$$

In (18), $M T_{\mathrm{reconf}}$ denotes the total time for M IRS reconfiguration (in ***Step 1*** in Sec. III-B) used in both RA and DPIC approaches, where T_{reconf} is the time for each IRS reconfiguration, and T_{feedback} denotes the time required for the feedback from the BS to the IRS (in ***Step 3***), which is different for the RA and DPIC approach. For the RA approach, the feedback time is

$$T_{\mathrm{feedback}} = \frac{\lceil \log_2 M \rceil}{W R_{\mathrm{feedback}}}, \tag{19}$$

where R_{feedback} (bits/s/Hz) is the unit data rate for the feedback link. For the DPIC approach, the feedback time is

$$T_{\mathrm{feedback}} = \frac{\lceil \log_2 M \rceil + M \lceil \log_2 K \rceil}{W R_{\mathrm{feedback}}}. \tag{20}$$

Lastly, T_{final} denotes the execution time of the final IRS reconfiguration (in ***Step 3***). If the selected index $m^\star[t]$ coincides with the last configuration in ***Step 1***, the IRS does not need to change the configuration, i.e., $T_{\mathrm{final}} = 0$; otherwise $T_{\mathrm{final}} = T_{\mathrm{reconf}}$.

To measure the average data rate during one coherence block with coherence time T_c, we define *effective data rate* as

$$R_{\mathrm{eff}}[t] = \frac{T_c - T_p}{T_c} W \log_2 \left(1 + \frac{P \|\mathbf{h}_{\mathrm{eff}}(\mathbf{q}_\star[t], t)\|_2^2}{\sigma^2}\right), \tag{21}$$

where $T_c - T_p$ is the actual data transmission time and $\mathbf{q}_\star[t] \in \mathcal{C}[t]$ is the selected codeword for the IRS configuration. The above metric captures the tradeoff between the data rate and the time overhead T_p. As M increases, the data rate may increase due to having larger number of reconfigurations, while T_p also increases leading to decreasing of $T_c - T_p$. In Sec. V, we evaluate the data rate and effective data rate under different M.

V. Numerical Evaluation and Discussion

In this section, we describe the simulation setup in Sec. V-A, and then present and discuss the simulation results in Sec. V-B.

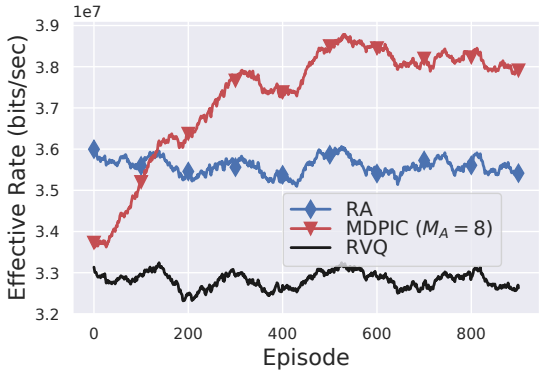
A. Simulation Setup

To emulate practical IRS reflection behavior, we recover $\Gamma(C, \theta)$ from the data in Fig. 4 and Table 1 of [9], where the ranges of C and θ are $(C_{\min}, C_{\max}) = (0.4, 2.7)$ pF and $(0^o, 90^o)$, respectively. We set $f = 5.195$ GHz and consider only azimuth coordinates as in [9]. We assume $T_c = 5$ ms, $N_{\text{BS}} = 5$ and $N_{\text{IRS}} = 200$. We consider a group control with the number of groups $N_{\text{G}} = 10$, where $N_{\text{IRS}}/N_{\text{G}} = 20$ meta-atoms are controlled by each common capacitance. The BS antenna spacing is $d_{\text{BS}} = \lambda/2$, and the IRS meta-atom spacing is $d_{\text{IRS}} = \lambda/10$ where $\lambda = c/f$ and $c = 3 \times 10^8$ m/s. The BS and IRS are located at $(0, 0)$ m and $(90, 30)$ m, respectively. The initial UE position is randomly generated within the circle with radius 5 m at $(100, 0)$ m. The UE is moving with the velocity $v = 3$ km/h and constant azimuth angle generated from $\mathcal{U}(0, 2\pi)$. We set $P = 20$ dBm, $\sigma^2 = -80$ dBm, $W = 10$ MHz, $R_{\text{feedback}} = 0.1$ bits/s/Hz, and $T_{\text{reconf}} = 100\mu s$ [11].[4]

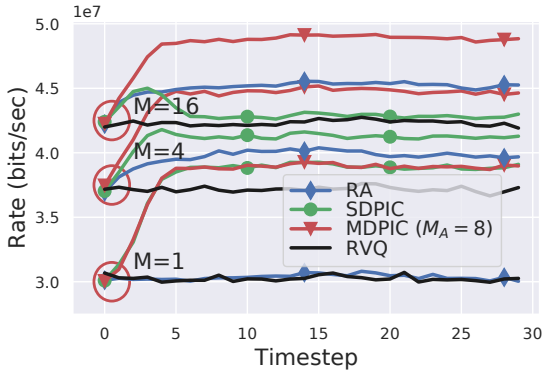
For RA, we set $\delta = (C_{\max} - C_{\min})/5$. For DPIC, we set $N_{\text{batch}} = 32$ and $|\mathcal{B}_m| = 5 \times 10^5$, $m \in \{1, ..., M_{\text{A}}\}$. For the DNNs, we consider a fully connected neural network with two hidden layers consisting of 400 and 300 neurons, respectively, with ReLU activation function. For the DNN policy, in the output layer the tanh function is employed, and the output is scaled by $\delta = (C_{\max} - C_{\min})/4$ to bound the actions. We set $|\mathcal{D}| = K = 2048$, where each codeword in $\mathcal{D}$ is constructed by RVQ [17] ranging within $[-\delta, \delta]^{N_{\text{G}}}$. We employ the Adam optimizer for training. For MDP, we normalize the values as $\mathbf{h}_{\text{eff}}(\cdot) \leftarrow \sqrt{P/(\sigma^2 N_{\text{BS}} N_{\text{G}})} \times \mathbf{h}_{\text{eff}}(\cdot)$ and $\mathbf{q}_m \leftarrow 10^{12} \times \mathbf{q}_m$ in (14), $\mathbf{a}_m \leftarrow 10^{13} \times \mathbf{a}_m$ in (15), and $R(\cdot) \leftarrow R(\cdot)/W$ in (16).

We adopt a multi-path geometric channel model [12] for the IRS-BS, UE-BS, and UE-IRS channels. We model (i) the IRS-BS channel as Rician channel with K-factor of 5 [12], 10 non line-of-sight (NLoS) signal paths, and a path loss exponent (PLE) of 2, (ii) the UE-BS channel with only NLoS signals of 10 paths and a PLE of 3.75, and (iii) the UE-IRS channel with only NLoS signals of 10 paths and a PLE of 2.2. The small scale fading factors evolve according to a first-order Gauss-Markov process [18] with a time correlation coefficient of 0.95 (corresponding to speed of 3 km/h of the UE/scatterers), and the angle of each path varies over every coherence time by an amount generated from $\mathcal{U}(-0.1^o, 0.1^o)$.

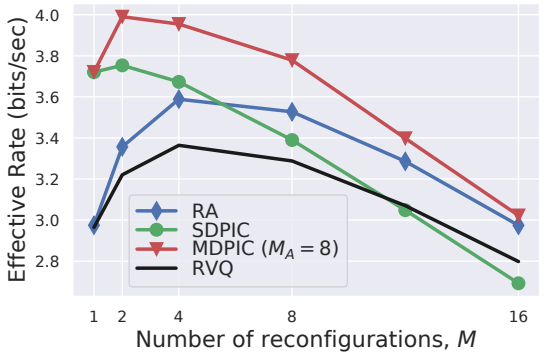
[4]The reconfiguration time of IRS is determined by the characteristics of the control board and the internal communication between the control board and the meta-surface. The reconfiguration speed is typically a few kHz [11].



(a) Effective data rate along episodes



(b) Data rate along timesteps



(c) Effective data rate along M

Fig. 3: Performance evaluation of our methodology. (a) corresponds to the training phase, while (b)-(c) correspond to the utilization phase.

B. Simulation Results and Discussion

We train $M_{\text{A}} = 8$ agents in the training phase with 1000 episodes, each with 500 timesteps (coherence blocks). Each episode has different realizations of the UE-IRS channel, UE-BS channel, IRS-BS channel, initial UE location, and UE moving direction. Our baseline is the RVQ codebook design [17]. Fig. 3(a) shows the effective data rate averaged over the timesteps. Each data point is a moving average over the

previous 100 episodes. The performance of MDPIC is improved over time since the agents (specifically, the DNN policies) are trained to conduct better codebook updates.

We then evaluate our proposed algorithms in the utilization phase with 2000 episodes each with 30 timesteps. Fig. 3(b) depicts the average data rate over the utilization episodes. Our proposed schemes – RA, SPDIC, and MDPIC – update the codebook adaptively by using previous observations (i.e., previously used codeword and end-to-end channel) to improve the data rate over time. Our methods obtain their peak performance within 4-5 timesteps. Overall, as the number of IRS reconfiguration M increases, a higher data rate is achieved. The MDPIC yields better data rate compared to those of the SDPIC and RA due to the advantage of using multiple agents.

Fig. 3(c) shows the effective data rate along M. The effective data rate in (21) captures the tradeoff between the data rate and the time overhead discussed in Sec. IV-C. The MDPIC method shows the best performance in terms of effective data rate for any M, and has the highest value at $M^\star = 2$. As M grows larger than 2, the increased time overhead outweighs the improvement of the data rate, leading to the decrease of the effective data rate.

VI. Conclusion

In this paper, we introduced a novel signal model incorporating the practical IRS reflection behavior. To address the design challenges faced with IRS control under multi-path dynamic channels and low-overhead feedback requirements, we proposed the codebook-based limited feedback protocol. We proposed two adaptive codebook designs: the random adjacency and the deep neural network policy-based IRS control. Through simulations, we showed that the data rate and effective data rate performances are improved by the proposed schemes.

References

[1] J. Kim, S. Hosseinalipour, A. C. Marcum, T. Kim, D. J. Love, and C. G. Brinton, "Learning-based adaptive IRS control with limited feedback codebooks," *arXiv preprint arXiv:2112.01874*, 2021.

[2] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, 2020.

[3] S. Hosseinalipour, C. G. Brinton, V. Aggarwal, H. Dai, and M. Chiang, "From federated to fog learning: Distributed machine learning over heterogeneous wireless networks," *IEEE Commun. Mag.*, vol. 58, no. 12, pp. 41–47, 2020.

[4] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, 2019.

[5] Y. Yang, B. Zheng, S. Zhang, and R. Zhang, "Intelligent reflecting surface meets OFDM: Protocol design and rate maximization," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4522–4535, 2020.

[6] L. Shao and W. Zhu, "Electrically reconfigurable microwave metasurfaces with active lumped elements: A mini review," *Front. Mater.*, vol. 8, p. 212, 2021.

[7] S. Abeywickrama, R. Zhang, Q. Wu, and C. Yuen, "Intelligent reflecting surface: Practical phase shift model and beamforming optimization," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5849–5863, 2020.

[8] X. Pei, H. Yin, L. Tan, L. Cao, Z. Li, K. Wang, K. Zhang, and E. Björnson, "RIS-aided wireless communications: Prototyping, adaptive beamforming, and indoor/outdoor field trials," *arXiv preprint arXiv:2103.00534*, 2021.

[9] W. Chen, L. Bai, W. Tang, S. Jin, W. X. Jiang, and T. J. Cui, "Angle-dependent phase shifter model for reconfigurable intelligent surfaces: Does the angle-reciprocity hold?" *IEEE Commun. Lett.*, 2020.

[10] J. Kim, S. Hosseinalipour, T. Kim, D. J. Love, and C. G. Brinton, "Multi-IRS-assisted multi-cell uplink MIMO communications under imperfect CSI: A deep reinforcement learning approach," in *IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, 2021, pp. 1–7.

[11] S. Abadal, T.-J. Cui, T. Low, and J. Georgiou, "Programmable metamaterials for software-defined electromagnetic control: Circuits, systems, and architectures," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 10, no. 1, pp. 6–19, 2020.

[12] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[13] R. Mudumbai, G. Barriac, and U. Madhow, "On the feasibility of distributed beamforming in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1754–1763, 2007.

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

[15] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

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

[17] C. K. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 458–462, 2007.

[18] B. Sklar *et al.*, *Digital communications: fundamentals and applications*, 2001.