MFTTS: A Mean-Field Transfer Thompson Sampling Approach for Distributed Power Allocation in Unsourced Multiple Access

Tien Thanh Le, *Graduate Student Member, IEEE*, Yusheng Ji, *Fellow, IEEE*, and John C. S. Lui, *Fellow, IEEE*

***Abstract*—Unsourced multiple access (UMA) is a novel approach to support a large number of devices in a massive Machine-Type Communication (mMTC) system. UMA enables devices to concurrently encode their data using the same codebook to transmit without being individually identified, resulting in reduced signaling and computational overhead at the base station. Hybrid-domain non-orthogonal multiple access (NOMA), which combines power-domain NOMA with code-domain NOMA, is another technique that enhances the spectral efficiency of mMTC. While the study of hybrid-domain NOMA has been conducted, its integration with UMA has not been thoroughly investigated. Considering that mMTC traffic primarily consists of sporadic short packets in the uplink direction, employing a fully distributed mMTC multiple access protocol can substantially decrease signaling overhead and latency. In this work, a multi-armed bandits (MAB) paradigm is adopted to create a distributed power selection policy for devices that using UMA. Particularly, an MAB algorithm called Thompson Sampling (TS) is used to allow mMTC devices to minimize the transmission power without violating the minimum receiving signal-to-noise constraint needed to correctly decode the UMA codewords back to the original messages. A mean-field modeling technique is used to approximate the learned policies. The knowledge gained from the approximated policies can be transferred to new devices by initializing their prior distribution, which is called Mean-field Transfer Thompson Sampling (MFTTS). Simulations show that the mean-field approximation is indeed accurate and effective. Interestingly, MFTTS performs better than TS without knowledge transfer as well as other distributed power allocation methods.**

***Index Terms*—Distributed power control, HD-NOMA, mab, mean-field approximation, mMTC.**

Manuscript received 28 August 2023; revised 13 February 2024; accepted 23 April 2024. Date of publication 13 May 2024; date of current version 5 November 2024. This work was supported in part by JSPS KAKENHI under Grant JP20H00592 and Grant JP24K02937, in part by JST ASPIRE under Grant JPMJAP2325, and in part by MIC SCOPE Project under Grant JP235006102. The work of John C. S. Lui was supported in part by RGC under Grant GRF 14202923. Recommended for acceptance by C. Xin. *(Corresponding author: Tien Thanh Le.)*

Tien Thanh Le is with the Department of Informatics, Graduate University for Advanced Studies, SOKENDAI, Tokyo 101-8430, Japan (e-mail: lethanh@nii.ac.jp).

Yusheng Ji is with the National Institute of Informatics and the Graduate University for Advanced Studies, Sokendai, Tokyo 101-8430, Japan.

John C. S. Lui is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong.

This article has supplementary downloadable material available at https://doi.org/10.1109/TMC.2024.3399876, provided by the authors.

Digital Object Identifier 10.1109/TMC.2024.3399876

I. Introduction

IN TODAY'S highly connected world, there is a growing trend of individuals and devices being online and connected via the internet, while the physical resources of the network remain limited. The exponential growth of connected devices is attributed to the increasing adoption of massive Machine-Type Communication (mMTC) applications, such as autonomous vehicles, industrial automation, smart healthcare, and environmental sensing, over the last few decades. According to the latest report from Ericsson, the number of mMTC connected devices was 13.2 billion in 2022 and the projected number of connections in 2028 is 34.7 billion [1]. The 6G Flagship program also predicts that future mMTC will support up to 100 mMTC connections per cubic meter [2]. As a result, future mMTC protocols must be *highly scalable* to support a higher overloading factor to utilize a smaller number of wireless resource blocks.

To address this challenge, grant-free non-orthogonal muliple access (GF-NOMA) technology has emerged as an effective solution for mMTC networks [3], [4]. In GF-NOMA, the base station (BS) acts as the centralized coordinator responsible for allocating the available resource configurations to each connected device. Subsequently, each connected device can select one resource unit from the available options, such as different power levels or user-specific signature sequences or codebooks. The receiver then exploits the chosen resource unit for multi-user data detection and message decoding [3], [4]. Hence, this method enables multiple devices to utilize the same physical channels, albeit at the cost of increased decoding complexity to mitigate error propagation caused by signal interference between devices. non-orthogonal muliple access (NOMA) can be categorized into three main types: power-domain NOMA (PD-NOMA), code-domain NOMA (CD-NOMA), and hybrid-domain NOMA (HD-NOMA). In PD-NOMA, devices can select from different power levels, and successive interference cancellation (SIC) is employed to separate concurrent devices [5]. Conversely, in CD-NOMA, devices can choose from a range of available codebooks, and encode their data using their respective codebook, and message passing algorithm (MPA) decoding is used to decode the received signal [6]. HD-NOMA combines the principles of both PD-NOMA and CD-NOMA within the same system. Yadav et al. conducted an empirical study to evaluate the performance of these three types of NOMA [7]. The study demonstrated that HD-NOMA achieves superior sum rate per-

formance compared to non-hybrid schemes when the number of devices is 50% greater than the number of wireless channels.

In their pioneering research, [8] proposed the fundamental concept of unsourced multiple access (UMA) to enhance the effectiveness of uplink traffic and short-packet scenarios for mMTC [8], [9]. UMA is a new paradigm that tackles the problem that still exists in GF-NOMA for scaling it to even larger scale mMTC networks. First, [8] formally pointed out in [8] that under a sharp increase of active users, by using GF-NOMA, the average channel capacity of users converges to zero even though the sum rate is increasing. In other words, GF-NOMA cannot guarantee reliable uplink data delivery for each specific device as the number of devices increases sharply. Additionally, as the number of users increases exponentially in mMTC networks, the problem of identifying and configuring resources for each individual device becomes increasingly challenging [8]. Furthermore, additional signaling overhead is still needed by GF-NOMA when users join the system or when the state of the system changes, requiring resource reconfiguration [10]. Therefore, the UMA paradigm excludes the user identification problem and forces devices to adopt the same signature (codebook) to send their messages uplink. In UMA, device can transmit its local data uplink directly without any request or handshake with the coordinating center.

In the literature of UMA, the design of the UMA codebook has been receiving a lot of attention [11], [12], [13]. But to the best of our knowledge, only Fengler et al. had studied the hybrid scheme of UMA and power domain NOMA [14]. One of their contributions is to minimize the energy usage of the mMTC devices such that the Signal to Interference & Noise Ratio (SINR) satisfies the minimum threshold required by UMA codebook. However, the authors only solve this problem using a centralized optimization method. The problem of fully distributed power control for each autonomous device remains an open problem.

This paper presents a new approach to address this distributed power allocation problem for UMA-based mMTC communication system via a multi-armed bandit (MAB) paradigm. The proposed MAB policy is designed to address three main requirements of the mMTC scenario. First, the policy needs to be simple and lightweight, because the policy will mostly be implemented on embedded Internet of Things (IoT) devices with limited computational capability. Second, the policy needs to adaptively and gradually minimize the transmitting power of devices. Also, if all devices in our multi-agent system employ our proposed policy, the behavior of the system will remain stable and predictable. The primary contributions of this paper are as follows:

- We propose a fully uncoordinated power allocation protocol for UMA. Each device acts as if it is an autonomous agent, and relies on limited-size broadcast messages from BS to decide its own power level to send it messages. No multi-step information exchange is required.
- Each device uses the Thompson Sampling (TS) policy to minimize its transmission power. The TS policy of each device is initially based on the learned TS policies of the device in the relevant network setting to accelerate the process. The mean-field technique is used to approximate the learned policies of a large number of devices. We denote the proposed method as Mean-Field Transfer Thompson Sampling (MFTTS).
- To quantify the merit of our proposed method under a large number of devices, we characterize our system using a mean-field technique [15]. We show that if the number of devices is large, the behavior of devices in our system converges to a system of ordinary differential equation (ODE). We also demonstrate that the ODE's behaviors are stable, and that our system's behavior converges to a stable equilibrium.

This paper is organized as follows. Section II reviews the related studies. Section III presents the considered network and mMTC traffic model, as well as the proposed uncoordinated power control protocol. In section IV, we describe the fully distributed power allocation problem in HD-NOMA and the proposed algorithm MFTTS. We present a simulation study in the last section.

II. Related Works

This section reviews the latest distributed power control methods, followed by an overview of recent multi-agent learning algorithms

A. Distributed Power Control

In large networks such as the mMTC networks, a distributed resource allocation algorithm is a necessary choice. We expect mMTC devices to act autonomously without relying on the intervention from the coordination center at the BS. In [16], authors enumerated several algorithms for distributed resource allocation in mMTC networks; namely, non-cooperative games, evolutionary games, mean-field games, mean-field bandit games, and mean-field auctions. They concluded that mean-field bandit games and mean-field auctions are the top two methods since they can handle uncertain and incomplete information, while the other methods failed to handle such conditions.

A study of uncoordinated power control using an evolutionary game is presented by [17] [17]. They considered a device pairing scheme, in which, two devices are assigned to an orthogonal wireless resource block, and adjust their power level based on the transmission success rate. Therefore, two devices could transmit on the same orthogonal resource block without collision. Nevertheless, the considered system limits the number of total devices is doubling the number of orthogonal resource blocks only. In addition, the energy usage of the system is highly dependent on the device pairing algorithm, which is dictated by the BS in a centralized manner.

In [18], mean-field games have been applied to design a distributed power control algorithm for device-to-device communication. Here, the chosen power level of every player given the average power level of other players is modeled by a Hamilton-Jacobi-Bellman (HJB), and the movement of the average power level is modeled by a Fokker-Planck-Kolmogorov (FPK) equation. The system of differential equations above is approximately solved using finite difference methods. The method also assumed that the device-to-device link gains do not change while that value is likely to change in a real-world

deployment. [19] proposed a mean-field game for joint resource block allocation and power control. Different from the previous study [18], this study [19] can handle the fluctuation of the link gains, since it requires devices to send their link gains to the BS. The BS aggregates the average transitory channel gains and broadcasts to all devices to compute the next action. However, the downsides of this approach are the additional overhead of gathering the channel gain of all devices and leaking the privacy of devices when sharing their accurate location to BS via the channel gains.

In contrast, MAB-based methods operate without any assumption of static link gains or forwarding the link gains to the BS. MAB has been adopted for distributed power allocation in device-to-device communication [20], [21]. In these studies, the authors conducted simulations to show that upper confidence bound (UCB) policies are more effective than other simple fully distributed power allocation heuristics. Their works did not provide game theoretic analysis or use field games to model a large-scale system with many devices.

There have been several attempts to improve HD-NOMA protocol by combining PD-NOMA and CD-NOMA [22] [23]. The most recent work by Benamor et al. [23] proposes a bi-level optimization approach for a distributed HD-NOMA protocol, called multi-armed bandits mean-field games (MABMFG). In this work, the UCB algorithm is employed to select the best CD-NOMA codebook and the power selection problem is formulated as a mean-field game, following the same approach as in their previous work [19]. However, a major limitation of this approach is that devices are required to send the channel coefficients to the base station and wait for the aggregated information, resulting in excessive signaling overhead compared to combining PD-NOMA with UMA. Moreover, UMA is widely recognized for its power efficiency, making the combination of PD-NOMA with UMA potentially more power efficient than the proposed method.

As far as we know, there is no study on distributed power allocation in combination with UMA. Also, previously distributed power control using MAB did not consider the TS technique [24], which displayed a strong empirical performance [25].

B. Multi-Agent Learning

A multi-agent learning deals with the problem of designing autonomous agents that interact with the environment of other agents to achieve their goals. Multi-agent learning is a more flexible model than the previous game theoretic methods since no prior knowledge about the system is required by multi-agent learners. It can be divided into two main categories: (1) multi-agent (deep) reinforcement learning and (2) multi-player MAB. Recently, multi-agent (deep) reinforcement learning has received much attention and it has been adapted for distributed resource allocation algorithms in cellular networks [26], [27]. However, this method is only suitable if agents are different BSs or devices with decent computational capability and battery because each agent uses deep neural networks to perform distributed optimization.

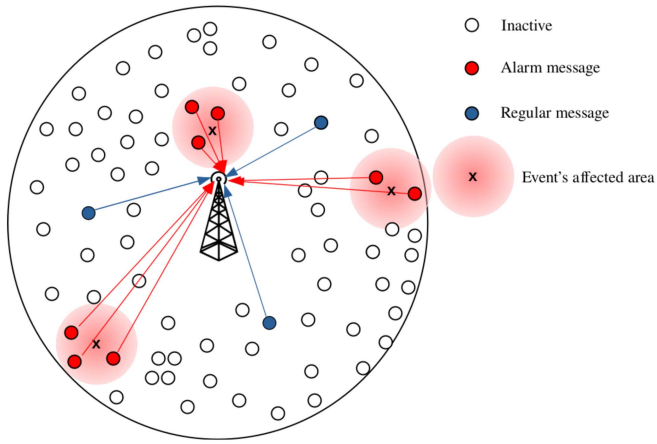


Fig. 1. Overview of the considered mMTC network and uplink traffic at a particular time slot.

On the other hand, multi-player MAB is a better candidate for mMTC scenario, which requires low complexity and energy usage. In [28], a mean-field model is used to analyze the behavior of a large-scale multi-player MAB, under the assumption of a binary reward function and state regeneration. An attempt has been made in order to apply their mean-field bandit games with binary reward model in distributed device association in multiple cell wireless networks [29]. In a recent advance in 2021, [30] formulated mean-field games to analyze the large-scale multi-player MAB under a continuous reward function. The study also provided a certain range of MAB parameters such that the states of the MAB players always converge to a unique equilibrium. To the best of our knowledge, there is no attempt to apply the mean-field model and TS algorithm for large-scale distributed power control.

III. System Model

In this section, we present the mMTC network setting, the traffic model, and the distributed power control for our hybrid domain NOMA scheme.

A. Network Model

We first consider a single-cell wireless system in Fig. 1. Here, the BS is positioned at the center and it needs to support a total of K^{tot} devices. We consider the system to be slotted in time, with time slots indexed by $n \in \{1, 2, \ldots\}$. During time slot n with the duration Δti, only a subset of active devices transmits their messages. Let K be the number of active devices at a time slot and $K \ll K^{tot}$. All active devices utilize a single shared wireless resource block for transmitting their messages. They share the block by encoding their messages with the same UMA codebook, such that the BS can decode messages from different devices. Besides, devices can send their encoded messages with different power levels [14]. Through the combination of code-domain and power-domain diversification, the HD-NOMA system enables an increased number of successfully

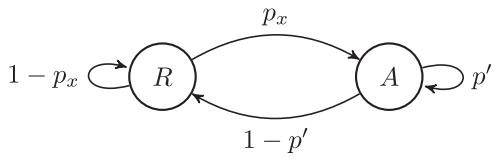


Fig. 2. State transition diagram in each device.

served messages, when compared to power-domain NOMA or code-domain NOMA approaches.

B. Traffic Model

Consider that the uplink traffic from mMTC devices adheres to the mMTC source traffic model in [31]. This model generates the positions of both devices and events over time using two separated Poisson point processes (PPP). Specifically, the locations of devices are distributed around the BS following a PPP Φ_D with a density of λ_D. Note that the devices are deployed at a fixed location. On the other hand, the location of critical events is distributed according to a different PPP Φ_E with a density of λ_E. Unlike device locations, the positions of critical events change over time and are re-sampled at each discrete time slot.

In this source traffic model for mMTC, the packet generation process of each device is modeled by the Gilbert-Elliot model with two states: regular state (R) and alarm state (A) (Fig. 2). In state R, the event of a device generating an uplink packet is distributed according to a Bernoulli distribution with parameter p_R. Likewise, the probability of generating an uplink packet in state A is Bernoulli distributed with parameter p_A. In general, p_A is much higher than p_R. The state of a device changes depending on the location of the devices and events. When the location of an event lies within the sensing perimeter of a device, the state of that device potentially moves from R to A. Then, the probability of a device at location x is triggered by an event at location y is modeled as $p_{xy} = exp(-\frac{d_{xy}}{\bar{d}})$, where d_{xy} is the Euclidean distance between sensor and event, and $\bar{d}$ is a constant to normalize the distance.

Let p_x be the probability of a device at location x be triggered by at least one event $y \in \Phi_E$, then:

$$p_x = 1 - \prod_{y \in \Phi_E} (1 - p_{xy}). \tag{1}$$

A device moves from state R to state A with probability p_x. That device stays in the state A for a duration geometrically distributed with parameter p', and moves back to a regular state with probability $1 - p'$.

C. Distributed Power Allocation Protocol

1) Message Structure: Similar to [14], we consider active devices transmit their messages over UMA codebook with blocklength of $n_p + n_d$ bits. Here, n_p is the length of a pilot sequence and n_d is the length of codewords for the device data. For handling the fast fading, the pilot sequence is transmitted alongside data symbols in uplink packets. The channel estimation error at the BS is quantified by its variance σ_q^2:

$$\sigma_q^2 = \frac{N_0 B}{N_0 B + n_p \pi_q}, \tag{2}$$

where N_0 (dBm/Hz) is the environmental noise spectral density, B (Hz) is the bandwidth of the considered wireless resource block, and π_q (dBm/bit) is the receiving power coefficient. Note that, UMA is expected to serve a large number of devices; therefore, it is infeasible to assign a unique pilot sequence to each participating device. Therefore, devices share a common pool of pilot sequences and each device randomly picks one from 2^{n_p} pilot sequences to send to the BS. In particular, similar to the well-known Birthday Paradox [32], the probability of at least one device selecting the same pilot as another device is given by:

$$1 - \frac{2^{n_p}!}{(2^{n_p} - K)! 2^{K n_p}}, \tag{3}$$

where K is the number of active users. One can choose the pilot sequence length n_p that is not too short to minimize the chance of two devices selecting the same pilot sequence (3), as well as n_p that is not too long to cause pilot signal overhead.

2) Power Selection Procedure: In our system, mMTC devices handle the distortion by slow fading in a distributed manner. In particular, each device estimates and adjusts its transmitting power coefficient to match the predefined receiving power coefficient at the BS [14]. First, let g_k be the path loss and shadowing of each device k. To enable the estimation of g_k, the BS broadcasts a beacon signal at a constant power to all devices. Then, each device k compares the power of the received beacon signal with the known transmit power of the beacon signal to compute its own g_k. The variable g_k is randomly distributed according to:

$$g_k[dB] = \alpha + \beta \log_{10}(d_k) + \sigma^2_{\text{shadow}} \zeta, \tag{4}$$

where d_k is the distance from the k^{th} device to the BS in kilometers and α, β are two constants to represent the offset and scale of the path loss, σ^2_{shadow} is the variance of the shadowing effect, and ζ is a standard Normal distribution random variable.

After the path loss and shadowing are obtained, device k transmits its data with power P_k such that $P_k = \pi_q g_k$ where π_q is the chosen receiving power coefficient. Let $\{\pi_1, \pi_2, \ldots, \pi_q, \ldots, \pi_Q\}$ be Q power levels that BS allows, and SIC is implemented at the BS to decode messages which are received at these levels.

In contrast with power-domain NOMA in which each device is assigned an exclusive power coefficient, this power allocation for UMA scheme allows many devices to share the same power level. Since each device can choose its level without receiving direct instruction by the BS, the power control scheme here is inherently distributed. Each device is expected to choose a power coefficient as small as possible to conserve energy. Also, the SINR of the chosen power level π_q should be higher than the minimum required power level to decode the UMA codewords with negligible error. Let the power levels be sorted in descending order such that $\pi_1 > \pi_2 > \ldots > \pi_q > \ldots > \pi_Q$.

The BS decodes from largest power coefficient (π_1) to the smallest power coefficient (π_Q). We consider a simplified SIC scheme [14], where the BS divides the active users into Q groups, based on their received power. Then all messages within one group are decoded and substracted in parallel starting from the group with the highest average power. In particular, the SINR calculation is given as follows:

$$\text{SINR}(q) = \frac{N^{\text{MIMO}}(1-\sigma_q^2)\pi_q}{N_0 B + \sum_{i=1}^{q-1}[(1-p_e)\sigma_i^2 + p_e]K^i\pi_i + \sum_{j=q}^{Q} K^j\pi_j}, \tag{5}$$

where N^{MIMO} is the number of multiple-input multiple-output (MIMO) receiver antennas. The nominator contains the power of the signal at the q^{th} level after taking into account the gain by N^{MIMO} array antennas and the loss because of the channel estimation error $(1-\sigma_q^2)$. K^q is the number of devices that selects the q^{th} power level. p_e is the maximum probability of decoding error by the UMA codebook. The denominator is the sum of noise and interference: (i) environmental noise $N_0 B$, (ii) interference channel estimation errors propagated from higher power levels $\sum_{i=1}^{q-1}[(1-p_e)\sigma_i^2 + p_e]K^i\pi_i$, and (iii) interference by equal or lower power levels $\sum_{j=q}^{Q} K^j\pi_j$.

3) Minimum SINR Threshold: The maximal achievable coding rate at a finite block-length regime is adopted to compute the minimum SINR threshold of the UMA codebook [33]. Specifically, given block length is n_d and the target maximum decoding error rate is p_e, the maximal achievable coding rate is approximated by:

$$\frac{M}{n_d} \approx C - \sqrt{\frac{V}{n_d}}\mathcal{Q}^{-1}(p_e), \tag{6}$$

where M is the number of bits per message, $\frac{M}{n_d}$ is the bit rate of the UMA coding. $C = \frac{1}{2}\log(1+\text{SINR})$ is the Shannon channel capacity. V is the channel dispersion, which measures the variability of the channel at finite block length n_d. V is defined as:

$$V = \frac{\text{SINR}}{2}\frac{(\text{SINR}+2)}{(\text{SINR}+1)^2}\log^2 e, \tag{7}$$

and $\mathcal{Q}^{-1}(.)$ is the inverted Gaussian tail Q-function. Basically, (6) provides a tighter bound than the general Shannon channel capacity for the maximum data rate of the short packet and reliable mMTC uplink transmission. Let the minimum value of SINR that the UMA codebook can support be SINR*. Following (6) and given a specific value for M, n_d, and p_e, one can search for SINR* by evaluating the following constraint on different values of SINR:

$$\mathcal{Q}\left(\frac{C - M/n_d}{\sqrt{V/n_d}}\right) \leq p_e. \tag{8}$$

Then SINR* is the minimum of SINR that satisfies (8).

The problem of distributed power allocation to minimize the transmission energy and satisfy SINR $\geq$ SINR* cannot be solved by a static allocation. If every device chooses the lowest power level π_Q, the interference by equal or lower power level $K^Q\pi_Q$ will be too large, which leads to a smaller SINR. In this case, the SINR may not be adequate to satisfy the constraint of the UMA codebook. Note that $(1-p_e)\sigma_i^2 + p_e < 1$. As a result, if a small fraction of the active devices choose a higher power level instead of selfishly choosing π_Q, then SINR of power level Q will be larger. Since the number of active devices is an unknown random variable, one cannot determine a fixed number of devices that need to choose the higher power level. Therefore, in the following sections, we formulate and solve the problem of distributed power allocation as an online multi-player MAB.

IV. Proposed Approach: Mean-Field Transfer Thompson Sampling

In this section, we present our approach to addressing the decentralized power control problem, which we define as a general optimization problem as well as a multi-player multi-armed bandit (MPMAB) problem, and propose a solution using the TS algorithm. We also use a mean-field technique to address the large number of devices in this system. Furthermore, when the system converges, the final TS policies obtained from the mean-field model can be transferred to new devices (i.e., devices that have just joined the system) to accelerate the convergence to an effective power allocation strategy.

A. Power Allocation Problem

The power allocation is formally formulated as an optimization problem:

$$\begin{aligned} \underset{q_k, \forall k \in \mathcal{K}_n}{\text{minimize}} \quad & \sum_{k=1}^{K} \pi_{q_k} g_k \\ \text{subject to} \quad & \text{SINR}(q) \geq \text{SINR}^*, \forall q \in \{1, \ldots, Q\} \end{aligned} \tag{9}$$

The centralized solver optimizes a vector of K variables, each variable q_k contains the power level for device k. The primary objective is to minimize the sum transmitting power of all active devices while ensuring the satisfaction of the SINR constraint. Since the set of active device $\mathcal{K}$ changes in every time slot n, the centralized solver must recompute the solution frequently.

B. Problem Formulation Using Multi-Players Multi-Armed Bandits

We propose a system model for mMTC devices, where each device is considered a player or an agent in the system. Multiple players interact with each other to compete for the available power level, which can be thought of as a free market for selecting power levels. The demand in this market is distributed according to the mMTC source traffic model. When a device is active, it can choose one among Q different receiving power levels (or Q arms) to send its message as it sees fit. For example, if the k^{th} player experiences a large path loss and shadowing, characterized by g_k, it has a tendency to select a lower receiving power level to minimize its transmitting power $P_k = g_k\pi_{q_k}$. On the other hand, if another player experiences a smaller path loss

Algorithm 1: Thompson Sampling Policy.

1: **Initialize:** $\mu_k(q_k) = 1, c_k(q_k) = 0\ \forall q_k \in \{1, \ldots, Q\}$
2: **for** timeslot $n \in \{1, \ldots, N\}$ **do**
3: Sample model $\theta_k(q_k) \sim \mathcal{N}(\mu_k(q_k), \frac{1}{c_k(q_k)+1})$
4: Select arm $q_k = argmax_{q_k}\theta_k(q_k)$
5: Receive reward $r_k(\mathbf{K}, q_k) \in [0, 1)$.
6: Update model

$$c_k(q_k) := c_k(q_k) + 1 \tag{11}$$

$$\mu_k(q_k) := \mu_k(q_k) + \frac{1}{c_k(q_k)}(r_k(\mathbf{K}, q_k) - \mu_k(q_k)) \tag{12}$$

7: **end for**

and shadowing, it can tolerate choosing a higher target power level without sacrificing its energy.

We now focus on the implementation details of the proposed protocol in a real-world system. At the end of each time slot, a small message is broadcasted to all players by the BS, which contains the number of active devices that have selected each power level $\mathbf{K} = \{K^1, \ldots, K^Q\}$. The broadcast message has a dimension of Q, and thus its size is small and independent of the number of active devices in the system. By decoding the message, each player k can compute its reward for selecting an arm q_k at time slot n as follows:

$$r_k(\mathbf{K}, q_k) = \begin{cases} 1 - \frac{\pi_{q_k} g_k}{P_{\max}} & \text{if SINR}(q) \geq \text{SINR}^*, \\ & \forall q \in \{1, \ldots, Q\} \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where $P_{\max}$ is the maximum transmit power, which is used to normalize the transmit power into the range of $[0, 1)$. Thus, the reward is bounded by $[0, 1)$. The SINR(q) is calculated by (5) and SINR* is the minimum SINR that satisfies (8). The proposed scheme operates as follows: If all power levels satisfy the SINR constraint, each player will receive a reward of one unit subtracted by its respective normalized transmit power. Conversely, if any of the chosen power levels fail to meet the minimum SINR constraint, no rewards will be granted to any of the players. The reward is designed to encourage cooperation among players and to ensure that no power level fails.

C. Thompson Sampling With Gaussian Reward Distribution

Our proposed system consists of K^{tot} devices, where each device is represented as a player utilizing a TS algorithm to explore and exploit the effective power level to transmit their uplink messages (Algorithm 1).

Firstly, the TS algorithm initializes the estimated reward distribution for each arm q of player k (line 1 of Algorithm 1). The algorithm stores: (i) $\mu_k(q_k)$ - the estimated average reward for player k, (ii) $c_k(q_k)$ - the number of times that arm q_k has been selected by player k. The estimated reward follows a Gaussian distribution with a mean of $\mu_k(q_k)$ and variance of $\frac{1}{c_k(q_k)+1}$. Initially, each player has limited knowledge about the system's state and the actions of other players, so they start by exploring the arm set with a large variance of $\frac{1}{c_k(q_k)+1} = 1$ at time step 1. As an arm gets selected more frequently, the uncertainty about its reward decreases, and the variance of the estimated reward distribution also reduces as $\lim_{n\to\infty} \frac{1}{c_k(q_k)+1} = 0$. We set the default initial mean parameter $\mu_k(q_k)$ to 1, following the optimistic initial values principle to encourage players to explore all arms equally. The average estimated reward gradually converges from 1 to a value less than 1 which resembles the actual reward that the player obtains while interacting with other players.

The central idea of the TS algorithm is to first sample the estimated Gaussian reward distribution for each arm, and then select the arm with the highest "sample value" (lines 3-4, Algorithm 1). The "sample value" contains information on both the average reward of the arm for exploitation and the variance that reflects the level of uncertainty in the estimation for exploration. By selecting the arm with the highest sample value, the player can balance between exploiting the arm with the highest reward and exploring potential high-reward arms.

At the end of each time slot, upon receiving the broadcast message from the BS, each player calculates its reward and updates the Gaussian distribution parameters using (11) and (12). These equations are the incremental implementations for updating the mean and variance of a Gaussian distribution as a new observed reward becomes available one item at a time.

D. Mean-Field Approximation of Multi-Player Thompson Sampling With Gaussian Reward Distribution

In this section, we present a mean-field model to approximate the behavior of our proposed distributed power allocation based on TS. The mean-field model is a mathematical tool commonly used in computer science and other fields to analyze interactions between individual components that are too numerous to analyze directly. Our system is inherently a large-scale, uncoordinated, and contention-based network, with numerous TS-based agents interacting with each other. By adopting mean-field analysis, we can develop a model to gain insight into and predict the average behavior of the agents when they converge. With this insight, we can improve the algorithm further by making a better initialization to the TS agents to be closer to the converged state.

To determine the average change in states $\mu_k(q_k)$ and $c_k(q_k)$ over time, we seek to model their dynamic behavior as outlined in Algorithm 1. Specifically, we model the system as a discrete-time Markov process with continuous states $\mu_k(q_k), c_k(q_k)$ for all $k \in \{1, \ldots, K^{tot}\}$ and $q \in \{1, \ldots, Q\}$. At each time slot, if the k^{th} device is active and arm q_k is selected, then the device's states are updated as (11) and (12). The probability of state changes for device k depends on its probability of being active P_{RA}, which can be calculated as in [31]. At the beginning of each time slot, the active player k selects a power level that corresponds to the highest sample value. On average, the probability of player k selecting a particular power level q is the product of the probability that player k's power level q has a higher sample value than the sample values of other arms:

$$P(q_k = q) := \prod_{j\in\{1,\ldots,Q\}, j\neq q} P(\theta_k(q) > \theta_k(j)) \tag{13}$$

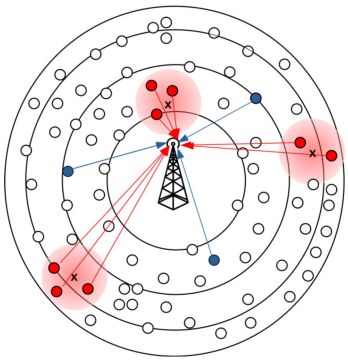


Fig. 3. Dividing devices into groups (or rings) with close path loss and shadowing.

To determine the preference of player k for a particular arm, we utilize the fact that the sample values are drawn from a Gaussian distribution. Specifically, the probability that player k prefers arm q over arm j can be calculated as the probability of a Gaussian variable $\mathcal{N}(\mu_k(q), \frac{1}{c_k(q)+1})$ being larger than another Gaussian variable $\mathcal{N}(\mu_k(j), \frac{1}{c_k(j)+1})$. This can be expressed as follows:

$$P(\theta_k(q) > \theta_k(j)) = \frac{1}{2}\text{erfc}\left(-\frac{\mu_k(q)-\mu_k(j)}{\sqrt{2\left(\frac{1}{c_k(q)+1}+\frac{1}{c_k(j)+1}\right)}}\right), \quad (14)$$

where $\text{erfc}(x)$ is the complementary error function. The erfc function represents the probability that a Gaussian distributed random variable with mean 0 and variance 1 is greater than x, and it is formally defined as:

$$\text{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^{\infty} e^{-t^2}dt \quad (15)$$

By combining (13) and (14), we can obtain the probability of player k selecting power level q at any time slot n.

In order to simplify the system, we can approximate the average value of the mean and action counter of all players, but this may not work well since each player has a different reward depending on their channel g_k. Another approach is to divide the K^{tot} devices into L groups with equal numbers of devices and with similar g_k values (Fig. 3). These L groups can then form non-overlapping rings around the base station. This allows us to keep track of only $2LQ$ different mappings of the states from a time slot to the next time slot. Each device in group l, for all $l \in \{1,\ldots,L\}$, has similar path loss and shadowing, which can be represented by g_l. We can then approximate the rewards that each device in that group is receiving with:

$$r(\mathbf{K}, q_l) = \begin{cases} 1 - \frac{\pi_{q_l} g_l}{P_{\max}} & \text{if SINR}(q) > \text{SINR}^*; \forall q \in \{1,\ldots,Q\} \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

We can formulate the deterministic differential equation for the continuous average states transition in each group l, $l \in \{1,\ldots,L\}$, as follows:

$$\begin{cases} \frac{d\mu_l(q)}{dt} & := p_{RA}P(a_l = q)\frac{1}{c_l(q)}(r(\mathbf{K}, g_l) - \mu_l(q)) \\ \frac{dc_l(q)}{dt} & := p_{RA}P(a_l = q) \end{cases} \quad (17)$$

We prove that the stochastic distributed power allocation system approximates the deterministic system in (17) by applying Kurtz's Theorem [15].

Theorem 1. (Kurtz's theorem) Suppose we have a density-dependent family:

$$F(x) = \sum_{l \in L} \beta_l(x), \quad (18)$$

where l is a transition, L is the set of possible component transitions, and $\beta_l(x)$ is a transition function. $F(x)$ is a function that map the current state x to the next state, and $F(x)$ satisfies the Lipschitz condition (condition 1):

$$|F(x) - F(y)| \leq M|x - y|. \quad (19)$$

Furthermore, suppose $\lim_{n\to\infty} X(0) = x_0$, and let X be the deterministic process:

$$X(t) = x_0 + \int_0^t F(X(u))du, t \geq 0. \quad (20)$$

Consider the path $\{X(u) : u \leq t\}$ for some fixed $t \geq 0$, and assume that there exists a neighborhood K around this path satisfying (condition 2):

$$\sum_{l \in L} |l| \sup_{x \in K} \beta_l(x) < \infty \quad (21)$$

Then

$$\lim_{n\to\infty} \sup_{u \leq t} |X_n(u) - X(u)\| = 0 \text{ almost surely.} \quad (22)$$

Here, t represents the continuous-time equivalent of discrete time slot n. As $n \to \infty$ and $t \to \infty$ and $K^{tot} \to \infty$, we have the following theorem:

Theorem 2. For each group of devices having the average path-loss and shadowing level g_l, the state $\mu_k(i)$ of each device asymptotically converges to the state $\mu_l(i)$ of that group specified by the system of ODEs in (17).

In our problem, $F(x)$ refers to the system of ODEs defined in (17), which must satisfy the conditions outlined in Kurtz's theorem. To satisfy condition 1 of Kurtz's theorem, we must demonstrate that (17) satisfies the Lipschitz condition in the following lemma:

Lemma 1. (17) satisfies the Lipschitz condition with $|F(x) - F(y)| \leq M|x - y|$, where:

$$M = Qp_{RA}\left[1 + \frac{(Q-1)}{\sqrt{2\pi e}}\left(\left|\frac{1}{\Delta\mu_{\min}}\right| + \left|\frac{1}{c_{\min}+1}\right|\right)\right], \quad (23)$$

where $\Delta\mu_{\min}$ is the minimum of $|\mu_l(q) - \mu_l(j)|$, and $c_{\min}$ is the number of times an eliminated arm with low performance has been selected.

It means that the accuracy of the model will become worse at any group l as the rewards of two arms of that group are equal, or there is some very low-performance arm that gets eliminated early and has low $c_q(l)$. Otherwise, if two arms have distinguishable expected rewards or there are no very low-performance arms then the approximation model will be more accurate. If the ODEs satisfies the Lipschitz condition, it means that they will not fluctuate too quickly. Thus, even after the stochastic process

and the deterministic process of the ODE are separate, they will remain close and have nearly the same behavior. Further details about the proof of Lemma 1 are provided in the supplemental material.

As the reward of our system is bounded within $[0, 1)$, the transition in (17) is also bounded, and there are finite transition equations. Therefore, condition 2 of Kurtz's theorem is also satisfied. By applying Kurtz's Theorem, we can show that the stochastic process of the TS algorithm that makes distributed power allocation decisions converges to the deterministic process in (17) almost surely.

E. Mean-Field Transfer Thompson Sampling

In our setting with a large number of players, the conventional TS algorithm may not be effective as it exhausts resources exploring all arms, even those with marginal reward gains [34]. Instead, we propose a strategy to enhance the learning process of each agent by providing better state initialization. Specifically, rather than initializing the reward distribution of arm q of agent k with an optimistic estimate of reward $\mu_k(q) = 1$ and high variance $\frac{1}{c_k(q)+1}$ where $c_k(q) = 0$, we initialize it using the solution of the agent that has learned from the existing system. Assuming we have trained a system in which all players use the TS algorithm, we can extract a tuple of $\{g_k, \boldsymbol{\mu}_k, \mathbf{c}_k\}$ for each agent k, where $\boldsymbol{\mu}_k = \{\mu_k(1), \ldots, \mu_k(Q)\}$ and $\mathbf{c}_k = \{c_k(1), \ldots, c_k(Q)\}$ represent the final parameters of the agent after training. If a new agent k' enters the system with path loss and shadowing coefficients $g_{k'}$, we initialize it with the parameter set corresponding to the agent with the smallest $\text{argmin}_k |g_k - g_{k'}|$. This approach accelerates the learning progress of the new agent and improves the overall system performance. We denote this approach as Transfer Thompson Sampling (TTS).

In designing a good network, one core principle is that the network should be transparent to devices, who should feel as though the network does not exist while they use it. We face the dilemma that collecting data on the states and path loss and shadowing of learned devices from past systems are difficult, as devices may not be willing to provide this information to network operators. This presents a challenge in accelerating the learning of new devices without collecting any data from past devices. To address this challenge, we propose to solve the system using the mean-field approximation model and use states at convergence from this model to initialize new devices and accelerate their learning process. Specifically, the new device k' is initialized as the approximated state of the nearest group l given by $\text{argmin}_l |g_l - g_{k'}|$. This approach enables us to initialize the agent more effectively using the approximated solution without collecting any device data. We denote this approach as MFTTS. While the initialization given by the mean-field model may not be as good as the initialization transferred from past device data due to the approximation error, we hypothesize that it will still be a better power-level selection policy than initializing from scratch.

F. Condition for Sub-Linear Regret

In MAB, regret is a metric used to quantify whether a player could effectively learn or not [35], [36]. It represents the difference between the optimal cumulative reward the player could have received and the actual cumulative reward it obtained. Having a sub-linear regret is important because this shows that the cumulative reward obtained by our algorithm will converge to the optimal cumulative reward as time (N) increases. In this section, we present the conditions under which a TS player can achieve sub-linear regret. Formally, the regret after a player interacts with the environment N time-slots is defined as:

$$\text{Regret}(N) := Nr^* - \sum_{n=1}^{N} r(q), \tag{24}$$

where r^* is the reward when the player selects the optimal arm. We assume that, over a sufficient number of time slots, the estimated reward distribution for each arm in each of the K existing players has converged around its mean, rendering the variance negligible. Consequently, these K existing players will predominantly select the optimal arm. Given that there are ΔK new players with their arm distribution not concentrated around their mean yet. If ΔK new players do not affect the converged arm selection distribution of K existing players, then ΔK new players can treat the existing K players as a stochastic bandit environment, and the regret bound for TS with Gaussian prior for a general stochastic multi-armed bandits environment can be applied.

In particular, the condition in which ΔK new players will not change the reward of K existing players from $1 - \frac{\pi_{q_l} g_l}{P_{\max}}$ to 0 is:

$$\begin{aligned}&(N^{\text{MIMO}}(1-\sigma_q^2)\pi_q)/(N_0 B \\ &\quad + \sum_{i=1}^{q-1}[(1-p_e)\sigma_i^2 + p_e](1+\Delta K)K^i\pi_i \\ &\quad + \sum_{j=q}^{Q}(1+\Delta K)K^j\pi_j) \le \text{SINR}^* \ \forall q \in \{1, \ldots, Q\}.\end{aligned} \tag{25}$$

Then, the bound for the maximum number of new players is given as:

$$\begin{aligned}\Delta K \le &\left(\frac{N^{\text{MIMO}}(1-\sigma_q^2)\pi_q}{\text{SINR}^*} - N_0 B\right) \\ &\frac{1}{\sum_{i=1}^{q-1}[(1-p_e)\sigma_i^2 + p_e]K^i\pi_i + \sum_{j=q}^{Q}K^j\pi_j} - 1, \\ &\forall q \in \{1, \ldots, Q\},\end{aligned} \tag{26}$$

If the condition in (26) is satisfied, then ΔK new players will converge to an optimal arm selection policy with sub-linear regret. The regret for Q-armed stochastic bandit problem with the learning horizon for a new agent is N time slots, for which TS using Gaussian priors has a bounded expected regret [37]:

$$\mathbb{E}[\text{Regret}(N)] \le \mathcal{O}(\sqrt{NQ\ln(Q)}). \tag{27}$$

TABLE I
SIMULATION PARAMETERS FOR OUR POWER CONTROL IN HD-NOMA

Parameters	Description	Value
K^{tot}	total #devices	$10,000$
$E[K]$	expected #device/time slot	600
$r_{\min}$	min distance from device to BS	0.25 km
$r_{\max}$	max distance from device to BS	1 km
λ_D	PPP density of devices	≈ 3184.0
λ_E	PPP density of events	2.0
p_R	packet generation rate in state R	0.03
p_A	packet generation rate in state A	1
p'	transition probability from A to R	0.5
$\bar{d}$	distance normalization constant	0.005
N_0	noise spectral density	-174 dBm/Hz
B	bandwidth	1.4 MHz
	maximum transmission power	0.2 Watts
$N_0 B$	total noise power	-112.54 dBm
α	path loss constant	128.1 dB
β	path loss constant	37.6 dB
σ^2_{shadow}	variance of shadowing	4.0
n_p	length of UMA pilot sequence	1052 bits
n_d	length of UMA codewords	2048 bits
π_Q	min receiving power	$-3N$ dBm
π_1	max receiving power	$3N$ dBm
N^{MIMO}	#receiving MIMO antennas	100
M	bit rate of UMA coding	100 bits/timeslot
p_e	max decoding error rate	10^{-5}
N	#simulation steps	$10,000$
Δt	duration of a time slot	10 ms
	#independent runs	10

In short, if the number of new players is not large enough to disrupt existing players, then new players will find the optimal power allocation policy with sub-linear regret.

V. PERFORMANCE EVALUATION

In this section, we begin by introducing the setup for the performance evaluation. We then compare the decision-making time in each time slot for our proposed method and the considered baselines. An ablation study is carried out to assess the efficacy of the modifications made to the MFTTS in contrast to the conventional TS algorithm. Additionally, we demonstrate the accuracy of our mean-field approximation scheme. Ultimately, we establish that MFTTS exhibits better power saving compared to other baselines when the number of mMTC devices is exceedingly large.

A. Parameter Settings

To evaluate the effectiveness of the proposed distributed power control policies, a numerical simulation was conducted using a discrete event simulator with parameters that closely resemble a real-world system (see Table I). For the traffic model, the device density λ_U was selected to achieve a total of approximately 10,000 devices, while the value of λ_E was chosen to allow for around 600 active devices at any given time. The parameters related to the UMA transmission were determined based on previous research [14], specifically the UMA coding scheme $(n_p, n_d, \pi_Q, \pi_1, N^{MIMO}, M, p_e)$ and the large scale fading of the system $(\alpha, \beta, \sigma^2_{\text{shadow}})$. The noise spectral density was set to a widely used value of $N_0 = -174$ dBm/Hz. Note that, under these conditions, the SINR constraint would not be satisfied if every device chose the smallest power level, and the maximum number of devices that a no-SIC system could support is around 450 devices. Under the setting with $n_p = 1152$ bit and the number of active devices is less than 1000 per time-slot, the possibility of pilot collision according to (3) is nearly zero. Each distributed power selection policy was run for 10,000 discrete time slots, and the experiment was replicated 10 times to aggregate their average performance.

B. Baseline Methods

Our proposed MFTTS is compared with both centralized and decentralized methods.

1) Centralized Approximation Algorithm: Due to the exponential nature of the search space of the given problem (9), which scales with the number of active devices K, the brute-force method cannot be feasibly employed for a system with a massive number of devices. Therefore, we aim to propose an approximate algorithm to solve problem (9) with a time complexity that does not exponentially increase as the number of active devices grows. Our primary aim in developing this algorithm is to establish an empirical upper bound for the distributed power allocation algorithms. Additionally, we can ascertain whether the centralized algorithm can execute quickly enough for the considered system and justify the need for proposing a lightweight distributed receiving power level allocation algorithm.

The main idea of the approximated algorithm is to sort and map devices with high path loss and shadowing to lower power levels to reduce the sum of their multiplication. Assuming that the BS has access to the channel information g_k for all active devices, it can sort the active devices based on their path-loss and shadowing. Specifically, the devices can be sorted such that $g_1 \geq g_2 \geq \ldots \geq g_K$, while the receiving power levels can be ordered as $\pi_Q < \pi_{Q-1} < \ldots < \pi_1$. Then, the set of active devices can be partitioned into Q groups. The optimization variable is the set of separation indices $k_1, \ldots, k_{Q-1}$, such that:

- $g_1, \ldots, g_{k_1}$ map to the lowest power level Q
- $g_{k_1}, \ldots, g_{k_2}$ map to the power level $Q-1$
- $\ldots$
- $g_{k_{Q-1}}, \ldots, g_{K_Q}$ map to the power level 1

This method requires $\mathcal{O}(K\log(K) + \frac{K!}{(Q-1)!(K-Q+1)!})$ operations to sort the devices and evaluate all separation indices, which is polynomial in terms of the number of active devices. However, the complexity of the algorithm can be further reduced by dividing the K active devices into L groups with an equal number of devices, such that devices with similar g_k are put into the same group. The separation indices $k_1, \ldots, k_{Q-1}$ can then be used to separate a smaller number of L device groups instead of separating all K devices. This approach significantly reduces the complexity of the algorithm to $\mathcal{O}(K\log(K) + \frac{L!}{(Q-1)!(L-Q+1)!})$.

2) Decentralized Baselines: For the decentralized baselines, we compare MFTTS with the following policies: hedge, TS, TTS, and MABMFG. The Hedge technique was presented in a previous paper on MPMAB for continuous reward [30]. The vanilla version of TS with the default initialization is represented by TS. TTS is the version of TS with initialization of the supposed user in a previously trained system. MABMFG is

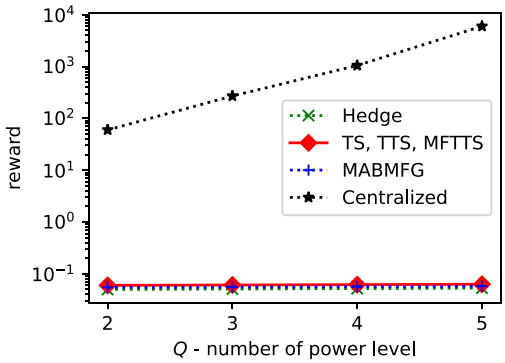


Fig. 4. The required time to make the decision per time slot for different algorithms in logarithmic scale.

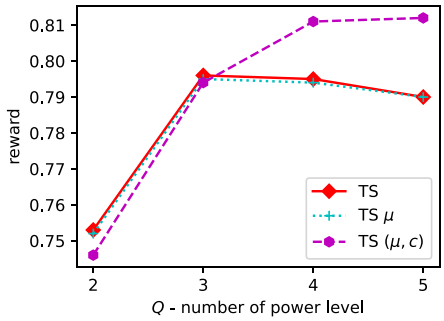


Fig. 5. Comparison no transfer, transfer only mean and transfer the arm counting variable, which represents the variance.

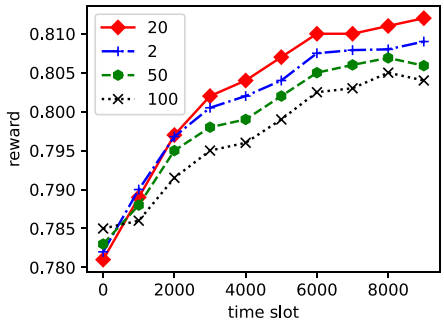


Fig. 6. Running average reward overtime when $Q = 5$, for MFTTS with different number of partitions.

the latest distributed joint resource and power allocation for HD-NOMA [23].

C. Time Complexity

In Fig. 4, we compare the total running time required by different power allocation algorithms on a logarithmic scale. We measured the average time to compute the power level of different methods on a single-threaded implementation in Python. Moreover, all algorithms are executed on the same type of CPU with a clock speed of 2.4 GHz.

It should be noted that the duration of each time slot is 10 ms, so the time to find the power level decision is expected to be much shorter than 10 ms. However, the centralized approximated algorithm spent a significantly longer time computing the result. As the number of power groups increased, the running time of the centralized algorithm increased exponentially.

On the other hand, distributed algorithms require less than one-tenth of a millisecond to complete. Note that TTS and MFTTS have the same time complexity as TS since the only difference lies in the initialization phase. Moreover, their running time remained almost the same as the number of arms increased. These results suggest that these distributed algorithms can be implemented to make instant power allocation decisions locally on IoT devices.

D. Ablation Study

This subsection presents two experiments related to the parameter selection of the proposed TTS and MFTTS algorithms.

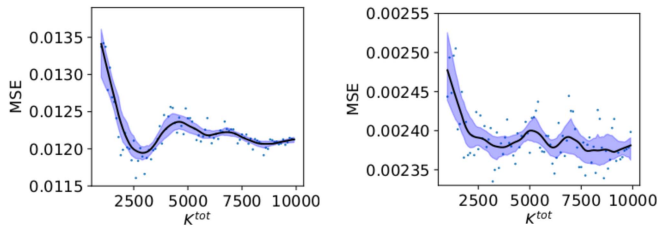
In the first experiment, we compared the performance of three variants of TS as the number of power levels (i.e., SIC levels) is increased (Fig. 5). The comparison metric is the average rewards over 10,000 time slots. We found that initializing only the mean of the sampling distribution for new users resulted in average rewards similar to those obtained with TS without transfer. However, when both the mean and variance of new users were copied from past solutions, the resulting variant of TTS significantly outperformed TS in terms of average reward over the entire time horizon of 10,000 time slots. The variance of the sampling distribution plays a crucial role in regulating the exploration level of TS and TTS, which may explain the superior performance of TTS. Based on these results, we chose to transfer all past parameters to new users in TTS.

The impact of the number of partitions L used in the mean-field approximation model of our proposed MFTTS algorithm was investigated in Fig. 6. In MFTTS, new users initialize their prior distribution parameters based on the partition with the nearest large-scale fading coefficient. If the number of partitions is too small (e.g., $L = 2$), the approximation error may become too large. Furthermore, due to the shadowing effect, the large-scale fading coefficient of users is random and not constant. Conversely, if the number of partitions is too large (e.g., $L = 50$ or $L = 100$), users may randomly belong to different groups, making the transfer policy less robust. In our experiment settings, $L = 20$ is the most suitable number of partitions that balances the trade-off between approximation accuracy and transfer policy robustness, leading to the best average reward performance.

E. Mean-Field Approximation Error

All agents in the simulation use a TS algorithm and the final parameter $\mu_k(q)$ and $n_k(q)$ are captured. An evaluation metric, the mean squared error (MSE) metric is calculated using the following equation:

$$MSE(\mu, \hat{\mu}) = \frac{1}{LQ} \sum_{l=1}^{L} \sum_{q=1}^{Q} (\bar{\mu}_l(q) - \hat{\mu}_l(q))^2, \qquad (28)$$



(a) MSE of Mean (b) MSE of Variance

Fig. 7. The scatter points in the figure represent the Mean Squared Error (MSE) between the state of the mean-field approximation and the state of the discrete event simulator, where the total number of users K^{tot} varies from 1000 to 10000. The smoothed path is generated using the LOcally WEighted Scatterplot Smoothing (LOWESS) technique. The shaded region corresponds to the 90% confidence interval.

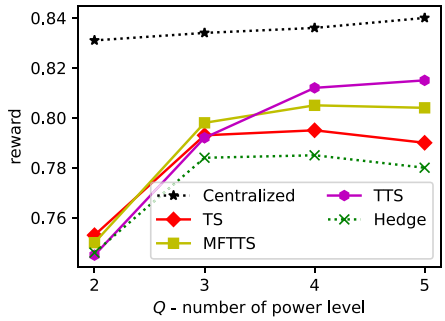


Fig. 8. The average reward overall time slot, with different numbers of power levels.

where $\bar{\mu}_l(q)$ is the average value of $\mu_l(q)$ of users in partition l. We also used a similar metric to measure the mean-field error of predicting the converged variance state.

In order to assess the accuracy of the mean-field approximation, we computed the MSE between the results obtained by the discrete event simulator and the approximated model. Specifically, we compared the final state of the discrete event simulator at the end of 10000 time slots with the predicted state from the approximation model. To test our approach, we varied the network size from $K^{tot} = 1000$ to $K^{tot} = 10000$. We adjusted the corresponding parameters of the UMA codebook to account for the size of K^{tot}, ensuring that the maximum number of users that the non-SIC scheme can support is less than the expected number of active users $E[K]$. In other words, we tested on the overloaded network with various sizes. For instance, when the total number of users $K^{tot} = \{1000, 1100, 1200, \ldots, 10000\}$, the number of MIMO antennas was set to $N^{MIMO} = \{10, 11, 12, \ldots, 100\}$. Consequently, the maximum number of users that the non-SIC scheme can support was less than the expected number of active users.

Fig. 7 depicts the error between the predicted and simulated mean and variance. As the smallest network size we could experiment with was already 1000 users, we observed only a modest 20% reduction in the MSE of the mean state as the network size increased. We found that the MSE of both the mean and variance only decreased as the network scale increased from 1000 users to 2500 users, and then plateaued thereafter. However, the downward trend in the average MSE remained visible, indicating that our approximation model improve accuracy when the number of users ranged from 1000 to 2500, and maintain the same level of accuracy when the number of users ranged from 2500 to 10000.

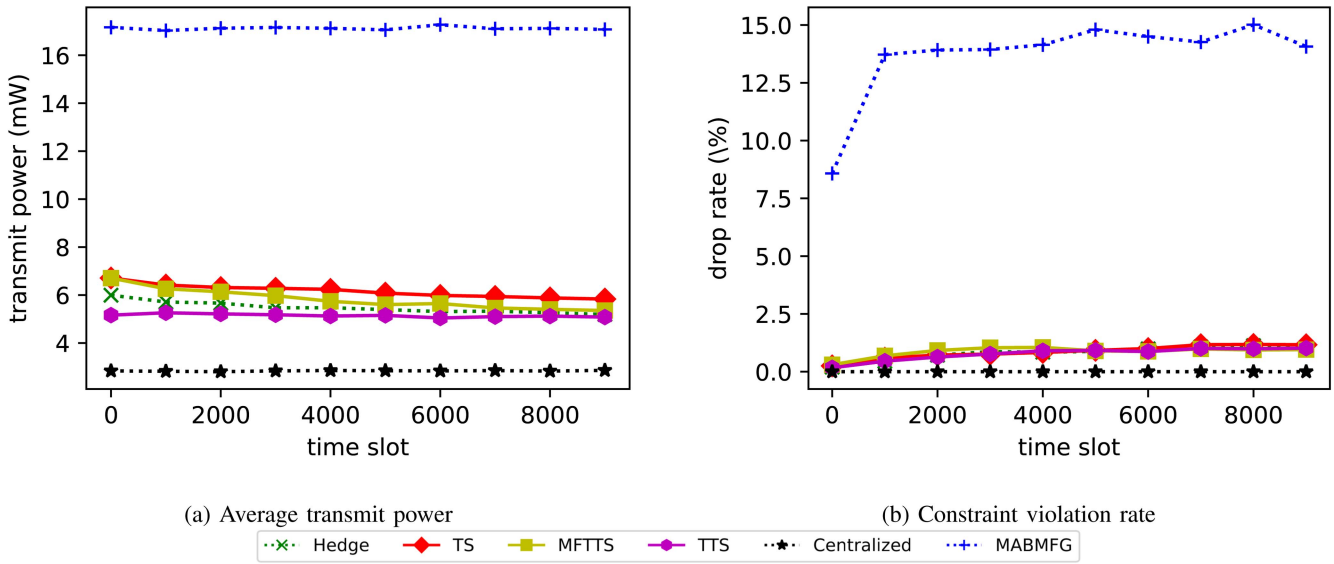
F. Comparison of the Average Rewards

In this study, we compare different power allocation methods for the hybrid domain NOMA technique in terms of the average reward in a large-scale scenario with 10,000 users. Fig. 8 shows the average reward of different algorithms across 10,000 time slots and 10 independent runs as the number of SIC power levels increases. The average reward is determined by a combination of the power users spend to transmit the message uplink and the outage probability, which is the event where the selected power level fails to meet the signal-to-interference-plus-noise ratio (SINR) constraint (10). The higher the reward, the fewer time slots the SINR is not satisfied, and the less total power users spend for sending data uplink.

As expected, the centralized approximation algorithm provides the best performance, while the random method yields the worst rewards. As the number of SIC levels increases, the reward for all algorithms generally increases.

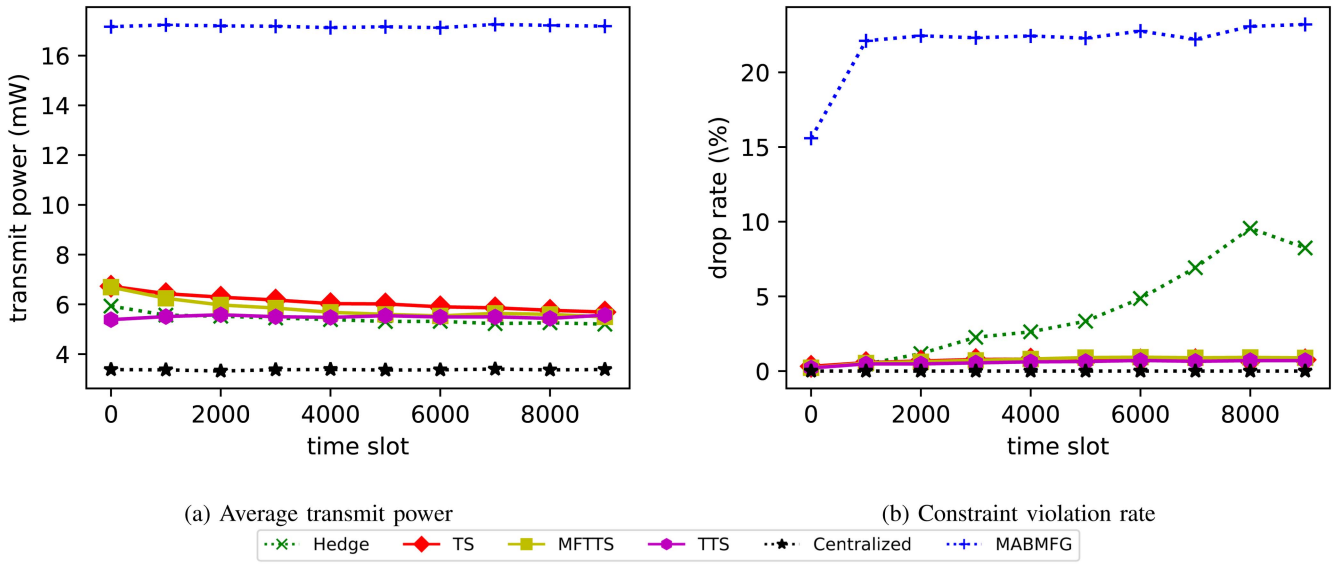
In terms of the MPMAB algorithms, TS outperforms the Hedge policy, as expected based on prior empirical studies [25]. For the case where the number of arms is 4 and 5, both TTS and MFTTS exhibit a significant improvement over TS. Among the distributed allocation algorithms, TTS provides the best performance, followed by MFTTS. These results confirm our hypothesis that transferring knowledge from previously learned players is beneficial. Although transferring knowledge from the mean-field approximation model in MFTTS exhibits a slightly worse performance compared to transferring directly from real data from learned players, it does not collect more data from mMTC devices and does not require additional overhead as in TTS.

Figs. 9, 10, and 11 present the running average of the reward across 10,000 time slots and across three scenarios: (1) low load, (2) medium load, and (3) high load. In all scenarios, MABMFG uses a grant-free NOMA, so it is less energy efficient than Hedge, TS, TTS, and MFTTS, as they are all based on UMA. Additionally, MABMFG is less reliable than other UMA-based methods, even in the low load scenario, as it facilitates more than 10% of SINR constraint violations. The centralized algorithm with unrealistic assumptions sets the performance bound for all algorithms. As the number of active users increases, the lowest achievable transmit power without SINR constraint violation also increases. Hedge reduces its transmit power over time in all scenarios. However, it reduces the transmit power at the cost of an increased amount of SINR constraint violations. Particularly, Hedge can only handle the low load scenario, while in the medium load and high load scenarios, it reduces the transmit power too much, leading to SINR constraint violations of 10% and 35% respectively. TS, TTS, and MFTTS can handle the constraint better than Hedge, even at high loading, achieving



(a) Average transmit power (b) Constraint violation rate

Fig. 9. Performance when average number of active users is 500 (low load).

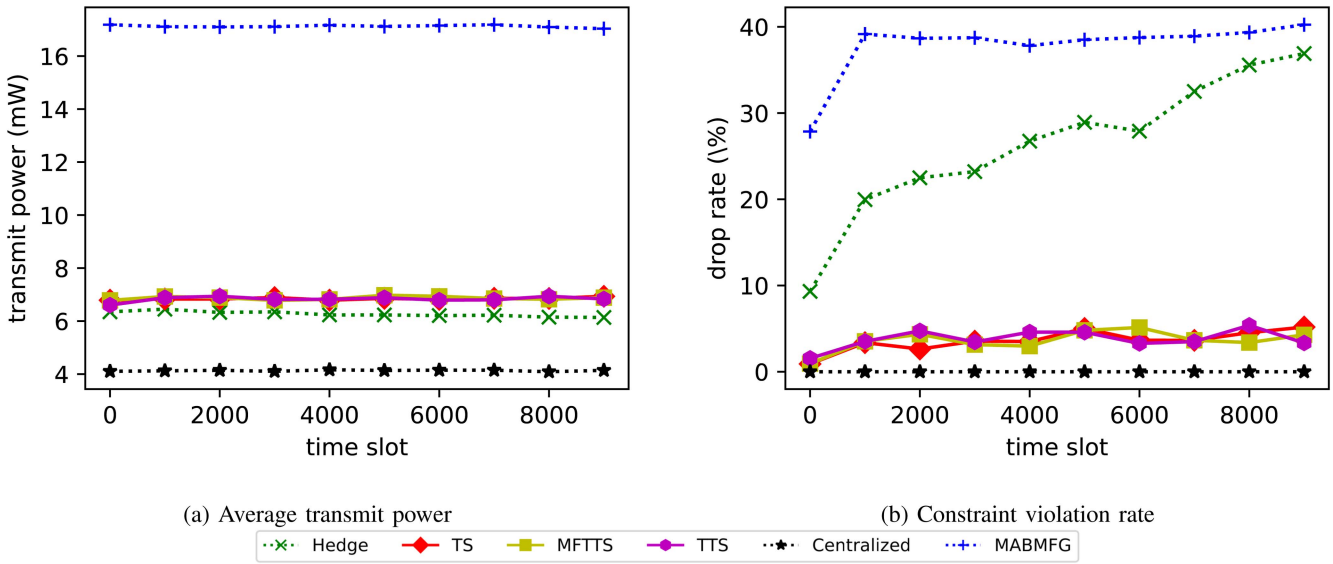


(a) Average transmit power (b) Constraint violation rate

Fig. 10. Performance when average number of active users is 700 (medium load).

only around 5% constraint violation. On the other hand, in the low and medium load scenarios, the proposed method MFTTS allows devices to reach the low transmit power state faster than learning from scratch with TS. MFTTS learns and reaches the same reward level as TTS at the end of 10,000 time slots, while TS is slower and unable to achieve the same level as MFTTS. In real-world terms, this corresponds to a period of approximately 1 minute and 40 seconds for MFTTS to reach the same performance as TTS. These results confirm our hypothesis that initializing players with the solution of the mean-field approximation model helps them converge to high-performing power allocation policies faster.

To understand why one algorithm performs better than another, we can observe their probability of selecting a certain option over time. In Fig. 12, we can see that the probability of selecting the smallest option is directly related to the algorithm's performance. When a centralized approximation algorithm is used to allocate almost all of the active users to the lower power level, the resulting reward is higher. The MPMAB algorithms have learned to allocate more active users to the lower power level over time, which improves their performance in terms of reward. The TTS and MFTTS algorithms are the best-performing algorithms among the distributed algorithms for power allocation. They can allocate the highest amount of active users to the lowest power level, but they still only learned to allocate around 50% of users to that level, compared to the centralized approach which can allocate 70%. However, the MFTTS algorithm has a significantly lower execution time than



(a) Average transmit power (b) Constraint violation rate

Fig. 11. Performance when average number of active users is 1000 (high load).

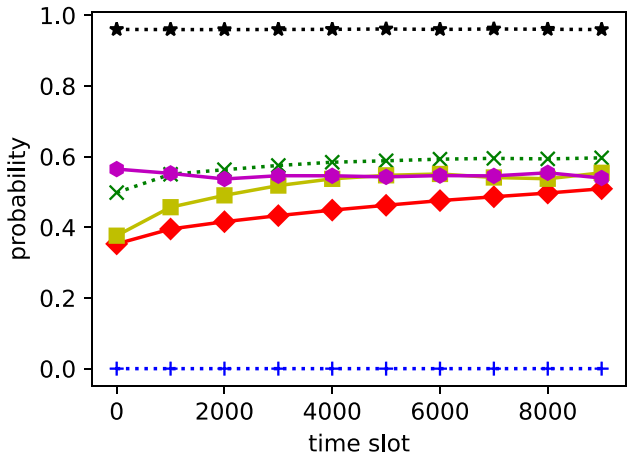


Fig. 12. Comparison of the probability the smallest power arm has been selected.

the centralized approximation algorithm, and it is feasible to implement MFTTS in real-world networks.

VI. Conclusion

In this paper, we proposed a novel approach called MFTTS to solve the distributed power control problem in the HD-NOMA system in the context of mMTC devices. MFTTS uses a MAB algorithm and the mean-field approximation model. Our approach addresses the key requirements of mMTC devices by being decentralized and lightweight, adaptively minimizing power consumption. Simulation studies demonstrated that our proposed MFTTS is more practical than the centralized solver, while achieving higher SINR constraint satisfaction and lower power consumption compared to existing distributed policies such as random selection, Hedge or Thompson Sampling with a standard prior distribution. The mean-field approximation for the large-scale mMTC network proved to be useful in predicting the steady state of the MPMAB mMTC system. Our proposed method MFTTS provided better performance by transferring initialization from the approximated state. In the future, we will examine if the same online learning framework can offer a promising solution for not only this specific case of distributed power allocation but also general large-scale distributed resource allocation problems.

References

[1] P. J. et al., "Ericsson mobility report," 2023. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report
[2] N. H. Mahmood et al., "White paper on critical and massive machine type communication towards 6G," in *Proc. 6G Res. Vis., No. 11*, 2020. [Online]. Available: http://urn.fi/urn:isbn:9789526226781
[3] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surv. Tut.*, vol. 22, no. 3, pp. 1805–1838, thirdquarter 2020.
[4] N. Docomo et al., "Uplink multiple access schemes for NR," in *Proc. R1-165174, 3GPP TSG-RAN WG1 Meeting*, 2016, vol. 85. [Online]. Available: https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_85/Docs/R1-165174.zip
[5] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, Secondquarter 2017.
[6] K. Au et al., "Uplink contention based SCMA for 5G radio access," in *Proc. IEEE Globecom Workshops*, 2014, pp. 900–905.
[7] A. Yadav, C. Quan, P. K. Varshney, and H. V. Poor, "On performance comparison of multi-antenna HD-NOMA, SCMA, and PD-NOMA schemes," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 715–719, Apr. 2021.
[8] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017, pp. 2523–2527.
[9] O. Ordentlich and Y. Polyanskiy, "Low complexity schemes for the random access Gaussian channel," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017, pp. 2528–2532.
[10] Y. Li et al., "Unsourced multiple access for 6g massive machine type communications," *China Commun.*, vol. 19, no. 3, pp. 70–87, 2022.
[11] K.-H. Ngo, A. Lancho, G. Durisi, and A. G. i Amat, "Unsourced multiple access with random user activity," *IEEE Trans. Inf. Theory*, vol. 69, no. 7, pp. 4537–4558, Jul. 2023.

[12] A. Fengler, P. Jung, and G. Caire, "SPARCs for unsourced random access," *IEEE Trans. Inf. Theory*, vol. 67, no. 10, pp. 6894–6915, Oct. 2021.
[13] V. K. Amalladinne, A. K. Pradhan, C. Rush, J.-F. Chamberland, and K. R. Narayanan, "Unsourced random access with coded compressed sensing: Integrating AMP and belief propagation," *IEEE Trans. Inf. Theory*, vol. 68, no. 4, pp. 2384–2409, Apr. 2022.
[14] A. Fengler, O. Musa, P. Jung, and G. Caire, "Pilot-based unsourced random access with a massive MIMO receiver, interference cancellation, and power control," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1522–1534, May 2022.
[15] M. Mitzenmacher, "The power of two choices in randomized load balancing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 12, no. 10, pp. 1094–1104, Oct. 2001.
[16] P. Semasinghe, S. Maghsudi, and E. Hossain, "Game theoretic mechanisms for resource management in massive wireless IoT systems," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 121–127, Feb. 2017.
[17] J. Choi and J.-B. Seo, "Evolutionary game for hybrid uplink NOMA with truncated channel inversion power control," *IEEE Trans. Commun.*, vol. 67, no. 12, pp. 8655–8665, Dec. 2019.
[18] C. Yang, J. Li, P. Semasinghe, E. Hossain, S. M. Perlaza, and Z. Han, "Distributed interference and energy-aware power control for ultra-dense D2D networks: A mean field game," *IEEE Trans. Wireless Commun.*, vol. 16, no. 2, pp. 1205–1217, Feb. 2017.
[19] A. Benamor, O. Habachi, I. Kammoun, and J.-P. Cances, "Mean field game-theoretic framework for distributed power control in hybrid NOMA," *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 10502–10514, Dec. 2022.
[20] S. Hashima, B. M. ElHalawany, K. Hatano, K. Wu, and E. M. Mohamed, "Leveraging machine-learning for D2D communications in 5G/beyond 5G networks," *Electronics*, vol. 10, no. 2, 2021, Art. no. 169.
[21] F.-C. Kuo, C. Schindelhauer, H.-C. Wang, W.-J. Lin, and C.-C. Tseng, "D2D resource allocation with power control based on multi-player multi-armed bandit," *Wireless Pers. Commun.*, vol. 113, pp. 1455–1470, 2020.
[22] B. Lin, X. Tang, and Z. Ghassemlooy, "A power domain sparse code multiple access scheme for visible light communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 61–64, Jan. 2020.
[23] A. Benamor, O. Habachi, I. Kammoun, and J.-P. Cances, "Multi-armed bandit framework for resource allocation in uplink NOMA networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2023, pp. 1–6.
[24] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3–4, pp. 285–294, 1933.
[25] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 2249–2257.
[26] T. T. Le, Y. Ji, and J. C. Lui, "TinyQMIX: Distributed access control for mMTC via multi-agent reinforcement learning," in *Proc. IEEE 96th Veh. Technol. Conf.*, 2022, pp. 1–6.
[27] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.
[28] R. Gummadi, R. Johari, S. Schmit, and J. Y. Yu, "Mean field analysis of multi-armed bandit games," 2013. [Online]. Available: https://ssrn.com/abstract=2045842
[29] S. Maghsudi and E. Hossain, "Distributed user association in energy harvesting dense small cell networks: A mean-field multi-armed bandit approach," *IEEE Access*, vol. 5, pp. 3513–3523, 2017.
[30] X. Wang and R. Jia, "Mean field equilibrium in multi-armed bandit game with continuous reward," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Z.-H. Zhou, Ed., Aug. 2021, pp. 3118–3124, doi: 10.24963/ijcai.2021/429.
[31] H. Thomsen, C. N. Manchón, and B. H. Fleury, "A traffic model for machine-type communications using spatial point processes," in *Proc. IEEE 28th Annu. Int. Symp. Personal, Indoor, Mobile Radio Commun.*, 2017, pp. 1–6.
[32] F. Mosteller, "Understanding the birthday problem," *The Math. Teacher*, vol. 55, no. 5, pp. 322–325, 1962.
[33] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
[34] D. J. Russo et al., "A tutorial on Thompson sampling," *Foundations Trends in Mach. Learn.*, vol. 11, no. 1, pp. 1–96, 2018.
[35] A. Slivkins et al., "Introduction to multi-armed bandits," *Foundations Trends Mach. Learn.*, vol. 12, no. 1-2, pp. 1–286, 2019.
[36] S. Bubeck et al., "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Foundations Trends Mach. Learn.*, vol. 5, no. 1 , pp. 1–122, 2012.
[37] S. Agrawal and N. Goyal, "Near-optimal regret bounds for thompson sampling," *J. ACM*, vol. 64, no. 5, pp. 1–24, 2017.

Tien Thanh Le (Graduate Student Member, IEEE) received the BE degree in information and communication technology in 2019, the MS degree in data science from the Hanoi University of Science and Technology, Vietnam in 2021. He is currently working toward the PhD degree in informatics with The Graduate University for Advanced Studies, SOKENDAI, Tokyo, Japan. His research interests include next-generation access control, game theory, and multiagent learning for machine-type communication system.

Yusheng Ji (Fellow, IEEE) received the BE, ME, and PhD degrees in electrical engineering from the University of Tokyo, Tokyo, Japan, in 1984, 1986, and 1989, respectively. She joined the National Center for Science Information Systems, Tokyo, Japan, in 1990. She is currently a professor and the director of Information Systems Architecture Science Research Division, National Institute of Informatics, Tokyo, Japan, and the Graduate University for Advanced Studies (SOKENDAI), Japan. Her research interests include network resource management and mobile computing. She was the TPC co-chair of IEEE INFOCOM 2023, IEEE VTC 2024-Spring, general co-chair of BigCom 2023, MSN 2020, ICT-DM 2018, symposium co-chair of IEEE ICC 2020, IEEE GLOBECOM 2012, 2014, and track co-chair of IEEE VTC 2016 Fall and 2017 Fall. She is a distinguished lecturer of IEEE Vehicular Technology Society.

John C. S. Lui (Fellow, IEEE) received the PhD degree in computer science from the University of California at Los Angeles in 1993. He is currently the Choh-Ming Li Chair professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His current research interests include quantum Internet, machine learning, online learning theories, future Internet architectures and protocols, network economics, and network/system security. He was the recipient of various departmental teaching awards and the CUHK Vice Chancellor's Exemplary Teaching Award. He is an Elected Member of the IFIP WG 7.3, a Fellow of ACM, a Senior Research Fellow of the Croucher Foundation, and the RGC Senior Research Fellow.