Adversarial Bandit Approach for Stand Alone RIS Operation

Messaoud Ahmed Ouameur, Dương Tuấn Anh Lê, Gwanggil Jeon,
Felipe A.P. De Figueiredo and Daniel Massicotte

Abstract— Even though, reconfigurable intelligent surfaces (RISs) are adopted in various scenarios to enable the implementation of a smart radio environment, there are still challenging issues for its real-time operation due to the need for a costly full dimensional channel estimation with offline exhaustive search or online exhaustive beam-training. The application of the deep learning (DL) tools is favored to enable feasible solutions. In this work, we propose two low training overhead and energy efficient adversarial bandit-based schemes with outstanding performance gains compared to reference DL based reflection beamforming methods. The resulting deep learning models are also discussed using state-of-the-art model quality prediction trends.

Index Terms— Reconfigurable intelligent surfaces, Reflection beamforming prediction, Deep learning, Adversarial bandit, exponential-weight algorithm for exploration and exploitation, follow the perturbed leader.

I. INTRODUCTION

Reconfigurable intelligent surfaces (RISs) enable the control of the wireless propagation environment by smartly controlling the signal reflections via its massive low-cost elements [1-2]. Unfortunately, due to the additional channel links between the RIS and its associated transmitter and intended receivers (Figure 10 in [1] and Figure 1 in [2]), the large gain is achieved at the expense of more overhead for the channel estimation [3- 6]. However, obtaining this channel knowledge, in practice, may require large and possibly prohibitive training overhead, which represents the main challenge for the real-time RIS operation. As such, machine learning (ML) is introduced and has started to be extensively used to enhance the implementation of various components within the 5G radio access network (RAN) [4]. In [5], a joint design of transmitting beamforming (BF) matrix at the base station and the phase shift matrix at the RIS is performed by leveraging policy-based deep deterministic policy gradient (DDPG). However, the phase shifts are assumed to be continuous. The authors in [6] present a novel RIS hardware architecture along with two solutions based on compressive sensing and deep reinforcement learning with negligible training overhead. Nevertheless, the

choice of the training scheme to build the dataset is not analyzed in depth. A detailed survey that introduces the interplay between AI and RIS is found in [7].

In this letter, we propose an efficient reinforcement learning based scheme to improve upon the proposed method in [6]. The main contributions are as follow

- We propose an adversarial bandit approach based on exponential-weight algorithm for exploration and exploitation (EXP3). To show the merits of the proposed scheme, we conduct extensive simulations using the publicly available accurate ray-tracing based DeepMIMO datasets [8], with the ‘O1’ scenario.
- To improve upon the computational complexity, the follow the perturbed leader (FPL) scheme is discussed.
- To compare the quality of the state-action deep neural network models used with the reference method [6] and with the proposed ones (EXP3 and FPL), we leverage state-of-the-art techniques such as the power low (PL) exponents [9].

II. SYSTEM MODEL AND PROBLEM FORMULATION

To enable the practical implementation of the RIS aided communication systems, new path loss models [2], and open-source channel models [2], [8] have been developed. As such to reproduce the results and perform a fair comparison, we will adopt the system and channel model in [6].

A. System model

The transmitter-receiver communication is aided by a RIS having M reconfigurable elements. For the sake of simplicity, we assume that both the transmitter and receive are equipped with a single antenna. For generalization one can adopt the signal model from [2]. An OFDM-based transmission with K subcarriers is adopted. The direct channel per subcarrier k between the transmitter and the receiver is denoted by $h_{TR,k} \in \mathbb{C}$ whereas links via the RIS are represented by $M \times 1$ complex valued vectors $\mathbf{h}_{T,k}, \mathbf{h}_{R,k} \in \mathbb{C}^{M \times 1}$. By neglecting the direct path, the received signal can be written as $y_k = \mathbf{h}_{R,k}^T \mathbf{\Psi}_k \mathbf{h}_{T,k} s_k + n_k$, where $\mathbf{\Psi}_k \in \mathbb{C}^{M \times M}$ is the RIS interaction diagonal matrix, s_k and n_k are the transmitted symbol per subcarrier k and the receive noise with zero mean and variance of σ_n . With P_T being the total transmit power, the following power per subcarrier

M. Ahmed Ouameur and D. Massicotte are with the Université du Québec à Trois-Rivières, Department of Electrical and Computer Eng., 3351 Boul. des Forges, Trois-Rivières, Qc, Canada, G9A 5H7. messaoud.ahmed.ouameur@uqtr.ca and daniel.massicotte@uqtr.ca
D.T.A. Lê is with the Faculty of Information Technology, VNU-HCM University of Science, Vietnam. 20c14001@student.hcmus.edu.vn.
G. Jeon is with the Department of Embedded Systems Engineering, College of Information Technology, Incheon National University, Incheon, South Korea. gjeon@inu.ac.kr
F.A.P. De Figueiredo is with the National Institute of Telecommunications, Santa Rita do Sapucaí - Minas Gerais, Brazil. felipe.figueiredo@inatel.br

TABLE 1. THE ADVERSARIAL BANDIT BASED SCHEME FOR REFLECTION BEAMFORMING VECTOR PREDICTION

Input: reflection beamforming codebook	
Initialize: $\gamma \in (0,1]$ and Weights $w_n(t=1)=1$ for $n=1,\dots,N$	
Phase I: Learning phase	
1	For $s = 1$ to S do (span over S channel coherent blocks)
2	RIS receives two pilots to estimate $\bar{\mathbf{h}}(s)$
3	For $t = 1$ to T do (go over T iterations)
4	Compute the probability of the n -th codebook $p_n(t) = (1-\gamma) \frac{w_n(t)}{\sum_{c=1}^N w_c(t)} + \frac{\gamma}{N}$
5	Draw n_t randomly according to the probabilities $p_1(t), p_2(t), \dots, p_N(t)$ and RIS uses Ψ_{n_t}
6	RIS receives $R_{n_t}(s)$
7	Compute the reward $r_{n_t} \in [0,1]$
8	For $c = 1$ to N do (update the weights)
9	$\hat{r}_c(t+1) = \begin{cases} r_c(t)/p_c(t) & \text{if } c = n_t \\ 0 & \text{otherwise} \end{cases}$
10	$w_n(t+1) = w_n(t) e^{\gamma \hat{r}_c(t)/N}$
11	Store the new entry in the learning data set $\Upsilon \leftarrow (\bar{\mathbf{h}}(s), \mathbf{p}(s))$ where $\mathbf{p}(s) = [p_1(T), p_2(T), \dots, p_N(T)]^T$
12	Train the DL model using the learning dataset Υ
Phase II: Prediction phase	
13	While True do (for every channel coherent block s')
14	RIS receives two pilots to estimate $\bar{\mathbf{h}}(s')$
15	Predict the probability vector $\mathbf{p}(s')$
16	RIS uses Ψ_{n^*} where $n^* = \arg \max_n [\mathbf{p}(s')]_n$

constraint is enforced $E(|s_k|^2) = P_T/K$. Herein $(\cdot)^T$ and $E(\cdot)$ denote the transpose and the expectation operations, respectively. If we rearrange the diagonal elements of the interaction matrix Ψ_k in a $M \times 1$ column vector Ψ_k , we refer to as the reflection BF vector, such that $\Psi_k = \text{diag}(\Psi_k)$, the received signal can also be expressed in more convenient way as $y_k = (\mathbf{h}_{R,k} \odot \mathbf{h}_{T,k})^T \Psi_k s_k + n_k$, where $\odot$ denotes the Hadamard product. Imposing a few practical implementation constraints of a nearly-passive RIS where the phase shifters apply the same phase shift over all subcarriers, the m -th element of Ψ_k is modeled as $[\Psi_k]_m = e^{j\phi_m}$.

B. Channel model and RIS design objective

As for the channel model a wideband geometric channel model for $\mathbf{h}_{T,k}$ and $\mathbf{h}_{R,k}$ is used [2][6]. Therefore, $\mathbf{h}_{T,k}$ and $\mathbf{h}_{R,k}$ can be expressed as a function

Figure 1. The achievable rate the number of active elements where the number of training data is set to 30K. Note that EXP3 achieves similar performance even with 5K training data.

Figure 2. The energy efficiency η versus the number of active elements M where the number of training data is set to 30K.

of the azimuth and elevation angles of arrival/departure of the ℓ^{th} path from a total of L paths [2] [6]. The RIS design objective is therefore to find out the reflection BF vector Ψ_k that maximizes the achievable rate at the

$$R = \frac{1}{K} \sum_{k=1}^K \log_2 \left(1 + \rho \left| (\mathbf{h}_{R,k} \odot \mathbf{h}_{T,k})^T \Psi_k \right|^2 \right),$$

where the signal to noise ratio is $\rho = P_T/K\sigma_n$. The maximization is done over a predefined discrete codebook P due to the fact that a practical radio frequency (RF) phase shifter uses quantized phase values. Unfortunately, maximizing the achievable rate entails an exhaustive search over the codebook P . Fortunately, the authors in [6] have proposed a novel hardware architecture along with a compressive sensing and DL based framework to tackle the issue with low training overhead. However, there is still a large room for improvement as we will discuss throughout this letter.

III. ADVERSARIAL BANDIT APPROACH VIA EXPONENTIAL-WEIGHT ALGORITHM FOR EXPLORATION AND EXPLOITATION

The authors in [6] use a DL-based approach to predict the reflection BF vector. Over a channel coherence block size S , the RIS sends two pilots to estimate a sampled

Figure 3. The spectral energy efficiency versus the number of training samples and the number of active elements $\bar{M}$ for the proposed FPL-based scheme.

channel vector $\bar{\mathbf{h}}(s) = \text{vec}([\bar{\mathbf{h}}_1(s), \bar{\mathbf{h}}_2(s), \dots, \bar{\mathbf{h}}_K(s)])$

where $\bar{\mathbf{h}}_k(s) \in \mathbb{C}^{\bar{M} \times 1}$ denotes the sampled combined channel vector, $\mathbf{h}_k = \mathbf{h}_{R,k} \odot \mathbf{h}_{T,k}$, for the k -th subcarrier at s -th channel coherent block using a fraction number of the RIS elements $\bar{M} \ll M$ that are assumed to be active elements (i.e., equipped with full RF and baseband processing chain). During the beam training, the RIS is configured using one reflection beam $\boldsymbol{\psi}$ (notice that the subscript k is removed because one reflection BF vector is available for all subcarriers) from the codebook P . Then, a dataset is constructed out of the tuples $\Upsilon \leftarrow (\bar{\mathbf{h}}(s), \mathbf{r}(s))$ where

$\mathbf{r}(s) = [R_1(s), R_2(s), \dots, R_N(s)]^T$ and $R_n(s)$ is the measured rate using the n -th codebook (N is the cardinality of the codebook P). Finally, a deep neural network is trained to use the dataset Υ .

A. Adversarial bandit approach via exponential-weight algorithm for exploration and exploitation

In this section, we propose an approach based on adversarial bandit wherein instead of spanning equally every element of the codebook P we adopt a scheme that favor the more likely optimal beams. Therefore, the dataset Υ will have more useful data to train with. Table 1 shows the proposed adversarial bandit based on exponential-weight algorithm for exploration and exploitation (EXP3) [10]. The adversarial bandit is a variant of the multi-armed bandit problem where a fixed and limited set of resources (phase shifters) must be assigned among alternative choices (reflection BF) in a way that maximizes their expected gain (achievable rate). In its basic form, EXP3 chooses a reflection BF vector $\boldsymbol{\psi}_n$ (steps 4 and 5 in Table 1) from the codebook P at random with probability $(1-\gamma)$ where it prefers choices with higher weights (exploit), or it selects with

Figure 4. The low power exponent α to compare the quality of the models used in the reference method [6], EXP3 and FPL algorithms.

probability γ to uniformly randomly explore. After receiving the rewards (steps 6 and 7 in Table 1), the weights are updated (in steps 9 and 10 in Table 1). The exponential growth significantly increases the weight of good reflection BF vectors. The key advantage of the proposed scheme is, instead of using the rates as the deep neural network outputs, we use the pull-probability vector $\mathbf{p}(s) = [p_1(T), p_2(T), \dots, p_N(T)]^T$ computed at step 4 using the updated weights which are in turn computed using the normalized reward (step 7). The normalized reward r_{n_i} is computed using the received rate $R_{n_i}(s)$ as $r_{n_i} = R_{n_i}(s)/R_{\max}$, where $R_{\max}$ is the maximum achievable rate.

B. Simulation results

The proposed EXP3 based learning scheme is evaluated using the outdoor ray tracing scenario O1 from the deep-MIMO dataset that is publicly available at [8]. For the sake of facilitating the comparison, similar setup is used in [6, section VII-A] as well. The results herein are also validated using channel data generated using the SimRIS tool [2]. The adopted RIS employs a uniform planar array (UPA) with 16-by-16 ($M=256$) antenna elements with 3dBi gain at the 28 GHz mmWave setup. The transmit power is set to 10 dBW while the receiver's noise figure is 5 dB. The codebook P is constructed using a 2D discrete Fourier transform (DFT) matrix.

The number of subcarriers involved in $\bar{\mathbf{h}}(s)$ is $(\bar{K}=64) \ll (K=512)$ which sets the input of the DL model equal to $2\bar{K}\bar{M}$. The sampled channel vector is normalized prior to the training phase. The DL models consists of four layers similar to the one used in [6] where the number of nodes in the hidden layers is $(2\bar{K}\bar{M}, 4M, 4M, M)$. The regular training and optimization parameters are: the batch size is set to 500 samples, the drop-out rate is 0.5 and the L_2 regularization factor is 0.0001. Of course, we do not attempt to optimize

the DL model but we will discuss its quality using the state-of-the-art techniques such as the power law exponents [9] in section IV.

Note shown here for space considerations (but will be relevant in section IV), the proposed EXP3 based scheme requires (substantially) 6 times less training data compared to DL reflection BF technique [6] owing to the optimal selection of the dataset which stresses that less likely reflection beams are given lower probability which excludes them during the exploitation phase of the EXP3 algorithm (Table 1-step 5). The reference DL reflection BF requires more active elements $\bar{M}$ to sustain competitive performance as shown in Figure 1 where EXP3-based learning schemes achieve 96% of the optimal achievable rate compared to 88% using the reference method in [6]. However, this will come at the expense of higher-power consumption.

We reformulate the energy efficiency as $\eta = W \times R/P_c$ measured in Mbit/J where W is the transmission bandwidth and P_c is the RIS power consumption. For the sake of space please refer to [6]. Figure 2 depicts the energy efficiency η as a function of the number of active elements $\bar{M}$. Like the reference DL reflection BF [6], the proposed method shows optimal but higher energy efficiency performance using 4 active elements only.

In light of these results, the EXP3-based adversarial bandit method demonstrates outstanding performance gains compared to state-of-the-art methods. So far, the adopted deep neural network architecture is similar to the one used in [6]. The reason being, one would be keen to see the effect of using a new learning scheme rather than proposing a new DL model. The other reason, which we will discuss in the next section, is that one will also be interested to compare the quality of the two networks trained to use $\Upsilon \leftarrow (\mathbf{h}(s), \mathbf{r}(s))$ for [6] and $\Upsilon \leftarrow (\mathbf{h}(s), \mathbf{p}(s))$ in the proposed method. However, let's first introduce a second computationally efficient adversarial bandit scheme based that uses the FPL algorithm.

IV. IMPROVING AND EVALUATION OF THE QUALITY OF PROPOSED APPROACHES

Even if the EXP3 algorithm has efficient theoretical guarantees, it is computationally expensive due to the calculation of the exponential terms [10]. The FPL algorithm is then introduced to alleviate such burden by following the reflection beam that has the best performance while adding exponential noise to it to provide exploration [11]. Even the base line FPL algorithm does not have appreciated theoretical guarantees, it is worth evaluating its performance in the scope of the current RIS' reflection BF prediction. In the FPL algorithm the arm n_t is drawn based on

$n_t = \arg \max_n [\hat{r}_n(t) + z_n(t)]$ where the exponential noise, $z_n(t) \sim \exp(\gamma)$, is added to provide exploration (step 5).

The updated reward in step 9 is replaced by $\hat{r}_c(t+1) = \hat{r}_n(t) + r_c(t)$ if $c = n_t$. Figure 3 shows that with considerably less training data, the FPL algorithm provides similar performances to the EXP3 algorithm at the expense of less "explainability" information such as the pull-probabilities and weights inherent in EXP3. However, how one can finally decide what algorithm is better beyond just comparing the achievable rates (accuracy)? Even if all algorithms have different approaches to build the training dataset Υ , they all share a similar model. The slight differences are in using the dropout layers to improve the regularization of the reference method and the use of the softmax activation for the model used with EXP3 to generate the pull-probabilities. Nevertheless, at the end these models are considered as black box that need to be compared.

It is beyond the scope of this paper to dig into explainability of DL models which can be found in [12]. We will rather use state-of-the-art tools from [9] and mainly the power law (PL) exponents to compare the quality of the DL models. Figure 4, depicts the PL exponents for the 4 layers. Indeed, we expect that a poorly trained model will lack good (i.e., small exponents α - see equation (12) in [9]) PL behavior in some layers whereas the EXP3 has, on average, smaller α values than the reference method, with all $\alpha \leq 6$ and with smaller mean/median α . It also has far fewer unusually large outlying α values than the reference method. The model used with FPL algorithm is rather showing the best training quality at the expense of less theoretical guarantees. The exponent values are obtained using the WeightWatcher tool [9]. For future investigation, this should also be contrasted with the behavior shown by scale-dependent metrics such as the Frobenius norm and the Spectral norm [9].

V. CONCLUSION

We have discussed two adversarial bandit-based schemes that provide substantial spectral and energy efficiency gains. We have also, discussed the associated DL models' quality using the PL exponents to show the training quality using the dataset generated from the proposed schemes. Our work contributes to shading light on the potential improvement that can be made in exploring the interplay between ML and RISs. As future works, one can investigate the proposed schemes under different channels and system parameters while adopting meta-learning approach to improve the online training performance. Different DL models can also be used along with these schemes wherein explainability shall be given a considerable attention to improve the trustworthiness of the DL-enabled solutions.

REFERENCES

- [1] M. Di Renzo *et al.*, "Smart radio environments empowered by reconfigurable intelligent surfaces: how it works, State of Research, and The Road Ahead," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 11, pp. 2450-2525, Nov. 2020.
- [2] E. Basar and I. Yildirim, "SimRIS channel simulator for reconfigurable intelligent surface-empowered communication systems," *IEEE Latin-American Conference on Communications (LATINCOM)*, pp. 1-6, 2020.
- [3] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communication*, vol. 18, no. 11, pp. 5394-5409, Nov. 2019.
- [4] Nokia Bell Labs, "Toward a 6G AI-Native Air Interface," White paper, retrieved on July 2021, online access: https://d1p0gxnqcu0lvz.cloudfront.net/documents/Nokia_Towards_a_6G_AI-Native_Air_Interface_Article_EN_final.pdf
- [5] C. Huang, R. Mo and C. Yuen, "Reconfigurable intelligent surface assisted multiuser miso systems exploiting deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1839-1850, Aug. 2020.
- [6] A. Taha, M. Alrabeiah and A. Alkhateeb, "Enabling large intelligent surfaces with compressive sensing and deep learning," *IEEE Access*, vol. 9, pp. 44304-44321, 2021.
- [7] S. Zhang, M. Li, M. Jian, Y. Zhao and F. Gao, "AIRIS: Artificial intelligence enhanced signal processing in reconfigurable intelligent surface communications," *China Communications*, vol. 18, no. 7, pp. 158-171, July 2021.
- [8] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2019, pp. 1-8, online access: www.deepmimo.net
- [9] C.H. Martin, T. Peng, and M.W. Mahoney, "Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data," *Nature Communications*, vol. 12, no 1, pp. 1-13, 2021.
- [10] Y. Seldin, C. Szepesvári, P. Auer, and Y. Abbasi-Yadkori, Y., "Evaluation and analysis of the performance of the exp3 algorithm in stochastic environments," *European Workshops on Reinforcement Learning (EWRL)*, December 2012, pp. 103-116.
- [11] M. Hutter, and J. Poland, "Adaptive online prediction by following the perturbed leader," *Journal of Machine Learning Research*, April 2005, pp.639-660.
- [12] W. Guo, "Explainable artificial intelligence for 6G: improving trust between human and machine," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 39-45, June 2020.
- [13] T.M. Hospedales, A. Antoniou, P. Micaelli and A.J. Storkey, "Meta-Learning in Neural Networks: A Survey," appear to *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3079209.