Two-Stage Resource Allocation in Reconfigurable Intelligent Surface Assisted Hybrid Networks via Multi-Player Bandits

Jingwen Tong, *Student Member, IEEE,* Hongliang Zhang, *Member, IEEE,* Liqun Fu, *Senior Member, IEEE,* Amir Leshem, *Senior Member, IEEE,* and Zhu Han, *Fellow, IEEE*

Abstract—This paper considers a resource allocation problem where several Internet-of-Things (IoT) devices send data to a base station (BS) with or without the help of the reconfigurable intelligent surface (RIS) assisted cellular network. The objective is to maximize the sum rate of all IoT devices by finding the optimal RIS and spreading factor (SF) for each device. Since these IoT devices lack prior information of the RISs or the channel state information (CSI), a distributed resource allocation framework with low complexity and learning features is required to achieve this goal. Therefore, we model this problem as a two-stage multi-player multi-armed bandit (MPMAB) framework to learn the optimal RIS and SF sequentially. Then, we put forth an exploration and exploitation boosting (E2Boost) algorithm to solve this two-stage MPMAB problem by combining the ϵ-greedy algorithm, Thompson sampling (TS) algorithm, and non-cooperation game method. We derive an upper regret bound for the proposed algorithm, i.e., $\mathcal{O}(\log_2^{1+\delta} T)$, increasing logarithmically with the time horizon T. Numerical results show that the E2Boost algorithm has the best performance among the existing methods and exhibits a fast convergence rate. More importantly, the proposed algorithm is not sensitive to the number of combinations of the RISs and SFs thanks to the two-stage allocation mechanism, which can benefit the high-density networks.

Index Terms—Reconfigurable intelligent surface (RIS), Internet-of-Things (IoT), multi-player multi-armed bandit (MPMAB), Thompson sampling (TS), exploration and exploitation boosting (E2Boost) algorithm.

I. INTRODUCTION

Reconfigurable intelligent surface (RIS), which enhances the communication quality by adjusting the amplitude and the phase shift of the incident signal on a 2D planar surface with massive low-cost passive reflecting elements, has drawn increasing attention in future communication networks [1]–[3]. There have existed some works accounting for this vision by studying the performance of the RIS-assisted cellular network [4], [5], RIS-assisted unmanned aerial vehicle network [6], and RIS-assisted secure wireless communications [7].

J. Tong and L. Fu are with the Department of Information and Communication Engineering, Xiamen University, Xiamen 361005, China (e-mails: tongjingwen@stu.xmu.edu.cn; liqun@xmu.edu.cn).

H. Zhang is with the Department of Electrical and Computer Engineering, Princeton University, NJ, USA (e-mail: hongliang.zhang92@gmail.com).

A. Leshem is with the Faculty of Engineering, Bar-Ilan University, Ramat Gan 52900, Israel (e-mail: leshema@biu.ac.il).

Z. Han is with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701, USA (e-mail: zhan2@uh.edu).

Meanwhile, the cellular Internet-of-Things (C-IoT) with RIS is regarded as one of the paradigms in future communication networks, providing the capabilities of low-cost, large-scale, and ultra-durable connectivity for everything [8]–[10]. By employing the LoRa (short for Long Range) technology, C-IoT can operate on the unlicensed band since the resulting signal has substantial anti-interference properties [11]. On the other hand, C-IoT can achieve the rate adaptation by employing the chirp spreading spectrum modulation at the physical layer with different spreading factors (SFs) [12]. However, the study of the network-level performance of these C-IoT devices in the RIS-assisted hybrid cellular network still needs more research.

In light of this, we consider a hybrid uplink network where several C-IoT devices transmit data to a base station (BS) by opportunistically accessing the RIS-assisted cellular network. The goal is to maximize the sum rate of all C-IoT devices by finding the best RIS and SF for each device. Although these C-IoT devices can directly send data to the BS, a higher SF that corresponds to a lower data rate will be assigned to combat the harsh channel environment or to enable a long-range transmission [13]. As pointed out in [9], the low data rate will result in high data latency and security problems. Therefore, these C-IoT devices may opportunistically access the vacant RISs to improve their data rate by reflecting their signal to the BS. However, finding the optimal RIS and collecting the exact channel state information (CSI) are challenging for these C-IoT devices. On the one hand, the C-IoT device has no information (e.g., the phase shifts) about the RISs since they are deployed for cellular users (UEs). On the other hand, there is no communication among C-IoT devices in such distributed network. These features render most traditional optimization methods infeasible in this resource allocation problem, such as the convex optimization methods [14] and the combinatorial optimization methods [15].

To overcome the above impediments, the learning theory has been considered in [8], [11], [16]–[18] to address this problem by sequentially exploring all actions and automatically exploiting the best action. Refs. [16], [17] study the distributed resource allocation problem in wireless networks by formulating ir as a Markov decision processing (MDP) problem. Then, the authors propose the multi-agent reinforcement learning (RL) based method to solve this MDP problem. Unfortunately, these solutions often suffer from the issues of the curse of dimensionality, lack of performance guarantee (*e.g.*, the unknown convergence rate), and high computational complexity [18]. As pointed out in [8], [11], low complexity and fast convergence resource allocation algorithms are crucial

for the energy-constrained IoT devices in future communication networks.

This inspires us to consider the multi-armed bandit (MAB) technique. MAB is a basic framework for the sequential decision-making problem [19]. In the classic MAB setting, at each time round, a decision-maker (or player) must select an arm from a set of arms (or arm space) with unknown distribution. After that, the player will observe a reward from the environment (or the unknown distribution). The goal is to minimize the pseudo-regret that is defined as the difference between the mean rewards of the optimal arm and the currently selected arm. During this process, the player faces an exploration and exploitation (EE) dilemma. On the one hand, the player needs to explore the arm space sufficiently to ensure its long-term performance (i.e., not miss the optimal arm); on the other hand, it needs to exploit the current best arm as many times as possible to maximize its total rewards. Compared with the other learning-based methods, MAB has a theoretical guarantee (i.e., regret bound) and low computational complexity, and it is easy to implement.

Recently, the multi-player MAB (MPMAB) framework has gained much attention in wireless communications [20]–[23]. Ref. [20] studies the SF allocation problem in the LoRa network by devising a fully distributed MPMAB framework. It solves this MPMAB problem by using the Exponential-weight algorithm for Exploration and Exploitation (Exp3) [24] algorithm. However, the solution of the Exp3 algorithm is selfish in that it cannot guarantee the optimal allocation for each device. The optimal MPMAB framework is considered in [21], where different players contend for the same set of channels in an ad-hoc network. Based on the Hungarian algorithm [25], the authors propose a probably approximately correct (PAC) based MPMAB algorithm to estimate the CSI matrix sequentially. However, the PAC-based MPMAB algorithm requires players to exchange messages, leading to extra signaling in the system. The fully distributed resource allocation framework with the optimal solution is investigated in [22] and [23]. Ref. [22] aims to maximize the sum rate of all users by combining the MAB algorithm and the auction algorithm. A more general version of the distributed MPMAB framework named the game-of-thrones (GoT) algorithm has been proposed in [23]. The authors intend to find the optimal assignment for each player by combining the MAB algorithm and the game theory. However, algorithms in [22] and [23] suffer from low convergence rate, especially when the arm space is large.

In this paper, we propose a two-stage MPMAB framework to attack this resource allocation problem in the hybrid uplink network. In this two-stage MPMAB framework, players are the IoT devices; arms are the RISs in the first stage and the SFs in the second stage, respectively. We assume that two or more players who select the same RIS will observe a collision and receive zero reward. This resource allocation problem is quite different from that in [21] and [22], because it not only needs to learn the CSI but also the phase shifts of the RISs. Moreover, the ascending order in the set of SFs and the corresponding descending order in the successful transmission probabilities enable us to devise a two-stage
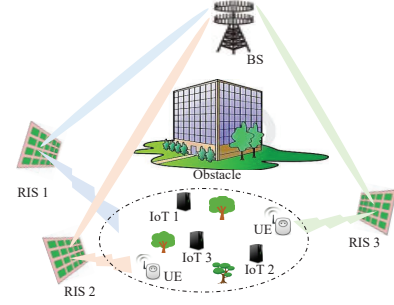

Fig. 1: A RIS-assisted hybrid uplink network.

MPMAB framework. To address this two-stage MPMAB problem, we put forth an exploration and exploitation boosting (E2Boost) algorithm by combining the game theory and the MAB algorithm. The E2Boost algorithm proceeds in epochs and has three phases, i.e., ϵ-greedy EE phase, non-cooperation game phase, and Thompson sampling (TS) EE phase. Each phase contains a specific mechanism to tradeoff the EE dilemma. That is why we call it the E2Boost algorithm. In addition, we derive an upper pseudo-regret bound for the E2Boost algorithm, i.e., $\mathcal{O}(\log_2^{1+\delta} T)$ where $0 \leq \delta < 1$, indicating that the per-round regret will trend to 0 when the time horizon T is sufficiently large. More importantly, this upper regret bound is about M times lower than that in the GoT algorithm, where M is the number of SFs. In other words, the proposed algorithm is not sensitive to the number of combinations of the RISs and SFs, which can benefit high-density networks.

The difference between this work and the existing ones and the main contributions of this work are summarized as follows.

- The E2Boost algorithm embeds the ϵ-greedy algorithm [26] in the first phase to reduce the regrets generated from the uniform exploration. Specifically, we use the Wasserstein distance (WD) [27] to measure the convergence rate of the second phase. In return, this measurement is regarded as a criterion to optimize the parameter ϵ.

- The E2Boost algorithm adopts the TS algorithm [28], [29] in the third phase to determine the best SF. Since the only observed information is the success or failure transmission feedback, the TS algorithm maintains a Beta distribution on the successful transmission probability of each SF. For the Bernoulli reward processing, the TS algorithm accounts for the best performance among the existing stochastic MAB algorithms [29].

- The E2Boost algorithm has a smaller arm space to explore than the GoT algorithm. Thanks to the two-stage allocation mechanism, the E2Boost algorithm only needs to explore the sets of the RISs or the SFs. In contrast, the GoT algorithm requires exploring the combinations of the RISs and SFs.

The remainder of this paper is organized as follows. In Section II, we introduce the channel model and the achievable data rate. The problem formulation is given in Section III. In Section IV, we introduce the two-stage MPMAB framework for this joint resource assignment problem. The E2Boost algorithm is presented in Section V. Numerical results are

given in Section VI to evaluate the proposed algorithm. This paper is concluded in Section VII.

II. SYSTEM MODEL

We consider a hybrid uplink cellular network, as shown in Fig. 1, where several UEs and N IoT devices are located in an area. Both UE and IoT device need to transmit data to the BS. Since there may exist some obstacles (*e.g.*, buildings) between UEs and the BS, the signal will experience deep-fading. Thus, K RISs are deployed to reflect the UEs' signals to the BS by adjusting the RISs' phase shifts. These RISs are operated over different frequencies[1]. The N IoT device has no information about these RISs, but it may opportunistically access these RISs to improve its data rate. In this hybrid network, UE is the legal user to communicate with the BS through the RIS; while the IoT device requires to perform spectrum sensing[2] before access to the vacant RIS. Time is slotted in $t = 1, 2, \ldots, T$. At each time slot, we assume that one RIS can serve multiple UEs but can be exploited by only one IoT device. The reason is that the BS requires the precoding and beamforming vectors to maintain communication quality in this multi-user RIS-assisted system [30]. These vectors often contain the information of the CSI and the RISs' phase shifts determined by the UEs. As a result, a RIS can only support one IoT device since the IoT device lacks these precoding and beamforming vectors.

A. Channel Models

There are two transmission patterns for each IoT device in this hybrid network. The first one is RIS-assisted transmission pattern (Pattern I), where the IoT device transmits to the BS through the RIS when the target RIS is detected in an idle state. The second one is non-RIS-assisted transmission pattern (Pattern II), where the IoT device directly transmits to the BS with a low data rate if the target RIS is detected in a busy state.

Pattern I: Assume that each element on the RIS is equipped with b PIN diodes, producing 2^b phase shifts in $[0, 2\pi)$ by controlling the ON/OFF state of each diode. Hence, the available phase shift at the (l_1, l_2)-th element is

$$\tau_{l_1, l_2} = \frac{\pi \rho_{l_1, l_2}}{2^{b-1}}, \tag{1}$$

where (l_1, l_2) is the index of the RIS elements' matrix and ρ_{l_1, l_2} is an integer in $[0, 2^b - 1]$. Let A_{l_1, l_2} be the reflection factor at the l_1-th row and l_2-th column of the RIS elements' matrix, which is defined as

$$A_{l_1, l_2} = Ae^{-j\tau_{l_1, l_2}}, \tag{2}$$

where A is a reflection amplitude with a constant value among $(0, 1]$[3].

By taking advantage of the directional reflections of the RIS, the BS - RIS - IoT device link is usually stronger

than other multi-path as well as the deep-fading direct link between the BS and the IoT device [7]. Therefore, we model the channel between the BS and the IoT device as a Ricean model. In this way, the BS - (RIS k) - (IoT device n) link acts as the dominant "LoS" component; while all the other paths together form the "non-LoS (NLoS)" component. Hence, the RIS-assisted channel model $h_{l_1, l_2}^{n,k}$ is defined as

$$h_{l_1, l_2}^{n,k} = \sqrt{\frac{\zeta}{\zeta + 1}} \tilde{h}_{l_1, l_2}^{n,k} + \sqrt{\frac{1}{\zeta + 1}} \hat{h}_{l_1, l_2}^{n,k}, \tag{3}$$

where $\tilde{h}_{l_1, l_2}^{n,k}$ and $\hat{h}_{l_1, l_2}^{n,k}$ are the LoS component and the NLoS component with the k-th RIS and the n-th IoT device through the (l_1, l_2)-th element, respectively. Symbol ζ is the Rician factor, indicating the ratio of the LoS component to the NLoS component. In the following, we omit the IoT device index n and the RIS index k in the superscript if no confusion occurs.

Let D_{l_1, l_2} be the distance between the BS and the (l_1, l_2)-th RIS element, and let d_{l_1, l_2} be the distance between the (l_1, l_2)-th RIS element and the IoT device. The transmission distance of BS - $((l_1, l_2)$-th RIS element) - (IoT device n) link is $L_{l_1, l_2} = D_{l_1, l_2} + d_{l_1, l_2}$. According to [4], the LoS component of this link is given by

$$\tilde{h}_{l_1, l_2} = \sqrt{GD_{l_1, l_2}^{-\iota} d_{l_1, l_2}^{-\iota}} e^{-j\frac{2\pi}{\lambda} L_{l_1, l_2}}$$
$$= \sqrt{G} \left[\sqrt{D_{l_1, l_2}^{-\iota}} e^{-j\frac{2\pi}{\lambda} D_{l_1, l_2}} \right] \left[\sqrt{d_{l_1, l_2}^{-\iota}} e^{-j\frac{2\pi}{\lambda} d_{l_1, l_2}} \right], \tag{4}$$

where ι is the path-loss exponent. Symbol G is the antenna gain and λ is the wave length of the signal. Meanwhile, the NLoS component is given by

$$\hat{h}_{l_1, l_2} = \sqrt{PL_{\mathrm{NLoS}}(L_{l_1, l_2})} g_{l_1, l_2}, \tag{5}$$

where g_{l_1, l_2} is the small-scale NLoS component, following the i.i.d. complex Gaussian distribution, i.e., $g_{l_1, l_2} \sim \mathcal{CN}(0, 1)$. Term $PL_{\mathrm{NLoS}}(\cdot)$ is the NLoS channel power gain that we adopt the urban macro (UMa) path-loss model[4] [32] in the simulation.

Pattern II: In non-RIS-assisted transmission pattern, the IoT device directly transmits to the BS without the help of the RIS. Since there are some obstacles, the signal may experience deep fading. Thus, we use the shadow fading model to describe the channel between IoT device n and the BS [33], i.e.,

$$h_n = \sqrt{\varrho_n} g_n, \tag{6}$$

where ϱ_n is the channel power gain, following the i.i.d. log-normal distribution with mean μ_n and standard deviation σ_n of $\ln(\varrho_n)$. The typical value of σ_n is between 6 and 12 dB for practical radio channels [32]. In addition, g_n is a small-scale NLoS component, following i.i.d. complex Gaussian distribution, i.e., $g_n \sim \mathcal{CN}(0, 1)$.

---

[1]The RIS can operate at different frequencies by changing the location and the wave-number of each element [5].

[2]If the received signal strength (RSS) exceeds a threshold, the IoT device marks this RIS with the busy state; otherwise, the state of the RIS is idle.

[3]The reflection amplitude can be a function of the phase shift as in [31].

[4]The calculation of $PL_{\mathrm{NLoS}}(\cdot)$ in dB form is $10 \log_{10} PL_{\mathrm{NLoS}}(d) = 13.54 + 39.08 \log_{10}(d) + 20 \log_{10}(f_c) - 0.6(h_{\mathrm{IoT}} - 1.5)$, where d is the Euclidean distance between the device and the BS, and h is the height of the device. Symbol f_c is the central frequency.

B. Signal Model and Achievable Data Rate

The received signal from IoT device n to the BS through RIS k is given by

$$\begin{cases} F_{n,k} = \sum_{l_1,l_2} A^k_{l_1,l_2} h^{n,k}_{l_1,l_2} \sqrt{\Omega_n} x + y + \omega, & \text{Pattern I,} \\ F_n = h_n \sqrt{\Omega_n} x + y + \omega, & \text{Pattern II,} \end{cases}$$
(7)

where x is the transmission signal with $|x|^2 = 1$ and y is the received interference signal[5] which can be modeled as a log-normal distribution [34], i.e., $y \sim Log\mathcal{N}(\mu_y, \sigma_y^2)$. In addition, $\omega \sim \mathcal{CN}(0, \sigma_\omega^2)$ is the i.i.d. additive complex Gaussian noise and Ω_n is the transmit power. Then, the received signal-to-interference-plus-noise ratio (SINR) can be calculated by

$$\begin{cases} \gamma_{n,k} = \dfrac{\Omega_n \left(\sum_{l_1,l_2} A^k_{l_1,l_2} \tilde{h}^{n,k}_{l_1,l_2} \sum_{l_1,l_2} \left(A^k_{l_1,l_2}\right)^* \left(\tilde{h}^{n,k}_{l_1,l_2}\right)^*\right)}{\exp\left(2\mu_y + 2\sigma_y^2\right) + \sigma_\omega^2}, \\ \gamma_n = \dfrac{\Omega_n h_n h_n^*}{\exp\left(2\mu_y + 2\sigma_y^2\right) + \sigma_\omega^2}, \end{cases}$$
(8)

where $(\cdot)^*$ is the conjugate operation.

In a practical system, each IoT device can only support a finite number of data rates according to the available SFs. Let $\mathcal{M} = \{c_1, c_2, \cdots, c_M\}$ and $\mathcal{S} = \{s_1, s_2, \cdots, s_M\}$ be the set of data rates and SFs, respectively. According to [12], the relationship between data rate and SF is given by

$$c_m = \frac{B s_m}{2^{s_m}} \times CR,$$
(9)

where B is the bandwidth in Hz and CR is the code rate. It can be seen that a higher SF is associated with a lower data rate. In other words, if $\mathcal{S}$ is in ascending order $s_1 < s_2 < \cdots < s_M$, $\mathcal{M}$ will be the descending order $c_1 > c_2 > \cdots > c_M$.

The achievable data rate not only corresponds to the selected modulation and coding scheme but also depends on the received SINR [35]. Thus, according to the instantaneous received SINR, the successful transmission probability of data rate c_m is given by

$$\begin{cases} \theta^n_{k,c_m} \triangleq \Pr\{\gamma'_{n,k} \geq \Psi_m\}, & \text{Pattern I,} \\ \theta^n_{c_m} \triangleq \Pr\{\gamma'_n \geq \Psi_m\}, & \text{Pattern II,} \end{cases}$$
(10)

where Ψ_m is the minimum required SINR for the BS to demodulate the received signal when the data rate is c_m. Note that SINR $\gamma'_{n,k}$ (or γ'_n) is a random variable with mean $\gamma_{n,k}$ (or γ_n) due to the small-scale NLoS components and the received interference signal. According to the Shannon formula, the better received SINR, the higher successful transmission probability when given a data rate. In other words, a descending data rates ($c_1 > c_2 > \cdots > c_M$) will lead to an ascending successful transmission probabilities ($\theta_{c_1} < \theta_{c_2} < \cdots < \theta_{c_M}$).

III. PROBLEM FORMULATION

The system's goal is to maximize the sum rate of all IoT devices at each time slot by finding the optimal RIS and SF for each device under Pattern I, as well as determining the optimal SF for each device under Pattern II. Let $\vec{\vartheta}^t =$

$\{\vartheta_1^t, \vartheta_2^t, \ldots, \vartheta_K^t\}$ be the state vector of the RISs at time slot t, where $\vartheta_k^t = 1$ means that the k-th RIS is vacant; otherwise, it is occupied. Note that this information is known prior to each IoT device with the spectrum sensing operation. Then, the resource allocation problem is given by

$$\max_{\phi^n_{k,c_m}, \psi^n_{c_m}} \quad \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{m=1}^{M} \left(\underbrace{\sum_{k=1}^{K} c_m \vartheta_k^t \theta^n_{k,c_m} \phi^n_{k,c_m}}_{\text{Pattern I}} + \underbrace{c_m \theta^n_{c_m} \psi^n_{c_m}}_{\text{Pattern II}} \right)$$

$$\text{s.t.} \quad \sum_{m=1}^{M} \sum_{k=1}^{K} \vartheta_k^t \phi^n_{k,c_m} + \sum_{m=1}^{M} \psi^n_{c_m} = 1, \ \forall n \in \mathcal{N},$$

$$\sum_{n=1}^{N} \phi^n_{k,c_m} \leq 1, \ \forall c_m \in \mathcal{M}, \ \text{and} \ \forall k \in \mathcal{K},$$
(11)

where ϕ^n_{k,c_m} and $\psi^n_{c_m}$ are the binary variables, where $\phi^n_{k,c_m} = 1$ denotes that IoT device n transmits on the k-th RIS with SF s_m to reflect its signal to the BS; otherwise, $\phi^n_{k,c_m} = 0$. The symbol $\psi^n_{c_m} = 1$ denotes that IoT device n directly transmits to the BS with SF s_m; otherwise, $\psi^n_{c_m} = 0$. Thus, the first constraint indicates that each IoT device either transmits on Pattern I or Pattern II. If IoT device n transmits on Pattern I, then $\sum_{m=1}^{M} \sum_{k=1}^{K} \vartheta_k^t \phi^n_{k,c_m} = 1$ means that each IoT device can only select a pair of RIS and SF; if IoT device n transmits on Pattern II, then $\sum_{m=1}^{M} \psi^n_{c_m} = 1$ denotes that each IoT device can only select a SF. The second constraint means that the number of IoT devices that select the k-RIS and the m-th SF is subject to 0 or 1. In addition, $\mathcal{N} = \{1, 2, \cdots, N\}$ and $\mathcal{K} = \{1, 2, \cdots, K\}$ are the sets of IoT devices and RISs, respectively. The symbol θ^n_{k,c_m} is the successful transmission probability that IoT device n transmits on the k-th RIS and the m-th SF; while $\theta^n_{c_m}$ is the successful transmission probability that IoT device n directly sends data to the BS with SF m.

It is difficult to solve problem (11) in this distributed hybrid network, especially in Pattern I. First, c_m and θ^n_{k,c_m} are discrete values[6], resulting in a non-convex problem. Second, it requires the exact value of θ^n_{k,c_m}. This information is difficult to obtain since the channel characteristic is determined by the UE-controlled RISs. Third, it needs some communications among IoT devices to share the information of θ^n_{k,c_m} so as to determine the optimal available RIS for each IoT device. In addition, problem (11) in Pattern II can be regarded as a rate adaptation problem [36] since the SF allocation is independent for each IoT device, but it still requires the exact CSI (or the value of $\theta^n_{c_m}$), which is hard to estimate from the time-varying channel.

To overcome these challenges, we adopt the online learning method to learn the values of $\theta^n_{c_m}$ and θ^n_{k,c_m} sequentially and to allocate the optimal RIS and SF to each IoT device adaptively. During this process, the IoT device not only needs to explore the combinations of the RISs and SFs sufficiently but also needs to exploit the current best RIS and SF as many times as possible at each time slot. To better tradeoff this EE

---

[5]Notice that the interference y may come from the neighboring cellular networks or the local UEs when the UEs' signals are missed detection by the IoT devices.

[6]According to (9), c_m is discrete since the number of SFs is limited in practice. In addition, according to (8) and (10), θ^n_{k,c_m} is a function of the RIS's phase shifts, which are discrete values in range $[0, 2\pi)$.

dilemma, we introduce the MPMAB framework to solve this problem, where players are the IoT devices, and arms are the combinations of the RISs and SFs.

However, the MPMAB framework still suffers from a slow convergence rate due to the large arm space (i.e., the combinations of the RISs and SFs) under Pattern I. Therefore, we decouple the MPMAB problem into a two-stage MPMAB framework to shrink the feasible arm space. The reason is that, on the one hand, a descending data rates will result in an ascending successful transmission probabilities; on the other hand, an IoT device with different data rates will experience the same channel-fading under a particular RIS. These features indicate that the average successful transmission probabilities of the ordered data rates over different RISs have the same trend. Therefore, we can explore these RISs by arbitrarily assigning a data rate to the IoT device. In other words, the SF allocation and the RIS allocation processes are independent of each other under Pattern I.

IV. TWO-STAGE MPMAB-BASED RESOURCE ALLOCATION FRAMEWORK

In this two-stage MPMAB framework, players are the IoT devices; arms are the RISs and the SFs in the first and second stages, respectively. The first-stage MPMAB problem is to determine the best RIS for each IoT device; while the second-stage MPMAB problem is to find the optimal SF based on the state of the determined RIS.

A. First-Stage MPMAB Framework

We first introduce the transmission feedback model and the collision model. In the transmission feedback model, the IoT device can receive the transmission feedback from the BS when it transmits on Pattern I or Pattern II. Specifically, let $I'_{n,t}$ be the selected arm by the n-th IoT device at time slot t. After transmitting on the $I'_{n,t}$-th arm, the IoT device n will receive a transmission feedback $X_{I'_{n,t}}(t)$ from the BS. If the transmission is successful, then $X_{I'_{n,t}}(t) = 1$; otherwise, $X_{I'_{n,t}}(t) = 0$.

The collision model only exists in the first stage, referring to that two or more IoT devices that choose the same RIS will receive no rewards. We assume that each IoT device can deduce this collision information by observing the timeout feedback flag. Specifically, let η be the collision indicator. If an IoT device does not receive any feedback from the BS in the current time slot, then a collision happens, i.e., $\eta = 0$; otherwise, $\eta = 1$. Therefore, IoT device can distinguish the collision and the transmission failure events by checking it whether or not receives transmission feedback from the BS. Moreover, this collision model also works in some extreme situations. For example, when the received SINR in Pattern I is too low to be recognized by the BS, the IoT device can always set $\eta = 0$ (i.e., the reward is 0) since the target RIS is suboptimal to it.

Denote $\boldsymbol{I}'_t = \{I'_{1,t}, I'_{2,t} \cdots, I'_{N,t}\}$ by the strategy profile at time slot t. The collision indicator of RIS k is defined as

$$\eta_k\left(\boldsymbol{I}'_t\right) = \begin{cases} 0, & |\mathcal{N}_k| > 1, \\ 1, & \text{otherwise}, \end{cases} \quad (12)$$

where $\mathcal{N}_k$ is the set of players that select the k-th RIS in strategy profile $\boldsymbol{I}'_t$. The reward that IoT device n transmits on the k-th RIS is given by

$$r_{n,I'_{n,t}=k}(t) \triangleq \eta_k\left(\boldsymbol{I}'_t\right) X_{I'_{n,t}=k}(t). \quad (13)$$

Then, the estimated average successful transmission probability that the n-th IoT device transmits on the k-th RIS is given by

$$\hat{\theta}_{n,k} = \mathbb{E}\left[r_{n,I'_{n,t}=k}(t)\right], \quad (14)$$

where $\mathbb{E}[\cdot]$ is the expectation operator.

B. Second-Stage MPMAB Framework

In this stage, each player has a targeted RIS after the first-stage allocation. Thus, the IoT device transmits directly or on a targeted RIS to the BS at each time slot. Note that there is no collision in this stage since two or more players can choose the same SF. Therefore, this stage can be regarded as a single-player MAB framework.

Let $I''_{n,t}$ be the currently selected arm at the second-stage allocation. The reward that the IoT device n chooses the m-th data rate is defined as

$$r_{n,I''_{n,t}=m}(t) \triangleq c_m \eta_k\left(\boldsymbol{I}'_t\right) X_{I''_{n,t}=m}(t), \quad (15)$$

where c_m is the m-th data rate and $\boldsymbol{I}'_t$ is the strategy profile of all players' target RISs at the first stage. The instantaneous rewards $r_{n,I''_{n,t}}(t)$ are independently and identically distributed w.r.t. player n and time slot t. Therefore, the estimated average reward is given by

$$\hat{\mu}_{n,m} = \mathbb{E}\left[r_{n,I''_{n,t}=m}(t)\right] = c_m \hat{\theta}_{n,m}, \quad (16)$$

where $\hat{\theta}_{n,m}$ is the estimated average successful transmission probability that the n-th IoT device transmits on the m-th data rate.

C. Performance Metric for the Two-Stage MPMAB Framework

In the following, we design a criterion to quantify the performance loss that players select the suboptimal arms rather than the optimal arm in this two-stage MPMAB problem. According to (11), the objective function consists of Pattern I and Pattern II. For Pattern I, we define the joint RIS and SF selection profile by $\boldsymbol{a} = \{a_1, a_2, \ldots, a_N\}$, where $a_n \in \mathcal{K} \otimes \mathcal{M}$ and $\otimes$ is the Cartesian product of the RIS set and the data rate set. However, for Pattern II, the selection profile $\boldsymbol{a}$ is the set of data rates, i.e., $a_n \in \mathcal{M}$. Therefore, the two-stage allocation aims to solve the following problem,

$$\begin{aligned} \boldsymbol{a}^* &= \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} \hat{\mu}_{n,a_n} \\ &= \arg\max_{\boldsymbol{a}} \sum_{n=1}^{N} \mathbb{E}\left[c_{\boldsymbol{a}} \eta_k\left(\boldsymbol{a}\right) X_{\boldsymbol{a}}(t)\right], \end{aligned} \quad (17)$$

where $\boldsymbol{a}^* = \{a_1^*, a_2^*, \cdots, a_N^*\}$ is the optimal strategy profile.
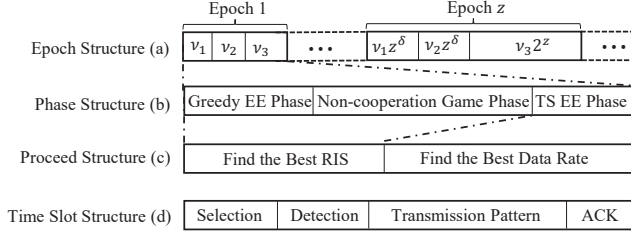
Fig. 2: The structure of the E2Boost algorithm.

Then, we define the difference between the optimal arm and the currently selected arm as the performance metric, also known as *regret*. According to [23], the expression of accumulated regrets is given by

$$\mathcal{R}eg \triangleq \sum_{t=1}^{T} \sum_{n=1}^{N} r_{n,a_n^*}(t) - \sum_{t=1}^{T} \sum_{n=1}^{N} r_{n,a_n}(t), \quad (18)$$

where $a_n^* \in \boldsymbol{a}^*$ and T is the total time slots. For mathematical analysis, we further define the *pseudo-regret* [19] *w.r.t.* the stochastic rewards and the randomly selected arms as

$$\overline{\mathcal{R}eg} = \sum_{n=1}^{N} \left(T \times \mu_{n,a_n^*} - \mathbb{E} \sum_{t=1}^{T} \mu_{n,a_n} \right)$$
$$= \begin{cases} \sum_{n=1}^{N} \sum_{i=1}^{K \times M} \Delta_{n,i} \mathbb{E}[W_{n,i}], & \text{Pattern I}, \\ \sum_{n=1}^{N} \sum_{i=1}^{M} \Delta_{n,i} \mathbb{E}[W_{n,i}], & \text{Pattern II}, \end{cases} \quad (19)$$

where $\Delta_{n,i} = \mu_{n,a_n^*} - \mu_{n,i}$ and $W_{n,i}$ is the number of times that arm i has been selected up to time T. Term $\mu_{n,i}$ is the real expected throughput of player n at arm i.

V. E2BOOST ALGORITHM

In this section, we propose an E2Boost algorithm to solve this two-stage MPMAB problem by combining the game theory and the MAB algorithms. The structure of the E2Boost algorithm is shown in Fig. 2. Since time horizon T is unknown to each player, the E2Boost algorithm proceeds in epochs (i.e., $z = 1, \cdots, Z$). Each epoch consists of three phases: ϵ-Greedy EE, non-cooperation game, and Thompson sampling EE phases. Each phase contains several time slots and specific mechanism to balance the EE dilemma.

A. The Exploration and Exploitation Boosting Algorithm

The E2Boost algorithm is shown in Algorithm 1. The first two phases are designed to find the optimal RIS for each IoT device by solving the first-stage MPMAB problem; while the last phase is to determine the best SF by solving the second-stage MPMAB problem. In the following, we elaborate on the above three phases in detail.

ϵ-**Greedy EE Phase**: There are $\nu_1 z^\delta$ rounds in this phase for epoch $z = 1, \cdots, Z$, where $\nu_1 > 0$ and $\delta > 0$ are two constants. It aims to estimate the average successful transmission probability of each RIS. The SF is randomly chosen from the set $\mathcal{S}$ when $z = 1$; otherwise, it uses the SF determined in the last epoch of the third phase. We adopt the ϵ-greedy algorithm to balance the EE dilemma. Specifically, if $z = 1$, we set $\epsilon = 1$ to uniformly explore all RISs; otherwise, we update the parameter ϵ according to *Lemma* 1, as given in

the next paragraph. Hence, when the players' strategy profile deviates from $\boldsymbol{a}^*$, Algorithm 1 tends to uniformly explore all actions; otherwise, it inherits the last epoch's action with a high probability.

Non-cooperation Game Phase: This phase has a length of $\nu_2 z^\delta$ rounds, which is the core step of Algorithm 1 to allocate the optimal RIS for each player. By adopting the estimated average successful transmission probability $\hat{\theta}_{n,k}^z$ in the first phase as a utility, players in this phase will play a non-cooperation game.

Specifically, let the utility of player n in strategy profile $\boldsymbol{I}'$ be

$$u_n(\boldsymbol{I}') \triangleq \eta_k(\boldsymbol{I}') \hat{\theta}_{n,k}^z, \quad \forall k \in \mathcal{K}, \quad (20)$$

where $\hat{\theta}_{n,k}^z$ is the estimated successful transmission probability that the n-th IoT device transmits on the k-th RIS from epoch 1 to z at the first phase. Let $u_{n,\max} = \max_{\boldsymbol{I}'} u_n(\boldsymbol{I}')$ be the maximum utility of player n. Assume that each player has a private state $ST_n = \{C, D\}, \forall n \in \mathcal{N}$, where C and D represent content and discontent state, respectively. In addition, each player maintains a baseline RIS $\bar{k}$. Then, a player chooses a RIS according to the following strategy:

- A content player has a very high probability to stay at the current baseline RIS:

$$P_{n,k} = \begin{cases} \frac{\varepsilon^\nu}{K-1}, & k \neq \bar{k}; \\ 1 - \varepsilon^\nu, & k = \bar{k}. \end{cases} \quad (21)$$

- A discontent player selects a RIS following a uniform distribution, i.e.,

$$P_{n,k} = \frac{1}{K}, \quad \forall k \in \mathcal{K}. \quad (22)$$

The transition between content state C and discontent state D is given by:

- If $k = \bar{k}$, $u_n > 0$, and $ST_n = C$, then a content player keeps state C with a probability of 1:

$$(\bar{k}, C) \to (\bar{k}, C). \quad (23)$$

- If $k \neq \bar{k}$ or $u_n = 0$ or $ST_n = D$, then transitions of baselines and states are given by

$$\left(\bar{k}, C/D \right) = \begin{cases} (k, C), & \frac{u_n}{u_{n,\max}} \varepsilon^{u_{n,\max} - u_n}; \\ (k, D), & 1 - \frac{u_n}{u_{n,\max}} \varepsilon^{u_{n,\max} - u_n}. \end{cases} \quad (24)$$

Eq. (24) indicates that, when a RIS records a collision or in a busy state (i.e., $\eta_k = 0$), the player will transfer to the discontent state D with a probability of 1 as $u_n = 0$. On the other hand, when a RIS is optimal to the player, it will transfer to the content state C with a probability of 1 as $u_n = u_{n,\max}$.

Assume that all players' actions and states constitute a strategy profile $\boldsymbol{a_1}$. Then, a strategy graph can be constructed at the end of this phase, where the vertex is the strategy profile, and an edge exists if the players can switch from one strategy to the other. Actually, this strategy graph forms a perturbed time-reversible Markov process over state space $\prod_{n=1}^{N} (\mathcal{K}_n \times (C, D))$. As pointed out in [23], [37], [38], there exists an optimal strategy profile that players will visit many

---

Algorithm 1 E2Boost Algorithm Run by Player n

1: **Initialization:** $\delta > 0, \varepsilon > 0$ and $\nu, \nu_1, \nu_2, \nu_3 > 0$. Let $\epsilon = 1, V_{n,k}(0) = 0, Q_{n,k}(0) = 0, \alpha_{n,m}(0) = 0, \beta_{n,m}(0) = 0, \ \forall k \in \mathcal{K}, \ \forall m \in \mathcal{M}$.

2: **for** each epoch $z = 1, 2, \cdots, Z$ **do**

3:     **i) ϵ-Greedy EE Phase:** For the next $\nu_1 z^\delta$ time slots.

4:         a) Pick up a data rate $I'_{n,t} = m$ uniformly from set $\mathcal{M}$ if $z = 1$, otherwise $I'_{n,t} = c_n^*$;

5:         b) Select a RIS $I'_{n,t} = k$ uniformly from set $\mathcal{K}$ with probability ϵ or $I'_{n,t} = k_n^*$ with probability $1 - \epsilon$;

6:         c) Detect the selected RIS: jump to **Phase iii** if busy; otherwise, continue the following steps:

7:         d) Observe the transmission feedback $X_{n,I'_{n,t}}$. Set $\eta_k = 0$ if timeout and $\eta_k = 1$ otherwise;

8:         e) If $\eta_k = 1$ then update $V_{n,I'_{n,t}}(t) = V_{n,I'_{n,t}}(t-1) + 1$ and $Q_{n,I'_{n,t}}(t) = Q_{n,I'_{n,t}}(t-1) + X_{n,I'_{n,t}}(t)$;

9:     At the end of this phase, compute the successful transmission probabilities of RISs by

$$\hat{\theta}_{n,k}^z = \frac{Q_{n,k}}{V_{n,k}}, \quad \forall k \in \mathcal{K}.$$

10:     **ii) Non-cooperation Game Phase:** For the next $\nu_2 z^\delta$ time slots, play with the dynamics. Set $ST_n = C$, and let $\bar{k}$ be the last RIS chosen in the $z - \lfloor \frac{z}{2} \rfloor - 1$ Game phase, or a random choice if $z = 1, 2$.

11:         a) If $ST_n = C$ choose a RIS $I'_{n,t}$ using (21) and if $ST_n = D$ select $I'_{n,t}$ at random (22);

12:         b) Detect the selected RIS: jump to **Phase iii** if busy; otherwise continue the following steps:

13:         c) If $I'_{n,t} \neq \bar{k}$ or $u_n = 0$ or $ST_n = D$ then set $ST_n = C$ or D according to (24);

14:         d) Record the number of times each RIS has been selected within the content state using (25);

15:         e) Adjust parameter ϵ according to *Lemma 1* when $z \geq 2$.

16:     At the end of this phase, determine the current best RIS by

$$k_n^* = \arg \max_{k \in \mathcal{K}} \sum_{j=0}^{\lfloor \frac{z}{2} \rfloor} F_n^{z-j}(k).$$

17:     **iii) Thompson Sampling EE Phase:** For the next $\nu_3 2^z$ time slots, run the Thompson sampling algorithm based on the current state of the best RIS, as well as the corresponding collision indicator.

18:         a) Draw $\hat{\theta}_{n,m} \sim \text{Beta}\left(\alpha_{n,m}(t) + 1, \beta_{n,m}(t) + 1\right)$;

19:         b) Select a data rate $I''_{n,t} = \arg \max_{m \in \mathcal{M}} c_m \times \hat{\theta}_{n,m}$;

20:         c) Detect the target RIS: device directly transmits to the BS if busy, otherwise continue the following steps:

21:         d) Transmit on the selected data rate and observe the random transmission feedback $X_{I''_{n,t}}(t)$;

22:         e) Posterior update: Set $\alpha_{n,I''_{n,t}}(t) = \alpha_{n,I''_{n,t}}(t-1) + X_{I''_{n,t}}(t)$ and $\beta_{n,I''_{n,t}}(t) = \beta_{n,I''_{n,t}}(t-1) + 1 - X_{I''_{n,t}}(t)$.

23:     At the end of this phase, determine the current best data rate by

$$c_n^* = \arg \max_{m \in \mathcal{M}} \frac{c_m \alpha_{n,m}}{\alpha_{n,m} + \beta_{n,m}}.$$

24: **end for**

---

times than other strategy profiles. As a result, each player can agree on its optimal arm by recording the number of times that each arm has been selected, i.e.,

$$F_n^z(k) \triangleq \sum_{t \in \mathcal{G}_z} \mathbb{I}\left(I'_{n,t} = k, ST_n = C\right), \ \forall k \in \mathcal{K}, \quad (25)$$

where $F_n^z(k)$ is the number of times that the k-th RIS has been played by the n-th player at the z-th epoch under state C. The symbol $\mathcal{G}_z$ represents the number of time slots in the z-th epoch and $\mathbb{I}(\cdot)$ is an indicator function. Finally, we can determine the best RIS by using the recent $\lfloor z/2 \rfloor + 1$ epochs' F_n^z, i.e.,

$$k_n^* = \arg \max_{k \in \mathcal{K}} \sum_{j=0}^{\lfloor \frac{z}{2} \rfloor} F_n^{z-j}(k). \quad (26)$$

In addition, F_n^z can be used to design a criterion to balance the EE dilemma in the first phase by adjusting the parameter ϵ. The reason is that it is unnecessary to uniformly explore

all RISs when the non-cooperation game phase asymptotically approaches the optimal RIS. This asymptotical behavior can be quantified by the distance between two adjacent vectors of F_n^z and F_n^{z-1}, which can be measured by the WD[7]. As a result, the distance between the probability mass functions (PMFs)[8] of F_n^z and F_n^{z-1} is regarded as a criterion to adjust the parameter ϵ. Thus, we have the following lemma.

Lemma 1: For the n-th player, given $z > 1$, the parameter of the ϵ-greedy algorithm in the first phase can be chosen according to

$$\epsilon \triangleq \min\{1, \mathrm{D_{WD}}\left(\mathbb{P}(F_n^z) \| \mathbb{P}(F_n^{z-1})\right)\},$$

where $\mathrm{D_{WD}}\left(\cdot \| \cdot\right)$ is the calculation of the WD in [39]. Terms

---

[7]Wasserstein distance, also known as earth mover's distance, is a measure to calculate the distance between two probability distributions on a metric space. In the simulation, we compute it using the corresponding function in Matlab.

[8]The PMF is computed by $\mathbb{P}(F_n^z(i)) = F_n^z(i)/\sum_i F_n^z(i), \ \forall i \in \mathcal{K}$.

$\mathbb{P}(F_n^z)$ and $\mathbb{P}(F_n^{z-1})$ are the PMFs of (25) at epochs z and $z-1$ of the second phase, respectively.

Thompson Sampling EE Phase: The last phase has a length of $\nu_3 2^z$ rounds where $\nu_3 > 0$ is a constant. The objective is to find the best SF for each player based on the busy or idle state of the determined RIS in the second phase. Note that there are two types of resource allocations in this phase for transmission Pattern I and Pattern II. The main difference between the two allocations is that, in Pattern I, it requires to jointly estimate θ_{k,c_m}^n and $\theta_{c_m}^n$ at each epoch; while in Pattern II, it only needs to estimate the $\theta_{c_m}^n$ at each time slot.

As mentioned before, the second-stage MPMAB problem can be regarded as a single-player MAB problem. Therefore, we can adopt the TS algorithm [29] to solve the second-stage MPMAB problem to track the Bernoulli distribution rewards (i.e., transmission success or failure). The TS algorithm first maintains a Beta prior distribution[9] for each SF. Thus, the objective of this phase is equivalent to estimating the parameter in the Beta distribution, which will converge to the true value of θ_{k,c_m}^n or $\theta_{c_m}^n$. Based on the transmission feedback, TS algorithm is able to update the posterior distribution by: $\alpha = \alpha + 1$ if transmission is successful, otherwise $\beta = \beta + 1$. Notice that the value function (i.e., $c_m \hat{\theta}_{n,m}$) is the current data rate, instead of the successful or failed transmission feedback. At the end of this phase, it can determine the current best SF for each IoT device by using the estimated parameters in the Beta distribution, i.e.,

$$c_n^* = \arg \max_{m \in \mathcal{M}} \frac{c_m \alpha_{n,m}}{\alpha_{n,m} + \beta_{n,m}}. \tag{27}$$

B. Complexity and Feasibility Analysis

We first give a brief discussion on the computational complexity of the proposed algorithm. In Algorithm 1, the computational complexity of the first phase is $\mathcal{O}(\nu_1 z^\delta L_{\text{ED}})$, where L_{ED} is the length of samples in the energy detector of the spectrum sensing operation. Meanwhile, the computational complexity of the second phase is $\mathcal{O}(\nu_2 z^\delta + zK \log_2 K)$, where the second term comes from the WD in the calculation of parameter ϵ [27]. In addition, the computational complexity of the third phase is $\mathcal{O}(\nu_3 2^z M \log_2 M)$, where the complexity comes from the 'argument maximum' operation in the TS algorithm [29]. Therefore, the total computational complexity is $\mathcal{O}(\nu_1 z^\delta L_{\text{ED}} + \nu_2 z^\delta + zK \log_2 K + \nu_3 2^z M \log_2 M)$, which increases linearly logarithmically with the number of RISs K and SFs M. As the time epoch z increases, the complexity of the third phase will become the dominant factor in the total complexity. Therefore, the total complexity of Algorithm 1 is about $\mathcal{O}(\nu_3 2^z M \log_2 M)$.

Next, we discuss the feasibility of the proposed algorithm in practical applications, e.g., B5G/6G networks. First, the proposed algorithm performs in real-time and automatically converges to the optimal solution (i.e., the online learning feature). Second, its distributed feature can reduce the communication overhead and make it easy to apply to the other

---

[9]Beta(α, β) is the beta distribution with probability density function (pdf): $f_{\alpha,\beta}(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{B'(\alpha,\beta)}, y \in [0,1]$, where $B'(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

---

network scenarios. Third, the complexity of the proposed algorithm increases linearly logarithmically with the number of RISs K and the SFs M. These features demonstrate that the proposed algorithm has great potential to be applied in B5G/6G networks with different requirements of the rate, delay, scalability, and reliability.

In Algorithm 1, we only consider the case that the number of IoT devices is less than the RISs, i.e., $K \geq N$. However, the proposed algorithm can also handle the case of $K < N$ by dividing N IoT devices into K clusters using the k-means clustering method [40] according to their geographic locations. We assume that the IoT devices in the same cluster prefer the same RIS and communicate with the BS using the round-robin method. Therefore, one of the clustering IoT devices can transmit on the RIS; while others directly transmit data to the BS at each time slot. In other words, Algorithm 1 still works in the case of $K < N$ by allocating the optimal RIS to each cluster rather than each IoT device.

Algorithm 2 Modified E2Boost Algorithm Run by Player n for the case $K < N$

1: **Initialize:** Parameters in Algorithm 1
2: **for** each time slot $t = 1, 2, \cdots, T$ **do**
3:     Check the round-robin flag
4:     If the flag is equal to 1, run the E2Boost algorithm in Algorithm 1
5:     Otherwise, run the TS algorithm in the third phase of Algorithm 1
6: **end for**

Therefore, we give a modified E2Boost algorithm to handle the case of $K < N$, as shown in Algorithm 2. At the beginning of each time slot, the IoT device n first checks its round-robin flag to determine its transmission patterns: If the flag is equal to 1, the IoT device runs the E2Boost algorithm in Algorithm 1 to find the optimal RIS and SF; otherwise, it runs the TS algorithm in the third phase of Algorithm 1 to find the optimal SF.

C. An Upper Pseudo-Regret Bound

We derive an upper pseudo-regret bound for the E2Boost algorithm. Since each IoT device has two transmission patterns, the pseudo-regret also consists of the RIS-enabled regret and the non-RIS-enabled regret parts.

For the RIS-enabled regret part, the performance analysis of the RIS-enabled regret is mainly built on [23] since the E2Boost algorithm shares the same architecture of the GoT algorithm. On the other hand, the Bernoulli reward processing in this work strictly meets the key condition of *Definition 1* in [23]. However, compared with the GoT algorithm, the proposed algorithm has the following features. First, it is a two-stage MPMAB framework that has a small arm space (i.e., $\mathcal{K}$) to explore. Its total pseudo-regret only depends on the number of RISs, instead of the whole arm space (i.e., $\mathcal{K} \otimes \mathcal{M}$) as that in the GoT algorithm. Second, we embed the ϵ-greedy algorithm into the first phase to further tradeoff the EE dilemma. Thus, the accumulated regrets of this phase

will trend to 0 when all players agree on their optimal RISs. Third, we incorporate the TS algorithm into the third phase to determine the best SF. Similarly, only a few accumulated regrets will be accrued in this phase when the optimal RIS is determined. Therefore, these features enable us to achieve a tighter pseudo-regret bound than the GoT algorithm.

For the non-RIS-assisted regret part, the E2Boost algorithm only has the third phase, i.e., the second-stage MPMAB problem. Since two or more players that select the same arm (or SF) will not collide, this MPMAB problem is reviewed as a single stochastic MAB problem. In Algorithm 1, we use the modified TS (MTS) algorithm [36] to solve this single stochastic MAB problem. Therefore, we adopt the theoretical results in [36] to derive the non-RIS-enabled regret.

To conclude, we have the following theorem.

Theorem 1: Let $\Gamma_{\max} = \max_{n,i} \mu_{n,i}$ be the maximum real expected rewards among all players' arms. For any hybrid uplink network, given $\nu_1 > 0, \nu_2 > 0, \nu_3 > 0, \delta \geq 0, 0 < \varpi < 1$ and a small enough ε, the total upper pseudo-regret bound obtained by the E2Boost algorithm is

$$\overline{Reg} \leq N\Gamma_{\max}(1 - P_a)\left(2(\nu_1 + \nu_2)\log_2^{1+\delta}\left(\frac{T}{\nu_3} + 2\right)\right.$$
$$\left. + (6NK + 1)\nu_3 \log_2\left(\frac{T}{\nu_3} + 2\right)\right) \quad (28)$$
$$+ P_a(1 + \varpi)\sum_{n=1}^{N}\sum_{a_n \in \mathcal{M}} \frac{\log_2 T}{D_{KL}(a_n, a_n^*)}\Delta_{n,a_n},$$

where $\log_2^{1+\delta}(T/\nu_3 + 2)$ denotes $\log_2(T/\nu_3 + 2)$ to the power of $(1 + \delta)$ and $D_{KL}(\cdot)$ is the Kukkback-Leibler divergence. Term P_a is the active probability of the UE.

Proof 1: See Appendix A.

Remark 1: The first two terms of the upper pseudo-regret bound accounts for the RIS-assisted regret part; while the third term is the non-RIS-assisted regret part. We can see that the weights of these two parts rely on the active probability of the UE.

Remark 2: The total upper pseudo-regret bound increases logarithmically with T, i.e., $\overline{Reg} = \mathcal{O}(\log_2^{1+\delta} T)$, indicating that Algorithm 1 will converge and the per-round regret approaches zero when T is sufficiently large.

Remark 3: The total upper pseudo-regret bound in the E2Boost algorithm is much tighter than the GoT algorithm. According to [23], the total upper pseudo-regret bound of the GoT algorithm is

$$\overline{Reg}_{GoT} \leq 4N\Gamma_{\max}(\nu_1 + \nu_2)\log_2^{1+\delta}\left(\frac{T}{\nu_3} + 2\right)$$
$$+ N\Gamma_{\max}(6NKM + 1)\nu_3 \log_2\left(\frac{T}{\nu_3} + 2\right) \quad (29)$$
$$= \mathcal{O}(\log_2^{1+\delta} T).$$

For example, when $\nu_1 = \nu_2 = \nu_3$, $\delta = 0$, and $P_a = 0$, we have

$$\overline{Reg} \leq (5 + 6NK)\,\nu_1 N\Gamma_{\max} \log_2\left(\frac{T}{\nu_3} + 2\right), \quad (30)$$

TABLE I: The transmission parameters for the C-IoT device

Spreading Factor (SF)	7	8	9	10	11	12
Data Rate (Mbps)	1.09	0.63	0.35	0.20	0.11	0.06
Threshold ($\times 10^3$)	4.5	4	3.5	3	2.5	2

and

$$\overline{Reg}_{GoT} \leq (9 + 6NKM)\,\nu_1 N\Gamma_{\max} \log_2\left(\frac{T}{\nu_3} + 2\right). \quad (31)$$

It can be seen that $\overline{Reg}$ is about M times lower than $\overline{Reg}_{GoT}$. This observation can be verified by the numerical results in the following section.

VI. SIMULATION RESULTS

We conduct extensive simulations to evaluate the performance of the proposed algorithms. The simulation parameters are chosen according to the 3GPP standard [32] and refs. [4], [5]. All results are obtained from 10^3 Monte Carlo (MC) trials.

A. Parameter Configuration and Baseline Algorithms

Parameter Configuration: The transmit power at each IoT device is $\Omega_n = 20$ dBm, $\forall n \in \mathcal{N}$. The background noise plus interference power is -95 dBm, and the wavelength λ is set according to the central carrier frequency 5.9 GHz. The bandwidth B is 40 MHz. The Rician factor is $\zeta = 4$, and the antenna gain G is set to 1. Each IoT device has 6 SFs to choose from, as shown in TABLE I. The data rates are determined by (9) and the thresholds are the minimum required SINR to demodulate the received signal. Assume that the active probability of each RIS (i.e., occupied by the legal UEs) is $P_a^k = 0.2$. In addition, we adopt the UMa model [32] to describe the path loss of both LoS and NLoS components.

The RIS is placed perpendicular to the ground, and the number of elements is 101×101. The direction of the RIS in the XY-plane is shown in Fig. 3. The angle $\angle\varphi$ and all elements' locations in RIS are determined according to Appendix B. Each element contains $b = 8$ PIN diodes with the reflection amplitude $A = 1$. We consider two types of phase shift settings, i.e., the *optimal phase shift* and the *constant phase shift*. For the *optimal phase shift* setting, each RIS's phase shifts are set to be optimal to the UEs (*see Proposition 2 of [4]*), i.e.,

$$\tau_{l_1, l_2} = \left\lfloor \left(\Pi - \frac{2\pi}{\lambda}L_{l_1,l_2}^k\right)\frac{2^b}{2\pi}\right\rfloor \frac{2\pi}{2^b}, \quad (32)$$

where Π is an arbitrary constant and L_{l_1,l_2}^k is the distance between the BS, and the UEs through the k-th RIS's (l_1, l_2) element. For the *constant phase shift* setting, we assume that the phase shifts on all RISs' elements are equal. That is, all integers ρ_{l_1,l_2} in (1) are simply set to a constant (we set to 170 in the following simulations). Note that ρ_{l_1,l_2} can be an arbitrary integer in the range of $[0, 2^b - 1]$.

Baseline Algorithms: We compare the E2Boost algorithm with the optimal solution, GoT algorithm, Q-learning method, random selection method, E2Boost without TS algorithm, and
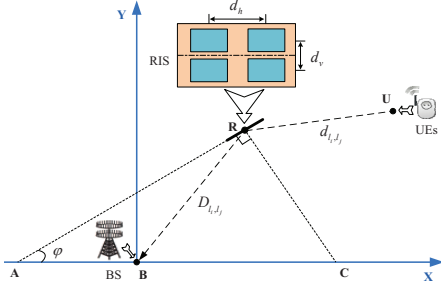
Fig. 3: A schematic diagram of the UE-RIS-BS link (top view).

Fig. 4: A fixed network scenario in a 200×200 m square area with $K = 3, N = 3$ (top view).

E2Boost without WD algorithm. Next, we introduce these baseline algorithms in detail.

- *Optimal Solution:* The optimal solution is obtained by solving the two-stage MPMAB problem in a centralized form. In the Pattern I, it allocates the optimal RIS and SF to each IoT device by using the Hungarian algorithm [15], where the only required information is θ_{k,c_m}^n and $\theta_{c_m}^n$. In the simulation, we obtain this information by recording the received SINR $\gamma_{n,k}$ and γ_n with the above simulation parameters over 10^5 MC trails. Then, $\hat{\theta}_{k,c_m}^n$ and $\hat{\theta}_{c_m}^n$ can be estimated by comparing these SINRs with a given threshold Ψ_m. Note that $\hat{\theta}_{k,c_m}^n$ and $\hat{\theta}_{c_m}^n$ can approach the true values of θ_{k,c_m}^n and $\theta_{c_m}^n$ arbitrarily as long as the number of MC trials is sufficiently large. Based on this information and the data rate c_m in Table I, the optimal RIS and SF of each IoT device can be obtained by using the Hungarian algorithm (i.e., the *munkres* function in Matlab). In the Pattern II, we determine the optimal SF for each IoT device using the genie-aided solution (i.e., from God's perspective) as the $\hat{\theta}_{c_m}^n$ and c_m are known.
- *GoT Algorithm:* The GoT algorithm in [23] is a fully distributed algorithm to solve the decentralized resource allocation problems. It has the same architecture of the proposed algorithm. However, it lacks the ϵ-greedy algorithm and the TS algorithm in the first and third phases to further balance the EE dilemma. In addition, it needs to explore the combinations of RISs and SFs; while the proposed algorithm explores the RISs and SFs separately.
- *Q-learning method:* For the Q-learning method, the state is the target RIS's busy or idle state. The state transition probability is the RIS's active or passive probability P_a^k or $1 - P_a^k$. The actions are the set of SFs $\mathcal{M}$ if the target RIS is in a busy state; otherwise, the actions are the combinations of SFs and RISs, i.e., $\mathcal{K} \otimes \mathcal{M}$.
- *Random Selection:* For the random selection method, each IoT device uniformly chooses an arm from the arm space $\mathcal{K} \otimes \mathcal{M}$ in Pattern I or the arm space $\mathcal{M}$ in Pattern II at each time slot. There is no EE mechanism inside.
- *E2Boost without TS:* Compared with the E2Boost algorithm, it removes the TS algorithm from the third phase. Moreover, it requires exploring the combinations of the RISs and SFs in the first phase with the ϵ-greedy algorithm.
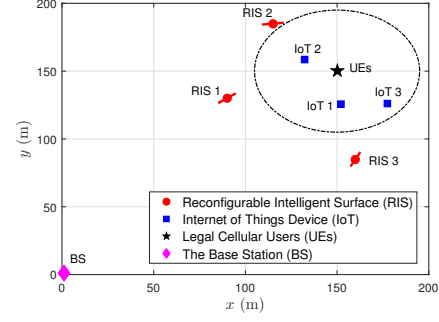
- *E2Boost without WD:* Compared with the E2Boost algorithm, the only difference is that it maintains a constant exploration rate ϵ for the ϵ-greedy algorithm, rather than adaptively adjusting ϵ in the E2Boost algorithm.

It is worth noting that there is a performance gap between the solutions of the two-stage MPMAB problem and the original problem (11). Specifically, the solution of problem (11) is to allocate the optimal available RIS and SF to each IoT device at each time slot t; while the solution (i.e., the optimal solution) of the two-stage MPMAB problem is to assign the optimal RIS and SF to each IoT device average over the time horizon T. As a result, the performance of the two-stage MPMAB problem is slightly poorer than that of problem (11). However, Theorem 1 shows that the proposed algorithm can converge to the optimal solution when T is sufficiently large. The following simulation results will also verify this.

B. Fixed Network Scenario

We first consider a fixed network scenario in a 200 m $\times$ 200 m square area, as shown in Fig. 4, where $N = 3$ cellular IoT devices are located in a 45 m $\times$ 45 m circle area. For simplicity, we assume that all UEs are centered in the point $(x, y) = (150, 150)$ m. Outside this circle are the BS and $K = 3$ RISs. The distances between the BS and the center of the RIS, as well as the IoT device and the center of the RIS, are calculated by the Euclidean distances *w.r.t.* their locations (i.e., D_{l_i, l_j} and d_{l_i, l_j} in Fig. 3), respectively. The RIS and the BS heights are 10 m and 20 m, respectively.

Fig. 5 shows the allocation results of the E2Boost algorithm for the *optimal phase shift* setting. The simulation parameters for the E2Boost algorithm are $\nu = 1.4$, $\delta = 0$, $\varepsilon = 0.01$, $Z = 10$, $\nu_1 = \nu_2 = 1000$ and $\nu_3 = 100$. The average throughput is computed by $\frac{1}{t} \sum_{t=1}^{T} \mu_{n, I_{n,t}}$. It can be seen from Fig. 5 (b-d) that each player will converge to its own optimal SF and RIS, i.e., player 1 $\to$ (RIS3, SF1), player 2 $\to$ (RIS1, SF1) and player 3 $\to$ (RIS2, SF1). All players prefer SF1 with the highest data rate of 1.09 Mbps in Table I. The average throughput of all players is 2.3859 Mbps, which is slightly less than the optimal solution's 2.4315 Mbps. In addition, IoT device 3 accounts for the lowest average throughput by 0.4971 Mbps due to the long transmission distance between IoT 3-RIS2-BS links. The IoT device 3 does not choose RIS3 because the direction (or the phase shifts) of
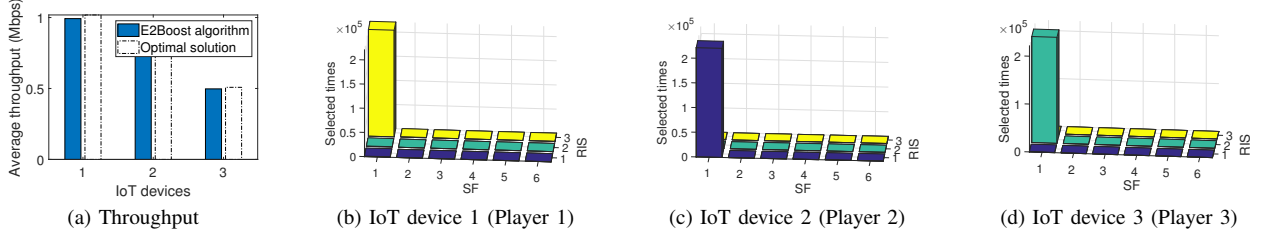
Fig. 5: (a) Average throughput of three IoT devices, (b-d) The number of selected times at each arm for each player, by running the E2Boost algorithm with *optimal phase shift* setting in the fixed network scenario Fig. 4.
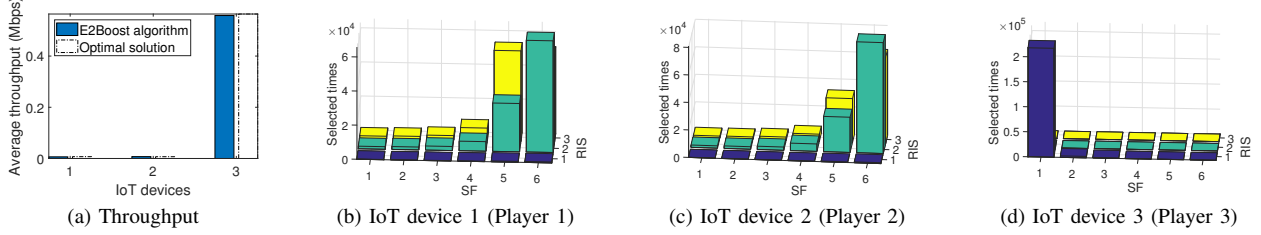

Fig. 6: (a) Average throughput of three IoT devices, (b-d) The number of selected times at each arm for each player, by running the E2Boost algorithm with *constant phase shift* setting in the fixed network scenario Fig. 4.

RIS2 is more suitable for IoT 3 than RIS3, i.e., RIS2-UEs-IoT3 in a line.

Similarly, Fig. 6 shows the allocation results of the E2Boost algorithm for the *constant phase shift* setting. We can see that the average throughput of all players is just about 0.5782, which is much lower than the *optimal phase shift* setting. In addition, player 1 and player 2 disagree on the optimal RIS since there is a collision between them. The reason is that the time horizon of the second phase in Algorithm 1 is too short (i.e., $\nu_2 = 1,000$) to resolve this collision. As a result, the highest SF for Pattern II is chosen frequently, resulting in low average throughput. To conclude, Figs. 5 and 6 demonstrate that the channel gains between the IoT device and the BS not only rely on the path-loss gain but also depend on the settings of phase shifts and direction in the RIS.

Fig. 7 depicts the total pseudo-regret of the E2Boost algorithm, the E2Boost algorithm without TS, and the GoT algorithm in the cases of $\nu_1 = \nu_2 = 1,000$ and $\nu_1 = \nu_2 = 2,000$, when $Z = 10$ under the *optimal phase shift* setting. Other parameters are the same as those in Fig. 5. We can see that the proposed algorithm has the lowest expected total pseudo-regret in both cases since it has a small arm space (i.e., due to the two-stage allocation mechanism) to explore. In addition, the total pseudo-regrets of all algorithms in the case of $\nu_1 = \nu_2 = 1,000$ are lower than those in the $\nu_1 = \nu_2 = 2,000$ case. The reason is that a larger value of ν_1 and ν_2 indicates that a longer time is needed to explore all arms, leading to more performance loss. However, when $\nu_1 = \nu_2 = 1,000$, the GoT algorithm and the E2Boost algorithm without TS will not converge since the value of ν_2 is too small for the second phase to resolve the collisions among IoT devices. More importantly, Fig. 7 validates our theoretical analysis of Theorem 1, where the total pseudo-
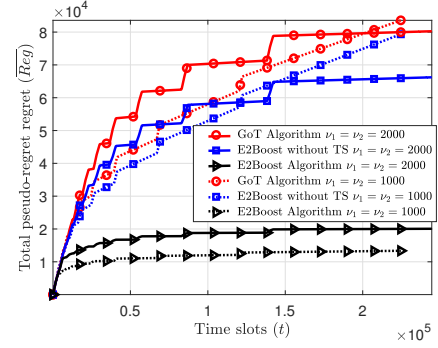

Fig. 7: The total pseudo-regret via time slot in the cases of $\nu_1 = \nu_2 = 1,000$ and $\nu_1 = \nu_2 = 2,000$ with *optimal phase shift* setting in the fixed network scenario Fig. 4.

regret of the E2Boost algorithm increases logarithmically *w.r.t.* the time horizon T and is about four times better than the GoT algorithm.

Fig. 8 compares the average total throughput of the E2Boost algorithm, the E2Boost algorithm with $\epsilon = 0$ and $\epsilon = 1$ (without WD), the E2Boost algorithm without TS, the GoT algorithm, and the random selection method in the *optimal phase shift setting* with $\nu_1 = \nu_2 = 2,000, Z = 10$. It can be seen that the proposed algorithm outperforms the other algorithms and is close to the optimal solution. By contrast, the proposed algorithm with $\epsilon = 0$ accounts for the worst performance since there is no exploration in the first phase. Meanwhile, the E2Boost algorithm with $\epsilon = 1$ and the E2Boost algorithm without TS is better than the GoT algorithm, indicating that the E2Boost algorithm with WD can effectively tradeoff the EE dilemma by sequentially optimizing the parameter ϵ. More importantly, since the two-stage allocation mechanism, the E2Boost algorithm has a faster
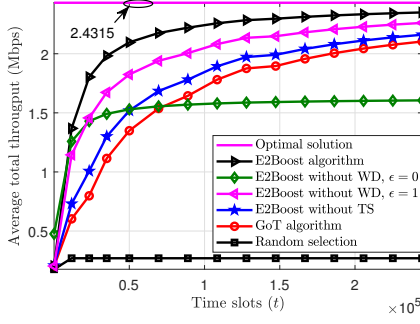
Fig. 8: The performance of different algorithms versus time slot with *optimal phase shift* setting when $\nu_1 = \nu_2 = 2,000, Z = 10$ in a fixed network scenario of Fig. 4.
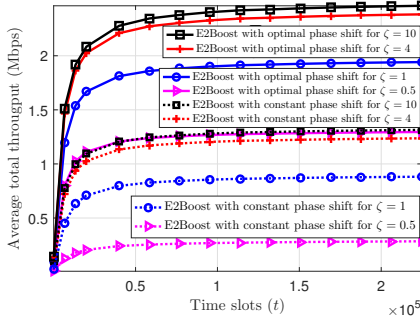

Fig. 9: The performance of E2Boost algorithm for different phase shift settings and Rice factors in a fixed network scenario of Fig. 4.

convergence rate than the GoT algorithm and the E2Boost algorithm without TS.

Next, we evaluate the impact of the RIS-enabled channel on the performance of the proposed algorithm. Fig. 9 depicts the performance of the E2Boost algorithm under the optimal and constant phase shift setting for different Rice factors ($\zeta = 0.5, 1, 4, 10$) when $\nu_1 = \nu_2 = 2,000, Z = 10$. We can see that the performance of the *optimal phase shift* setting is much better than the *constant phase shift* setting for different Rice factors. This is because the optimal phase shift is designed for the centralized UEs. Thus, IoT devices close to the UEs will also have better performance. On the other hand, a bigger ζ will result in a higher average total throughput. This phenomenon can be explained by (3), where a big ζ means that the channel gain is dominated by the LoS component, i.e., the directional reflection link of IoT-RIS-BS. Therefore, the channel gain is dominated by the RIS when ζ trends to $+\infty$; while ζ trends to 0 mean that the IoT device only transmits on Pattern II.

C. Random Network Scenario

In the following, we evaluate the proposed algorithm under the random network scenario. At each MC trial, we regenerate the locations of the IoT devices uniformly in the circle area of Fig. 4. Meanwhile, the distance of any two devices is subject to no less than 5 m. The locations of RISs, UEs, and BS are set the same as those in Fig. 4.

Fig. 10 compares the average total throughput of different algorithms in the *optimal phase shift* setting with $\nu_1 = \nu_2 =$
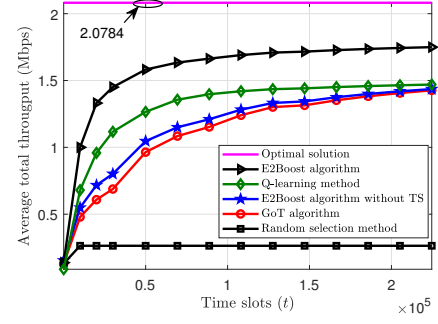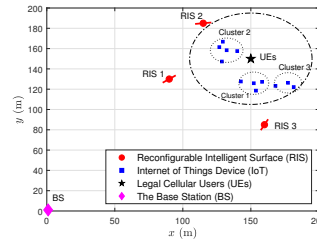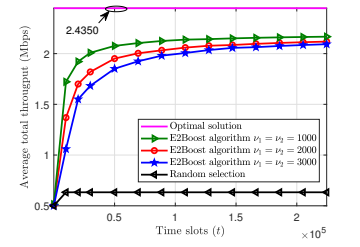

Fig. 10: The performance of different algorithms versus time slot with *optimal phase shift* setting when $\nu_1 = \nu_2 = 2000, Z = 10$ over 10^3 random network scenarios.

(a) A random network scenario   (b) Performance comparison at left network scenario

Fig. 11: (a) A random network scenario in a 200×200 m square area with $K = 3, N = 11$. (b) The performance of different algorithms versus time slot with *optimal phase shift* setting over 10^3 random network scenarios.

$2,000, Z = 11$ over 10^3 random network scenarios. It can be observed that the performance of all algorithms except the random selection method increases with time slot t. Again, the E2Boost algorithm has the best performance and a fast convergence rate. The Q-learning method also exhibits a fast convergence rate, but it suffers from some performance loss due to the lack of the non-cooperation game phase to resolve the collisions among players. Moreover, the gaps between the optimal solution and these algorithms increase compared with Fig. 8 in the fixed network case. The reason is that these algorithms fail to find the optimal RIS for each player under some extreme network scenarios with the constant parameter ν_2 and the limited time horizon T.

Moreover, we study the performance of the proposed algorithm by considering the case that the number of IoT devices is larger than that of RISs, i.e., $N > K$. We first generate a new random network scenario, as shown in Fig. 11a, where $N = 11$, $K = 3$, and the other parameters are the same as those in Fig. 4. We can see from Fig. 11a that $N = 11$ IoT devices are divided into three clusters by using the k-means clustering method according to their geographic locations. Fig. 11b compares the performance of the modified E2Boost algorithm (i.e., Algorithm 2) with different settings of $\nu_1 = \nu_2 = \{1000, 2000, 3000\}$, and the random selection method in the *optimal phase shift* setting over 10^3 random network scenarios of Fig. 11a. It can be seen that the modified E2Boost algorithm with $\nu_1 = \nu_2 = 1,000$
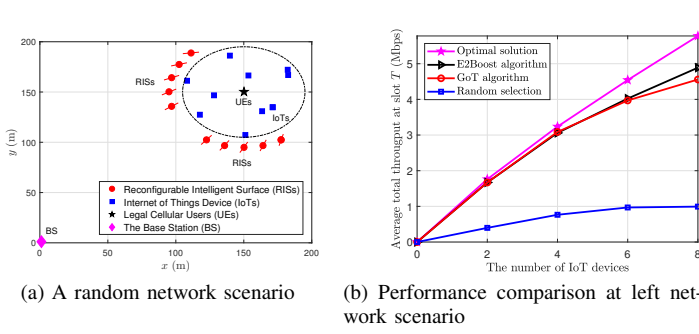
(a) A random network scenario

(b) Performance comparison at left network scenario

Fig. 12: (a) A random network scenario in a 200 m × 200 m square area with $K = 10$. (b) The performance of the optimal solution, the E2Boost algorithm, the original GoT algorithm, and the random selection method versus the number of IoT devices at 10^3 random network scenarios of the left figure.

has the best performance, and all the algorithms except for the random selection method can converge to the optimal allocation. Compared with the results in Fig. 10, the average total throughput in the network scenario of Fig. 11a is about 2.4350 Mbps, which is slightly better than 2.0784 Mbps in Fig. 4. This demonstrates that, although the number of IoT devices increases, the performance gain from the non-RIS-assisted transmission pattern is insignificant.

At last, we investigate the influence of the number of IoT devices on the proposed algorithm. The number of RISs is set to 10 and is placed on a semicircle with a radius of 55 m from $3\pi/4$ to $5\pi/4$, as shown in Fig. 12a. The distances between two neighboring RISs are equal except for the two pairs that are located in the middle and both ends. Fig. 12b shows the performance of the optimal solution, the E2Boost algorithm, the original GoT algorithm, and the random selection method versus the number of IoT devices in the *optimal phase shift* setting where $\nu_1 = \nu_2 = 2,000, K = 10, Z = 10$ over 10^3 random network scenarios of Fig. 12a. It can be seen that the performance of these algorithms increases with the number of IoT devices. However, the proposed algorithm is better than the GoT algorithm and the random selection method since it has a small arm space to explore. In addition, the performance gaps between the optimal solution and these algorithms increase with the number of IoT devices. The reason is that collision probabilities among players increase with the number of IoT devices, resulting in more performance loss.

VII. CONCLUSION AND DISCUSSION

This paper studied the resource allocation problem in a RIS-assisted hybrid uplink network where several IoT devices transmit data to the BS. The objective is to maximize the sum rates of all IoT devices by finding the optimal RIS and SF for each device. We modeled this problem as a two-stage MPMAB framework, where the first stage is to find the optimal RIS, and the second stage is to find the optimal SF. Then, we proposed an E2Boost algorithm to tackle this problem by combining the ϵ-greedy algorithm, TS algorithm, and non-cooperation game method. Therefore,

it can efficiently balance the EE dilemma. Furthermore, we provided an upper regret bound for the E2Boost algorithm, i.e., $\mathcal{O}(\log_2^{1+\delta} T)$, indicating that the per-round regret will trend to 0 when T is sufficiently large. In addition, simulation results demonstrated the effectiveness of the proposed algorithm. More importantly, it is not sensitive to the joint arm space thanks to the two-stage allocation mechanism, which can benefit practical applications.

In the system model, we assume that different RISs use different frequencies, and one RIS can serve at most one IoT device. A more general scenario is that RIS can reuse these frequencies and serve multi-IoT devices. Then, two interesting problems are how to design a mechanism that the UE' and IoT device' signals can coexist in the same RIS and how to design the RIS-assisted multi-IoT system by estimating the exact information of the RIS and CSI. These are important yet challenging problems for future study.

APPENDIX A
PROOF OF THEOREM 1

At each time slot, IoT device transmits on either the Pattern I or the Pattern II. For the Pattern I, the total pseudo-regret term $\overline{Reg}^{(1)}$ can be expanded by investigating $\overline{Reg}_z$, where z is the epoch. Thus, we begin to bound $\overline{Reg}_z$ by computing the probability of event E_z, which is the event that the optimal assignment $\boldsymbol{a}^*$ is not adopted in the third phase at epoch z. We have

$$\Pr(E_z) = \Pr\left(E_z^{k^*}, E_z^{m^*}\right) + \Pr\left(E_z^{k^*}, \overline{E_z^{m^*}}\right) + \Pr\left(\overline{E_z^{k^*}}, E_z^{m^*}\right)$$
$$= \Pr\left(E_z^{k^*}\right) + \Pr\left(\overline{E_z^{k^*}}, E_z^{m^*}\right),$$
(33)

where $E_z^{k^*}$ is the event that the optimal RIS is not used at the end of the z-th epoch of the second phase, and $E_z^{m^*}$ is the event that the optimal SF is not used at the end of the z-th epoch of the third phase.

First, we bound the probability that event $E_z^{k^*}$ holds, i.e.,

$$\Pr\left(E_z^{k^*}\right) \le \Pr\left(\bigcup_{j=0}^{\lfloor \frac{z}{2}\rfloor} P_{e,z-j}\right) + P_{g,z},$$
(34)

where $P_{e,z}$ is the probability that the optimal assignment is different from $\boldsymbol{a}^*$ in the first phase at epoch z, and $P_{g,z}$ is the probability that the frequently visited strategy profile is not $\boldsymbol{a}^*$ in the last $\lfloor z/2 \rfloor + 1$ non-cooperation game phases. Then, we need to calculate the probabilities of $P_{e,z}$ and $P_{g,z}$ before bounding $\overline{Reg}_z$. In the first phase, we estimate the average successful transmission probabilities $\hat{\theta}_{n,k}$ of all RISs. Assume i.i.d. rewards $X_{n,k}$ and each player uniformly explores all K RISs when event $E_z^{k^*}$ holds. By adopting the result in [23] (see *Lemma 8*), we have

$$P_{e,z} \le 2NKe^{-w\nu_1\left(\frac{z}{2}\right)^\delta z} + NKe^{-\frac{\nu_1\left(\frac{z}{2}\right)^\delta}{36K^2}z},$$
(35)

where w is a predefined positive constant. Therefore,

$$\Pr\left(\bigcup_{j=0}^{\lfloor\frac{z}{2}\rfloor} P_{e,z-j}\right) \le \frac{2NKe^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^\delta z}}{1-e^{-w\nu_1\left(\frac{z}{4}\right)^\delta}} + \frac{NKe^{-\frac{\nu_1\left(\frac{z}{4}\right)^\delta}{72K^2}z}}{1-e^{-\frac{\nu_1}{36K^2}\left(\frac{z}{2}\right)^\delta}}. \tag{36}$$

In the second phase, we investigate the probability that the optimal strategy profile is not visited frequently. Let $v^{z*} = [\boldsymbol{a}^{k*}, C^N]$ be the optimal strategy profile in the z-th game phase and $F_z(v^*)$ be the number of times the optimal strategy profile has been visited at the last $\lfloor\frac{z}{2}\rfloor + 1$ game phases. According to [23] (see *Lemma 16*), we have

$$F_z(v^*) \triangleq \sum_{i=z-\lfloor\frac{z}{2}\rfloor}^{z} \sum_{t\in\mathcal{G}_z} \mathbb{I}\left(v(t) = v^{i*}\right), \ \forall k \in \mathcal{K}. \tag{37}$$

Denote the stationary distribution of the optimal strategy profile by $\pi_{v^*} = \min_{z-\lfloor\frac{z}{2}\rfloor\le j\le z} \pi_{v^{i*}}$. If $0 < \eta < \frac{1}{2}$, then $\pi_{v^*} > \frac{1}{2(1-\eta)}$ for a sufficiently large z, we have

$$P_{g,z} \triangleq \Pr\left(F_z(v^*) \le \frac{1}{2}\sum_{i=z-\lfloor\frac{z}{2}\rfloor}^{z} \nu_2 i^\delta\right)$$
$$\le \left(C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^\delta}\right)^z, \tag{38}$$

where C_0 is a constant and independent of z, π_{v^*} and η.

Second, we bound the probability that event $\left(\overline{E_z^{k*}}, E_z^{m*}\right)$ holds. The method is based on the regret analysis of the TS algorithm [29]. Here, event $\overline{E_z^{k*}}$ means that the player found the optimal RIS at the end of the z-th game phase. Let $P_{t,z}^n$ be the probability that player n fails to find the best SF. Since players can find the optimal SF in the third phase only when event $\overline{E_z^{k*}}$ holds, we have

$$\Pr\left(E_z^{m*}|\overline{E_z^{k*}}\right) \triangleq \sum_{n=1}^{N}\Pr\left(\sum_{m=1,m\ne m^*}^{M}\sum_{j=1}^{z}W_{n,m}^j \ge \frac{1}{2}\sum_{i=1}^{z}\nu_3 2^i\right)$$
$$\overset{(a)}{\le} \sum_{n=1}^{N} 2^{-D_{kl}\left(\left(1-\frac{c_m^*\theta_{n,m}^*}{\sum_{m=1}^{M}c_m\theta_{n,m}}\right)\|\frac{1}{2}\right)\sum_{i=1}^{z}\nu_3 2^i}$$
$$\overset{(b)}{\le} \sum_{n=1}^{N} 2^{-2\left(\frac{c_m^*\theta_{n,m}^*}{\sum_{m=1}^{M}c_m\theta_{n,m}}-\frac{1}{2}\right)^2\left(2^{z+1}-2\right)\nu_3}$$
$$\overset{(c)}{\le} N2^{-\frac{(M-2)^2(2^z-1)\nu_3}{M^2}}, \tag{39}$$

where D_{kl} is the KL-divergence and W_{n,m^*}^j is the number of times that the m-th suboptimal SF has been selected by player n up to the j-th epoch. Inequality (a) holds by using the large deviation theory [41]. Inequality (b) follows from the Pinsker's inequality, i.e., $D_{kl}(p\|q) \ge 2(p-q)^2$. Inequality (c) holds due to $Mc_m^*\theta_{n,m}^* \ge \sum_{m=1}^{M}c_m\theta_{n,m}$, considering the worst case that each SF has the same probability of being selected. Therefore, by using $\Pr\left(\overline{E_z^{k*}}, E_z^{m*}\right) = \Pr\left(E_z^{m*}|\overline{E_z^{k*}}\right)\Pr\left(\overline{E_z^{k*}}\right)$, we have (40) which is given in the top of next page.

Then, we continue to bound $\overline{Reg}_z$ based on (36), (38) and

(40). For $z > z_0$, we have

$$\overline{Reg}_z \le N\Gamma_{\max}\nu_2 z^\delta + \Pr(E_z)N\Gamma_{\max}\nu_1 z^\delta + \Pr(E_z)N\Gamma_{\max}\nu_3 2^z$$
$$\le N\Gamma_{\max}\nu_2 z^\delta + N\Gamma_{\max}\left(\nu_1 z^\delta + \nu_3 2^z\right)\left(\frac{2NKe^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^\delta z}}{1-e^{-w\nu_1\left(\frac{z}{4}\right)^\delta}}\right.$$
$$+ \frac{NKe^{-\frac{\nu_1\left(\frac{z}{4}\right)^\delta}{72K^2}z}}{1-e^{-\frac{\nu_1}{36K^2}\left(\frac{z}{2}\right)^\delta}} + \left(C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^\delta}\right)^z\right) +$$
$$N^2\Gamma_{\max}\left(\nu_1 z^\delta + \nu_3 2^z\right)2^{-\frac{(M-2)^2(2^z-1)\nu_3}{M^2}}\left(1 - \frac{2NKe^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^\delta z}}{1-e^{-w\nu_1\left(\frac{z}{4}\right)^\delta}}\right.$$
$$+ \frac{NKe^{-\frac{\nu_1\left(\frac{z}{4}\right)^\delta}{72K^2}z}}{1-e^{-\frac{\nu_1}{36K^2}\left(\frac{z}{2}\right)^\delta}} + \left(C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^\delta}\right)^z\right)$$
$$\le N\Gamma_{\max}\left(\frac{\nu_1}{2} + 3NK + \nu_2\right)z^\delta + N\Gamma_{\max}(6NK+1)\nu_3$$
$$+ N^2\Gamma_{\max}\left(\nu_1 z^\delta + \nu_3 2^z\right)2^{-\nu_3(2^z-1)}. \tag{41}$$

where $\Gamma_{\max} = \max_{n,i}\mu_{n,i}$ is the maximum real expected reward among all players' arms. The first inequality holds since we consider the worst case that each player contributes the maximum regret $\Gamma_{\max}$. The second inequality follows by using (36) and (38). The last inequality establishes on the facts that, for $z > z_0$,

$$\max\left\{C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v^*}-\frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^\delta}, e^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^\delta}, e^{-\frac{\nu_1\left(\frac{z}{4}\right)^\delta}{72K^2}}\right\}$$
$$< \frac{1}{2} \tag{42}$$

and

$$2^{-\frac{(M-2)^2(2^z-1)\nu_3}{M^2}} \le 2^{-\nu_3(2^z-1)}. \tag{43}$$

Finally, let Z be the total number of epochs. The total pseudo-regret $\overline{Reg}^{(1)}$ in Pattern I can be bounded as

$$\overline{Reg}^{(2)} \overset{(a)}{\le} \sum_{z=1}^{Z}\overline{Reg}_z \overset{(b)}{\le} N\Gamma_{\max}\sum_{z=1}^{z_0}\left((\nu_1+\nu_2)z^\delta + \nu_3 2^z\right)$$
$$+ N\Gamma_{\max}\sum_{z=z_0+1}^{Z}\left(N\Gamma_{\max}\left(\frac{\nu_1}{2} + 3NK + \nu_2\right)z^\delta\right.$$
$$+ N\Gamma_{\max}(6NK+1)\nu_3 + N^2\Gamma_{\max}\left(\nu_1 z^\delta + \nu_3 2^z\right)2^{-\nu_3(2^z-1)}\Big)$$
$$\overset{(c)}{\le} N\Gamma_{\max}\sum_{z=1}^{z_0}\nu_3 2^z + N\Gamma_{\max}\sum_{z=1}^{Z}(\nu_1+\nu_2)z^\delta + ZN\Gamma_{\max}(6NK+1)\nu_3$$
$$\overset{(d)}{\le} N\Gamma_{\max}(\nu_1+\nu_2)\log_2^{1+\delta}\left(\frac{T}{\nu_3}+2\right) + N\Gamma_{\max}\nu_3 2^{z_0+1}$$
$$+ N\Gamma_{\max}(6NK+1)\nu_3\log_2\left(\frac{T}{\nu_3}+2\right) = O(\log_2^{1+\delta}T), \tag{44}$$

where (b) holds since (41) for $z > z_0$ and the worst case of $\Gamma_{\max}$ pre-round regret for $z \le z_0$. Inequality (d) follows from $\sum_{z=1}^{Z}z^\delta \le Z^{1+\delta}$ and the fact that $T \ge \sum_{z=1}^{Z-1}\nu_3 2^z \ge \nu_3(2^Z-2)$, which gives $Z^{1+\delta} \le \log_2^{1+\delta}(T/\nu_3+2)$.

For the Pattern II, the total pseudo-regret $\overline{Reg}^{(2)}$ can be bounded according to the regret analysis of the MTS algorithm

$$\Pr\left(\overline{E_z^{k*}}, E_z^{m*}\right) \le N2^{-\frac{(M-2)^2(2^z-1)\nu_3}{M^2}} \left(1 - \frac{2NKe^{-\frac{w}{2}\nu_1\left(\frac{z}{4}\right)^\delta z}}{1 - e^{-w\nu_1\left(\frac{z}{4}\right)^\delta}} + \frac{NKe^{-\frac{\nu_1\left(\frac{z}{4}\right)^\delta}{72K^2}z}}{1 - e^{-\frac{\nu_1}{36K^2}\left(\frac{z}{2}\right)^\delta}} + \left(C_0 e^{-\frac{\nu_2\eta^2}{144T_m\left(\frac{1}{8}\right)}\left(\pi_{v*} - \frac{1}{2(1-\eta)}\right)\left(\frac{z}{2}\right)^\delta}\right)^z\right). \quad (40)$$

in [28], i.e.,

$$\overline{Reg}^{(2)} \le P_a(1+\varpi) \sum_{n=1}^{N} \sum_{a_n \in \mathcal{M}} \frac{\log_2 T}{\mathrm{D_{KL}}(a_n, a_n^*)} \Delta_{n,a_n}, \quad (45)$$

where $\varpi \in (0,1)$ and $\mathrm{D_{KL}}(\cdot)$ is the KL-divergence. Term P_a is the active probability of the UE.

To sum up, the total pseudo-regret $\overline{Reg}$ of Algorithm 1 is given by

$$\begin{aligned}
\overline{Reg} =&\overline{Reg}^{(1)} + \overline{Reg}^{(2)} \\
\le& N\Gamma_{\max}(1 - P_a)\left(2(\nu_1 + \nu_2)\log_2^{1+\delta}\left(\frac{T}{\nu_3} + 2\right)\right. \\
&\left. + (6NK+1)\nu_3 \log_2\left(\frac{T}{\nu_3} + 2\right)\right) \\
&+ P_a(1+\varpi) \sum_{n=1}^{N} \sum_{a_n \in \mathcal{M}} \frac{\log_2 T}{\mathrm{D_{KL}}(a_n, a_n^*)} \Delta_{n,a_n}.
\end{aligned} \quad (46)$$

APPENDIX B
RIS'S DIRECTION AND ELEMENT'S LOCATION

We first determine the direction of the RIS in XY-plane by computing the angle $\angle\varphi$ between X-axis and the RIS, as shown in Fig. 3. Given the coordinates of B $= (x_\mathrm{B}, y_\mathrm{B})$, R $= (x_\mathrm{R}, y_\mathrm{R})$, U $= (x_\mathrm{U}, y_\mathrm{U})$, we have two vectors $\overrightarrow{\mathrm{RB}} = (x_\mathrm{B} - x_\mathrm{R}, y_\mathrm{B} - y_\mathrm{R})$ and $\overrightarrow{\mathrm{RU}} = (x_\mathrm{U} - x_\mathrm{R}, y_\mathrm{U} - y_\mathrm{R})$. According to the plane analytical geometry theory, we can obtain the direction vector $\overrightarrow{\mathrm{RC}}$, i.e., the bisector of angle $\angle\mathrm{BRU}$, as

1) If $\cos\langle \overrightarrow{\mathrm{RB}}, \overrightarrow{\mathrm{RU}}\rangle \ge 0$, then

$$\begin{aligned}
\overrightarrow{\mathrm{RC}} = (x_\mathrm{RC}, y_\mathrm{RC}) &= -\frac{\overrightarrow{\mathrm{RB}}}{|\overrightarrow{\mathrm{RB}}|} + \frac{\overrightarrow{\mathrm{RU}}}{|\overrightarrow{\mathrm{RU}}|} \\
&= \left(\frac{x_\mathrm{R} - x_\mathrm{B}}{|\overrightarrow{\mathrm{RB}}|} + \frac{x_\mathrm{U} - x_\mathrm{R}}{|\overrightarrow{\mathrm{RU}}|}, \frac{y_\mathrm{R} - y_\mathrm{B}}{|\overrightarrow{\mathrm{RB}}|} + \frac{y_\mathrm{U} - y_\mathrm{R}}{|\overrightarrow{\mathrm{RU}}|}\right);
\end{aligned}$$

2) If $\cos\langle \overrightarrow{\mathrm{RB}}, \overrightarrow{\mathrm{RU}}\rangle < 0$, then

$$\begin{aligned}
\overrightarrow{\mathrm{RC}} = (x_\mathrm{RC}, y_\mathrm{RC}) &= \frac{\overrightarrow{\mathrm{RB}}}{|\overrightarrow{\mathrm{RB}}|} + \frac{\overrightarrow{\mathrm{RU}}}{|\overrightarrow{\mathrm{RU}}|} \\
&= \left(\frac{x_\mathrm{B} - x_\mathrm{R}}{|\overrightarrow{\mathrm{RB}}|} + \frac{x_\mathrm{U} - x_\mathrm{R}}{|\overrightarrow{\mathrm{RU}}|}, \frac{y_\mathrm{B} - y_\mathrm{R}}{|\overrightarrow{\mathrm{RB}}|} + \frac{y_\mathrm{U} - y_\mathrm{R}}{|\overrightarrow{\mathrm{RU}}|}\right).
\end{aligned}$$

Thus, the direction of the RIS in XY-plane (i.e., the normal vector $\overrightarrow{\mathrm{AR}}$ of line RC) is $\overrightarrow{\mathrm{AR}} = (-y_\mathrm{RC}, x_\mathrm{RC})$. It is easy to obtain the angle $\angle\varphi$ by

$$\angle\varphi = -\arctan\left(\frac{x_\mathrm{RC}}{y_\mathrm{RC}}\right). \quad (47)$$

Next, based on the angle $\angle\varphi$, we can compute the location of each element in the RIS, i.e.,

$$\begin{cases}
x(l_1, l_2) = (l_1 - 51)\,d_v\cos\angle\varphi + x_\mathrm{R}, \\
y(l_1, l_2) = (l_1 - 51)\,d_v\sin\angle\varphi + y_\mathrm{R}, \\
z(l_1, l_2) = (l_2 - 51)\,d_h + 10,
\end{cases} \quad (48)$$

where $d_v = d_h = 0.01$ are the offsets in RIS's horizontal and vertical planes, respectively. Constant 51 is the 51-th row or column elements in the RIS and constant 10 is the height of the RIS. Symbol (l_1, l_2) are the integers in $[0, 101]$, standing for the index ceil of the RIS elements' matrix.

REFERENCES

[1] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Feb. 2019.

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

[3] M. A. ElMossallamy, H. Zhang, L. Song, K. G. Seddik, Z. Han, and G. Y. Li, "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities," *IEEE Trans. on Cogn. Commun. and Net.*, vol. 6, no. 3, pp. 990–1002, Mar. 2020.

[4] H. Zhang, B. Di, L. Song, and Z. Han, "Reconfigurable intelligent surfaces assisted communications with limited phase shifts: How many phase shifts are enough?" *IEEE Trans. Veh. Technol.*, vol. 64, no. 4, pp. 4498–4502, Apr. 2020.

[5] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," vol. 38, no. 8, pp. 1809–1822, Aug. 2020.

[6] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. Di Renzo, "Reconfigurable intelligent surface assisted UAV communication: Joint trajectory design and passive beamforming," vol. 9, no. 5, pp. 716–720, May 2020.

[7] H. Zhang, B. Di, L. Song, and Z. Han, *Reconfigurable Intelligent Surface-Empowered 6G*. Springer, 2021.

[8] O. Liberg, M. Sundberg, E. Wang, J. Bergman, and J. Sachs, *Cellular Internet of things: technologies, standards, and performance*. Academic Press, 2017.

[9] Q. Qi and X. Chen, "Wireless powered massive access for cellular Internet of Things with imperfect SIC and nonlinear EH," *IEEE Internet of Things J.*, vol. 6, no. 2, pp. 3110–3120, Feb. 2018.

[10] S. Dama, V. Sathya, K. Kuchi, and T. V. Pasca, "A feasible cellular Internet of Things: Enabling edge computing and the iot in dense futuristic cellular networks," *IEEE Consumer Electron. Mag.*, vol. 6, no. 1, pp. 66–72, Jan. 2016.

[11] M. Elsaadany, A. Ali, and W. Hamouda, "Cellular LTE-A technologies for the future Internet-of-Things: Physical layer features and challenges," *IEEE Commun. surveys Tuts.*, vol. 19, no. 4, pp. 2544–2572, Apr. 2017.

[12] A. Waret, M. Kaneko, A. Guitton, and N. El Rachkidy, "LoRa throughput analysis with imperfect spreading factor orthogonality," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 408–411, Oct. 2018.

[13] J. Lyu, D. Yu, and L. Fu, "Achieving max-min throughput in LoRa networks," in *Int. Conf. on Comput., Net. and Commun.*, Big Island, HI, Feb. 2020.

[14] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[15] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

[16] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Select. Areas in Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[17] B. Gu, X. Zhang, Z. Lin, and M. Alazab, "Deep multiagent reinforcement-learning-based resource allocation for Internet of controllable things," *IEEE Int. of Things J.*, vol. 8, no. 5, pp. 3066–3074, May 2020.

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

[19] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations and Trends® in Mach. Learning*, vol. 5, no. 1, pp. 1–122, Dec. 2012.

[20] D.-T. Ta, K. Khawam, S. Lahoud, C. Adjih, and S. Martin, "LoRa-MAB: Toward an intelligent resource allocation approach for LoRaWAN," in *Proc. IEEE Glob. Telecom. Conf.*, Waikoloa, HI, Dec. 2019.

[21] H. Tibrewal, S. Patchala, M. K. Hanawal, and S. J. Darak, "Distributed learning and optimal assignment in multiplayer heterogeneous networks," in *Proc. IEEE INFOCOM*, Pairs, France, Jun. 2019.

[22] S. Zafaruddin, I. Bistritz, A. Leshem, and D. Niyato, "Distributed learning for channel allocation over a shared spectrum," *IEEE J. Select. Areas in Commun.*, vol. 37, no. 10, pp. 2337–2349, Aug. 2019.

[23] I. Bistritz and A. Leshem, "Distributed multi-player bandits-a game of thrones approach," in *Advances in Neural Inform. Process. Syst.*, Montréal, Canada, Dec. 2018.

[24] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "The nonstochastic multi-armed bandit problem," *SIAM Journal on Computing*, vol. 32, no. 1, pp. 48–77, Jan. 2002.

[25] R. Jonker and T. Volgenant, "Improving the Hungarian assignment algorithm," *Operations Res. Letters*, vol. 5, no. 4, pp. 171–175, Apr. 1986.

[26] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learning*, vol. 47, no. 2-3, pp. 235–256, Feb. 2002.

[27] B. Arras, E. Azmoodeh, G. Poly, and Y. Swan, "A bound on the Wasserstein-2 distance between linear combinations of independent random variables," *Stochast. Processes and Their Appl.*, vol. 129, no. 7, pp. 2341–2375, Jul. 2019.

[28] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-complexity, low-regret link rate selection in rapidly-varying wireless channels," in *Proc. IEEE INFOCOM*, Honolulu, HI, Apr. 2018.

[29] S. Agrawal and N. Goyal, "Further optimal regret bounds for Thompson sampling," in *Artificial Intell. and Statist.*, Scottsdale, AZ, Apr. 2013.

[30] L. You, J. Xiong, Y. Huang, D. W. K. Ng, C. Pan, W. Wang, and X. Gao, "Reconfigurable intelligent surfaces-assisted multiuser MIMO uplink transmission with partial CSI," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5613–5627, Sep. 2017.

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

[32] 3GPP TR 38.901, "Study on channel model for frequencies from 0.5 to 100 GHz (release 14)," *3GPP*, Jan. 2018.

[33] J. Tong, M. Jin, Q. Guo, and Y. Li, "Cooperative spectrum sensing: A blind and soft fusion detector," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2726–2737, Apr. 2018.

[34] J. Tong, M. Jin, Q. Guo, and L. Qu, "Energy detection under interference power uncertainty," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1887–1890, Aug. 2017.

[35] J. Choi, C. Joo, J. Zhang, and N. B. Shroff, "Distributed link scheduling under SINR model in multihop wireless networks," *IEEE/ACM Trans. Networking.*, vol. 22, no. 4, pp. 1204–1217, Aug. 2014.

[36] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-complexity, low-regret link rate selection in rapidly-varying wireless channels," in *Proc. IEEE INFOCOM*, Honolulu, HI, Apr. 2018.

[37] A. Menon and J. S. Baras, "Convergence guarantees for a decentralized algorithm achieving Pareto optimality," in *American Control Conf.*, Washington, DC, Jun. 2013.

[38] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving Pareto optimality through distributed learning," *SIAM J. on Control and Optimization*, vol. 52, no. 5, pp. 2753–2770, May 2014.

[39] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. of Comput. Vision*, vol. 40, no. 2, pp. 99–121, Nov. 2000.

[40] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, Feb. 2003.

[41] T. M. Cover and J. A. Thomas, *Elements of information theory.* John Wiley & Sons, 2012.
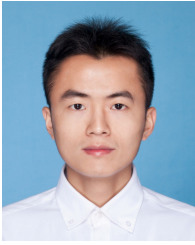
Jingwen Tong received the B.E. degree in electrical engineering from China Jiliang University, Hangzhou, China, in 2015, and the M.S. degree in electrical engineering from Ningbo University, Ningbo, China, in 2018. He is currently pursuing the Ph.D. degree in department of communication engineering with Xiamen University, Xiamen, China. His current research interests include multi-armed bandit and game theory.

Hongliang Zhang (S'15-M'19) received the B.S. and Ph.D. degrees at the School of Electrical Engineering and Computer Science at Peking University, in 2014 and 2019, respectively. Currently, he is a Postdoctoral Associate in the Department of Electrical and Computer Engineering at Princeton University, New Jersey. His current research interest includes reconfigurable intelligent surfaces, aerial access networks, optimization theory, and game theory. He received the best doctoral thesis award from Chinese Institute of Electronics in 2019. He is an exemplary reviewer for IEEE Transactions on Communications in 2020. He is also the recipient of 2021 IEEE Comsoc Heinrich Hertz Award for Best Communications Letters and 2021 IEEE ComSoc Asia-Pacific Outstanding Paper Award. He has served as a TPC Member for many IEEE conferences, such as Globecom, ICC, and WCNC. He is currently an Editor for IEEE Communications Letters, IET Communications, and Frontiers in Signal Processing. He has also served as a Guest Editor for several journals, such as IEEE Internet of Things Journal, Journal of Communications and Networks, etc.
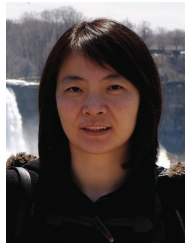
Liqun Fu (S'08-M'11-SM'17) is a Full Professor of the School of Informatics at Xiamen University, China. She received her Ph.D. Degree in Information Engineering from The Chinese University of Hong Kong in 2010. She was a post-doctoral research fellow with the Institute of Network Coding of The Chinese University of Hong Kong, and the ACCESS Linnaeus Centre of KTH Royal Institute of Technology during 2011-2013 and 2013-2015, respectively. She was with ShanghaiTech University as an Assistant Professor during 2015-2016.

Her research interests are mainly in communication theory, optimization theory, and learning theory, with applications in wireless networks. She is on the editorial board of IEEE Access and the Journal of Communications and Information Networks (JCIN). She served as the Technical Program Co-Chair of the GCCCN Workshop of the IEEE INFOCOM 2014, the Publicity Co-Chair of the GSNC Workshop of the IEEE INFOCOM 2016, and the Web Chair of the IEEE WiOpt 2018. She also serves as a TPC member for many leading conferences in communications and networking, such as the IEEE INFOCOM, ICC, and GLOBECOM.

Amir Leshem received the B.Sc. degree (cum laude) in mathematics and physics, the M.Sc. degree (cum laude) in mathematics, and the Ph.D. degree in mathematics from Hebrew University, Jerusalem, Israel, in 1986, 1990, and 1998, respectively. He is currently a Professor and one of the Founders of the Faculty of Engineering, Bar-Ilan University, where he is also the Head of the Signal Processing Track. From 2003 to 2005, he was the Technical Manager of the U-BROAD Consortium developing technologies to provide 100 Mb/s and beyond over copper lines. His main research interests include multichannel wireless and wireline communication, applications of game theory to dynamic and adaptive spectrum management of communication networks, array and statistical signal processing with applications to multiple element sensor arrays and networks, wireless communications, radio astronomical imaging and brain research, set theory, and logic and foundations of mathematics. He was an Associate Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2008 to 2011. He was the Leading Guest Editor for Special Issues on signal processing for astronomy and cosmology in the IEEE Signal Processing Magazine and the IEEE JOURNAL OF SELECTED TOPICS.

Zhu Han (S'01–M'04-SM'09-F'14) received the B.S. degree in electronic engineering from Tsinghua University, in 1997, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. His research interests include wireless resource allocation and management, wireless communications and networking, game theory, big data analysis, security, and smart grid. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015-2018, AAAS fellow since 2019 and ACM distinguished Member since 2019. Dr. Han is 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of 2021 IEEE Kiyo Tomiyasu Award, for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."