IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 24, 2021, accepted April 23, 2021, date of publication April 30, 2021, date of current version May 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076771

Two-Sided Learning for NOMA-Based Random Access in IoT Networks

JINHO CHOI, (Senior Member, IEEE)
School of Information Technology, Deakin University, Geelong, VIC 3220, Australia
e-mail: jinho.choi@deakin.edu.au

This work was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme under Grant DP200100391.

ABSTRACT In the Internet-of-Things (IoT), different types of devices can co-exist within a network. For example, there can be cheap but inflexible devices and flexible devices in terms of radio frequency (RF) capabilities. Thus, in order to support different types of devices in different ways and improve throughput, we propose a multichannel random access scheme based on power-domain non-orthogonal multiple access (NOMA), where each flexible or dynamic device (DD) can dynamically choose one of multiple channels when it has a packet to send. In addition, since DDs need to learn the channel selection probabilities to maximize the throughput of DDs, we consider two-sided learning based on a multi-armed bandit (MAB) formulation where rewards are decided by learning outcomes at a base station (BS) to improve learning speed at DDs. Simulation results confirm that two-sided learning can help improve learning speed at DDs and allows the proposed NOMA-based random access approach to achieve near maximum throughput.

INDEX TERMS IoT, random access, NOMA, learning.

I. INTRODUCTION

In the Internet-of-Things (IoT), a large number of devices including sensors and actuators are to be connected to networks for a number of applications including smart cities and factories [1], [2]. To allow devices to be connected, IoT connectivity plays a crucial role in the IoT and a number of solutions are studied including WiFi, cellular IoT, low-power wide area networking (LPWAN), and so on [3].

Within a certain geographical area, a number of devices can be deployed to form an IoT network with a base station (BS). In this IoT network, for wireless communications, dedicated licensed bands or unlicensed bands can be used [3], [4] to form multiple channels in order to support a number of devices that can transmit their packets simultaneously. In [5], with multichannel ALOHA for random access, learning algorithms to access multiple channels are considered when two different types of devices co-exist, namely static devices (SD) and dynamic devices (DD). A SD is a low-cost device that only transmits through a pre-determined channel due to limited radio frequency (RF) capability. On the other hand, a DD is a more flexible and capable device than SD, and it can choose a channel from multiple channels. Since DDs are capable to select channels for transmissions, multiarmed bandit (MAB) algorithms [6], [7] are considered so that each DD can maximize its average reward when multiple channels are regarded as multiple arms.

As mentioned in [5], MAB algorithms have been studied for resource allocation in wireless networks [8]–[10], where MAB is generalized with multiple players due to multiple users that access multiple channels. Unlike the problem in [5], however, the problems in [8]–[10] assume the case that the number of users is smaller than that of channels. Thus, the solution can be characterized by stable matchings [11]. On the other hand, in IoT networks, it is expected that there will be far more devices than channels. Thus, each device cannot be associated with a specific channel, and has to randomly select one of multiple channels with the risk of collision. As a result, the setting in [5], where a large number of devices exist with a limited number of channels, is practical, and the proposed approach is important in IoT networks.

Power-domain non-orthogonal multiple access (NOMA) has been extensively investigated for cellular networks as it can improve the spectral efficiency [12], [13]. In [14], the notion of NOMA is applied to random access in order to improve throughput. For access control in NOMA-based random access, a game-theoretic approach is adopted in [15]. As shown in [14], [15], since NOMA can increase the throughput

The associate editor coordinating the review of this manuscript and approving it for publication was Alessandro Pozzebon.

of multichannel ALOHA, it seems promising to support a lot of devices by utilizing NOMA in IoT networks like the one shown in [5].

In this paper, we apply power-domain NOMA to random access for an IoT network that consists of one BS and a large number of devices with a limited number of channels. In addition, MAB is considered so that devices can learn channel selection probabilities to maximize the throughput as in [5]. Two different MAB approaches are derived. One is similar to that in [5], while the other approach requires learning in both the BS and devices or two-sided learning. In two-sided learning, the BS is to learn the optimal channel selection probabilities and decide rewards based on them. In addition, each device is to learn channel selection probabilities through a MAB formulation. Thanks to the rewards based on optimal channel selection probabilities that are decided by the BS, the learning speed at devices can be improved. Note that in [16], the learning based MAB is analyzed as an evolutionary game, no two-sided learning is studied.

In summary, the aim of the paper is to improve the performance of IoT networks by introducing power-domain NOMA for DDs that can dynamically change access channels and design a learning scheme for both smarter devices and BS to improve the throughput. The novelty is the two-sided learning scheme that allows smarter devices and BS to interact so that key parameters can be adjusted to maximize the throughput (note that the learning in [5] is carried out at DDs, not BS). As a result, the main contribution of the paper becomes two-fold: *i)* a NOMA-based random access approach is proposed for the IoT network with different types of devices as in [5] to improve the throughput; *ii)* two-sided learning is proposed for the NOMA-based random access to improve learning speed.

The rest of the paper is organized as follows. In Section II, the system model for IoT networks is presented with different types of devices. To show the throughput improvement by power-domain NOMA, the throughput is analyzed in Section III. For the proposed NOMA-based random access approach, MAB is considered with two-sided learning in Section IV. We present simulation results in Section V and conclude the paper with remarks in Section VI.

NOTATION

Matrices and vectors are denoted by upper- and lower-case boldface letters, respectively. The superscript T denotes the transpose. For a set $\mathcal{A}$, $|\mathcal{A}|$ represents the cardinality of $\mathcal{A}$. $\mathbb{E}[\cdot]$ and $Var(\cdot)$ denote the statistical expectation and variance, respectively. $\mathcal{N}(a, R)$ represents the distribution of Gaussian random vectors with mean vector a and covariance matrix R.

II. SYSTEM MODEL

In this section, we present the system model based on [5]. Throughout the paper, it is assumed that a system consists of a number of devices and a BS. In addition, we assume that there are L orthogonal resource blocks (RBs) or channels (i.e., throughout the paper, the terms RBs and channels are interchangeable).

A. CO-EXISTING DIFFERENT TYPES OF DEVICES

In this paper, we assume two groups of devices. One group consists of SDs of a low duty cycle or access probability. In addition, the other group consists of DDs that are more capable than SDs in the following ways:

1) each DD can choose a channel dynamically;
2) the transmit power of DDs is higher than that of SDs so that power-domain NOMA can be employed.

Power-domain NOMA can effectively create multiple channels within a RB to improve the spectral efficiency. Assuming that there are two different power levels, we can see that SDs access the channels of low power, while DDs access the channels of high power. Thus, the presence of DDs may not significantly degrade the throughput of SDs thanks to NOMA.

As in [5], each SD has a fixed channel (among L channels) to communicate with the BS as it is less flexible due to poor RF capability. Thus, the number of SDs that access through channel l, denoted by S_l, is assumed to be a constant. On the other hand, DDs are more flexible and able to dynamically choose a channel out of L channels according to channel selection probabilities that can be learned. In Fig. 1, we illustrate the system model with L RBs to support both SDs and DDs using power-domain NOMA.

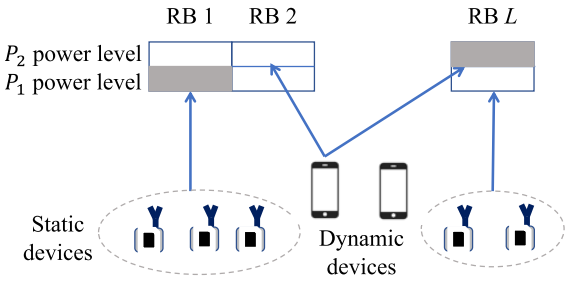


FIGURE 1. An illustration of the system model with L RBs to support both SDs and DDs using power-domain NOMA.

In addition, as mentioned above, each DD is to transmit a higher power than SDs to avoid collision with SDs by exploiting power-domain NOMA [14]. While this feature is not considered in [5], it plays a crucial role in not only improving performance, but also learning as will be explained later.

B. NOMA-BASED RANDOM ACCESS FOR TWO DIFFERENT TYPES OF DEVICES

In this subsection, power-domain NOMA is considered to support both SDs and DDs in an IoT network.

Let $\mathcal{K}_l$ and $\bar{\mathcal{K}}_l$ denote the index sets of active SDs and DDs that transmit their signals to channel l, respectively. Let $K_l = |\mathcal{K}_l|$ and $\bar{K}_l = |\bar{\mathcal{K}}_l|$. The signals from the kth active SD and DDs are denoted by s_k and $\bar{s}_k$, respectively. Then, the received signal through channel l is given by

$$y_l = \sqrt{P_1} \sum_{k \in \mathcal{K}_l} s_k + \sqrt{P_2} \sum_{k \in \bar{\mathcal{K}}_l} \bar{s}_l + n_l, \tag{1}$$

where $n_l \sim \mathcal{N}(0, \sigma^2)$ denotes the background noise, and P_1 and P_2 represent the power levels of SDs and DDs, respectively. We assume that $Var(s_k) = Var(\bar{s}_k) = 1$ for normalization with $\mathbb{E}[s_k] = \mathbb{E}[\bar{s}_k] = 0$ for all k. Thus, if $|\bar{\mathcal{K}}_l| = \bar{K}_l = 1$, the signal-to-interference-plus-noise ratio (SINR) of DDs is given by

$$\mathsf{SINR}_2 = \frac{P_2}{K_l P_1 + \sigma^2}. \tag{2}$$

It is assumed that the signals from one DD through channel l is decodable if

$$\mathsf{SINR}_2 \geq \Gamma, \tag{3}$$

where $\Gamma > 0$ denotes the SINR threshold for successful decoding. Provided that there is only one DD in channel l, when $\Gamma \leq \frac{P_2}{P_1+\sigma^2}$, the signal from the DD is decodable if there is at most one active SD in the same channel. Of course, in this case, the signal of the active SD is also decodable after successive interference cancellation (SIC) [12], [13].

It is noteworthy that the packet collision with multiple DDs in a channel results in decoding failure of the SD in the same channel due to error propagation [14]. Thus, it has to be assumed that the probability of packet collision with DDs is sufficiently low. In other words, the average number of active DDs, denoted by λ, has to be lower than the number of channels, L. Throughout the paper, therefore, we assume that $\lambda \leq L$.

Note that the total number of SDs is $M_1 = \sum_{l=1}^{L} S_l$. Let p_1 denote the access probability of SDs that are active independently. In general, the total number of DDs, denoted by M_2, is also finite. Denote by p_2 the access probability of DDs that also become active independently. Thus, we have

$$\lambda = \mathbb{E}[N_2] = M_2\, p_2, \tag{4}$$

where $N_2 = \sum_{l=1}^{L} \bar{K}_l$ is the number of active DDs. For convenience, with a sufficiently large M_2 and a low p_2, N_2 is assumed to be a Poisson random variable, i.e.,

$$N_2 \sim Poiss(\lambda) \text{ or } \Pr(N_2 = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \tag{5}$$

which is an approximation [17].

III. THROUGHPUT ANALYSIS

In this section, we focus on the throughput analysis in order to demonstrate that the performance can be improved by exploiting the notion of power-domain NOMA in supporting SDs and DDs in different ways (or powers) as discussed in Subsection II-B.

A. PERFORMANCE WITHOUT NOMA

For comparisons, we discuss the throughput of the conventional random access approach in [5], where no power-domain NOMA is considered.

In each channel, both SDs and DDs transmit signals with power $P = P_1 = P_2$ as NOMA is not used. Then, provided that there are N_2 active DDs, the conditional probability that one DD can successfully transmit its packet through channel l is given by

$$\mathbb{P}_{2,l}(N_2) = (1-p_1)^{S_l} \binom{N_2}{1} q_l\, (1-q_l)^{N_2-1}, \tag{6}$$

where q_l is the probability that an active DD chooses channel l or DD's selection probability of channel l. Thus, the throughput of DDs, which is the average number of successfully transmitted packets by DDs, is given by

$$\begin{aligned}\eta_{conv,2}(q) &= \mathbb{E}\left[\sum_{l=1}^{L} \mathbb{P}_{2,l}(N_2)\right] \\ &= \sum_{l=1}^{L} (1-p_1)^{S_l} \sum_{k=0}^{\infty} k q_l\, (1-q_l)^{k-1}\, \frac{e^{-\lambda}\lambda^k}{k!} \\ &= \sum_{l=1}^{L} (1-p_1)^{S_l}\, \lambda q_l e^{-\lambda q_l}. \end{aligned} \tag{7}$$

Let $\mathbb{P}_{1,l}(N_2)$ denote the conditional probability that an active SD in channel l can successfully transmit its packet provided that there are N_2 active DDs, which is given by

$$\mathbb{P}_{1,l}(N_2) = S_l p_1\, (1-p_1)^{S_l-1}\, (1-q_l)^{N_2}. \tag{8}$$

Then, the throughput of SDs can also be found as

$$\begin{aligned}\eta_{conv,1}(q) &= \mathbb{E}\left[\sum_{l=1}^{L} \mathbb{P}_{1,l}(N_2)\right] \\ &= \sum_{l=1}^{L} S_l p_1\, (1-p_1)^{S_l-1}\, e^{-\lambda q_l}. \end{aligned} \tag{9}$$

For a given channel l, the throughput of SDs per channel is a decreasing function of q_l as shown in (9). On the other hand, as shown in (7), the throughput of DDs per channel is an increasing function of q_l when $q_l \leq \frac{1}{\lambda}$. The relationship between the throughput (per channel) and DD's access probability q_l is illustrated in Fig. 2. As a result, the increase of throughput of DDs leads to the decrease of throughput of SDs.

B. PERFORMANCE WITH NOMA

In this subsection, we focus on the throughput when NOMA is employed.

Let ω_l denote the probability that an active DD can successfully transmit its packet when it is only one active DD. Then, with $\Gamma = \frac{P_2}{P_1+\sigma^2}$, from (2) and (3), it can be given by

$$\begin{aligned}\omega_l &= \Pr(\mathsf{SINR}_2 \geq \Gamma \mid \bar{K}_l = 1) \\ &= (1-p_1)^{S_l} + \binom{S_l}{1} p_1 (1-p_1)^{S_l}. \end{aligned} \tag{10}$$

As shown in (10), ω_l only depends on the number of active SDs in channel l. Provided that there are N_2 active DDs, the conditional probability that an active DD choosing channel l can successfully transmit its packet is given by

$$\mathbb{P}_{2,l}(N_2) = \omega_l \binom{N_2}{1} q_l (1-q_l)^{N_2-1}. \tag{11}$$

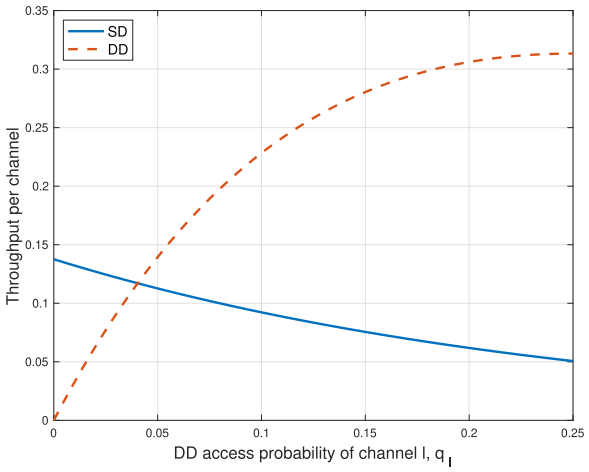


FIGURE 2. Throughput of the conventional approach per channel as a function of DD access probability, q_l, for SDs and DDs when $p_1 = 0.01$, $S_l = 16$, and $\lambda = 4$.

Thus, the throughput of DDs is given by

$$\eta_2(q) = \mathbb{E}\left[\sum_{l=1}^{L} \omega_l N_2 q_l (1-q_l)^{N_2-1}\right] = \lambda \sum_{l=1}^{L} \omega_l q_l e^{-q_l \lambda}. \quad (12)$$

To decode the signal from an active SD in a channel, it is necessary to decode any active DD and perform SIC. Thus, the throughput of SDs can be given by

$$\eta_1(q) = \mathbb{E}\Bigg[\sum_{l=1}^{L} S_l p_1 (1-p_1)^{S_l-1} \times \left(N_2(1-q_l)^{N_2-1} + (1-q_l)^{N_2}\right)\Bigg] = \sum_{l=1}^{L} S_l p_1 (1-p_1)^{S_l-1} e^{-q_l\lambda}(1+\lambda q_l). \quad (13)$$

Comparing (13) and (9), we can conclude that the proposed random access approach with NOMA has a higher throughput of SDs than the conventional random access approach in [5] at the cost of high transmit power of DDs. In other words, the presence of DDs has less impact on the performance of SDs when power domain NOMA is used. Furthermore, if $q_l = \frac{1}{L}$, we have

$$e^{-q_l\lambda}(1+\lambda q_l) = e^{-\frac{\lambda}{L}}\left(1+\frac{\lambda}{L}\right) \leq 0.7358,$$

as $\lambda \leq L$. This demonstrates that the throughput of SDs can be degraded by a factor of up to 0.7358 due to the presence of DDs. On the other hand, as shown in (9), without NOMA, the throughput of SDs can be degraded by a factor of up to $e^{-1} = 0.3679$ due to the presence of DDs.

We also have the following result to show that the throughput of the proposed approach with NOMA is higher than that of the conventional one [5].

Lemma 1:

$$\max_q \eta_1(q) \geq \max_q \eta_{conv,1}(q)$$
$$\max_q \eta_2(q) \geq \max_q \eta_{conv,2}(q)$$
$$\max_q \eta_1(q) + \eta_2(q) \geq \max_q \eta_{conv,1}(q) + \eta_{conv,2}(q). \quad (14)$$

Proof: The result can be easily obtained from (9), (7), (13), and (12). Thus, we omit the proof. ■

In Fig. 3, the relationship between the throughput (per channel) and DD's channel selection probability q_l is illustrated with the same values of the parameters as those in Fig. 2. Comparing Figs. 2 and 3, it is clear that the proposed approach with NOMA can provide a higher throughput than the conventional one for both SDs and DDs.

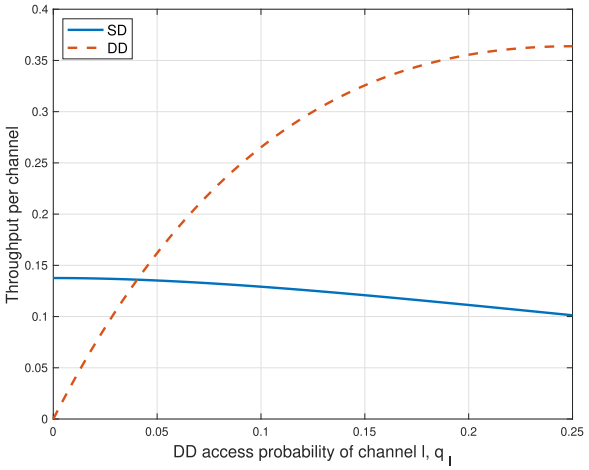


FIGURE 3. Throughput of the proposed approach per channel as a function of DD selection probability, q_l, for SDs and DDs when $p_1 = 0.01$, $S_l = 16$, and $\lambda = 4$.

In Fig. 4, we also compare the conventional random access approach and the proposed random access one with NOMA in terms of the total throughput when λ varies from 0 to L with $L = 10$, $S_l = S = 10$, $p_1 = 0.1$, and $q_l = \frac{1}{L}$ for all l. It is clear that the proposed one has a higher throughput than the conventional one thanks to power-domain NOMA.

C. OPTIMAL CHANNEL SELECTION PROBABILITY

As shown in (13), a salient feature of the proposed random access approach is that the throughput of SDs is less dependent on q_l as long as λq_l is sufficiently low, which might be the case that $\lambda < L$. This is a desirable result as DDs are to be opportunistic in accessing channels. That is, the presence of smart DDs should not have a serious impact on poorly capable SDs. Based on this, we can consider the following optimization problem:

$$q^* = \operatorname*{argmax}_q \eta_2(q) \quad \text{subject to} \sum_{l=1}^{L} q_l = 1,\ q_l \geq 0. \quad (15)$$

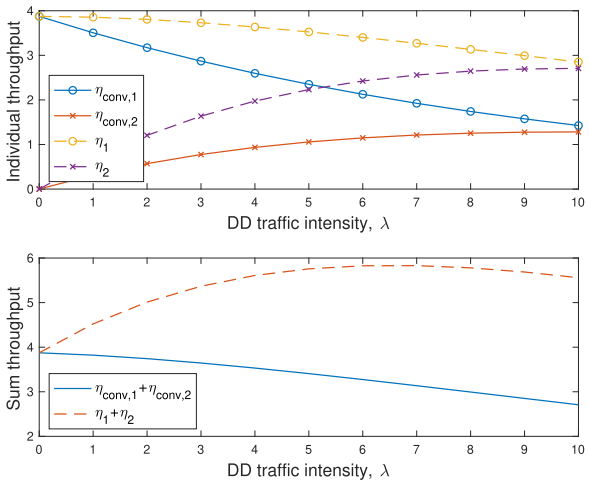


FIGURE 4. Performance of the conventional random access approach and the proposed one with NOMA in terms of throughput for various values of λ when $L = 10$, $S_l = S = 10$, $p_1 = 0.1$ and $q_l = \frac{1}{L}$ for all l.

In other words, in optimizing q, we only need to consider the performance of DDs provided that the presence of DDs does not have a serious impact on SDs.

Lemma 2: The solution of the problem in (15) is given by

$$q_l^* = \begin{cases} \dfrac{1 - \mathbb{W}\left(\frac{\beta e}{\omega_l}\right)}{\lambda}, & \text{if } \beta \le \omega_l \\ 0, & \text{o.w.,} \end{cases} \tag{16}$$

where $\mathbb{W}(\cdot)$ is the Lambert W function and β is the Lagrange multiplier. The value of the Lagrange multiplier has to be decided to satisfy $\sum_{l=1}^{L} q_l^* = 1$.

Proof: From (15), the following unconstrained optimization problem can be considered:

$$\min_{q} \sum_{l=1}^{L} \omega_l q_l e^{-q_l \lambda} - \beta \sum_{l=1}^{L} q_l. \tag{17}$$

By taking the derivative with respect to q_l and setting it to 0, we have

$$e^{-q_l\lambda}(1 - q_l\lambda) = \frac{\beta}{\omega_l}, \text{ for all } l. \tag{18}$$

Then, (18) is re-written as

$$u_l e^{u_l} = z_l, \tag{19}$$

where $u_l = 1 - q_l\lambda$ and $z_l = \frac{\beta e}{\omega_l}$. Since $u = \mathbb{W}(z)$ when $ue^u = z$, we have

$$u_l = \mathbb{W}(z_l) \text{ or } q_l = \frac{1 - \mathbb{W}(z_l)}{\lambda}. \tag{20}$$

In addition, since $\mathbb{W}(z)$ is an increasing function of $z \ge 0$ and becomes 1 when $z = e$, for $z_l > e$ or $\beta > \omega_l$, q_l has to be 0. This results in (16), which completes the proof. ■

Note that finding β that satisfies $\sum_l q_l^* = 1$ is straightforward as $1 - \mathbb{W}(z_l)$ is a nonincreasing function of β. It can be found by any simple numerical technique, e.g., the bisection method [18].

IV. TWO-SIDED LEARNING

In this section, we discuss learning for the proposed random access approach. In particular, two-sided learning is considered where the BS and DDs perform learning to maximize the throughput of DDs.

Throughout this section, it is assumed that SDs and DDs become independently active to transmit their packets in each time slot. Thus, we denote by $N_1(t)$ and $N_2(t)$ the total numbers of active SDs and DDs at time slot t, respectively, which are assumed to be independent and identically distributed (iid).

A. BS's LEARNING FOR SDs' ACTIVITIES

As shown in (16), the DD's optimal channel selection probabilities, $\{q_l^*\}$, depend on $\{\omega_l\}$. Thus, it is necessary for the BS to learn or estimate the ω_l's that are assumed to be fixed. To this end, the BS needs to learn SDs' activities.

From (10), we can show that

$$\omega_l = \Pr(\mathcal{A}_{1,l}), \tag{21}$$

where $\mathcal{A}_{1,l}$ denotes the event that the number of active SDs in channel l is 0 (idle) or 1. Unfortunately, the event $\mathcal{A}_{1,l}$ cannot be observed if there are more than 1 active DDs in channel l due to error propagation. On the other hand, if there is none or one active DD (no collision between DDs), the BS is able to observe the event $\mathcal{A}_{1,l}$ and update ω_l. In particular, an on-line estimate of ω_l at time t can be updated as follows:

$$\hat{\omega}_l(t) = \begin{cases} \dfrac{(t-1)\hat{\omega}_l(t-1)}{t} + \dfrac{\mathbb{1}(\mathcal{A}_{1,l}(t))}{t}, & \text{if } \bar{K}_l \le 1 \\ \hat{\omega}_l(t-1), & \text{o.w.,} \end{cases} \tag{22}$$

where $\mathcal{A}_{1,l}(t)$ is the event $\mathcal{A}_{1,l}$ at time slot t. Let $X_l(t) = \mathbb{1}(\mathcal{A}_{1,l}(t)) \in \{0, 1\}$. If the BS can observe $\mathcal{A}_{1,l}(t)$ regardless of DD collision, it can be seen that $\hat{\omega}_l(t)$ is the sample mean of the $X_l(t)$'s, which are iid. Thus, in this case, as $t \to \infty$, $\hat{\omega}_l(t)$ converges to ω_l w.p. 1 [19].

However, as mentioned earlier, the BS may not be able to see some events $\mathcal{A}_{1,l}(t)$ due to collision between active DDs in channel l. In the presence of DD collision, to see whether or not $\hat{\omega}_l(t)$ can converge to ω_l, consider an example. Suppose that DD collision happens at $t = 3$ in channel l. Thus, from (22), the estimate of ω_l at $t = 4$, is given by

$$\begin{aligned} \hat{\omega}_l(4) &= \frac{X_l(1) + X_l(2) + \frac{X_l(1)+X_l(2)}{2} + X_l(4)}{4} \\ &= \frac{3}{8}X_l(1) + \frac{3}{8}X_l(2) + \frac{1}{4}X_l(4). \end{aligned}$$

Thus, in general, $\hat{\omega}_l(t)$ can be written as

$$\hat{\omega}_l(t) = \sum_{\tau \in \mathcal{T}(t)} V(\tau; t) X_l(\tau), \tag{23}$$

where $\mathcal{T}(t)$ denotes the index set of the time slots of no DD collision up to time slot t and $V(\tau; t)$ is the weight for $X_l(\tau)$ that depends on the events of DD collision. Due to the

normalization, we have

$$\sum_{\tau \in \mathcal{T}(t)} V(\tau; t) = 1.$$

Since the activity of SDs is independent of that of DDs, $X_l(t)$ is independent of DD collision, i.e., $V(\tau; t)$. In addition, since the events of DD collision are independent, as $t \to \infty$, it can be shown that $V(\tau; t) = O(1/t)$ if the probability of DD collision is not 1. Note that the probability of DD collision in channel l is $\rho = 1 - e^{-q_l \lambda}(1 + q_l \lambda) < 1$. Thus, as in [20], we expect that $\hat{\omega}_l(t) \to \omega_l$ w.p. 1 as $t \to \infty$.

Thanks to power-domain NOMA, as shown above, active DDs do not significantly interfere with learning to estimate the ω_l's, and reliable estimates of the ω_l's become available after a sufficient number of slots. Then, with the estimates, the BS is able to obtain q_l^* as in (16).

B. DDs' LEARNING VIA MAB FORMULATION

Prior to deriving a learning approach for DDs using the online estimates of the ω_l's at the BS, we consider a straightforward extension of the learning algorithm used in [5] to the proposed random access approach with NOMA in this subsection.

As in [5], MAB can be used for learning at DDs. Suppose that a DD has L arms or channels. When a DD becomes active to transmit its packet, it chooses one of L arms and receives a reward from the BS. If rewards are iid for selected arms in multiple plays, the DD can learn or estimate the mean reward, denoted by μ_l for arm l, after a number of plays. Then, it can select the best arm, i.e., $l^* = \text{argmax}_l \hat{\mu}_l$, where $\hat{\mu}_l$ is an estimate of μ_l.

For MAB, Thompson sampling [21] can be used. For each arm, a beta distribution $Beta(a_{l;m}(t), b_{l;m}(t))$, where $a_{l;m}(t)$ and $b_{l;m}(t)$ are the shape parameters, is assumed at DD m, where $m \in \{1, \ldots, M_2\}$. For uniform prior, it can be assumed that $a_{l;m}(0) = b_{l;m}(0) = 1$. After each play, DD m receives a reward, $r_{l;m}(t) \in \{-1, 1\}$. Here, $r_{l;m}(t) = -1$ or 1 implies that transmission through channel l by DD m is unsuccessful or successful, respectively. The shape parameters can be updated as follows:

$$\begin{aligned}
&\text{if } r_{l;m}(t) = 1 \\
&\quad a_{l;m}(t) = a_{l;m}(t-1) + 1,\ b_{l;m}(t) = b_{l;m}(t-1) \\
&\text{if } r_{l;m}(t) = -1 \\
&\quad b_{l;m}(t) = b_{l;m}(t-1) + 1,\ a_{l;m}(t) = a_{l;m}(t-1).
\end{aligned} \tag{24}$$

If DD m does not play at time t (i.e., DD m is not an active DD), the shape parameters are not updated, i.e., $a_{l;m}(t) = a_{l;m}(t-1)$ and $b_{l;m}(t) = b_{l;m}(t-1)$.

At time slot t, if DD m has a packet to send, it can have samples from the Beta posterior as follows:

$$Z_{l;m}(t) \sim Beta(a_{l;m}(t), b_{l;m}(t)). \tag{25}$$

Then, from the samples, the selected arm or channel to send a packet is given by

$$l^* = \underset{l \in \{1, \ldots, L\}}{\text{argmax}}\ Z_{l;m}(t), \tag{26}$$

which is a randomized selection policy.

As discussed in [5], although the approach in (26) does not consider multiple players, i.e., other DDs, interacting with the same set of arms, i.e., L channels, it may provide reasonable performance after a number of plays. However, there are a few drawbacks as follows:

- The BS does not exploit the estimates of the ω_l's or the channel selection probabilities, $\{q_l\}$, in making the rewards, although they are available.
- A DD can receive a reward only when it is active. Thus, when p_2 is low, the time to learn for each DD is limited. Furthermore, since an active DD can choose one arm at a time, with a large L, it may take a long time to learn or have reliable shape parameters for each DD.

C. A MODIFIED LEARNING APPROACH FOR DDs

In this subsection, we propose a two-sided learning approach where the BS learns the channel selection probabilities using the estimates of the ω_l's as discussed in Subsection IV-A, and DDs learn using designed rewards from the BS.

At time t, the BS expects to receive $\Sigma_l(t) = \lambda \sum_{i=1}^{t} q_l(i)$ packets from DDs through channel l, where $q_l(t)$ denotes the selection probability of channel l at time t obtained using $\{\hat{\omega}_l(t)\}$. Note that if $\hat{\omega}_l(t) \to \omega_l$, we have $q_l(t) \to q_l^*$ as t increases. Thus, the BS can make a reward, which is $r_l(t) = 1$, for the active DDs that send packets through channel l to increase the selection probability if

$$\hat{\Sigma}_l(t) < \Sigma_l(t) - \epsilon(t), \tag{27}$$

where $\hat{\Sigma}_l(t)$ represents the accumulated number of active DDs that transmit packets through channel l up to time t and $\epsilon(t)$ is a positive increasing function of t. On the other hand, if

$$\hat{\Sigma}_l(t) > \Sigma_l(t) + \epsilon(t), \tag{28}$$

the BS makes a different reward, which is $r_l(t) = -1$, for the active DDs that send packets through channel l to decrease the selection probability.

If an active DD, say DD m, receives $r_l(t)$, it can update the shape parameters as in (24). Note that if

$$|\Sigma_l(t) - \hat{\Sigma}_l(t)| \leq \epsilon(t), \tag{29}$$

the reward becomes 0, i.e., $r_l(t) = 0$. In this case, the active DDs do not update their shape parameters.

Note that the same rewards, $r_l(t)$, are shared by all the active DDs that transmit packets through channel l. Thus, the BS broadcasts $r_l(t)$. In Algorithm 1, the algorithm to make the rewards is summarized.

Consequently, we expect that $\hat{\Sigma}_l(t)$ can follow $\Sigma_l(t)$ as illustrated in Fig. 5. We can also consider asymptotic behaviors. Regardless of DD's learning of the channel selection probabilities, it can be shown that

$$\Sigma_l(t) \to t q_l^* \lambda,\ t \to \infty,$$

because the activity of SDs is independent of that of DDs. Suppose that $\epsilon(t) = \epsilon t$ with a sufficiently small $\epsilon > 0$. Then,

Algorithm 1 Making Rewards at the BS

Data: $\{\hat{\omega}_l(t)\}$, $\{\hat{\Sigma}_l(t)\}$, and $\epsilon(t)$
Result: $\{r_l(t)\}$
update $\{q_l(t)\}$ from $\{\hat{\omega}_l(t)\}$ using (16);
find $\{\Sigma_l(t)\}$ from $\{q_l(t)\}$;
for $l = 1 : L$ **do**
  if $\hat{\Sigma}_l(t) < \Sigma_l(t) - \epsilon(t)$ **then**
    $r_l(t) = 1$;
  else if $\hat{\Sigma}_l(t) > \Sigma_l(t) + \epsilon(t)$ **then**
    $r_l(t) = -1$;
  else
    $r_l(t) = 0$;
  end
end

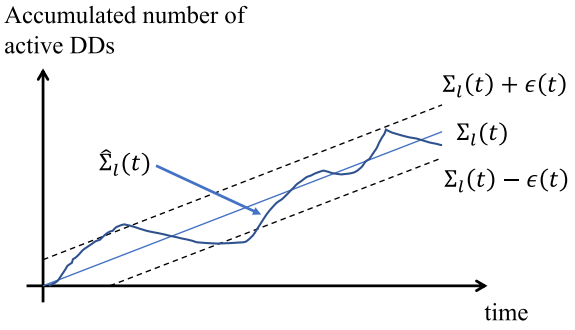


FIGURE 5. An illustration of trajectory of $\hat{\Sigma}_l(t)$.

for a large t, we have

$$\frac{\hat{\Sigma}_l(t) - \Sigma_l(t)}{t} = \lambda \left(\frac{\hat{\Sigma}_l(t)}{t} - q_l^* \right). \tag{30}$$

If the sample mean is not close to λq_l^*, the BS sends appropriate rewards to update the shape parameters, which may happen during the early phase of learning. However, when the sample mean $\frac{\hat{\Sigma}_l(t)}{t}$ is sufficiently close to λq_l^*, i.e., $\lambda|\frac{\hat{\Sigma}_l(t)}{t} - q_l^*| \leq \epsilon$, the shape parameters are not updated according to Algorithm 1. Thus, as DDs have learned the channel selection probabilities well through the shape parameters $\{a_{l;m}(t), b_{l;m}(t)\}$, the sample mean will converge to the average number of active DDs in channel l, which is λq_l, as $t \to \infty$.

While the MAB approach in Subsection IV-B, which will be referred to as MAB 1 for convenience, has rewards that depend on the *instantaneous outcomes* of DD's collision, the approach based on two-sided learning in this subsection, which will be referred to as MAB 2, provides the rewards based on the *accumulated number* of active DDs, $\hat{\Sigma}_l(t)$, which results from BS's learning. Thus, the rewards in MAB 2 can be seen as smoothed versions of those in MAB 1, which may lead to faster learning at DDs (as confirmed by simulation results in Section V).

As mentioned earlier, each active DD chooses only one channel to transmit its packet and receives a reward $r_l(t)$ for the selected channel l. As a result, it may take a long time to learn or estimate the selection probabilities for all L channels. Thus, in order to obtain reliable estimates of the channel selection probabilities, each DD may use a fraction of L channels, i.e., B channels, where $B < L$. Let $\mathcal{B}_m$ denote the index set of B selected channels at DD m. Clearly, $|\mathcal{B}_m| = B$ and $\mathcal{B}_m \subseteq \{1, \ldots, L\}$. In this case, each DD only needs to learn the selection probabilities of B channels. In Algorithm 2, we summarize the learning process at DD m for the received reward, $r_l(t)$, when it becomes active and sends a packet through channel $l \in \mathcal{B}_m$.

Algorithm 2 Updating Shape Parameters at DD m

Data: $r_l(t)$, for a $l \in \mathcal{B}_m$
Result: $\{a_{l;m}(t), b_{l;m}(t)\}$
if $r_l(t) = 1$ **then**
  $a_{l;m}(t) = a_{l;m}(t-1) + 1$, $b_{l;m}(t) = b_{l;m}(t-1)$;
else if $r_l(t) = -1$ **then**
  $b_{l;m}(t) = b_{l;m}(t-1) + 1$, $a_{l;m}(t) = a_{l;m}(t-1)$;
else
  $a_{l;m}(t) = a_{l;m}(t-1)$, $b_{l;m}(t) = b_{l;m}(t-1)$;
end

Note that B has to be greater than 1. If $B = 1$, each DD needs to choose only one fixed channel. In other words, it becomes an SD. In addition, it is required that B is not too small. To see this, suppose that $B = 2$ and assume that $\mathcal{B}_m = \{1, 2\}$ for DD m. In this case, if $q_1^* = q_2^* = 0$, DD m cannot transmit any packets. Thus, B is sufficiently large so that the sum of channel selection probabilities is greater than 0.

V. SIMULATION RESULTS

In this section, we present simulation results under the setting that is similar to that in [5]. In particular, we assume that $L = 10$ and the SDs are distributed over $L = 10$ channels as follows:

$$(S_1, \ldots, S_L) = (0.3, 0.2, 0.1, 0.1, 0.05, 0.05, 0.02, 0.08, 0.01, 0.09) \times M_1,$$

where the total number of SDs is set to $M_1 = 1000$. The total number of DDs, M_2, is set to 100 for all simulations. Recall that the MAB approaches in Subsections IV-B and IV-C are referred to as MABs 1 and 2, respectively. That is, MAB 2 requires two-sided learning, while MAB 1 only requires learning at DDs. For MAB 2, we assume that $B = L$ unless it is stated otherwise and $\epsilon(t) = \epsilon t$ with $\epsilon = 0.01$.

Note that as illustrated in Fig. 4, since the conventional random access approach that does not use NOMA cannot have better performance than the proposed approach with NOMA, we only present simulation results of the proposed random access approach and focus on learning aspects.

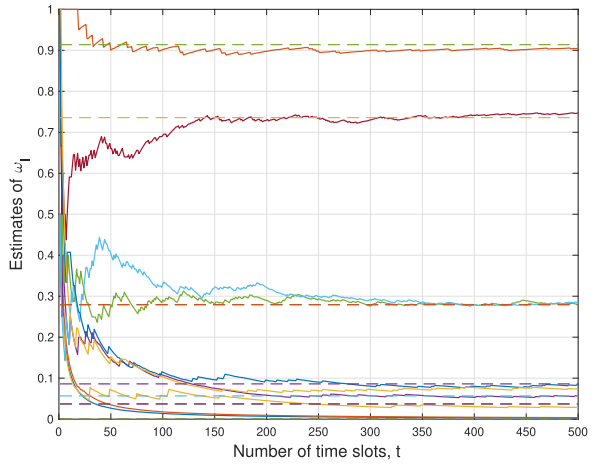


FIGURE 6. On-line estimation of ω_l when $p_1 = 0.05$ and $p_2 = 0.06$. In the figure, the solid lines represent the on-line estimates and the dash lines represent the true values of the ω_l's.

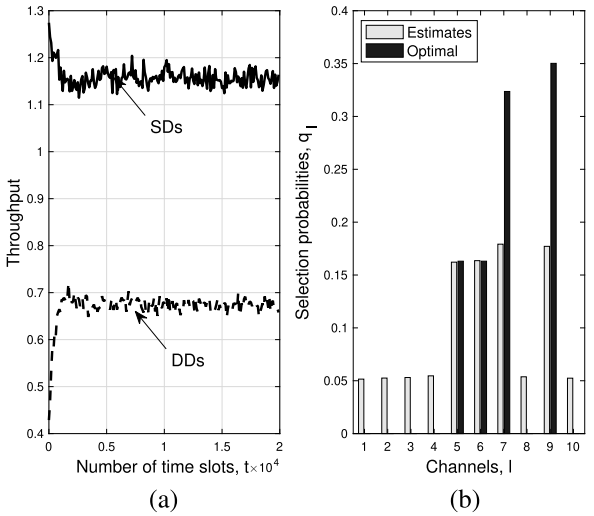


FIGURE 8. Performance of MAB 2 with $p_1 = 0.05$ and $p_2 = 0.02$; (a) throughput as a function of time; (b) the estimates of the channel selection probabilities, q_l^*, at DDs.

In Fig. 6, we show the estimates of the ω_l's that are obtained by (22) when $p_1 = 0.05$ and $p_2 = 0.06$. As expected, we can see that $\hat{\omega}_l(t)$ can approach ω_l as t increases.

In Figs. 7 and 8, the simulation results of MABs 1 and 2 are shown in terms of the throughput over time and the estimates of the channel selection probabilities, q_l, at DDs when $p_1 = 0.05$ and $p_2 = 0.02$. The estimate of q_l at DDs is obtained by taking the average of $\frac{a_{l;m}(t)}{a_{l;m}(t)+b_{l;m}(t)}$ over M_2 DDs. Comparing Figs. 7 and 8, we can see that with MAB 2, DDs can learn the channel selection probabilities faster than MAB 1. As mentioned earlier, in MAB 2, the BS provides rewards according to the optimal q_l's that are obtained using the estimates of ω_l's. Thus, DDs' learning is carried out under more stationary settings (once the ω_l's are reliably estimated as in Fig. 6), which results in faster learning outcomes in MAB 2 than those in MAB 1.

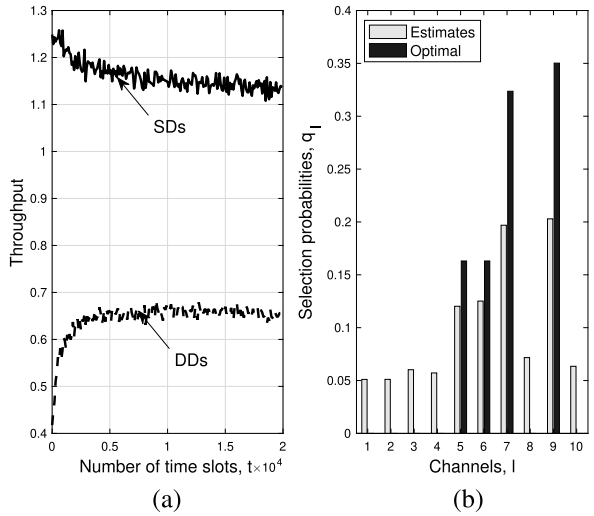


FIGURE 7. Performance of MAB 1 with $p_1 = 0.05$ and $p_2 = 0.02$; (a) throughput as a function of time; (b) the estimates of the channel selection probabilities, q_l^*, at DDs.

Fig. 9 shows the average throughput of DDs over 2000 time slots as functions of p_1 when $p_2 = 0.06$ (in Fig. 9 (a)) and $p_2 = 0.01$ (in Fig. 9 (b)). For each average throughput, 100 runs are used. It is shown that the throughput decreases with p_1, since ω_l decreases with p_1. It is also observed that the performance of MAB 1 differs from that of MAB 2 when p_2 is low. When p_2 or λ is high, the channel selection probability tends to be even (i.e., each q_l approaches $\frac{1}{L}$). On the other hand, as p_2 or λ decreases, the channel selection probabilities are different. Thus, learning becomes more important when p_2 or λ is low. Consequently, as shown in Fig. 9, the performance difference between MABs 1 and 2 is not significant with $p_2 = 0.06$, while MAB 2 performs better than MAB 1 with $p_1 = 0.01$.

In Fig. 10, we show the average throughput of DDs over 2000 time slots as functions of p_2 when $p_1 = 0.1$. For each average throughput, 100 runs are used. Since MAB 2 can learn the channel selection probabilities faster than MAB 1, it is shown that the performance of MAB 2 is better than that of MAB 1. As mentioned earlier, for a large p_2, the selection probability tends to be even. Thus, the performance difference between MABs 1 and 2 diminishes as p_2 increases.

Finally, we consider the case that each DD uses only B out L channels, where $B < L$, and learn the selection probabilities of B channels in MAB 2. This makes learning faster with the drawbacks mentioned earlier. In Fig. 11, we present simulation results when $p_1 = 0.05$ and $p_2 = 0.01$. If $B = 1$, DDs become SDs and cannot dynamically choose channels. As a result, the performance of MAB 2 is poor when $B = 1$, while the performance is improved as B increases. However, a large B requires a longer learning time. As a result, it seems there is an optimal B, which is $B = 7$ under the setting according to Fig. 11.

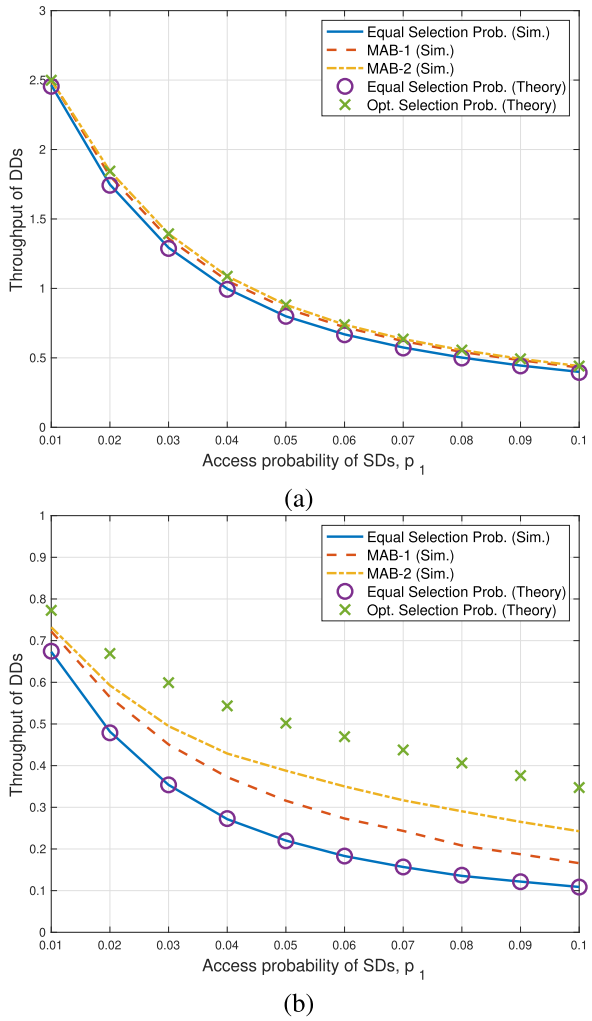


FIGURE 9. Average throughput of DDs over 2000 time slots as functions of p_1: (a) $p_2 = 0.06$ or $\lambda = 6$; (b) $p_2 = 0.01$ or $\lambda = 1$.

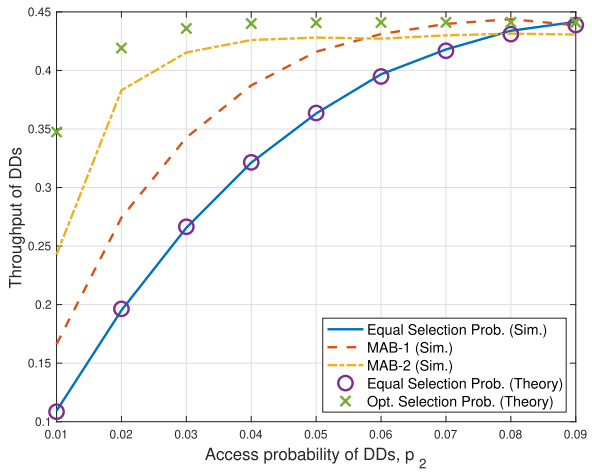


FIGURE 10. Average throughput of DDs over 2000 time slots as functions of p_2 when $p_1 = 0.1$.

It is noteworthy that the estimates of the channel selection probabilities at DDs are slightly different from the actual ones, as shown in Fig. 7 (b) and Fig. 8 (b) as MAB is a randomized selection policy. That is, even if $q_l^* = 0$ for some l, the selection probability of this arm or channel at a DD may not be zero. Thus, a better approach can be obtained by allowing to remove some unused channels. This may be combined with dynamic selection of the subset of channels for $\mathcal{B}_m$ by each DD, which might be a further work.

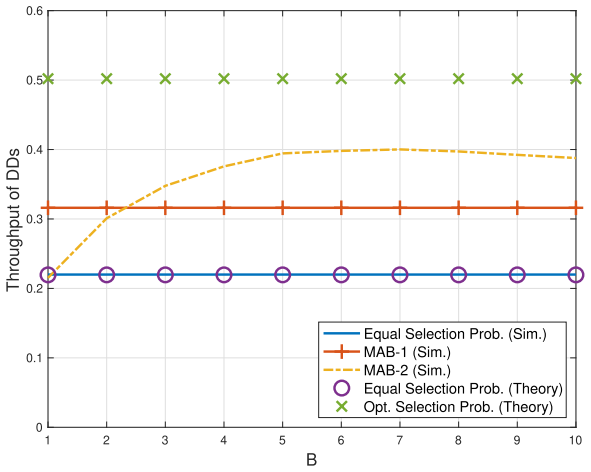


FIGURE 11. Average throughput of DDs over 2000 time slots as a function of B for MAB 2 when $p_1 = 0.05$ and $p_2 = 0.01$.

VI. CONCLUDING REMARKS

In this paper, we proposed a NOMA-based random access approach for IoT networks where SDs and DDs co-exist. It was shown that the proposed random access approach can provide a higher throughput than the conventional random access approach that does not use NOMA. For DDs that are flexible enough to choose one of multiple channels, we proposed two-sided learning where the learning outcomes at the BS are used to make rewards for DDs' learning based on a MAB formulation. Thanks to the rewards decided by the BS, the resulting MAB approach can improve learning speed at DDs compared to a conventional MAB approach.

REFERENCES

[1] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," *Future Gener. Comput. Syst.*, vol. 29, no. 7, pp. 1645–1660, Sep. 2013.

[2] J. Kim, J. Yun, S.-C. Choi, D. N. Seed, G. Lu, M. Bauer, A. Al-Hezmi, K. Campowsky, and J. Song, "Standard-based IoT platforms interworking: Implementation, experiences, and lessons learned," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 48–54, Jul. 2016.

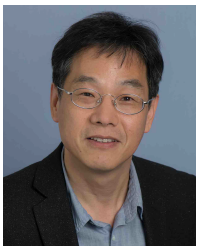
[3] J. Ding, M. Nemati, C. Ranaweera, and J. Choi, "IoT connectivity technologies and applications: A survey," *IEEE Access*, vol. 8, pp. 67646–67673, 2020.

[4] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: The rising stars in the IoT and smart city scenarios," *IEEE Wireless Commun.*, vol. 23, no. 5, pp. 60–67, Oct. 2016.

[5] R. Bonnefoi, L. Besson, C. Moy, E. Kaufmann, and J. Palicot, "Multi-armed bandit learning in IoT networks: Learning helps even in non-stationary settings," in *Cognit. Radio Oriented Wireless Netw.*, (Cham), pp. 173–185, Springer International Publishing, 2018.

[6] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2nd ed., 2018.

[7] A. Slivkins, "Introduction to multi-armed bandits," *Found. Trends Mach. Learn.*, vol. 12, nos. 1–2, pp. 1–286, 2019.
[8] L. Lai, H. Jiang, and H. V. Poor, "Medium access in cognitive radio networks: A competitive multi-armed bandit framework," in *Proc. 42nd Asilomar Conf. Signals, Syst. Comput.*, Oct. 2008, pp. 98–102.
[9] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Trans. Signal Process.*, vol. 58, no. 11, pp. 5667–5681, Nov. 2010.
[10] O. Avner and S. Mannor, "Multi-user lax communications: A multi-armed bandit approach," in *Proc. IEEE INFOCOM 35th Annu. IEEE Int. Conf. Comput. Commun.*, Apr. 2016, pp. 1–9.
[11] D. Gusfield and R. Irving, "The stable marriage problem: Structure and algorithms," *Foundations of Computing*. Cambridge, MA, USA: MIT Press, 1989.
[12] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.
[13] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 179–189, 2020.
[14] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, Dec. 2017.
[15] J. Choi, "Multichannel NOMA-ALOHA game with fading," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4997–5007, Oct. 2018.
[16] J. Choi, "On evolutionary game of dynamic devices in NOMA-based IoT networks," *IEEE Trans. Cognit. Commun. Netw.*, early access, Mar. 17, 2021, doi: 10.1109/TCCN.2021.3066191.
[17] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probability Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[18] S. Boyd and L. Vandenberghe, *Stochastic Processes*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[19] S. Ross, *Stochastic Processes*. New York, NY, USA: Wiley, 1996.
[20] B. Dae Choi and S. Hak Sung, "Almost sure convergence theorems of weighted sums of random variables," *Stochastic Anal. Appl.*, vol. 5, no. 4, pp. 365–377, Jan. 1987.
[21] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3–4, pp. 285–294, Dec. 1933.



JINHO CHOI (Senior Member, IEEE) was born in Seoul, South Korea. He received the B.E. degree *(magna cum laude)* in electronics engineering from Sogang University, Seoul, in 1989, and the M.S.E. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1991 and 1994, respectively. Prior to joining Deakin University in 2018, he was with Swansea University, U.K., as a Professor/the Chair in wireless, and the Gwangju Institute of Science and Technology (GIST), South Korea, as a Professor. He is currently working as a Professor with the School of Information Technology, Deakin University, Burwood, VIC, Australia. He authored two books published by Cambridge University Press in 2006 and 2010. His research interests include the Internet of Things (IoT), wireless communications, and statistical signal processing. He received a number of best paper awards, including the 1999 Best Paper Award for Signal Processing from EURASIP. He is on the list of World's Top 2% Scientists by Stanford University. He served as an Associate Editor or an Editor of other journals, including IEEE COMMUNICATIONS LETTERS, *Journal of Communications and Networks* JCN, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and *ETRI Journal*. He is currently an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE WIRELESS COMMUNICATIONS LETTERS and a Division Editor of *Journal of Communications and Networks* (JCN).