BLIND BEAMFORMING FOR INTELLIGENT REFLECTING SURFACE: A REINFORCEMENT LEARNING APPROACH

Wenhai Lai and Kaiming Shen

School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), China
E-mail: wenhailai@link.cuhk.edu.cn, shenkaiming@cuhk.edu.cn

ABSTRACT

The beamforming problem of intelligent reflecting surface (IRS) has been extensively considered from an optimization perspective assuming that channel state information (CSI) is available. However, the reality is that the existing prototypes seldom follow this model-based approach because channel estimation is technically difficult and costly for the network protocols and hardware to date. A recent trend is to perform beamforming blindly without channel knowledge, e.g., the so-called CSM method [1, 2]. This work looks at blind beamforming from a reinforcement learning point of view. We first show that CSM boils down to a special case of the greedy algorithm in the reinforcement learning context. We analyze the resulting cumulative regret, and further propose an upper approximation to facilitate the optimization of the exploration probability. Moreover, we show that a gradient sampling scheme can improve the efficiency of reinforcement learning as compared to the uniform sampling scheme used in CSM. Finally, we validate the performance advantage of the proposed methods in a prototype system.

1. INTRODUCTION

A major use of intelligent reflecting surface (IRS) is to coordinate phase shifts across its reflective elements (REs), i.e., *passive beamforming*, in order to maximize the received signal power at the target position. The existing approaches to passive beamforming typically entail channel state information (CSI), whereas two recent works [1, 2] suggest optimizing phase shifts blindly without channel acquisition. We further develop this new idea by means of reinforcement learning:

- We show that blind beamforming [1, 2] is equivalent to an ϵ-greedy algorithm for the multi-armed bandit problem.
- We improve the efficiency of blind beamforming by replacing uniform samples with gradient samples.

In the literature, passive beamforming has been considered extensively from a model-based optimization perspective. When CSI is available, the passive beamforming problem of IRS can be addressed by the standard methods, e.g., semidefinite relaxation (SDR) [3–7] and fractional programming (FP) [8–13]. But estimating channels for an IRS-assisted network is by no means trivial; a variety of sophisticated strategies have been proposed. In [14–16], the channel estimation is conducted assuming that each RE can be turned off, i.e., every OFF RE absorbs all signals falling on it. Another line of works [17–19] require introducing extra pilots according to the discrete Fourier transform (DFT) matrix. Assuming a certain type of sparse model for channels, [20] suggests a two-step method that first estimates the common column of AoD steering vectors and then solves the sparse matrix recovery problem. In particular, we mention that all the above methods entail full cooperation between the IRS and the base station.

However, the existing works [1, 2, 21, 22] that pursue prototype implementations of IRS rarely consider channel estimation, for the following three reasons. First, each reflected channel alone can be easily overwhelmed by the much stronger direct channel, and hence is difficult to measure. Second, the existing channel estimation methods for IRS are incompatible with the current networking protocols and hardware. Third, channel estimation requires computing the inverse of a coefficient matrix whose size equals the number of REs, so it becomes costly when dealing with a huge IRS.

The above issues can be completely resolved by the blind beamforming method in [1, 2]. Its main idea follows: Choose phase shifts for all REs at random, then use the average performance conditioned on a particular choice of phase shift for the nth RE to characterize the goodness of this phase shift decision, and finally choose phase shift for each RE n to maximize the corresponding conditional average performance. While [1, 2] give the statistical motivation behind blind beamforming, this work shows that blind beamforming can be interpreted as an ϵ-greedy algorithm [23] for solving the multi-armed bandit problem in the reinforcement learning field; the key step is to figure "cumulative regret" in our problem case. Following this line, we further use the gradient bandit method [24] to improve the efficiency of blind beamforming. Intuitively, rather than trying out new samples uniformly, we propose trying out those more "promising" samples with higher priorities so as to attain the optimum more quickly.

This work was supported by NSFC under Grant 92167202.

2. SYSTEM MODEL

Consider an IRS-assisted network in which the IRS comprises N REs. Denote by $h_0 \in \mathbb{C}$ the direct channel from the transmitter to the receiver, $h_{n,t} \in \mathbb{C}$ the channel from the transmitter to RE n, and $h_{n,r} \in \mathbb{C}$ the channel from RE n to the receiver, for $n = 1, \ldots, N$. The cascaded reflected channel $h_n \in \mathbb{C}$ of RE n is then given by $h_n = h_{n,r} \times h_{n,t}$. Assume that the configuration of IRS can be updated on a per block basis; let $\theta_n^t \in [0, 2\pi)$ be the phase shift of RE n in block t. Assume also that there are a total of T blocks and the channels are invariant throughout. In practice, the value of each phase shift is restricted to a prescribed discrete set

$$\Phi_K = \{0, \omega, \ldots, (K-1)\omega\} \;\text{ with }\; \omega = \frac{2\pi}{K}, \tag{1}$$

where the set size $K \geq 2$. For the transmit signal power P and the background noise power σ^2, the SNR in block t can be computed as a function of the corresponding phase shift array $\boldsymbol{\theta}^t = (\theta_1^t, \ldots, \theta_N^t)$ as

$$\mathrm{SNR}(\boldsymbol{\theta}^t) = \frac{P\left|h_0 + \sum_{n=1}^N h_n e^{j\theta_n^t}\right|^2}{\sigma^2}. \tag{2}$$

Roughly speaking, our goal is to maximize SNR for all the T blocks, which will be mathematically formalized in the context of reinforcement learning. Finally, we emphasize that the channels are unknown *a priori*.

3. A REINFORCEMENT LEARNING PERSPECTIVE

Should we alter the phase shifts when entering a new block? This question becomes trivial if CSI is already known. Since the channels are constant, we can just optimize the phase shift for the first block and fix the solution thereafter. Regarding the SNR maximization with CSI available, a common idea [25–28] is to rotate every h_n to the closest position to h_0, i.e.,

$$\theta_n^\star = \min_{\varphi \in \Phi_K} \left| \mathsf{Arg}\left(\frac{h_n e^{j\varphi}}{h_0}\right)\right|, \tag{3}$$

where Arg is the principal argument, namely the closest point projection (CPP).

But the reality is that the CSI is unknown. It turns out that the CPP solution can be recovered implicitly via random sampling. Specifically, as proposed in [1,2], we try out a new sample $\boldsymbol{\theta}^t$ with each θ_n^t drawn from Φ_K i.i.d. and then record the corresponding $\mathrm{SNR}(\boldsymbol{\theta}^t)$. After T_0 samples, we compute the conditional sample mean (CSM)

$$Q_n^{T_0}(\varphi) = \frac{\sum_{i=1}^{T_0} \mathrm{SNR}(\boldsymbol{\theta}^i) \cdot \mathbb{1}_{\theta_n^i = \varphi}}{\sum_{i=1}^{T_0} \mathbb{1}_{\theta_n^i = \varphi}} \tag{4}$$

for each $\varphi \in \Phi_K$. The blind beamforming method in [1, 2] is to choose each phase shift θ_n to maximize the CSM, i.e., $\theta_n = \arg\max_{\varphi \in \Phi_K} Q_n^{T_0}(\varphi)$. The main result in [1,2] follows:

Proposition 1 (Theorem 2 in [2]). *As $T_0 \to \infty$, the solution $\theta_n = \arg\max_{\varphi \in \Phi_K} Q_n^{T_0}(\varphi)$ tends to the CPP solution $\theta_n^\star$ in* (3).

But we cannot let $T_0 \to \infty$ since the total number of blocks T is finite. A practical realization of blind beamforming in our case is to use $T_0 < T$ blocks to learn the CSM values $\{Q_n^{T_0}(\varphi)\}$, namely training. For the rest blocks, we decide $\boldsymbol{\theta}^t$ by the blind beamforming method [1, 2] with the estimated $\{Q_n^{T_0}(\varphi)\}$. Thus, the above blind beamforming method can be recognized as a reinforcement learning approach: *explore* new solutions for the first T_0 blocks and *exploit* the CSM so far for the remaining $T - T_0$ blocks.

Furthermore, our problem can be thought of as a multi-armed bandit problem with each possible solution $\boldsymbol{\theta}^t \in \Phi_K^N$ treated as an arm and the received SNR under $\boldsymbol{\theta}^t$ is treated as reward. Following this line, we can generalize the CSM blind beamforming method [1,2] as an ϵ-greedy algorithm—which is a classical approach to the multi-armed bandit problem:

$$\theta_n^t = \begin{cases} \arg\max_{\varphi \in \Phi_K} Q_n^t(\varphi) \text{ with probability } 1-\epsilon; \\ \varphi \sim \mathrm{Uniform}(\Phi_K) \text{ with probability } \epsilon. \end{cases} \tag{5}$$

In contrast to the CSM blind beamforming method in [1, 2] that uses the first T_0 blocks for exploration, we now decide the use of each block t, exploration or exploitation, randomly.

The above interpretation further leads us to the definition of *regret* as considered in the reinforcement learning field. Since the CPP solution $\{\theta_n^\star\}$ is the ultimate goal and it can be obtained from $Q_n^{T_0}(\varphi)$ with $T_0 \to \infty$, we may consider the following regret for the action $\boldsymbol{\theta}^t$ in block t:

$$G_t = \sum_{n=1}^N \left(Q_n^\infty(\theta_n^\star) - Q_n^\infty(\theta_n^t)\right). \tag{6}$$

It can be shown that G_t must be nonnegative and it equals 0 if $\theta_n^t = \theta_n^\star$ for every n. We seek the optimal $\{\boldsymbol{\theta}^t\}$ to minimize the expected cumulative regret over the T blocks:

$$\underset{\{\boldsymbol{\theta}^t\}}{\text{minimize}} \quad \mathbb{E}\left[\sum_{t=1}^T G_t\right] \tag{7a}$$

$$\text{subject to} \quad \theta_n^t \in \Phi_K, \text{ for each } (n,t). \tag{7b}$$

The ϵ-greedy algorithm in (5) constitutes the standard approach to the above problem in the reinforcement learning field. But we raise two questions: What is the optimal value of ϵ in (5)? Is there a better way of exploration than the uniform sampling?

4. PROPOSED BLIND BEAMFORMING METHODS

4.1. Parameter Tuning for ϵ-Greedy Algorithm

There is a trade-off in deciding the value of $0 < \epsilon < 1$. If $\epsilon \to 0$, then the distortion between the real CSM $Q_n^\infty(\varphi)$ and

the estimated CSM $Q_n^t(\varphi)$ can be large, so the resulting blind beamforming solution is of low quality. Conversely, if $\epsilon \to 1$, then most blocks are used for exploration and consequently SNR should be fairly low for most of the time.

But optimizing ϵ directly in (7) is quite difficult. The following proposition provides an upper approximation of regret that facilitates the optimization significantly.

Proposition 2. *For the ϵ-greedy method, the expected value of the cumulative regret over T blocks is upper bounded as*

$$\mathbb{E}\left[\sum_{t=1}^{T} G_t\right] \leq \rho T N \left(\frac{2K}{T^4} + \epsilon\right) + 2N\rho(1-\epsilon)\sqrt{\frac{2KT\log T}{\epsilon}} \tag{8}$$

given any parameter $0 < \epsilon < 1$, where $\rho = \sum_{n=0}^{N} |h_n|^2$.

Proof. We begin with an exploitation block t. Defining an event for each $\varphi \in \Phi_K$ as

$$\mathcal{E}(\varphi) = \left\{ \left| Q_n^t(\varphi) - Q_n^\infty(\varphi) \right| > \rho\sqrt{\frac{2K\log T}{\epsilon \cdot T}} \right\}, \tag{9}$$

we show that

$$\begin{aligned}
&\mathbb{P}\left\{ \left| Q_n^t(\varphi) - Q_n^\infty(\varphi) \right| \leq \rho\sqrt{\frac{2K\log T}{\epsilon \cdot T}},\ \forall \varphi \in \Phi_K \right\} \\
&= 1 - \mathbb{P}\Big\{ \bigcup\nolimits_{\varphi \in \Phi_K} \mathcal{E}(\varphi) \Big\} \\
&\overset{(a)}{\geq} 1 - \sum_{\varphi \in \Phi_K} \mathbb{P}\left\{ \mathcal{E}(\varphi) \right\} \\
&\overset{(b)}{\geq} 1 - \frac{2K}{T^4},
\end{aligned} \tag{10}$$

where (a) follows by the union bound while (b) follows by Hoeffding's inequality. Moreover, the exploitation in (5) guarantees that

$$Q_n^t(\theta_n^t) \geq Q_n^t(\theta_n^\star). \tag{11}$$

Combining (10) and (11), we have

$$\mathbb{P}\left\{ Q_n^\infty(\theta_n^\star) - Q_n^\infty(\theta_n^t) \leq 2\rho\sqrt{\frac{2K\log T}{\epsilon \cdot T}} \right\} \geq 1 - \frac{2K}{T^4}. \tag{12}$$

We now consider a general block $t = 1, \ldots, T$. Clearly, it always holds that

$$Q_n^\infty(\theta_n^\star) - Q_n^\infty(\theta_n^t) \leq \rho. \tag{13}$$

Thus, the expected value of $Q_n^\infty(\theta_n^\star) - Q_n^\infty(\theta_n^t)$ can be bounded from above as

$$\begin{aligned}
&\mathbb{E}\left[Q_n^\infty(\theta_n^\star) - Q_n^\infty(\theta_n^t) \right] \\
&\leq \epsilon\rho + (1-\epsilon)\left(\left(1 - \frac{2K}{T^4}\right) 2\rho\sqrt{\frac{2K\log T}{\epsilon \cdot T}} + \frac{2K}{T^4}\rho \right) \\
&\leq \epsilon\rho + 2(1-\epsilon)\rho\sqrt{\frac{2K\log T}{\epsilon \cdot T}} + \frac{2K}{T^4}\rho.
\end{aligned} \tag{14}$$

Summing both sides of (14) gives rise to (8). □

We now use the upper approximation in (8) to replace the objective function in (7). The resulting new problem turns out to be convex in ϵ, so it suffices to solve the first-order condition:

$$\sqrt{\epsilon^3 T^3} - T(1+\epsilon)\sqrt{2K\log T} = 0. \tag{15}$$

Further, it can be shown that the above first-order equation has only one real root

$$\epsilon^\star = \sqrt{\frac{2K\log T}{9T}} + \sqrt[3]{a + \sqrt{a^2 + b^3}} + \sqrt[3]{a - \sqrt{a^2 + b^3}}, \tag{16}$$

where $a = \left(1 + \frac{4K\log T}{27T}\right)\sqrt{\frac{K\log T}{2T}}$ and $b = \frac{2K\log T}{9T}$.

4.2. Gradient Sampling

Our discussion thus far is based on uniform sampling—which is questionable since it does not take the past performance into account when choosing a new $\boldsymbol{\theta}^t$ for exploration. For instance, if using a particular phase shift $\varphi \in \Phi_K$ for RE n always results in poor performance, then it is advisable to avoid this setting in future explorations.

The gradient bandit method [24] can address the above issue, as stated in what follows. Introduce a *numerical preference* variable $H_n^t(\varphi)$ for each (n, φ) in block t to characterize the "goodness" of setting $\theta_n^t = \varphi$. With the probabilities

$$\mathbb{P}\left\{\theta_n^t = \varphi\right\} \triangleq \frac{e^{H_n^t(\varphi)}}{\sum_{\varphi' \in \Phi_K} e^{H_n^t(\varphi')}} \triangleq \pi_n^t(\varphi), \tag{17}$$

we make a soft decision for each θ_n^t as

$$\theta_n^t = \varphi \text{ with probability } \pi_n^t(\varphi). \tag{18}$$

It remains to specify how each $H_n^t(\varphi)$ is obtained. By convention, every $H_n^0(\varphi)$ is set to zero. Sequentially, each $H_n^t(\varphi)$ is computed based on the previous $\{H_n^{t-1}(\cdot), \pi_n^{t-1}(\varphi)\}$ and the past rewards as

$$H_n^t(\theta_n^t) = H_n^{t-1}(\theta_n^t) + \gamma\delta_t(1 - \pi_n^{t-1}(\theta_n^t)), \tag{19a}$$

$$H_n^t(\varphi) = H_n^{t-1}(\varphi) - \gamma\delta_t\pi_n^{t-1}(\varphi),\ \forall \varphi \neq \theta_n^t, \tag{19b}$$

where $\gamma > 0$ is the learning rate and

$$\delta_t = \mathrm{SNR}(\boldsymbol{\theta}^t) - \frac{1}{t}\sum_{i=1}^{t} \mathrm{SNR}(\boldsymbol{\theta}^i) \tag{20}$$

is the difference between the latest reward (which is the SNR value) and the average reward so far. Note that we no longer distinguish between exploration and exploitation when applying the gradient bandit method.

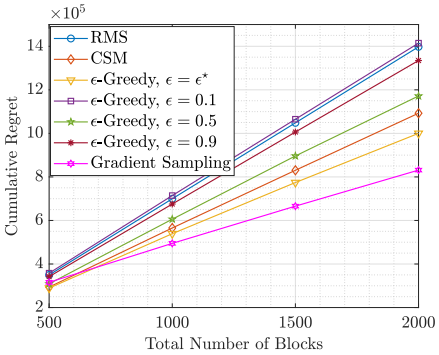

Fig. 1. Cumulative regret.

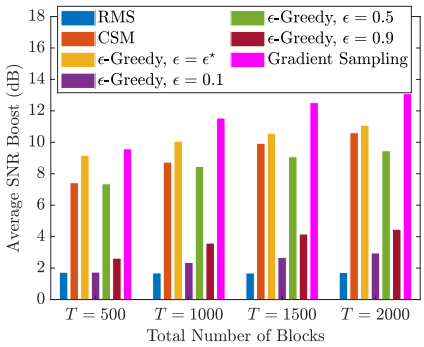

Fig. 2. Average SNR boost.

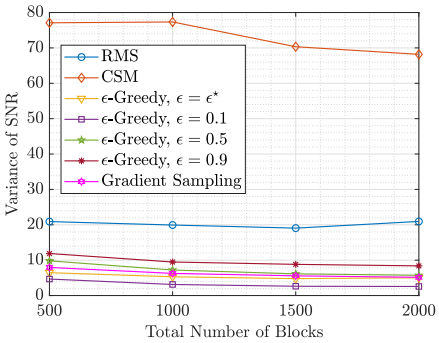

Fig. 3. Variance of SNR.

Fig. 4. Field test site.

Table 1. The values of $\epsilon^\star$ under different T.

T	500	1000	1500	2000
$\epsilon^\star$	0.478	0.374	0.325	0.295

5. FIELD TESTS

We conducted the field tests in an indoor scenario as shown in Fig. 4. The transmit power is -5 dBm; the carrier frequency is 2.6 GHz; antenna gain is 14.88 dBi. Our IRS has 294 REs with $\Phi_K = \{0, \pi\}$. Directional antennas are deployed at both the transmitter and the receiver. We set the learning rate $\gamma = 0.05$ for the gradient sampling method in Section 4.2. Regarding the ϵ-greedy algorithm in Section 4.1, we obtain $\epsilon^\star$ for the different T values according to (16) as shown in Table 1; we shall also try out other values of ϵ for the comparison purpose. For the CSM method in Section 3, we let $T_0 = \lceil \epsilon^\star T \rceil$. We also consider a baseline method called *random max sampling (RMS)*—which replaces the exploitation step in (5) by using the best phase shift vector so far, i.e., $\boldsymbol{\theta}^t = \arg\max_{\boldsymbol{\theta} \in \{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^{t-1}\}} \mathrm{SNR}(\boldsymbol{\theta})$. Moreover, we take $Q_n^{3000}(\varphi)$ as an approximation of $Q_n^{\infty}(\varphi)$.

Fig. 1 shows the cumulative regrets achieved by the various methods. Observe that the gradient sampling method outperforms the rest methods significantly; the advantage increases with T. Observe also that the ϵ-greedy with the optimized $\epsilon^\star$ has much better performance than the other ϵ cases. It is worth pointing out that ϵ-greedy is slightly better than CSM when they are given the same optimized exploration probability $\epsilon^\star$.

Fig. 2 displays the average SNR boost (as compared to the case without IRS) across T blocks for the different methods. It shows that the gradient sampling yields the highest SNR, e.g., its SNR is approximately 11 dB higher than that of RMS when $T = 2000$. Again, we observe that the performance of ϵ-greedy is sensitive to the choice of ϵ. When $\epsilon = 0.1$, ϵ-greedy is almost as poor as RMS. But when $\epsilon^\star$ is adopted, ϵ-greedy outperforms CSM.

Moreover, Fig. 3 shows the variance of SNR over time. We see that the variance increases with ϵ for ϵ-greedy; clearly, the SNR becomes less stable if we perform exploration more frequently. RMS is less stable than ϵ-greedy and gradient sampling. In particular, observe that CSM leads to the highest instability of SNR, even though it can asymptotically converge to the CPP solution. This is due to the fact that it does not randomize the exploration and the exploitation blocks.

6. CONCLUSION

Blind beamforming for IRS without channel estimation is of great practical importance. This work considers blind beamforming from a reinforcement learning perspective, showing that the existing CSM method [1, 2] can be recognized as a special case of the ϵ-greedy algorithm for the multi-armed bandit problem. We further propose an upper approximation of the cumulative regret and thereby optimize the exploration probability ϵ in closed form. Moreover, we suggest a gradient sampling scheme that is more efficient than the uniform sampling in the ϵ-greedy method. Field tests at 2.6 GHz demonstrate these reinforcement learning approaches.

7. REFERENCES

[1] V. Arun and H. Balakrishnan, "RFocus: beamforming using thousands of passive antennas," in *USENIX Symp. Netw. Sys. Design Implementation (NSDI)*, Feb. 2020, pp. 1047–1061.

[2] S. Ren, K. Shen, Y. Zhang, X. Li, X. Chen, and Z.-Q. Luo, "Configuring intelligent reflecting surface with performance guarantees: Blind beamforming," *IEEE Trans. Wireless Commun.*, vol. 22, no. 5, pp. 3355–3370, May 2023.

[3] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, Apr. 2010.

[4] G. Zhou, C. Pan, H. Ren, K. Wang, M. Di Renzo, and A. Nallanathan, "Robust beamforming design for intelligent reflecting surface aided MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1658–1662, Jun. 2020.

[5] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 234–238, Jan. 2020.

[6] H. Xie, J. Xu, and Y.-F. Liu, "Max-min fairness in IRS-aided multi-cell MISO systems with joint transmit and reflective beamforming," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1379–1393, Feb. 2020.

[7] S. Huang, Y. Ye, M. Xiao, H. V. Poor, and M. Skoglund, "Decentralized beamforming design for intelligent reflecting surface-enhanced cell-free networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 673–677, Jun. 2020.

[8] K. Shen and W. Yu, "Fractional programming for communication systems—part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[9] K. Feng, X. Li, Y. Han, S. Jin, and Y. Chen, "Physical layer security enhancement exploiting intelligent reflecting surface," *IEEE Commun. Lett.*, vol. 25, no. 3, pp. 734–738, Mar. 2020.

[10] J. Zhu, Y. Huang, J. Wang, K. Navaie, and Z. Ding, "Power efficient IRS-assisted NOMA," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 900–913, Feb. 2020.

[11] T. Shafique, H. Tabassum, and E. Hossain, "Optimization of wireless relaying with flexible UAV-borne reflecting surfaces," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 309–325, Jan. 2020.

[12] Z. Zhang, L. Dai, X. Chen, C. Liu, F. Yang, R. Schober, and H. V. Poor, "Active RIS vs. passive RIS: Which will prevail in 6G?" *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1707–1725, Mar. 2022.

[13] Z. Zhang and L. Dai, "A joint precoding framework for wideband reconfigurable intelligent surface-aided cell-free network," *IEEE Trans. Signal Process.*, vol. 69, pp. 4085–4101, Jun. 2021.

[14] D. Mishra and H. Johansson, "Channel estimation and low-complexity beamforming design for passive intelligent surface assisted MISO wireless energy transfer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2019, pp. 4659–4663.

[15] 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, Jul. 2020.

[16] S. Lin, B. Zheng, G. C. Alexandropoulos, M. Wen, M. Di Renzo, and F. Chen, "Reconfigurable intelligent surfaces with reflection pattern modulation: Beamforming design and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 741–754, Feb. 2021.

[17] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 518–522, Apr. 2019.

[18] T. L. Jensen and E. De Carvalho, "An optimal channel estimation scheme for intelligent reflecting surfaces based on a minimum variance unbiased estimator," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2020.

[19] G. T. de Araújo, A. L. De Almeida, and R. Boyer, "Channel estimation for intelligent reflecting surface assisted MIMO systems: A tensor modeling approach," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 3, pp. 789–802, Apr. 2021.

[20] J. Chen, Y.-C. Liang, H. V. Cheng, and W. Yu, "Channel estimation for reconfigurable intelligent surface aided multi-user mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, 2023, to be published.

[21] 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," *IEEE Trans. Commun.*, vol. 69, no. 12, pp. 8627–8640, Dec. 2021.

[22] P. Staat, S. Mulzer, S. Roth, V. Moonsamy, M. Heinrichs, R. Kronberger, A. Sezgin, and C. Paar, "IRShield: A countermeasure against adversarial physical-layer wireless sensing," in *IEEE Symp. Secur. Priv. (SP)*, May 2022.

[23] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, King's College, 1989.

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

[25] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1838–1851, Mar. 2020.

[26] C. You, B. Zheng, and R. Zhang, "Channel estimation and passive beamforming for intelligent reflecting surface: Discrete phase shift and progressive refinement," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2604–2620, Nov. 2020.

[27] 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, Sep. 2020.

[28] P. Wang, J. Fang, X. Yuan, Z. Chen, and H. Li, "Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 14 960–14 973, Dec. 2020.