Incremental Massive Random Access Exploiting the Nested Reed-Muller Sequences

Jue Wang, *Student Member, IEEE,* Zhaoyang Zhang, *Member, IEEE,* Caijun Zhong, *Senior Member, IEEE,* Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Massive machine-type communication (mMTC) is expected to provide reliable and low-latency connectivity for an enormous number of devices, which turn active sporadically or frequently. In this highly dynamic situation, it is crucial to design efficient random access (RA) procedures to cope both with the flood of simultaneous access requests and with the potential access failures. In this paper, by exploiting the large sequence space, the excellent correlation property and especially the elegant nested structure of Reed-Muller (RM) sequences, we propose a new RA scheme, which facilitates both instantaneous access for newly active users and incremental access for existing users who suffer from detection failures. In particular, when a failure occurs, the user continues accessing the channel employing an expanded RM sequence, which is combined with the previously received ones at the access point (AP) to form a longer sequence so as to attain potentially better detection probability. Furthermore, a recursive detection algorithm is designed for jointly detecting the resultant RM sequences and the channel coefficients of both the newly active users and the existing ones. The performance of the proposed algorithm is analyzed in detail. Our simulation results validate the analysis and show the scheme's superior access probability, access latency and computational complexity.**

***Index Terms*—mMTC, massive random access, Reed-Muller sequences, multi-user detection**

I. Introduction

Supporting reliable and low-latency Massive Machine Type Communication (mMTC) has become one of the major design goals of next generation cellular systems [2]. The main challenge of mMTC is how to support a large number of machine-type devices which may become active sporadically or frequently, using rather limited radio resources [3].

In the state-of-the-art wireless systems which mainly aim for human communications, a grant-based random access (RA) protocol is used for providing reliable and instantaneous connectivity for multiple users [4], which identifies different users exploiting the relatively limited Zadoff-Chu (ZC) sequence space [5] and establishes each connection following a sophisticated procedure with four phases of message exchange. However, in the mMTC use cases, the number of devices to be connected may range from dozens or even hundreds of times higher than that of the person-to-person communication scenario, which calls for much larger sequence spaces and much better detection efficiency in the RA procedure. Moreover, the tele-traffic is often in the form of short packets in mMTC, and thus the signaling overhead introduced in the conventional four-phase RA procedure may no longer be negligible. Most importantly, such kind of procedure also cannot meet the stringent requirements of many typical mMTC applications which are rather delay-sensitive and/or energy-limited. Therefore, it is crucial to design new RA protocols for the mMTC scenarios to support the escalating number of users while achieving better efficiency, lower access delay and less signaling overhead.

Many alternative methods have been proposed for avoiding the performance degradation in mMTC scenarios. Some of them propose to dynamically disperse the access requests. For example, Extended Access Barring (EAB), which is considered by 3GPP as the most feasible baseline solution [6], allocates different barring factors to MTC devices depending on their delay-tolerance and grants delay-constrained devices priority in the case of congestion. An analytical model of the 3GPP EAB mechanism is presented in [7]. In Dynamic Access Barring (DAB), the access parameters, such as the number of preamble transmissions and the size of back-off windows, are dynamically adjusted according to the network state. Duan *et al.* [8] estimate the number of backlogged MTC devices with the aid of the real-time traffic information and then use this estimate to update the access factors. The prioritized random access with dynamic access barring (PRADA) scheme proposed by Lin *et al.* [9] combines DAB and RA resource separation to increase the access probability. These approaches can indeed reduce the collision rate to some degree, but the extra access delay cannot be neglected.

Other researches seek to improve the preamble design for performance enhancement. On the one hand, it is intuitive that extending the sequence space can help to reduce the collision rate. To this end, Prata *et al.* [10] present a code-expanded RA mechanism which assembles several RA slots into a virtual frame. Kim *et al.* [11] construct virtual preambles by combining the existing preambles with physical random access channel (PRACH) indices. Jang *et al.* [12] propose to realize space extension by partitioning a cell coverage into multiple groups and reducing the cyclic shift size of ZC. On the other hand, some authors attempt to optimize the preamble allocation strategy. The relationship between the cell size and

Part of this work was presented at ICC 2020 [1]. This work was supported in part by National Natural Science Foundation of China under Grant 61725104, U20A20158 and 61922071, and National Key R&D Program of China under Grant 2020YFB1807101 and 2018YFB1801104.

J. Wang (e-mail: juew@zju.edu.cn), Z. Zhang (Corresponding Author, e-mail: ning_ming@zju.edu.cn) and C. Zhong (e-mail: caijunzhong@zju.edu.cn) are with College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and with International Joint Innovation Center, Zhejiang University, Haining 314400, China, and also with Zhejiang Provincial Key Laboratory of Info. Proc., Commun. & Netw. (IPCAN), Hangzhou 310027, China. L. Hanzo (e-mail: lh@ecs.soton.ac.uk) is with the Department of Electronics and Computer Science, University of Southampton, UK.

the availability of orthogonal preambles is investigated by Kalalas and Alonso-Zarate [13], who propose a cell-planning and ZC root sequence allocation mechanism for minimizing the inter-cell interference. A partial preamble-sharing scheme was employed in [14] for avoiding cross-tier interference in two-tier small-cell networks. The configuration of ZC root sequences in small cell deployment was investigated by Ilori *et al.* [15] to provide increased RA opportunities.

The second-order Reed-Muller (RM) sequences have wide applications in compressed sensing [16]–[19]. For example, S. D. Howard *et al.* [16] construct the deterministic compressed sensing matrix exploiting RM sequences and conceive a fast chirp reconstruction algorithm. L. Zhang *et al.* [19] utilize on-off RM signatures to achieve full-duplex compressed neighbor discovery. Recently, RM sequences have also been shown to constitute promising candidates for massive access in mMTC as a benefit of their vast sequence space and the potential to simplify the RA process [20]. R. Calderbank *et al.* incorporate the algorithm of [16] into a patching and slotting framework to facilitate unsourced multiple access [21]. The elegant nested structure of RM sequences is investigated in [22] and based on that, a joint active user detection and channel estimation algorithm is proposed.

Considering the highly dynamic nature of mMTC, there exists potential access failures. In this case, the commonly used solution is to repeatedly retransmit the sequence, until reaching the maximum affordable number of transmissions [4]. However, continual retransmission in this way would aggravate the congestion, since the detection capability degrades with the sequences of the same length from the existing active users and the newly active ones continuously accumulating.

Against the above background, we design a novel incremental RA scheme exploiting the nested structure of RM sequences, which enables both instantaneous access for the newly active users and incremental access for the existing active users who suffer from access failures. Explicitly, when access failures occur, the existing active users expand their RM sequences following a specific expansion rule before sending them to the access point (AP). The AP then executes the proposed recursive detection algorithm to progressively detect the existing active users and the newly active ones. To summarize, our novel contributions are listed as follows:

- An RM expansion rule utilizing the nested structure and the correlation property of RM sequences is designed for the existing active users that suffer from access failures. By combining the expanded RM sequences with the previously transmitted ones, longer RM sequences are formed for progressively enhancing the access probability.
- At the AP, a recursive detection algorithm is proposed for jointly detecting the RM sequences and the channel coefficients. By exploiting the discrepancy in the sequence length, the proposed detection algorithm can separate different batches of active users and detect them progressively, and thus can be well adapted to the incoming of newly active users as the system smoothly operates.
- The theoretical successful detection probability of the proposed algorithm is analyzed mathematically. Our simulation results show that the proposed scheme fully exploits the extremely large RM sequence space to support efficient massive access, whilst guaranteeing superior access probability and access latency at a low computational complexity.

The rest of the paper is organized as follows. Section II describes the system model and reviews the RA procedure using RM sequences as well as the critical nested structure of RM sequences. The proposed incremental retransmission scheme is elaborated on in Section III. Section IV analyzes the theoretical access probability of the proposed algorithms. Our simulation results are provided in Section V, while Section VI concludes the paper.

Notations: Throughout this paper, scalars are represented in lowercase letters. Boldface lowercase and uppercase letters denote column vectors and matrices, respectively. Given a vector $\mathbf{y} \in \mathbb{C}^{l\times 1}$, its sub-vector is defined as $\mathbf{y}_{[a:b]} \triangleq [y_a, y_{a+1}, \ldots, y_b]^T,\ 1 \le a < b \le l$. The result of appending the column vector $\mathbf{y}_2 \in \mathbb{C}^{l_2\times 1}$ to another one $\mathbf{y}_1 \in \mathbb{C}^{l_1\times 1}$ is denoted as $\mathbf{y} = [\mathbf{y}_1; \mathbf{y}_2] \in \mathbb{C}^{(l_1+l_2)\times 1}$. The notations $(\mathbf{y})'$ and $(\mathbf{y})''$ represent the first and the second half of the sequence $\mathbf{y}$, respectively. The symbol $|\cdot|$ represents the absolute value of a scalar, while $\|\cdot\|$ is the Euclidean norm of a vector. The symbols $\odot$ and $\oplus$ represent element-wise product and modulo-2 addition, respectively. We use $(\cdot)^T$ to denote the transpose of a matrix or vector and $(\cdot)^*$ to represent the complex conjugate. The vector $\mathbf{a}_{j-1}^m$ is the m-bit binary representation of the decimal number $(j-1)$, while conversely we have that $j = \langle \mathbf{a}_{j-1}^m \rangle$, i.e., the operation $\langle\cdot\rangle$ converts the binary vector to the corresponding decimal number and then increases the result by 1. The representation $x \sim \mathcal{N}(\mu, \sigma^2)$ indicates that x is a Gaussian random variable (r.v.) with a mean of μ and variance of σ^2, and its probability density function (PDF) is

$$f_{\mathcal{N}}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

while $x \sim \mathcal{CN}(\mu, \sigma^2)$ denotes that the r.v. x follows circular symmetric complex Gaussian distribution with its PDF being

$$f_{\mathcal{CN}}(x;\mu,\sigma^2) = \frac{1}{\pi\sigma^2} \exp\left(-\frac{|x-\mu|^2}{\sigma^2}\right).$$

II. System Model

The massive next-generation uplink connectivity scenario is considered in this paper. Specifically, a single AP is serving a huge pool of N potential users and each one of them has a unique user identity (ID) $n, 0 \le n \le N-1$. All the users are assumed to be synchronized and they decide whether to access the system in a specific RA slot independently. We assume that the number of newly active users in the slot t, denoted as K_t, obeys the Possion distribution having the parameter λ. Combined with the restriction of the number of potential users, the probability distribution of K_t is specified as

$$P(K_t = k) = \begin{cases} \frac{\lambda^k e^{-\lambda}}{k!}, & 0 \le k \le N \\ 0, & k > N \end{cases}.$$

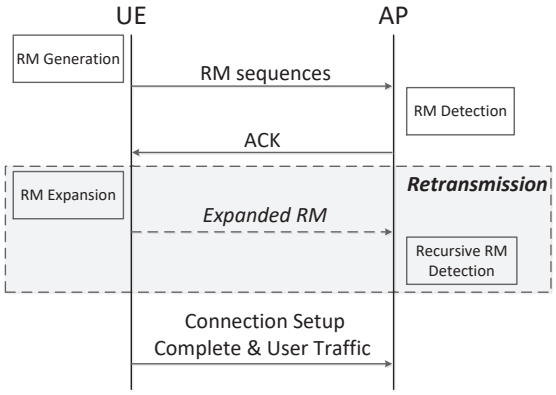


Fig. 1: *The illustration of the proposed incremental massive random access procedure using nested RM sequences.*

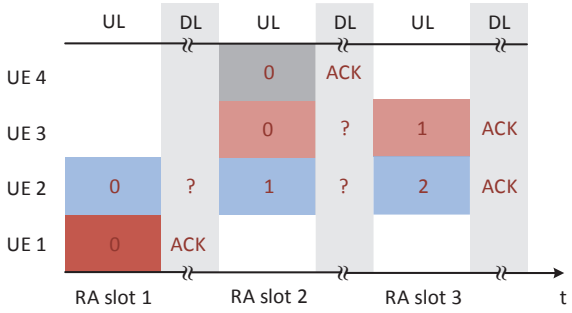


Fig. 2: *An example of the transmissions of RM sequences. The coloured blocks with the mark of "0" denote the RM sequences transmitted by the active users on their first access attempts, while those marked with "r", $1 \leq r \leq R_{max}$, represent the expanded RM sequences sent on the r-th retransmissions.*

The IDs of these active users are recorded in the set κ_t. Since the traffic pattern in the mMTC scenarios is typically sporadic, we have $\lambda \ll N$.

The proposed incremental RA procedure using RM sequences is illustrated in Fig. 1. To access the system, all the active users in the same slot send their RM sequences on the PRACH simutaneously. After receiving the signal, the AP first jointly detects the active users and estimates their channel coefficients, and then sends acknowledgements (ACKs) to the users detected. The successfully detected users complete their connection setup and proceed to transmit their data through the physical uplink shared channel (PUSCH), while those who fail to be detected embark on retransmissions. Explicitly, as long as the maximum number of retransmissions $R_{\max}$ is not reached, the active users who suffer from access failures continue making their access attempts with the expanded RM sequences instead of the original ones, like UE "2" and "3" in Fig. 2. The AP then performs the recursive RM detection algorithm for progressively detecting both the existing active users and the newly active ones.

For the sake of completeness, the processes of RM generation and detection, as well as the nested structure of RM sequences are briefly reviewed first.

A. RM Generation

To access the system, each active user first generate an RM sequence according to its ID. Since a length-2^m RM sequence is determined by an $(m \times m)$-element symmetric binary matrix and an $(m \times 1)$-element binary vector, the active user first maps its ID n to the matrix-vector pair $\left\{\mathbf{P}_n^m \in \mathbb{Z}_2^{m\times m}, \mathbf{b}_n^m \in \mathbb{Z}_2^{m\times 1}\right\}$. The *ID-to-pair mapping rule* adopted here is as follows [20]:

1) Convert the user ID n to the length-$2m$ binary array;
2) Take the least significant m bits to form the vector $\mathbf{b}_n^m$;
3) Map the remaining m bits to the matrix $\mathbf{P}_n^m$ in the Kerdock set [17].

Next, the user generates the RM sequence by substituting the mapped matrix-vector pair into the following function [16]:

$$c_{n,j}^m = i^{(2\mathbf{b}_n^m + \mathbf{P}_n^m \mathbf{a}_{j-1}^m)^T \mathbf{a}_{j-1}^m}, j = 1, \ldots, 2^m, \quad (1)$$

where i is the imaginary unit that satisfies $i^2 = -1$, and $\mathbf{a}_{j-1}^m \in \mathbb{Z}_2^{m\times 1}$ is the m-bit binary expression of $(j-1)$.

Several factors have to be taken into consideration when choosing the value of m. Firstly, following the above mapping rule, the size of the constructed sequence space equals to 2^{2m}, which has to be not less than the number of potential users N because of the one-to-one correspondence between the users' IDs and the RM sequences. Secondly, as analyzed in Section IV, the detection capability improves with greater m, but at the price of the increasing detection complexity. Hence, the value of m is also determined by the tradeoff between the access performance and the computational complexity.

B. The Nested Structure of Reed-Muller Sequences

For the ease of exposition, we omit the subscript of the matrix-vector pair in this subsection and then express it in the recursive form. Specifically, given an $(m \times m)$-element symmetric binary matrix $\mathbf{P}^m$, its lower right $(m-1) \times (m-1)$ sub-matrix is denoted as $\mathbf{P}^{m-1}$, while its first column $[p_{m,m}, p_{m,m-1}, \ldots, p_{m,1}]^T$ is decomposed into $[\rho_m, \boldsymbol{\alpha}^m]^T$, with $\rho_m \triangleq p_{m,m}$ and $\boldsymbol{\alpha}^m \triangleq [p_{m,m-1}, \ldots, p_{m,1}]^T$. As for the $(m\times 1)$-element binary vector $\mathbf{b}^m = [b_m, b_{m-1}, \ldots, b_1]^T$, we represent its low-order $(m-1)$ bits as the sub-vector $\mathbf{b}^{m-1}$. To summarize, the recursive representation of $\{\mathbf{P}^m, \mathbf{b}^m\}$ is given by

$$\mathbf{P}^m = \begin{bmatrix} \rho_m & (\boldsymbol{\alpha}^m)^T \\ \boldsymbol{\alpha}^m & \mathbf{P}^{m-1} \end{bmatrix}, \quad \mathbf{b}^m = \begin{bmatrix} b_m \\ \mathbf{b}^{m-1} \end{bmatrix}. \quad (2)$$

Let $\mathbf{c}^m$ and $\mathbf{c}^{m-1}$ denote the RM sequences generated by the pairs $\{\mathbf{P}^m, \mathbf{b}^m\}$ and $\left\{\mathbf{P}^{m-1}, \mathbf{b}^{m-1}\right\}$, respectively, and then the nested structure of RM sequences is expressed as [22]:

$$\mathbf{c}^m = \left[\mathbf{c}^{m-1}; \mathbf{c}^{m-1} \odot \mathbf{v}^{m-1}\right], \quad (3)$$

where the vector $\mathbf{v}^{m-1}$ is the length-2^{m-1} Walsh sequence parameterized by $\{\boldsymbol{\alpha}^m, \rho_m, b_m\}$, with its j-th entry being

$$v_j^{m-1} = i^{(\rho_m + 2b_m)} \cdot (-1)^{(\boldsymbol{\alpha}^m)^T \mathbf{a}_{j-1}^{m-1}}. \quad (4)$$

C. RM Detection

When there are only newly active users, the signal received in the RA slot t is expressed as

$$y_j^{(t)} = \sum_{n\in\kappa_t} h_n \cdot c_{n,j}^m + e_j^{(t)},\ j = 1, \ldots, 2^m. \tag{5}$$

In (5), $h_n \sim \mathcal{CN}(0,1)$ represents the Rayleigh fading channel coefficient between the active user n and the AP, and it is assumed to be constant during the access period of this user. The term $e_j^{(t)}$ is the complex additive white Gaussian noise (AWGN) obeying the distribution $\mathcal{CN}(0, N_0)$.

After receiving the signal $\mathbf{y}^{(t)}$, the AP performs the joint active user detection and channel estimation algorithm proposed in [22]. Although the algorithm of [22] was conceived for real RM sequences, it can be easily generalized to the complex-valued scenario. Furthermore, the matrices generated following the given ID-to-pair mapping rule belong to the Kerdock set and their structural properties can help to simplify the algorithm with no loss of performance, which will be elaborated on in Section III.B.

III. Incremental Retransmission Scheme based on the Nested RM Sequences

In this section, we elaborate on our proposed incremental massive RA scheme, including the RM expansion rule adopted at the transmitter and the recursive RM detection algorithm performed at the AP.

A. RM Expansion

As shown in (1), the RM sequence $\mathbf{c}^m$ is determined by the matrix-vector pair $\{\mathbf{P}^m, \mathbf{b}^m\}$. Hence, the expansion of $\mathbf{c}^m$ is essentially the expansion of $\{\mathbf{P}^m, \mathbf{b}^m\}$. Following the ID-to-pair mapping rule in Section II.A, $\mathbf{P}^m$ belongs to the Kerdock set, which has the following properties:

1) It is a special kind of Hankel matrix [23], whose elements on each ascending skew-diagonal are the same. This elegant structural property can be utilized to reduce the complexity of RM detection, which will be specified in Section III.B.
2) The inner product of two length-2^m RM sequences generated by different matrices $\mathbf{P}^m$ and $(\mathbf{P}^m)'$ satisfies [16]

$$|\chi^m| = \begin{cases} \sqrt{2^{2m-\mathcal{R}_m}}, & 2^{\mathcal{R}_m} \text{ times} \\ 0, & 2^m - 2^{\mathcal{R}_m} \text{ times} \end{cases}, \tag{6}$$

where $\mathcal{R}_m \triangleq \operatorname{rank}\left(\mathbf{P}^m - (\mathbf{P}^m)'\right)$. Since we have $\mathcal{R}_m = m$ for any two matrices in the Kerdock set, the magnitude of the inner product $|\chi^m|$ is ensured to be no more than $\sqrt{2^m}$ and its variance reaches its minimum, which can help to alleviate the multi-user interference and improve the detection capability as elaborated on in Section IV.A.

Considering the above benefits for RM detection, we expect to preserve these properties in the expansion. In the sequel, we first illustrate the expansion process with a toy example and then conclude the general rule. The matrix-vector pair to be expanded is given as

$$\mathbf{P}^3 = \begin{bmatrix} \rho_3 & p_{3,2} & p_{3,1} \\ p_{3,2} & \rho_2 & p_{2,1} \\ p_{3,1} & p_{2,1} & \rho_1 \end{bmatrix} = \begin{bmatrix} \cancel{\rho_3} & \cancel{p_{3,2}} & \cancel{\rho_2} \\ \cancel{p_{3,2}} & \cancel{\rho_2} & \cancel{p_{2,1}} \\ \cancel{\rho_2} & \cancel{p_{2,1}} & \cancel{\rho_1} \end{bmatrix},\quad \mathbf{b}^3 = \begin{bmatrix} b_3 \\ b_2 \\ b_1 \end{bmatrix}.$$

Firstly, following the structural character of the Hankel matrix, it is straightforward to have that $p_{4,2} = \rho_3$ and $p_{4,1} = p_{3,2}$. Next, to meet the rank requirement that $\mathcal{R}_4 = 4$ premised on $\mathcal{R}_3 = 3$, the expanded columns of two matrices have to be distinct. Considering that the first column $[\rho_3, p_{3,2}, p_{3,1}]^T$ varies from matrix to matrix, the most direct way is to let $p_{4,3} = p_{3,1} = \rho_2$. Thus, we arrive at

$$\left(\boldsymbol{\alpha}^4\right)^T = [p_{3,1}, \rho_3, p_{3,2}] = \mathbf{RCS}\left\{\left[\rho_3, \left(\boldsymbol{\alpha}^3\right)^T\right]\right\},$$

i.e., $\left(\boldsymbol{\alpha}^4\right)^T$ is the result of applying a right circular shift (**RCS**) to $\left[\rho_3, \left(\boldsymbol{\alpha}^3\right)^T\right]$ by one position. On this basis, the values of ρ_4 and b_4 can be arbitrary, and we can simply set $\rho_4 = \alpha_1^4 \oplus \alpha_2^4 \oplus \alpha_3^4$ and $b_4 = b_1 \oplus b_2 \oplus b_3$.

To summarize, when the active user n makes its r-th attempt to re-access the system, where $1 \le r \le R_{\max}$, the *expansion rule* it follows is specified as:

$$\begin{cases} \left(\boldsymbol{\alpha}_n^{m+r}\right)^T = \mathbf{RCS}\left\{\left[\rho_{n,m+r-1}, \left(\boldsymbol{\alpha}_n^{m+r-1}\right)^T\right]\right\} \\ \rho_{n,m+r} = \alpha_{n,1}^{m+r} \oplus \cdots \oplus \alpha_{n,m+r-1}^{m+r} \\ b_{n,m+r} = b_{n,1} \oplus \cdots \oplus b_{n,m+r-1} \end{cases}. \tag{7}$$

With the elements in (7), the user generates the latter half part of the expanded RM sequence according to the nested structure given in (3) and (4) as

$$\begin{aligned} \left(c_{n,j}^{m+r}\right)'' =& c_{n,j}^{m+r-1} \cdot v_{n,j}^{m+r-1} \\ =& c_{n,j}^{m+r-1} \cdot i^{(\rho_{n,m+r}+2b_{n,m+r})} \cdot (-1)^{\left(\boldsymbol{\alpha}_n^{m+r}\right)^T \mathbf{a}_{j-1}^{m+r-1}}, \end{aligned} \tag{8}$$

where $1 \le j \le 2^{m+r-1}$. After that, the sequence $\left(\mathbf{c}_n^{m+r}\right)''$ is transmitted to the AP.

Note that the maximum number of retransmissions $R_{\max}$ is restricted by several factors. First, the channel coefficient of one active user has to remain constant during its access period to enable the detection at the AP, and thus $R_{\max}$ is limited by the coherence time. Besides, as shown in (8), the length of the transmitted sequences doubles after each expansion, and hence $R_{\max}$ is also constrained by the size of the RA slot.

B. Recursive RM Detection

In the sequel, we describe our recursive RM detection algorithm in a slot-by-slot manner followed by a summary.

*1) **RA Slot 1**:* The received signal here is expressed as

$$y_j^{(1)} = \sum_{n\in\kappa_1} h_n \cdot c_{n,j}^m + e_j^{(1)},\ 1 \le j \le 2^m.$$

The algorithm proposed in [22] can be executed to detect these users after being generalized to the complex-valued scenario. Besides, it can be further simplified by exploiting the structural properties of Kerdock matrices. In the sequel, we describe the enhanced algorithm in the single user scenario and the received signal is rewritten as $y_j^m = h \cdot c_j^m + e_j^m$ for better clarification.

Firstly, the received signal is evenly divided into two sub-sequences $(\mathbf{y}^m)' \triangleq \mathbf{y}^{(1)}_{[1:2^{m-1}]}$ and $(\mathbf{y}^m)'' \triangleq \mathbf{y}^{(1)}_{[2^{m-1}+1:2^m]}$. According to the nested structure in (3), their elements are specified as:

$$\begin{cases} \left(y_j^m\right)' = hc_j^{m-1} + e_j^m \\ \left(y_j^m\right)'' = hc_j^{m-1} \cdot v_j^{m-1} + e^m_{j+2^{m-1}} \end{cases}, 1 \le j \le 2^{m-1}. \quad (9)$$

Then the element-wise conjugate multiplication is performed following

$$\tilde{y}_j^m = \left(y_j^m\right)' \cdot \left(\left(y_j^m\right)''\right)^* = |h|^2 \cdot \left(v_j^{m-1}\right)^* + \tilde{\varepsilon}_j^m, \quad (10)$$

where the equivalent noise $\tilde{\varepsilon}_j^m$ is defined as

$$\tilde{\varepsilon}_j^m \triangleq hc_j^{m-1}(e^m_{j+2^{m-1}})^* + e_j^m \left(hc_j^{m-1} v_j^{m-1}\right)^* + e_j^m (e^m_{j+2^{m-1}})^*. \quad (11)$$

As mentioned in (4), $\mathbf{v}^{m-1}$ is a Walsh sequence parameterized by $\{\boldsymbol{\alpha}^m, \rho_m, b_m\}$. Hence, the elements $\{\boldsymbol{\alpha}^m, \rho_m, b_m\}$ can be recovered by performing a Walsh-Hadamard Transformation (WHT) on $\tilde{\mathbf{y}}^m$. The result is denoted as $\mathbf{V}^{m-1} = \mathbf{WHT}\{\tilde{\mathbf{y}}^m\}$ with the w-th element being

$$\begin{aligned} V_w^{m-1} &= \sum_{j=1}^{2^{m-1}} \tilde{y}_j^m \cdot (-1)^{\left(\mathbf{a}_{w-1}^{m-1}\right)^T \mathbf{a}_{j-1}^{m-1}} \\ &= |h|^2 (-i)^{(\rho_m + 2b_m)} \sum_{j=1}^{2^{m-1}} (-1)^{\left(\boldsymbol{\alpha}^m + \mathbf{a}_{w-1}^{m-1}\right)^T \mathbf{a}_{j-1}^{m-1}} + \varepsilon_w^{m-1}, \end{aligned} \quad (12)$$

where $\varepsilon_w^{m-1} \triangleq \sum_{j=1}^{2^{m-1}} (-1)^{\left(\mathbf{a}_{w-1}^{m-1}\right)^T \mathbf{a}_{j-1}^{m-1}} \cdot \tilde{\varepsilon}_j^m$. It is revealed in (12) that when $\mathbf{a}_{w-1}^{m-1} = \boldsymbol{\alpha}^m$, i.e., $w = \langle \boldsymbol{\alpha}^m \rangle$, the magnitude of V_w^{m-1} reaches its maximum, while its sign and whether it is real- or complex-valued depend on $\{\rho_m, b_m\}$. Based on this observation, we construct the vector

$$\boldsymbol{\Lambda}^{m-1} = \left[V_{1,\Re}^{m-1}, \ldots, V_{2^{m-1},\Re}^{m-1}, V_{1,\Im}^{m-1}, \ldots, V_{2^{m-1},\Im}^{m-1}\right]^T, \quad (13)$$

where the subscripts "$\Re$" and "$\Im$" represent the real and imaginary component, respectively. Next, we search through $\boldsymbol{\Lambda}^{m-1}$ for the element with the maximum magnitude and denote its index as $w_{\max}^m$, namely,

$$w_{\max}^m = \arg\max_w \left\{\left|\Lambda_w^{m-1}\right|\right\}, w = 1, \ldots, 2^m. \quad (14)$$

Then we have the estimates $[\hat{\rho}_m; \hat{\boldsymbol{\alpha}}^m] = \mathbf{a}^m_{w^m_{\max}-1}$, while $\hat{b}_m$ is determined by checking the sign of $\Lambda^{m-1}_{w^m_{\max}}$. By substituting $\{\hat{\boldsymbol{\alpha}}^m, \hat{\rho}_m, \hat{b}_m\}$ into (4), the sequence $\mathbf{v}^{m-1}$ is reconstructed.

Combining the sub-sequences $(\mathbf{y}^m)'$ and $(\mathbf{y}^m)''$ in (9) with the estimate $\hat{\mathbf{v}}^{m-1}$, we obtain the length-2^{m-1} signal as

$$\begin{aligned} y_j^{m-1} =& \frac{1}{2}\left[\left(y_j^m\right)' + \left(\hat{v}_j^{m-1}\right)^* \cdot \left(y_j^m\right)''\right] \\ =& hc_j^{m-1} + e_j^{m-1}, j = 1, \ldots, 2^{m-1}, \end{aligned} \quad (15)$$

where $e_j^{m-1} \triangleq \frac{1}{2}\left(e_j^m + \left(\hat{v}_j^{m-1}\right)^* \cdot e^m_{j+2^{m-1}}\right)$. Based on (15), the elements $\left\{\boldsymbol{\alpha}^{m-1}, \rho_{m-1}, b_{m-1}\right\}$ are recovered by repeating the above steps, except that the search range of (14) can be narrowed. Specifically, since $\mathbf{P}^m$ is a Kerdock matrix, whose elements on each ascending skew-diagonal are the same, we have $\left[\rho_{m-1}; \boldsymbol{\alpha}^{m-1}_{[1:m-3]}\right] = \boldsymbol{\alpha}^m_{[2:m-1]}$. With this prior knowledge, we only have to compare the two elements at the locations $\langle[\hat{\boldsymbol{\alpha}}^m_{[2:m-1]}; 0]\rangle$ and $\langle[\hat{\boldsymbol{\alpha}}^m_{[2:m-1]}; 1]\rangle$ for determining the index $w_{\max}^{m-1}$, thus dramatically reducing the search complexity.

The process above continues until we obtain the estimates $\{\hat{\boldsymbol{\alpha}}^s, \hat{\rho}_s, \hat{b}_s\}, 2 \le s \le m$ and the sequence $\mathbf{y}^1 = [y_1^1, y_2^1]^T$. The last step is to recover the elements $\{\rho_1, b_1\}$ and to estimate the channel coefficient h. Since the elements of $\mathbf{y}^1$ are specified as

$$\begin{cases} y_1^1 = hc_1^1 + e_1^1 = h + e_1^1 \\ y_2^1 = hc_2^1 + e_2^1 = h \cdot (i)^{\rho_1 + 2b_1} + e_2^1 \end{cases}, \quad (16)$$

we compute that

$$\boldsymbol{\Lambda}^0 = [y_1^1, y_2^1] \cdot \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & -i & i \end{bmatrix}.$$

With $w_{\max}^1 = \arg\max_w \left\{|\Lambda_w^0|\right\}$, we have $\left[\hat{\rho}_1; \hat{b}_1\right] = \mathbf{a}^2_{w^1_{\max}-1}$ and the channel coefficient is estimated following

$$\hat{h} = \frac{1}{2}\left(y_1^1 + (-i)^{\hat{\rho}_1 + 2\hat{b}_1} \cdot y_2^1\right).$$

Finally, combining all the estimates $\{\hat{\boldsymbol{\alpha}}^s, \hat{\rho}_s, \hat{b}_s\}$, $1 \le s \le m$, the matrix-vector pair $\{\hat{\mathbf{P}}^m, \hat{\mathbf{b}}^m\}$ as well as the RM sequence $\hat{\mathbf{c}}^m$ can be reconstructed.

When $K_1 > 1$, multiple sequences and channel coefficients are detected in an iterative manner and the details can be found in [22], which is termed as **Alg. $\mathbf{\Phi}^{(1)}$** in the rest of the paper.

*2) **RA Slot 2**:* After receiving ACKs, the successfully detected users would reply with the *Connection Setup Complete* messages to the AP. The AP records their IDs in the set ς_1 and calculates the residual signal according to

$$y^{(1)}_{\text{res},j} = y_j^{(1)} - \sum_{n \in \varsigma_1} \hat{h}_n \cdot \hat{c}^m_{n,j}.$$

Meanwhile, after receiving the signal $\mathbf{y}^{(2)}$, the AP has to find out whether there are any sequences either from the existing active users with access failures, i.e., $n \in \kappa_1 \setminus \varsigma_1$, or from the newly active users $n \in \kappa_2$, so that it can choose a suitable way of accomplishing the detection task. This is achieved by comparing the energy of $\mathbf{y}^{(1)}_{\text{res}}$ and $\mathbf{y}^{(2)}$ to the noise threshold γ that used to differentiate between effective signals and thermal noise. In the sequel, we discuss every possible case in detail, respectively.

a) $\left\|\mathbf{y}^{(1)}_{\text{res}}\right\|^2 \le \gamma$ *and* $\left\|\mathbf{y}^{(2)} - \mathbf{y}^{(1)}_{\text{res}}\right\|^2 \le \gamma$: In this case, no users transmit useful signals and the AP just waits to receive the signal in the next slot.

b) $\left\|\mathbf{y}^{(1)}_{\text{res}}\right\|^2 \le \gamma$ *and* $\left\|\mathbf{y}^{(2)} - \mathbf{y}^{(1)}_{\text{res}}\right\|^2 > \gamma$: It suggests that there only exist newly active users, and the AP can directly perform Alg. $\mathbf{\Phi}^{(1)}$ on the received signal $\mathbf{y}^{(2)}$ to detect these users.

c) $\left\|\mathbf{y}^{(1)}_{\text{res}}\right\|^2 > \gamma$ *and* $\left\|\mathbf{y}^{(2)} - \mathbf{y}^{(1)}_{\text{res}}\right\|^2 \le \gamma$: It indicates that no newly active users are requesting access, while the existing ones who suffer from access failures in RA slot 1 have sent their expanded RM sequences. Assuming that the channel

estimates of the successfully detected users are perfect, the residual signal $\mathbf{y}_{\text{res}}^{(1)}$ and the newly arrived signal $\mathbf{y}^{(2)}$ are specified as

$$\begin{aligned} y_{\text{res},j}^{(1)} &= \sum_{n\in\kappa_1\setminus\varsigma_1} h_n \cdot c_{n,j}^m + e_j^{(1)}, \\ y_j^{(2)} &= \sum_{n\in\kappa_1\setminus\varsigma_1} h_n \cdot c_{n,j}^m v_{n,j}^m + e_j^{(2)}. \end{aligned} \tag{17}$$

Comparing (17) with (3), we find that by stitching these two signals together, the outcome $\left[\mathbf{y}_{\text{res}}^{(1)};\mathbf{y}^{(2)}\right]$ is the same as the result of overlapping the complete RM sequences $\mathbf{c}_n^{m+1}$. Thus, the users $n \in \kappa_1 \setminus \varsigma_1$ can be detected by applying Alg. $\mathbf{\Phi}^{(1)}$ to the signal $\left[\mathbf{y}_{\text{res}}^{(1)};\mathbf{y}^{(2)}\right]$.

d) $\left\|\mathbf{y}_{\text{res}}^{(1)}\right\|^2 > \gamma$ *and* $\left\|\mathbf{y}^{(2)} - \mathbf{y}_{\text{res}}^{(1)}\right\|^2 > \gamma$: In this case, the signal $\mathbf{y}^{(2)}$ contains the sequences from both the existing active users and the newly active ones. For the ease of exposition, we describe the detection process with a toy example, while the algorithmic details of the general case are given in **Algorithm 2** in the Appendix.

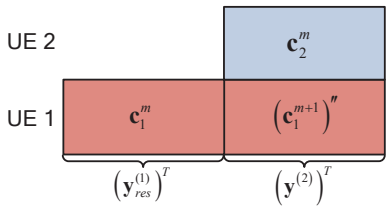


Fig. 3: *The composition of the concatenated signal* $\left[\mathbf{y}_{res}^{(1)};\mathbf{y}^{(2)}\right]^T$.

Assuming that a single existing active user "1" and a single newly active user "2" coexist, the AP detects these two active users iteratively. Specifically, as depicted in Fig. 3, $\mathbf{y}^{(2)}$ is first appended to $\mathbf{y}_{\text{res}}^{(1)}$ to form the signal $\mathbf{y}^{m+1} = \left[\mathbf{y}_{\text{res}}^{(1)};\mathbf{y}^{(2)}\right]$, and the estimates of these two active users' RM sequences are initialized to be all zeros. Before detecting UE "1", the signal of UE "2" is seen as the interference and removed, with the resultant residual signal expressed as

$$\begin{cases} \left(y_{\text{res},j}^{m+1}\right)' = y_{\text{res},j}^{(1)} \\ \left(y_{\text{res},j}^{m+1}\right)'' = y_j^{(2)} - \hat{h}_2 \cdot \hat{c}_{2,j}^m \end{cases}, \ 1 \le j \le 2^m.$$

Next, Alg. $\mathbf{\Phi}^{(1)}$ is executed on $\mathbf{y}_{\text{res}}^{m+1}$ to recover the expanded matrix-vector pair and the channel coefficient of UE "1", namely, $\left\{\hat{\mathbf{P}}_1^{m+1}, \hat{\mathbf{b}}_1^{m+1}, \hat{h}_1\right\}$, and then the expanded RM sequence $\hat{\mathbf{c}}_1^{m+1}$ can be reconstructed. Similarly, we have to eliminate the disturbance from UE "1" for detecting UE "2". To this end, the residual signal is updated following

$$y_{\text{res},j}^{m+1} = y_j^{m+1} - \hat{h}_1 \cdot \hat{c}_{1,j}^{m+1}$$

before its latter half part $\left(\mathbf{y}_{\text{res}}^{m+1}\right)''$ is input to Alg. $\mathbf{\Phi}^{(1)}$ for recovering the triplet $\left\{\hat{\mathbf{P}}_2^m, \hat{\mathbf{b}}_2^m, \hat{h}_2\right\}$. This process is repeated until the detection results remain unchanged or the maximum number of iterations is reached. This detection procedure is denoted as **Alg.** $\mathbf{\Phi}^{(2)}$ in the rest of the paper.

3) ***RA Slot 3****:* The AP first adds the IDs of the successfully detected users into the set ς_1 or ς_2 depending on the slot they turned active in, and then removes their signals from $\mathbf{y}^{m+1}$ to obtain the residual signal $\mathbf{y}_{\text{res}}^{(2)}$. The newly arrived signal in RA slot 3 is denoted as $\mathbf{y}^{(3)}$.

If $\left\|\left(\mathbf{y}_{\text{res}}^{(2)}\right)'\right\|^2 \le \gamma$, all the users that turned active in RA slot 1 have been detected successfully. In this case, the signal $\left(\mathbf{y}_{\text{res}}^{(2)}\right)''$ only contains the RM sequences of the users $n \in \kappa_2 \setminus \varsigma_2$, while the expanded RM sequences transmitted by these users are overlapped with the RM sequences of the newly active users $n \in \kappa_3$ in the signal $\mathbf{y}^{(3)}$. Thus, the composite signal $\left[\left(\mathbf{y}_{\text{res}}^{(2)}\right)'';\mathbf{y}^{(3)}\right]$ exhibits the same structure as the signal depicted in Fig. 3 and Alg. $\mathbf{\Phi}^{(2)}$ can be executed to detect the users $n \in (\kappa_2 \setminus \varsigma_2) \cup \kappa_3$.

In the other case of $\left\|\left(\mathbf{y}_{\text{res}}^{(2)}\right)'\right\|^2 > \gamma$, three batches of active users coexist in the system. As usual, we elaborate on the detection algorithm $\mathbf{\Phi}^{(3)}$ assuming that $\kappa_1 \setminus \varsigma_1 = \{1\}$, $\kappa_2 \setminus \varsigma_2 = \{2\}$ and $\kappa_3 = \{3\}$ here, with its generalized version summarized in **Algorithm 1** in the Appendix.

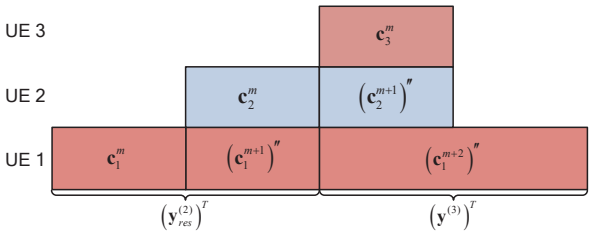


Fig. 4: *The composition of the concatenated signal* $\left[\mathbf{y}_{res}^{(2)};\mathbf{y}^{(3)}\right]^T$.

As shown in Fig. 4, there is only the signal sent by UE "1" in the latter half part of the signal $\mathbf{y}^{(3)}$, whose elements are given by

$$\left(y_j^{(3)}\right)'' = h_1 c_{1,j}^m v_{1,j}^m v_{1,j+2^m}^{m+1} + \left(e_j^{(3)}\right)'', \ 1 \le j \le 2^m. \tag{18}$$

After substituting

$$\begin{cases} c_{1,j}^m = i^{(2\mathbf{b}_1^m + \mathbf{P}_1^m \mathbf{a}_{j-1}^m)^T \mathbf{a}_{j-1}^m} \\ v_{1,j}^m = i^{(\rho_{1,m+1} + 2b_{1,m+1})} (-1)^{(\boldsymbol{\alpha}_1^{m+1})^T \mathbf{a}_{j-1}^m} \\ v_{1,j+2^m}^{m+1} = i^{(\rho_{1,m+2} + 2b_{1,m+2})} (-1)^{\alpha_{1,1}^{m+2} + (\boldsymbol{\alpha}_{1,[2:m+1]}^{m+2})^T \mathbf{a}_{j-1}^m} \end{cases}$$

into (18), we have that $\left(y_j^{(3)}\right)'' = \tilde{h}_1 \tilde{c}_{1,j}^m + \left(e_j^{(3)}\right)''$, where the biased channel coefficient is

$$\tilde{h}_1 = (-1)^{\left(b_{1,m+1} + b_{1,m+2} + \alpha_{1,1}^{m+2}\right)} \cdot i^{(\rho_{1,m+1} + \rho_{1,m+2})} \cdot h_1, \tag{19}$$

and the RM sequence $\tilde{\mathbf{c}}_1^m$ is generated by the matrix $\mathbf{P}_1^m$ and the biased vector specified as

$$\tilde{\mathbf{b}}_1^m = \mathbf{b}_1^m \oplus \boldsymbol{\alpha}_1^{m+1} \oplus \boldsymbol{\alpha}_{1,[2:m+1]}^{m+2}. \tag{20}$$

Therefore, the biased estimates $\left\{\hat{\mathbf{P}}_1^m, \hat{\tilde{\mathbf{b}}}_1^m, \hat{\tilde{h}}_1\right\}$ can be obtained by applying Alg. $\mathbf{\Phi}^{(1)}$ to the sequence $\left(\mathbf{y}^{(3)}\right)''$. Then, the elements $\{\boldsymbol{\alpha}_1^{m+1}, \boldsymbol{\alpha}_1^{m+2}, \rho_{1,m+1}, \rho_{1,m+2}, b_{1,m+1}, b_{1,m+2}\}$ are recovered by expanding $\hat{\mathbf{P}}_1^m$ and used for correcting the biased estimates according to (19) and (20).

Given the corrected matrix-vector pair and the channel coefficient, the signal $\hat{h}_1 \cdot \hat{c}_{1,j}^{m+2}, 1 \le j \le 2^{m+2}$, is reconstructed and subtracted from the signal $\left[\mathbf{y}_{\text{res}}^{(2)}; \mathbf{y}^{(3)}\right]$. The resultant residual signal is depicted in Fig. 5, which turns out to have the same structure as the signal in Fig. 3, and thus UE "2" and "3" can be detected by executing Alg. $\mathbf{\Phi}^{(2)}$.

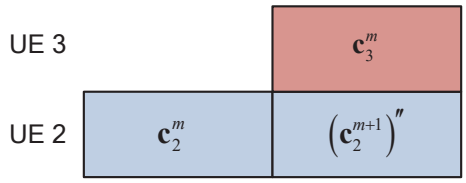


Fig. 5: *The structure of the residual signal obtained by subtracting the signal of UE "1" from* $\left[\mathbf{y}_{res}^{(2)}; \mathbf{y}^{(3)}\right]$.

Naturally, the detection processes in the subsequent RA slots follow this recursive pattern as well. To conclude, the AP first appends the newly arrived signal to the residual signal. Then, it separates different batches of active users exploiting the discrepancy in their sequence length and detects all of them progressively. Again, the detailed algorithm is given in the Appendix.

IV. Performance Analysis

A. Detection Capability of Algorithm $\mathbf{\Phi}^{(1)}$

1) Single User: When executing Alg. $\mathbf{\Phi}^{(1)}$ for detecting a single RM sequence $\mathbf{c}^m$, the matrix-vector pair is recovered layer by layer. Hence, we first discuss the correct recovery probability of each layer $\{\boldsymbol{\alpha}^s, \rho_s, b_s\}, 2 \le s \le m$, which is summarized in the following theorem.

Theorem 1: Assume that the input signal of Alg. $\mathbf{\Phi}^{(1)}$ is

$$y_j^m = h \cdot c_j^m + e_j^m,\ 1 \le j \le 2^m,$$

where the channel coefficient $h \sim \mathcal{CN}(0,1)$ and the noise $e_j^m \sim \mathcal{CN}(0, N_0)$. The notation $A^s, 2 \le s \le m$, represents the event that $\{\boldsymbol{\alpha}^s, \rho_s, b_s\}$ are recovered successfully, while $A^{m,\ldots,s+1}$ denotes the event that the estimates $\{\hat{\boldsymbol{\alpha}}^m, \hat{\rho}_m, \hat{b}_m\}, \ldots, \{\hat{\boldsymbol{\alpha}}^{s+1}, \hat{\rho}_{s+1}, \hat{b}_{s+1}\}$ are perfect. Then, the probability of the event A^s premised on $A^{m,\ldots,s+1}$ and the given h can be approximated as

$$\begin{aligned}&\Pr\left(A^s | A^{m,\ldots,s+1}, h\right)\\ &= \int_0^{+\infty} \left[g\left(x; 2^{s-1}|h|^2, \sigma_{V^{s-1}}^2\right)\right]\left[G\left(x; 0, \sigma_{V^{s-1}}^2\right)\right]^{\varrho} dx.\end{aligned} \tag{21}$$

In (21), the function $g\left(x; \mu, \sigma^2\right)$ and $G\left(x; \mu, \sigma^2\right)$ are respectively defined as

$$\begin{aligned}g\left(x; \mu, \sigma^2\right) &\triangleq f_{\mathcal{N}}\left(x; \mu, \sigma^2\right) + f_{\mathcal{N}}\left(x; -\mu, \sigma^2\right),\\ G\left(x; \mu, \sigma^2\right) &\triangleq \frac{1}{2}\left[\text{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right) + \text{erf}\left(\frac{x+\mu}{\sqrt{2\sigma^2}}\right)\right],\end{aligned}$$

where $\text{erf}(\cdot)$ is the error function. The variance $\sigma_{V^{s-1}}^2$ and the superscript ϱ are given by

$$\sigma_{V^{s-1}}^2 = 2^{s-2}\left[2|h|^2 \frac{N_0}{2^{m-s}} + \left(\frac{N_0}{2^{m-s}}\right)^2\right],$$

$$\varrho = \begin{cases} 2^m - 1, & s = m \\ 1, & 2 \le s \le m-1 \end{cases}.$$

Proof: First, to recover $\{\boldsymbol{\alpha}^m, \rho_m, b_m\}$ correctly, which is denoted as the event A^m, the magnitude of the element $\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}$ has to be the largest among the vector $\mathbf{\Lambda}^{m-1}$, i.e., the event A^m occurs with the probability of

$$\begin{aligned}\Pr(A^m) &= \Pr\left(\left|\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}\right| > \left|\Lambda_w^{m-1}\right|, \forall w \ne \langle[\rho_m;\boldsymbol{\alpha}^m]\rangle\right)\\ &= \Pr\left(\left|\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}\right| > \max\left\{\left|\Lambda_w^{m-1}\right|, \forall w \ne \langle[\rho_m;\boldsymbol{\alpha}^m]\rangle\right\}\right).\end{aligned} \tag{22}$$

Thus, the key of deriving the explicit expression of $\Pr(A^m)$ is to examine the distribution of $\left|\Lambda_w^{m-1}\right|$. To this end, we start to track the distribution and the evolution process of the input signal following the steps of Alg.$\mathbf{\Phi}^{(1)}$.

After dividing the input signal evenly to two subsequences, the result of their element-wise conjugate multiplication is shown in (10). Since e_j^m are independent identically distributed (i.i.d.) r.v. obeying $\mathcal{CN}(0, N_0)$, given the channel coefficient h, the equivalent noise $\tilde{\varepsilon}_j^m$ in (11) can be approximated as an r.v. following $\mathcal{CN}\left(0, 2|h|^2 N_0 + N_0^2\right)$. On this basis, the term ε_w^{m-1} in (12) obeys the distribution of $\mathcal{CN}(0, 2\sigma_{V^{m-1}}^2)$ with $\sigma_{V^{m-1}}^2 = 2^{m-2}\left(2|h|^2 N_0 + N_0^2\right)$. Thus, Λ_w^{m-1} can be seen as a Gaussian r.v. with the variance of $\sigma_{V^{m-1}}^2$, while its mean depends on the location w. Explicitly,

a) When $w = \langle[\rho_m;\boldsymbol{\alpha}^m]\rangle$: According to (12), the mean of $\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}$ equals to $(-1)^{b_m} \cdot 2^{m-1}|h|^2$. Under the assumption that b_m equals to 0 or 1 with equal probability, the PDF of $\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}$ has the form of

$$\begin{aligned}f_{\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}}(x) = \frac{1}{2}\Big[&f_{\mathcal{N}}\left(x; 2^{m-1}|h|^2, \sigma_{V^{m-1}}^2\right) +\\ &f_{\mathcal{N}}\left(x; -2^{m-1}|h|^2, \sigma_{V^{m-1}}^2\right)\Big],\end{aligned}$$

and the PDF of the magnitude of $\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}$ is given by

$$\begin{aligned}&f_{\left|\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}\right|}(x)\\ =&f_{\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}}(x) + f_{\Lambda_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}^{m-1}}(-x)\\ =&f_{\mathcal{N}}\left(x; 2^{m-1}|h|^2, \sigma_{V^{m-1}}^2\right) + f_{\mathcal{N}}\left(x; -2^{m-1}|h|^2, \sigma_{V^{m-1}}^2\right)\\ =&g\left(x; 2^{m-1}|h|^2, \sigma_{V^{m-1}}^2\right).\end{aligned} \tag{23}$$

b) When $w \ne \langle[\rho_m;\boldsymbol{\alpha}^m]\rangle$: Since any two different Walsh sequences are orthogonal, the result of (12) turns out to be $V_w^{m-1} = \varepsilon_w^{m-1}$ in this case. Thus, the cumulative distribution function (CDF) of $\left|\Lambda_w^{m-1}\right|$ is formulated as

$$F_{\left|\Lambda_w^{m-1}\right|}(x) = \text{erf}\Big(\frac{x}{\sqrt{2\sigma_{V^{m-1}}^2}}\Big) = G\left(x; 0, \sigma_{V^{m-1}}^2\right).$$

Assuming that $\left|\Lambda_w^{m-1}\right|, \forall w \neq \langle[\rho_m; \boldsymbol{\alpha}^m]\rangle$, are independent, the CDF of the maximum value among them has the form of

$$F_{\max\{|\Lambda_w^{m-1}|\}}(x) = \left[G\left(x; 0, \sigma^2_{V^{m-1}}\right)\right]^{2^m-1}. \tag{24}$$

Substituting (23) and (24) into (22), we arrive at the conclusion that

$$\begin{aligned}\Pr\left(A^m|h\right) &= \int_0^{+\infty} f_{\left|\Lambda^{m-1}_{\langle[\rho_m;\boldsymbol{\alpha}^m]\rangle}\right|}(x)\, F_{\max\{|\Lambda_w^{m-1}|\}}(x)dx \\ &= \int_0^{+\infty} \left[g\left(x; 2^{m-1}|h|^2, \sigma^2_{V^{m-1}}\right)\right]\left[G\left(x; 0, \sigma^2_{V^{m-1}}\right)\right]^{2^m-1}dx,\end{aligned} \tag{25}$$

which is in accordance with (21). Since $\hat{b}_m$ is determined by simply checking the sign, we assume that it is always true.

Based on the correct triplet $\{\hat{\boldsymbol{\alpha}}^m, \hat{\rho}_m, \hat{b}_m\}$, the sequence $\mathbf{y}^{m-1}$ is calculated following (15). Given the knowledge that $e_j^m \sim \mathcal{CN}(0, N_0)$, it is easy to derive that $e_j^{m-1} \sim \mathcal{CN}\big(0, \frac{N_0}{2}\big)$. The following analysis is similar to the process of deriving (25), except that we only have to compare $\left|\Lambda^{m-2}_{\langle[\hat{\boldsymbol{\alpha}}^m_{[2:m-1]},0]\rangle}\right|$ with $\left|\Lambda^{m-2}_{\langle[\hat{\boldsymbol{\alpha}}^m_{[2:m-1]},1]\rangle}\right|$ when recovering $\{\hat{\boldsymbol{\alpha}}^{m-1}, \hat{\rho}_{m-1}, \hat{b}_{m-1}\}$ and thus $\varrho = 1$. The derivations when $s = m-2, \ldots, 2$ are similar and will not be repeated here. ■

The RM sequence is reconstructed successfully iff all the estimates $\{\hat{\boldsymbol{\alpha}}^s, \hat{\rho}_s, \hat{b}_s\}, 1 \leq s \leq m$ are correct. Considering that the recovery of $\{\rho_1, b_1\}$ is straightforward, we can readily assume that it is always correct. On this basis, given the channel parameter h, the successful detection probability of Alg. $\boldsymbol{\Phi}^{(1)}$ in the single sequence scenario is formulated as

$$\bar{P}^{(1)}\left(m, N_0|h\right) = \prod_{s=2}^{m} \Pr\left(A^s|A^{m,\ldots,s+1}, h\right). \tag{26}$$

Taking the channel coefficient h into consideration, we have

$$\bar{P}^{(1)}\left(m, N_0\right) = \int_h f_{\mathcal{CN}}\left(h; 0, 1\right) \prod_{s=2}^{m} \Pr\left(A^s|A^{m,\ldots,s+1}, h\right) dh.$$

Remark. It is revealed in **Theorem 1** that the probability $\Pr\left(A^s|A^{m,\ldots,s+1}, h\right)$ is associated with three factors, namely s, m and N_0 (SNR). Although the explicit expressions are difficult to derive, it is easy to draw the trend of $\Pr\left(A^s|A^{m,\ldots,s+1}, h\right)$ when one of the three factors changes and the other two are fixed. Based on that, we have the following observations.

- Given m and SNR, the probability $\Pr\left(A^s|A^{m,\ldots,s+1}, h\right)$ keeps growing as s changes from m to 2, which means that if the first few layers are obtained with no error, the remaining layers can be recovered successfully with a high probability. Thus, it is important to improve the correct recovery probability of the first few layers, which can be realized by adopting the list detection algorithm in [22] and increasing the number of lists in the first few layers.
- When s and N_0 are fixed, $\Pr\left(A^s|A^{m,\ldots,s+1}, h\right)$ grows with m, i.e., the detection capability of Alg. $\boldsymbol{\Phi}^{(1)}$ can be enhanced by increasing the length of RM sequences. This observation verifies that it is feasible to improve the access probability by expanding the RM sequences in the case of access failures, which constitutes the fundamental spirit of our proposed incremental RA scheme.
- $\Pr\left(A^s|A^{m,\ldots,s+1}, h\right)$ improves as the level of SNR increases, which is in line with intuition.

2) Multiple Users: In the multi-user case, the detection capability of Alg. $\boldsymbol{\Phi}^{(1)}$ is concluded in the following lemma and theorem.

Lemma 1: Assume that the input signal of Alg. $\boldsymbol{\Phi}^{(1)}$ is the superposition of K active users' length-2^m RM sequences, and they are detected following the order $(n_1, n_2, \ldots, n_K)$. Let B_{n_k} indicate the event that the user n_k is successfully detected, while $B_{n_1,\ldots,n_{k-1}}$ denotes the event that the users $n_1, \ldots, n_{k-1}$ are all detected and their signals are removed perfectly. Then, when the channel coefficients $\mathbf{h} = [h_1, \ldots, h_K]^T$ are given, the probability that the event B_{n_k} occurs premised on the event $B_{n_1,\ldots,n_{k-1}}$ is approximated as

$$\begin{aligned}&\Pr\left(B_{n_k}|B_{n_1,\ldots,n_{k-1}}, \mathbf{h}\right) \\ =&\Pr\left(A^m_{n_k}|\mathbf{h}\right) \cdot \bar{P}^{(1)}\left(m-1, N_0^{m-1}|\mathbf{h}\right).\end{aligned} \tag{27}$$

In (27), $\Pr\left(A^m_{n_k}|\mathbf{h}\right)$ denotes the successful recovery probability of the elements $\{\boldsymbol{\alpha}^m_{n_k}, \rho_{m,n_k}, b_{m,n_k}\}$ under the given $\mathbf{h}$, and it is specified as

$$\begin{aligned}\Pr\left(A^m_{n_k}|\mathbf{h}\right) = \int_0^{+\infty} \Bigg\{ & g\left(x; 2^{m-1}|h_{n_k}|^2, \sigma^2_{V^{m-1}}\right) \\ & \cdot \left[G\left(x; 0, \sigma^2_{V^{m-1}}\right)\right]^{2^m-(K-k+1)} \\ & \cdot \Big[\prod_{n=n_{k+1}}^{n_K} G\left(x; 2^{m-1}|h_n|^2, \sigma^2_{V^{m-1}}\right)\Big]\Bigg\} dx,\end{aligned} \tag{28}$$

where the variance $\sigma^2_{V^{m-1}} = \sum_{n=n_k}^{n_K} \sum_{\substack{n'=n_k \\ n'\neq n}}^{n_K} |h_n^* h_{n'}|^2 \sigma^2_{\chi^{m-1}} + 2^{m-2}\Big[2\sum_{n=n_k}^{n_K}|h_n|^2 N_0 + N_0^2\Big]$, and $\sigma^2_{\chi^{m-1}}$ is the variance of the inner product of two length-2^{m-1} RM sequences. The second term $\bar{P}^{(1)}(m-1, N_0^{m-1}|\mathbf{h})$ in (27) represents the probability that the matrix-vector pair $\left\{\mathbf{P}^{m-1}_{n_k}, \mathbf{b}^{m-1}_{n_k}\right\}$ is recovered correctly with the given $\mathbf{h}$, where the function $\bar{P}^{(1)}(\cdot)$ is specified in (26) and the parameter $N_0^{m-1} = \frac{1}{2}\big(\sum_{n=n_{k+1}}^{n_K}|h_n|^2 + N_0\big)$.

Theorem 2: In the scenario where the input signal of Alg. $\boldsymbol{\Phi}^{(1)}$ contains K length-2^m RM sequences from the active users in the set $\kappa = \{1, 2, \ldots, K\}$, given the channel coefficients $\mathbf{h}$, the successful detection probability of the active user n, $1 \leq n \leq K$, is formulated in (29), where the set $\aleph$ contains all permutations of $(k-1)$ IDs in $\kappa \setminus \{n\}$. Taking the channel coefficients into account, the average successful detection probability of Alg. $\boldsymbol{\Phi}^{(1)}$ in the multi-sequence case

$$\bar{\bar{P}}_n^{(1)}(\kappa,m,N_0|\mathbf{h})=\Pr\left(B_n|\mathbf{h}\right)+\sum_{k=2}^{K}\left\{\sum_{(n_1,\ldots,n_{k-1})\in\aleph}\left[\left(\prod_{k'=1}^{k-1}\Pr\left(B_{n_{k'}}|B_{n_1,\cdots,n_{k'-1}},\mathbf{h}\right)\right)\cdot\Pr\left(B_n|B_{n_1,\cdots,n_{k-1}},\mathbf{h}\right)\right]\right\} \tag{29}$$

$$\begin{aligned}\Lambda_w^{m-1}=&\Big\{\sum_{n=1}^{2}|h_n|^2\cdot(-i)^{\left(\rho_{n,m}+2b_{n,m}\right)}\sum_{j=1}^{2^{m-1}}(-1)^{\left(\boldsymbol{\alpha}_n^m+\mathbf{a}_{w-1}^{m-1}\right)^T\mathbf{a}_{j-1}^{m-1}}+h_1^*h_2\sum_{j=1}^{2^{m-1}}\Big(\left(c_{1,j}^{m-1}v_{1,j}^{m-1}\right)^*c_{2,j}^{m-1}\\&\cdot(-1)^{\left(\mathbf{a}_{w-1}^{m-1}\right)^T\mathbf{a}_{j-1}^{m-1}}\Big)+h_1h_2^*\sum_{j=1}^{2^{m-1}}\left(c_{1,j}^{m-1}\left(c_{2,j}^{m-1}v_{2,j}^{m-1}\right)^*\cdot(-1)^{\left(\mathbf{a}_{w-1}^{m-1}\right)^T\mathbf{a}_{j-1}^{m-1}}\right)+\varepsilon_w^{m-1}\Big\}_{\Re/\Im}\end{aligned} \tag{30}$$

is approximated as

$$\begin{aligned}&\bar{\bar{P}}_{\text{ave}}^{(1)}(\kappa,m,N_0)\\=&\frac{1}{K}\int_{\mathbf{h}}\Big\{\prod_{n=1}^{K}f_{\mathcal{CN}}(h_n;0,1)\cdot\sum_{n=1}^{K}\bar{\bar{P}}_n^{(1)}(\kappa,m,N_0|\mathbf{h})\Big\}d\mathbf{h}.\end{aligned}$$

Proof: For better understanding, we prove the above conclusion assuming that the input signal of Alg. $\mathbf{\Phi}^{(1)}$ contains the RM sequences from the pair of users in the set $\kappa=\{1,2\}$. In this case, after performing the WHT, the equivalent noise ε_w^{m-1} in the element V_w^{m-1} is approximated to obey the distribution of $\mathcal{CN}\big(0,2^{m-1}\big(2\sum_{n=1}^{2}|h_n|^2N_0+N_0^2\big)\big)$ when the channel coefficients $\mathbf{h}=[h_1,h_2]^T$ are given. For the ease of expression, we drop its subscript in the following. On this basis, the elements in the vector $\mathbf{\Lambda}^{m-1}\in\mathbb{R}^{2^m}$ is specified in (30), where "$\Re/\Im$" means that the real component is taken when $1\le w\le 2^{m-1}$ while the imaginary part is taken when $2^{m-1}+1\le w\le 2^m$. To proceed, we rewrite the second term in (30) as

$$\begin{aligned}&\sum_{j=1}^{2^{m-1}}\left(c_{1,j}^{m-1}v_{1,j}^{m-1}\right)^*c_{2,j}^{m-1}\cdot(-1)^{\left(\mathbf{a}_{w-1,[2:m]}^{m}\right)^T\mathbf{a}_{j-1}^{m-1}}\\=&(-i)^{\left(\rho_{1,m}+2b_{1,m}\right)}\sum_{j=1}^{2^{m-1}}c_{\tilde{1},j}^{m-1}\cdot c_{2,j}^{m-1},\end{aligned} \tag{31}$$

where $\mathbf{c}_{\tilde{1}}^{m-1}$ represents the RM sequence generated by the matrix-vector pair $\left\{\mathbf{P}_1^{m-1},\mathbf{b}_1^{m-1}\oplus\boldsymbol{\alpha}_1^m\oplus\mathbf{a}_{w-1,[2:m]}^m\right\}$. The term $\sum_{j=1}^{2^{m-1}}c_{\tilde{1},j}^{m-1}\cdot c_{2,j}^{m-1}$ in (31) turns out to be the inner product of the sequences $\mathbf{c}_{\tilde{1}}^{m-1}$ and $\mathbf{c}_2^{m-1}$, hence denoted as $\chi_{\tilde{1},2}^{m-1}$. Similarly, the third term in (31) can be rewritten as

$$\begin{aligned}&\sum_{j=1}^{2^{m-1}}c_{1,j}^{m-1}\left(c_{2,j}^{m-1}v_{2,j}^{m-1}\right)^*\cdot(-1)^{\left(\mathbf{a}_{w-1,[2:m]}^{m}\right)^T\mathbf{a}_{j-1}^{m-1}}\\=&(-i)^{\left(\rho_{2,m}+2b_{2,m}\right)}\cdot\chi_{1,\tilde{2}}^{m-1}.\end{aligned}$$

Based on the ID-to-pair mapping rule in Section II.A and our proposed RM expansion rule in Section III.A, we have that $\mathcal{R}_s=s$ when $s\ge m$. Hence, according to (6), the value of χ^s is taken from $\left\{\sqrt{2^s},-\sqrt{2^s},\sqrt{2^s}i,-\sqrt{2^s}i\right\}$ with equal probability, and the real or imaginary component of χ^s, i.e., $\chi_{\Re/\Im}^s$, has a mean of 0 and variance of $\sigma_{\chi^s}^2=2^{s-1}$. However, the sub-matrices of $\mathbf{P}^m$ no longer possess this rank property and the distribution of χ^s when $s<m$ is related to the weight distribution of RM codes [24] [25]. Since it is always symmetrically distributed, the mean of $\chi_{\Re/\Im}^s$ equals to 0, while its variance is obtained via simulation here.

On this basis, when $w\ne\langle[\rho_{n,m};\boldsymbol{\alpha}_n^m]\rangle, n=1,2$, the result of (30) becomes:

$$\begin{aligned}\Lambda_w^{m-1}=&\Big\{(h_1^*h_2)\cdot(-i)^{(\rho_{1,m}+2b_{1,m})}\chi_{\tilde{1},2}^{m-1}\\&+(h_1h_2^*)\cdot(-i)^{(\rho_{2,m}+2b_{2,m})}\chi_{1,\tilde{2}}^{m-1}+\varepsilon^{m-1}\Big\}_{\Re/\Im},\end{aligned}$$

which approximately follows the distribution $\mathcal{N}\left(0,\sigma_{V^{m-1}}^2\right)$ with the variance of

$$\begin{aligned}\sigma_{V^{m-1}}^2=&\left(\left|h_1^*h_2\right|^2+\left|h_1h_2^*\right|^2\right)\sigma_{\chi^{m-1}}^2+\\&2^{m-2}\left[2\left(|h_1|^2+|h_2|^2\right)N_0+N_0^2\right].\end{aligned}$$

Hence, the CDF of $\left|\Lambda_w^{m-1}\right|$ is given as

$$F_{\left|\Lambda_w^{m-1}\right|}(x)=\text{erf}\Big(\frac{x}{\sqrt{2\sigma_{V^{m-1}}^2}}\Big)=G\left(x;0,\sigma_{V^{m-1}}^2\right). \tag{32}$$

As for the case where $w=\langle[\rho_{n,m};\boldsymbol{\alpha}_n^m]\rangle,\ n=1,2$, the mean of $\Lambda_{\langle[\rho_{n,m};\boldsymbol{\alpha}_n^m]\rangle}^{m-1}$ changes to $(-1)^{b_{n,m}}\cdot2^{m-1}|h_n|^2$. Hence, the PDF and the CDF of its magnitude are

$$\begin{aligned}&f_{\left|\Lambda_{\langle[\rho_{n,m},\boldsymbol{\alpha}_n^m]\rangle}^{m-1}\right|}(x)=g\left(x;2^{m-1}|h_n|^2,\sigma_{V^{m-1}}^2\right),\\&F_{\left|\Lambda_{\langle[\rho_{n,m},\boldsymbol{\alpha}_n^m]\rangle}^{m-1}\right|}(x)=G\left(x;2^{m-1}|h_n|^2,\sigma_{V^{m-1}}^2\right).\end{aligned} \tag{33}$$

We assume that these two active users are detected following the order $(n_1,n_2)\in\{(1,2),(2,1)\}$. In contrast to the single-sequence scenario, to recover the elements $\left\{\boldsymbol{\alpha}_{n_1}^m,\rho_{n_1,m},b_{n_1,m}\right\}$ correctly first, $\left|\Lambda_{\langle[\rho_{n_1,m};\boldsymbol{\alpha}_{n_1}^m]\rangle}^{m-1}\right|$ has to be not only larger than the terms $\left|\Lambda_w^{m-1}\right|, w\ne\langle[\rho_{n,m};\boldsymbol{\alpha}_n^m]\rangle$, but also larger than the interference term imposed by the active user n_2, i.e., $\left|\Lambda_{\langle[\rho_{n_2,m};\boldsymbol{\alpha}_{n_2}^m]\rangle}^{m-1}\right|$. Thus, combining (32) and (33), the correct probability of $\{\hat{\boldsymbol{\alpha}}_{n_1}^m,\hat{\rho}_{n_1,m},\hat{b}_{n_1,m}\}$ is expressed as

$$\begin{aligned}\Pr\left(A_{n_1}^m|\mathbf{h}\right)=&\int_0^{+\infty}\Big\{g\left(x;2^{m-1}|h_{n_1}|^2,\sigma_{V^{m-1}}^2\right)\cdot\\&\left[G\left(x;0,\sigma_{V^{m-1}}^2\right)\right]^{2^{m-2}}\cdot G\left(x;2^{m-1}|h_{n_2}|^2,\sigma_{V^{m-1}}^2\right)\Big\}dx,\end{aligned}$$

which is consistent with (28) in **Lemma 1**.

After recovering the sequence $\mathbf{v}_{n_1}^{m-1}$ with the estimates $\{\hat{\boldsymbol{\alpha}}_{n_1}^m, \hat{\rho}_{n_1,m}, \hat{b}_{n_1,m}\}$, the next step is to calculate the signal $\mathbf{y}^{m-1}$ following

$$\begin{aligned} y_j^{m-1} =& \frac{1}{2}\left[\left(y_j^m\right)' + \left(v_{n_1,j}^{m-1}\right)^* \cdot \left(y_j^m\right)''\right] \\ =& h_{n_1} \cdot c_{n_1,j}^{m-1} + e_j^{m-1}, \end{aligned}$$

where the equivalent noise $e_j^{m-1} \triangleq \frac{1}{2}h_{n_2}c_{n_2,j}^{m-1}\left(1 + v_{n_2,j}^{m-1}\left(v_{n_1,j}^{m-1}\right)^*\right) + \frac{1}{2}\left(e_j^m + e_{j+2^{m-1}}^m\left(v_{n_1,j}^{m-1}\right)^*\right)$ can be seen as a complex Gaussian r.v. with the mean of 0 and variance of $N_0^{m-1} = \frac{1}{2}(|h_{n_2}|^2 + N_0)$. On this basis, the following detection procedure is equivalent to recovering the pair $\left\{\mathbf{P}_{n_1}^{m-1}, \mathbf{b}_{n_1}^{m-1}\right\}$ under the noise variance of N_0^{m-1}, with the successful probability of $\bar{P}^{(1)}(m-1,\, N_0^{m-1}|\mathbf{h})$ according to (26). The active user n_1 is detected successfully iff $\{\hat{\boldsymbol{\alpha}}_{n_1}^m, \hat{\rho}_{n_1,m}, \hat{b}_{n_1,m}\}$ and $\{\mathbf{P}_{n_1}^{m-1}, \mathbf{b}_{n_1}^{m-1}\}$ are both correct, hence reaching the conclusion that

$$\Pr\left(B_{n_1}|\mathbf{h}\right) = \Pr\left(A_{n_1}^m|\mathbf{h}\right) \cdot \bar{P}^{(1)}\left(m-1, N_0^{m-1}|\mathbf{h}\right),$$

which is in accordance with (27) in **Lemma 1**. Next, assuming that the sequence $\mathbf{c}_{n_1}^m$ is perfectly removed from the signal $\mathbf{y}^m$, the AP proceeds to detect $\mathbf{c}_{n_2}^m$, with the successful detection probability of $\Pr\left(B_{n_2}|B_{n_1},\mathbf{h}\right) = \bar{P}^{(1)}\left(m, N_0|h_{n_2}\right)$.

Now, we are ready to compute the average successful detection probability of these two active users. UE "1" can be detected in the following two cases:

1) UE "1" is detected successfully first, which occurs with the probability of $\Pr\left(B_1|\mathbf{h}\right)$;
2) UE "2" is detected first with the probability of $\Pr\left(B_2|\mathbf{h}\right)$, and then UE "1" is detected from the residual signal with the probability of $\Pr\left(B_1|B_2,\mathbf{h}\right)$.

To sum up, the successful detection probability of UE "1" equals to

$$\bar{\bar{P}}_1^{(1)}\left(\kappa, m, N_0|\mathbf{h}\right) = \Pr\left(B_1|\mathbf{h}\right) + \Pr\left(B_2|\mathbf{h}\right) \cdot \Pr\left(B_1|B_2,\mathbf{h}\right),$$

and the situation for UE "2" is similar. Finally, the average successful detection probability is given by

$$\begin{aligned} \bar{\bar{P}}_{\text{ave}}^{(1)}\left(\kappa, m, N_0\right) = \frac{1}{2}\int_{\mathbf{h}} \Big\{ \prod_{n=1}^{2} f_{\mathcal{CN}}(h_n; 0, 1) \cdot \\ \left(\bar{\bar{P}}_1^{(1)}\left(\kappa, m, N_0|\mathbf{h}\right) + \bar{\bar{P}}_2^{(1)}\left(\kappa, m, N_0|\mathbf{h}\right)\right)\Big\} d\mathbf{h}, \end{aligned}$$

which agrees with the conclusion in **Theorem 2**. ∎

Remark 1. The algorithm proposed in [22] detects multiple sequences iteratively, which is capable of enhancing the detection performance. The theoretical results derived here correspond to the case where the number of iterations is 1 and can be used as the lower bound.

Remark 2. Compared to the single user scenario, there are two additional factors that may affect the performance of Alg. $\mathbf{\Phi}^{(1)}$ when detecting multiple users. Specifically,

1) the number of RM sequences

More RM sequences of the same length superimposed in the input signal would cause a decline in the detection capability, because the variance $\sigma_{V^{m-1}}^2$ in (28) and the parameter N_0^{m-1} in (27) both increase in this case and they are negatively correlated with the successful detection probability $\Pr\left(B_{n_k}|B_{n_1,\ldots,n_{k-1}},\mathbf{h}\right)$. It is the reason why retransmitting the original RM sequences repeatedly when access failures occur is not efficient enough. Inspired by this, in our incremental RA scheme, we reduce the number of fully-overlapped sequences in the input signal by separating different batches of active users utilizing the discrepancy in their sequence length and detect them progressively, thus improving the detection performance.

2) the variance of the inner product of RM sequences

As shown in **Lemma 1**, the variance $\sigma_{V^{m-1}}^2$ is also positively related to the variance of the inner product of two RM sequences $\sigma_{\chi^{m-1}}^2$. Hence, $\sigma_{\chi^{m-1}}^2$ has to be as small as possible for better detection capability. To this end, we limit the matrix $\mathbf{P}^m$ in the Kerdock set, which ensures that $\sigma_{\chi^{m-1}}^2$ reaches its minimum, thus enhancing the performance of Alg. $\mathbf{\Phi}^{(1)}$.

B. Detection Capability of Algorithm $\mathbf{\Phi}^{(2)}$

In this subsection, we analyze the performance of Alg. $\mathbf{\Phi}^{(2)}$ in the scenario where UE "1" is making a second access attempt while UE "2" is newly active. The input signal is illustrated in Fig. 3. In the first iteration, the element-wise conjugate multiplication is calculated following

$$y_{\text{res},j}^{(1)} \cdot \left(y_j^{(2)}\right)^* = |h_1|^2\left(v_{1,j}^m\right)^* + h_1(h_2)^* \cdot c_{1,j}^m\left(c_{2,j}^m\right)^* + \tilde{\varepsilon}_j^{m+1}, \tag{34}$$

where the equivalent noise

$$\begin{aligned} \tilde{\varepsilon}_j^{m+1} =& e_j^{(1)}\left(h_1 \cdot c_{1,j}^m v_{1,j}^m + h_2 \cdot c_{2,j}^m\right)^* + h_1 \cdot c_{1,j}^m\left(e_j^{(2)}\right)^* + \\ & e_j^{(1)}\left(e_j^{(2)}\right)^* \sim \mathcal{CN}\left(0, \left(2|h_1|^2 + |h_2|^2\right)N_0 + N_0^2\right). \end{aligned}$$

After performing the WHT on (34), the w-th element of the resultant vector $\mathbf{\Lambda}^m$ has the form of

$$\begin{aligned} \Lambda_w^m =& \\ & \Big\{(-i)^{\left(\rho_{1,m+1}+2b_{1,m+1}\right)}|h_1|^2 \sum_{j=1}^{2^m}(-1)^{\left(\boldsymbol{\alpha}_1^{m+1} + \mathbf{a}_{w-1,[2:m+1]}^{m+1}\right)^T \mathbf{a}_{j-1}^m} \\ & + h_1 h_2^* \cdot \sum_{j=1}^{2^m} c_{1,j}^m\left(c_{2,j}^m\right)^*(-1)^{\left(\mathbf{a}_{w-1,[2:m+1]}^{m+1}\right)^T \mathbf{a}_{j-1}^m + \varepsilon^m}\Big\}_{\Re/\Im}. \end{aligned}$$

Similar to (31), $\sum_{j=1}^{2^m} c_{1,j}^m\left(c_{2,j}^m\right)^*(-1)^{\left(\mathbf{a}_{w-1,[2:m+1]}^{m+1}\right)^T \mathbf{a}_{j-1}^m}$ can be rewritten as the inner product of $\mathbf{c}_1^m$ and the RM sequence generated by $\{\mathbf{P}_2^m, \mathbf{b}_2^m \oplus \mathbf{a}_{w-1,[2:m+1]}^{m+1}\}$. As before, we denote it by the r.v. χ^m, whose real or imaginary part has a mean of 0 and variance of 2^{m-1}. Then, following the analysis of Alg. $\mathbf{\Phi}^{(1)}$, the correct probability of the estimates $\{\hat{\boldsymbol{\alpha}}_1^{m+1}, \hat{\rho}_{1,m+1}, \hat{b}_{1,m+1}\}$ is

$$\begin{aligned} \Pr\left(A_1^{m+1}|\mathbf{h}\right) = \int_0^{+\infty} \Big\{ & g\left(x; 2^m|h_1|^2, \sigma_{V^m}^2\right) \\ & \cdot \left[G\left(x; 0, \sigma_{V^m}^2\right)\right]^{2^{m+1}-1}\Big\} dx. \end{aligned} \tag{35}$$

where the variance is given by

$$\sigma_{V^m}^2 = 2^{m-1}\left|h_1 h_2^*\right|^2 + 2^{m-1}\left[\left(2|h_1|^2 + |h_2|^2\right)N_0 + N_0^2\right].$$

After recovering $\hat{\mathbf{v}}_1^m$ with the aid of these estimates, we arrive at

$$\begin{aligned} y_j^m =& h_1 \cdot c_{1,j}^m + \frac{1}{2}h_2 \cdot c_{2,j}^m \left(\hat{v}_{1,j}^m\right)^* + e_j^m \\ =& h_1 \cdot c_{1,j}^m + h_{\tilde{2}} \cdot c_{\tilde{2},j}^m + e_j^m, \end{aligned} \tag{36}$$

which means that $\mathbf{y}^m$ can be viewed as the result of overlapping the signal of UE "1" with that of a single virtual user, say, "$\tilde{2}$", whose channel coefficient is $h_{\tilde{2}} = \frac{1}{2}h_2 \cdot (-i)^{(\rho_{1,m+1}+2b_{1,m+1})}$ and the RM sequence $\mathbf{c}_{\tilde{2}}^m$ is generated by the pair $\left\{\mathbf{P}_2^m, \mathbf{b}_2^m \oplus \boldsymbol{\alpha}_1^{m+1}\right\}$. The noise term is specified as $e_j^m = \frac{1}{2}\left(e_j^{(1)} + \left(\hat{v}_{1,j}^m\right)^* e_j^{(2)}\right)$ and obeys the distribution $\mathcal{CN}\big(0, \frac{N_0}{2}\big)$. Next, we expect to recover $\{\boldsymbol{\alpha}_1^m, \rho_{1,m}, b_{1,m}\}$ from (36). The analysis of its successful recovery probability is similar to the process of deriving (28), except that given the knowledge of $\hat{\boldsymbol{\alpha}}_1^{m+1}$, $[\rho_{1,m}; \boldsymbol{\alpha}_1^m]$ is chosen from $\left[\hat{\boldsymbol{\alpha}}_{1,[2:m]}^{m+1}; 0\right]$ and $\left[\hat{\boldsymbol{\alpha}}_{1,[2:m]}^{m+1}; 1\right]$, hence avoiding the interference of UE "$\tilde{2}$". Specifically, we have

$$\begin{aligned} \Pr\left(A_1^m | A_1^{m+1}, \mathbf{h}\right) = \int_0^{+\infty} \Big\{ & g\left(x; 2^{m-1}|h_1|^2, \left(\sigma_{V^{m-1}}^2\right)'\right) \\ & \cdot \left[G\Big(x; 0, \left(\sigma_{V^{m-1}}^2\right)'\Big)\right]\Big\} dx, \end{aligned} \tag{37}$$

with the variance of $\left(\sigma_{V^{m-1}}^2\right)' = \left(|h_1^* h_{\tilde{2}}|^2 + \left|h_1 h_{\tilde{2}}^*\right|^2\right)\sigma_{\chi^{m-1}}^2 + 2^{m-2}\Big[2\left(|h_1|^2 + |h_{\tilde{2}}|^2\right)\frac{N_0}{2} + \left(\frac{N_0}{2}\right)^2\Big]$.

Given the correct $\{\hat{\boldsymbol{\alpha}}_1^{m+1}, \hat{\rho}_{1,m+1}, \hat{b}_{1,m+1}\}$ and $\{\hat{\boldsymbol{\alpha}}_1^m, \hat{\rho}_{1,m}, \hat{b}_{1,m}\}$, the remaining process is equivalent to detecting the single sequence $\mathbf{c}_1^{m-1}$ under the noise variance $\left(N_0^{m-1}\right)' = \frac{1}{2}\left(|h_{\tilde{2}}|^2 + \frac{N_0}{2}\right)$. According to (26), its success probability equals to $\bar{P}^{(1)}(m-1, \left(N_0^{m-1}\right)'|\mathbf{h})$. Together with (35) and (37), the probability that UE "1" is detected successfully is

$$\begin{aligned} \Pr(B_1|\mathbf{h}) =& \Pr\left(A_1^{m+1}|\mathbf{h}\right)\Pr\left(A_1^m | A_1^{m+1}, \mathbf{h}\right) \\ & \cdot \bar{P}^{(1)}\big(m-1, \left(N_0^{m-1}\right)'|\mathbf{h}\big). \end{aligned} \tag{38}$$

The detection of UE "2" may be carried out under the following two situations:

1) If UE "1" is detected and its signal is removed perfectly, the residual signal only contains the signal of UE "2". Then, according to (26), UE "2" can be detected successfully with the probability of $\Pr(B_2|B_1, \mathbf{h}) = \bar{P}^{(1)}(m, N_0|h_2)$.
2) If some error occurred in the process of detecting UE "1", removing the detected signal is equivalent to adding an extra active user to the system. However, the channel coefficient of the wrongly detected user is normally so small according to the analysis in [22] that we can simply ignore its impact on detecting UE "2". Hence, we have that $\Pr(B_2|\bar{B}_1, \mathbf{h}) = \bar{\bar{P}}_2^{(1)}(\kappa, m, N_0|\mathbf{h})$, where $\bar{B}_1$ indicates the event that UE "1" fails to be detected and the specific expression of $\bar{\bar{P}}_n^{(1)}(\cdot)$ is given in (29).

To conclude, the successful detection probability of UE "2" is approximately $\Pr(B_2|\mathbf{h}) = \Pr(B_1|\mathbf{h})\Pr(B_2|B_1, \mathbf{h}) + \big(1 - \Pr(B_1|\mathbf{h})\big)\Pr(B_2|\bar{B}_1, \mathbf{h})$. Combined with $\Pr(B_1|\mathbf{h})$ in (38), the average successful detection probability of Alg. $\boldsymbol{\Phi}^{(2)}$ in the first iteration has the form of

$$\begin{aligned} & P_{\text{ave}}^{(2)}(\kappa, m, N_0) = \\ & \frac{1}{2}\int_{\mathbf{h}} f_{\mathcal{CN}}(h_1; 0, 1) f_{\mathcal{CN}}(h_2; 0, 1)\left(\Pr(B_1|\mathbf{h}) + \Pr(B_2|\mathbf{h})\right) d\mathbf{h}. \end{aligned} \tag{39}$$

With the iteration progressing, the success probability is expected to be improved, and hence the result given in (39) is the lower bound of the performance of Alg. $\boldsymbol{\Phi}^{(2)}$. In the general case where there are more active users in these two slots, the analysis process is similar, except that more multi-user interference has to be considered. We are not going to elaborate on it here.

C. Analysis of the Computational Complexity

In this subsection, we analyze the computational complexity of the proposed detection algorithm, which is quantified by the number of multiplication and search operations.

Theorem 3: The computational complexity of Alg. $\boldsymbol{\Phi}^{(1)}$ when applied to a length-2^m RM sequence equals to $\mathcal{O}\big[(m+1)\cdot 2^m + \frac{1}{2}m^2\big]$.

Proof: In Alg. $\boldsymbol{\Phi}^{(1)}$, the WHT operations dominate the computational tasks. Recovering $\{\boldsymbol{\alpha}^s, \rho_s, b_s\}, 2 \le s \le m$ requires on the order of $\mathcal{O}\left[(s-1)\cdot 2^{s-1}\right]$ multiplication operations when the fast WHT is adopted. If we further take into consideration the multiplications in (4), (10) and (15), the complexity of recovering $\{\boldsymbol{\alpha}^s, \rho_s, b_s\}$ is $\mathcal{O}\left[s\cdot 2^{s-1}\right]$. Meanwhile, the number of searches in (14) equals to 2^s when $s = m$ and to 2 otherwise. To sum up, the total complexity of applying Alg. $\boldsymbol{\Phi}^{(1)}$ to a length-2^m RM sequence is given in **Theorem 3**. ■

On this basis, let us consider the active user n who retransmitted R_n $(0 \le R_n \le R_{\max})$ times before being detected successfully. During the r-th re-detection, the length of the RM sequence that entered into Alg. $\boldsymbol{\Phi}^{(1)}$ equals to

$$\ell_r = \begin{cases} 2^{m+r}, & r = 1 \\ 2^{m+r-2}, & 2 \le r \le R_n \end{cases},$$

By substituting ℓ_r into the result in **Theorem 3**, the total complexity of detecting this user is

$$\begin{cases} \mathcal{O}\left[(m+1)\cdot 2^m + \frac{1}{2}m^2\right], & R_n = 0, \\ \mathcal{O}\left[(3m+5)\cdot 2^m + m^2\right], & R_n = 1, \\ \sum_{r=2}^{R_n} \mathcal{O}\Big[(m+r-1)\cdot 2^{m+r-2} + \\ \quad \frac{1}{2}(m+r-2)^2\Big] + \mathcal{O}\left[(3m+5)\cdot 2^m + m^2\right] & R_n \ge 2. \end{cases} \tag{40}$$

V. Simulation Results

In this section, we first examine the detection capability of Alg. $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$. Then, the performance of our proposed incremental RA scheme is evaluated in terms of access probability, computational complexity and access latency. In the sequel, the transmit power of each active user is normalized to 1 and the received signal-to-noise ratio (SNR) at the AP is given by

$$\text{SNR} \triangleq \frac{\|\mathbf{y}^{(t)} - \mathbf{e}^{(t)}\|^2}{\|\mathbf{e}^{(t)}\|^2}.$$

The successful detection probability is defined as the average ratio between the number of successfully detected active users and the total number of active users in the system.

A. Performance of Algorithms $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$

We focus on Alg. $\mathbf{\Phi}^{(1)}$ in the single user scenario first. It can be seen from Fig. 6(a) that despite the Gaussian approximation in the derivation, the theoretical results still tightly match the simulation results in both cases of $m = 8$ and $m = 10$. Besides, it also shows that the detection performance improves upon increasing the sequence length, which confirms the feasibility of our intention to improve the access probability by expanding the sequence before retransmission.

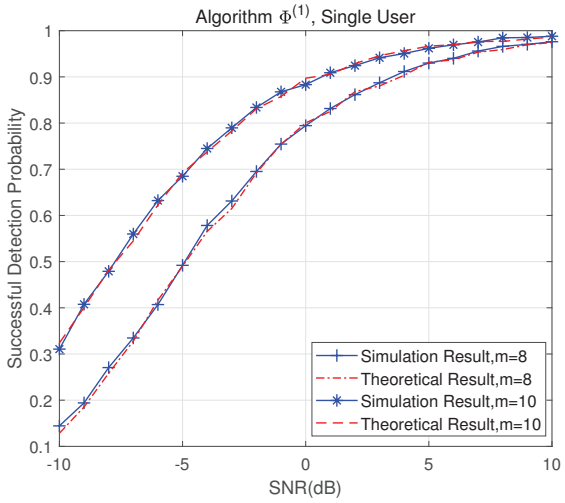
Next, we examine the performance of $\mathbf{\Phi}^{(1)}$ when it is executed to detect multiple users. The scenarios where 2 and 3 active users coexist are considered in Fig. 6(b). The number of iterations is set to be 1 and $m = 8$. Fig. 6(b) shows that the theoretical results are consistent with the simulation results. Furthermore, the detection performance degrades as the sequences of the same length accumulate, which reveals that retransmitting the RM sequences without expansion repeatedly in the case of access failures is not efficient enough and may cause serious congestion.

Fig. 7 shows the performance of Alg. $\mathbf{\Phi}^{(2)}$ given the input signal in Fig. 3. We set $m = 8$ and the number of iterations is 1. The theoretical result given in (39) is also depicted for comparison. It can be seen that the theoretical result is in strong agreement with the simulation result, thus validating our theoretical derivation.

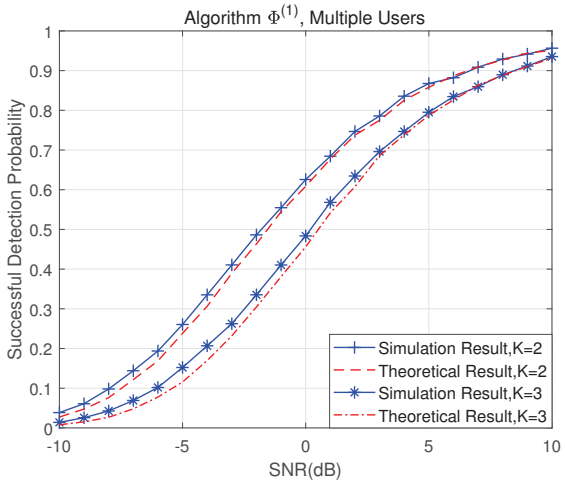
B. Performance of the Incremental RA Scheme Using RM Sequences

We evaluate the performance of our proposed incremental RA scheme ("*RM, proposed*") in the Rayleigh fading channel. The commonly used method ("*RM*"), where RM sequences without expansion are retransmitted if access failures occur, is used for comparison. The RA scheme using ZC sequences ("*ZC*") is regarded as a benchmark. The simulation parameters are listed in Table I and the threshold γ equals to the noise power.

Firstly, we consider the simpler case where no new users become active since RA slot 2. Fig. 8 compares the successful access probability of different retransmission schemes. It can be seen that when there is no retransmission ($R_{\max} = 0$), the detection capability of the RM detection algorithm $\mathbf{\Phi}^{(1)}$ is much better than that of the ZC sequences if the number



(a) single-user scenario



(b) multi-user scenario

Fig. 6: *The successful detection probability of* $\mathbf{\Phi}^{(1)}$ *versus SNR.*

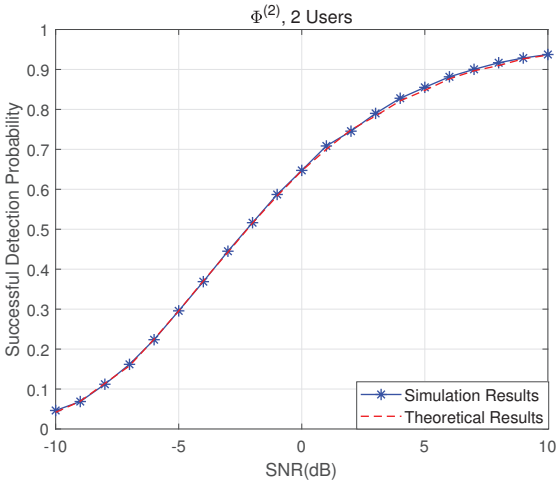


Fig. 7: *The successful detection probability of* $\mathbf{\Phi}^{(2)}$ *versus SNR with input signal shown in Fig. 3.*

TABLE I: *Simulation Parameters*

RM Sequence Length, $\ell_{\text{RM}} = 2^m$	256
The Number of ZC Root Sequences, N_{ZC}	250
ZC Sequence Length, ℓ_{ZC}	251
The Number of Iterations in $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$	5
The Noise Variance N_0 (dB)	-10
Simulation Times	5000

of active users is no more than 15. However, it starts to degrade when the number of active users continues increasing. If retransmission is allowed and limited to at most 3 times ($R_{\max} = 3$), the performance of "*RM*" improves to some degree. Nonetheless, when the number of active users is large, e.g., more than 20, "*RM*" still cannot achieve satisfactory performance. By contrast, the proposed incremental RA scheme ("*RM, proposed*") attains a significant improvement. As shown in the enlarged image, its successful detection probability is quite close to 1 and even slightly better than that of "*ZC*". The detection performance enhancement results from the longer RM sequences obtained by combining the expanded RM sequences with the previously transmitted ones, which is in accordance with the phenomenon shown in Fig. 6(a).

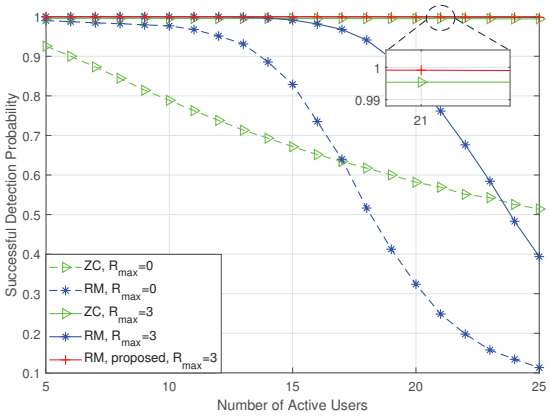


Fig. 8: *The successful detection probability versus the number of active users when $m = 8$.*

Next, we proceed to compare the computational complexity of these retransmission methods. The complexity of detecting the active user n who retransmitted R_n times in total is listed in Table II. An example under the simulation parameters in Table I and $R_n = 3$ is also given here. It can be seen that the complexity of the RM detection algorithm is much lower than that of the correlation detection algorithm for ZC. It is intuitive that the complexity of "*RM, proposed*" is higher than that of "*RM*" since the lengths of the sequences increase gradually. Combined with Fig. 8, it reveals that the overall performance of "*RM*" is superior when the number of active users is less than 15 since it can achieve the successful detection probability similar to that of "*RM, proposed*" with lower complexity. However, as the number of active users continues growing, "*RM, proposed*" defeats "*RM*" because of the distinct advantage in the detection capability. Besides, an insightful observation is that although "*RM, proposed*" and "*ZC*" exhibit similar detection capabilities when $R_{\max} = 3$ in Fig. 8, "*RM, proposed*" imposes much lower complexity, which validates the superiority of our proposed scheme.

TABLE II: *Computational Complexity Comparison of Different Retransmission Schemes.*

Scheme	Complexity	Example
ZC	$\mathcal{O}\left[2N_{\text{ZC}}\ell_{\text{ZC}}R\right]$	502000
RM	$\mathcal{O}\left[\left((m+1)\cdot 2^m + \frac{1}{2}m^2\right)R\right]$	9344
RM, proposed	(40)	14976

Finally, the general case where the number of newly active users in each RA slot follows the Possion distribution with a mean of λ is considered. Fig. 9 quantifies the successful access probability versus λ in the period of $T = 20$ RA slots. It shows that the performance of "*ZC*" erodes almost linearly upon increasing the number of active users. Although relaxing the retransmission limit to at most 5 times does improve the access probability, it still remains unsatisfactory. Upon using RM sequences, the access probability is significantly improved when the number of active users is mediocre. However, as λ increases, simply retransmitting the RM sequences repeatedly ("*RM*") results in a sharp access probability reduction since the detection capability of Alg. $\mathbf{\Phi}^{(1)}$ degrades upon increasing the number of fully overlapped sequences in the input signal, which agrees with the result of Fig. 6(b). Comparing "*RM, $R_{max} = 2$*" to "*RM, $R_{max} = 5$*", it is unexpected to find that increasing $R_{\max}$ degrades the access performance instead of improving it. The random backoff ("*RM, backoff*") can improve the successful detection probability since the access requests are dispersed. Different from "*RM*", increasing $R_{\max}$ here has a positive but little effect. By contrast, the proposed incremental RA scheme ("*RM, proposed*") exhibits great robustness vs λ. Besides, its advantage over "*RM*" becomes increasingly obvious with the growing λ, and its performance when $R_{\max} = 2$ is even better than that of "*RM, backoff*" when $R_{\max} = 5$.

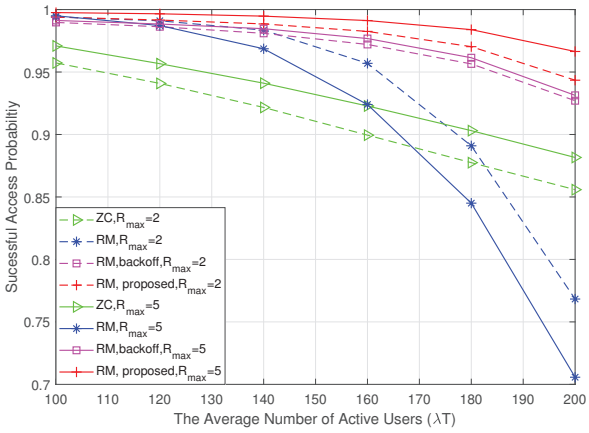


Fig. 9: *The successful access probability versus the number of active users when $m = 8$.*

In Fig. 10, we compare the access latency of different schemes, which is defined as the average number of RA slots

that active users require for their connection setup. It is shown that the access latency of "*ZC*" grows linearly with the increase of λT. The access efficiency of "*RM*" is fine when the number of active users is relatively small. However, with the increase of user activity, the access latency rises rapidly. Compared to "*RM*", the proposed RA scheme ("*RM, proposed*") has a much more gradual latency growth rate, and it also exhibits higher access efficiency than "*RM, backoff*", thus again validating the superior performance of our proposed scheme.

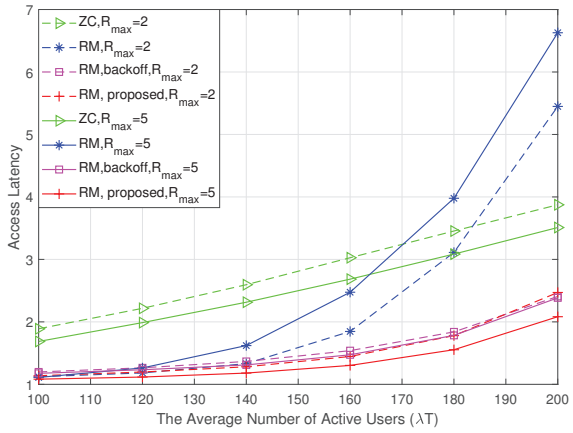


Fig. 10: *The access latency versus the number of active users when* $m = 8$.

VI. Conclusions

An incremental RA scheme exploiting the nested RM sequences was proposed to cope with a massive number of access requests and with the potential access failures caused by the highly dynamic nature of mMTC. Specifically, the existing active users suffering from access failures continue accessing the system with the aid of the expanded RM sequences. At the AP, a recursive detection algorithm is proposed. By combining the expanded RM sequences and the previously transmitted ones, the detection probability of the existing active users can be progressively improved. Furthermore, the discrepancy in the sequence lengths can be exploited for separating different batches of active users and then for detecting them progressively. The theoretical performance of the proposed algorithms was analyzed. Our simulation results verified the analysis and confirmed the superior performances of our RA scheme both in terms of enhanced detection capability and reduced access latency.

Appendix

In RA slot $t, 2 \leq t \leq T$, the AP appends the newly received signal to the residual signal as $\mathbf{y} = \left[\mathbf{y}_{\text{res}}^{(t-1)}; \mathbf{y}^{(t)}\right]$. Note that when $t > R_{\max}+1$, the residual signal has to be processed at first. Explicitly, the AP decomposes the signal $\mathbf{y}_{\text{res}}^{(t-1)}$ following

$$\bar{\mathbf{y}}_{\text{res}}^{(\tau)} = \begin{cases} \mathbf{y}_{\text{res},[1:2^m]}^{(t-1)}, & \tau = 1, \\ \mathbf{y}_{\text{res},[2^{m+\tau-2}+1:2^{m+\tau-1}]}^{(t-1)}, & 2 \leq \tau \leq R_{\max}+1, \end{cases}$$

and then updates $\mathbf{y}_{\text{res}}^{(t-1)} = \left[\bar{\mathbf{y}}_{\text{res}}^{(2)}; \left(\bar{\mathbf{y}}_{\text{res}}^{(3)}\right)'; \ldots; \left(\bar{\mathbf{y}}_{\text{res}}^{(R_{\max}+1)}\right)'\right]$. Then, the concatenated signal $\mathbf{y}$ is input to Alg. $\mathbf{\Phi}^{(\iota)}$, where $\iota = \min\{t, R_{\max}+1\}$. The algorithms are summarized as follows.

Algorithm 1 : $\mathbf{\Phi}^{(\iota)}$, $3 \leq \iota \leq R_{\max}+1$

Input: the signal $\mathbf{y}$
Output: the triplet set $\Xi^{(\iota)}$

Decompose $\mathbf{y}$ as illustrated in Fig. 11, i.e.,

$$\bar{\mathbf{y}}^{(\tau)} = \begin{cases} \mathbf{y}_{[1:2^m]}, & \tau = 1, \\ \mathbf{y}_{[2^{m+\tau-2}+1:2^{m+\tau-1}]}, & 2 \leq \tau \leq \iota. \end{cases}$$

if $\left\|\bar{\mathbf{y}}^{(1)}\right\|^2 \leqslant \gamma$ **then**
  Construct the sequence $\left[\bar{\mathbf{y}}^{(2)}; \left(\bar{\mathbf{y}}^{(3)}\right)'; \ldots; \left(\bar{\mathbf{y}}^{(\iota)}\right)'\right]$ and execute Alg. $\mathbf{\Phi}^{(\iota-1)}$ on it. Then let $\Xi^{(\iota)} = \Xi^{(\iota-1)}$.
else
  Perform Alg. $\mathbf{\Phi}^{(1)}$ on $\left(\bar{\mathbf{y}}^{(\iota)}\right)''$ to obtain the estimate triples $\left\{\hat{\mathbf{P}}_l^{m+\iota-3}, \hat{\hat{\mathbf{b}}}_l^{m+\iota-3}, \hat{\hat{h}}_l\right\}$, $1 \leqslant l \leqslant L$.
  Expand $\hat{\mathbf{P}}_l^{m+\iota-3}$ to get the elements $\left\{\hat{\boldsymbol{\alpha}}_l^{m+\iota-1}, \hat{\boldsymbol{\alpha}}_l^{m+\iota-2}, \hat{\rho}_{l,m+\iota-1}, \hat{\rho}_{l,m+\iota-2}, \hat{b}_{l,m+\iota-1}, \hat{b}_{l,m+\iota-2}\right\}$.
  Correct the biased vectors and channel estimates following

$$\hat{\mathbf{b}}_l^{m+\iota-3} = \hat{\hat{\mathbf{b}}}_l^{m+\iota-3} \oplus \hat{\boldsymbol{\alpha}}_{l,[2:m+\iota-2]}^{m+\iota-1} \oplus \hat{\boldsymbol{\alpha}}_l^{m+\iota-2},$$
$$\hat{h}_l = \hat{\hat{h}}_l \cdot (-1)^{(\hat{b}_{l,m+\iota-1}+\hat{b}_{l,m+\iota-2}+\hat{\alpha}_{l,1}^{m+\iota-1})} \cdot (-i)^{(\hat{\rho}_{l,m+\iota-1}+\hat{\rho}_{l,m+\iota-2})}.$$

  Add the triplets $\left\{\hat{\mathbf{P}}_l^m, \hat{\mathbf{b}}_l^m, \hat{h}_l\right\}$ into the set $\Xi^{(\iota)}$ and reconstruct the RM sequences $\hat{\mathbf{c}}_l^{m+\iota-1}$.
  Update the signal following

$$y_j = y_j - \sum_{l=1}^{L} \hat{h}_l \cdot \hat{c}_{l,j}^{m+\iota-1}.$$

  Decompose $\mathbf{y}$ as shown in Fig. 11 and then input $\left[\bar{\mathbf{y}}^{(2)}; \left(\bar{\mathbf{y}}^{(3)}\right)'; \ldots; \left(\bar{\mathbf{y}}^{(\iota)}\right)'\right]$ to $\mathbf{\Phi}^{(\iota-1)}$. Let $\Xi^{(\iota)} = \Xi^{(\iota)} \cup \Xi^{(\iota-1)}$.
end if
return $\Xi^{(\iota)}$

Algorithm 2 : $\mathbf{\Phi}^{(2)}$

Input: the signal $\mathbf{y}$, the maximum number of iterations T_{iter}, the maximum number of detected users L_1 and L_2, the threshold γ
Output: the triplet set $\Xi^{(2)}$

Decompose $\mathbf{y}$ into two subsequences $\mathbf{y}'$ and $\mathbf{y}''$ as shown in Fig. 12.
if $\|\mathbf{y}'\|^2 \leqslant \gamma$ and $\|\mathbf{y}'' - \mathbf{y}'\|^2 \leqslant \gamma$ **then**
  $\Xi^{(2)} = \emptyset$.
else if $\|\mathbf{y}'\|^2 \leqslant \gamma$ and $\|\mathbf{y}'' - \mathbf{y}'\|^2 > \gamma$ **then**
  Input $\mathbf{y}''$ to Alg. $\mathbf{\Phi}^{(1)}$ and let $\Xi^{(2)} = \Xi^{(1)}$.
else if $\|\mathbf{y}'\|^2 > \gamma$ and $\|\mathbf{y}'' - \mathbf{y}'\|^2 \leq \gamma$ **then**
  Input $\mathbf{y}$ to Alg. $\mathbf{\Phi}^{(1)}$ and let $\Xi^{(2)} = \Xi^{(1)}$.
else
  Initialize $\hat{h}_{l_2} \cdot \hat{c}_{l_2,j}^m = 0$, where $1 \leq j \leq 2^m$ and $1 \leq l_2 \leq L_2$.
  for $t_{\text{iter}} = 1 : T_{\text{iter}}$ **do**
    Calculate the residual signal following

$$\begin{cases} (y_{\text{res},j})' = y_j' \\ (y_{\text{res},j})'' = y_j'' - \sum_{l_2=1}^{L_2} \hat{h}_{l_2} \cdot \hat{c}_{l_2,j}^m \end{cases}.$$

12: Perform Alg. $\mathbf{\Phi}^{(1)}$ on the signal $\mathbf{y}_{\text{res}}$ to detect the triplets $\left\{\hat{\mathbf{P}}_{l_1}^{m+1}, \hat{\mathbf{b}}_{l_1}^{m+1}, \hat{h}_{l_1}\right\}, 1 \le l_1 \le L_1$, and then reconstruct the sequences $\hat{\mathbf{c}}_{l_1}^{m+1}$.

13: Update the residual signal as

$$y_{\text{res},j} = y - \sum_{l_1=1}^{L_1} \hat{h}_{l_1} \cdot \hat{c}_{l_1,j}^{m+1},$$

and input the latter half part of it, namely, $(\mathbf{y}_{\text{res}})''$, to Alg. $\mathbf{\Phi}^{(1)}$ to obtain the estimated triplets $\left\{\hat{\mathbf{P}}_{l_2}^{m}, \hat{\mathbf{b}}_{l_2}^{m}, \hat{h}_{l_2}\right\}$ before recovering the sequence $\hat{\mathbf{c}}_{l_2}^{m}$.

14: **end for**

15: Add all the detected triplets $\left\{\hat{\mathbf{P}}_{l_1}^{m}, \hat{\mathbf{b}}_{l_1}^{m}, \hat{h}_{l_1}\right\}$ and $\left\{\hat{\mathbf{P}}_{l_2}^{m}, \hat{\mathbf{b}}_{l_2}^{m}, \hat{h}_{l_2}\right\}$ into the set $\Xi^{(2)}$.

16: **end if**

17: **return** $\Xi^{(2)}$

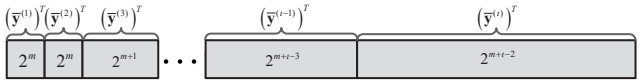


Fig. 11: *The illustration of the signal decomposition in Alg. 1. The number in each block denotes the length of the subsequence.*

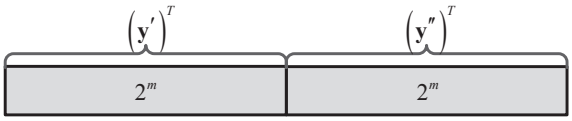


Fig. 12: *The illustration of the signal decomposition in Line 1 of Alg. 2.*

REFERENCES

[1] J. Wang, Z. Zhang, *et al.*, "Incremental Random Massive Access Exploiting Nested Reed-Muller Sequences," *IEEE ICC 2020*, accepted, Dublin, Jun. 2020.

[2] H. Shariatmadari *et al.*, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10-17, Sep. 2015.

[3] C. Bockelmann *et al.*, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59-65, Sep. 2016.

[4] 3GPP TS 36.321 V9.3.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control," Jun. 2010.

[5] 3GPP TS 36.211 V10.4.0, "Evolved Universal Terrestrial Radio Access (E-UTRA); physical channels and modulation," Dec. 2011.

[6] 3GPP TR 37.868 V11.0.0, "Study on RAN improvements for machine-type communications," Sep. 2011.

[7] R. Cheng, J. Chen, D. Chen and C. Wei, "Modeling and analysis of an extended access barring algorithm for machine-type communications in LTE-A networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 2956-2968, Jun. 2015.

[8] S. Duan, V. Shah-Mansouri, Z. Wang and V. W. S. Wong, "D-ACB: adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847-9861, Dec. 2016.

[9] T. Lin, C. Lee, J. Cheng and W. Chen, "PRADA: prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467-2472, Jun 2014.

[10] N. K. Pratas, H. Thomsen, C. Stefanovic, and P. Popovski, "Code-expanded random access for machine-type communications," in *IEEE Globecom Workshops (GC Wkshps)*, Anaheim, 2012, pp. 1681-1686.

[11] J. S. Kim, S. Lee, and M. Y. Chung, "Efficient random-access scheme for massive connectivity in 3GPP low-cost machine-type communications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6280-6290, Jul. 2017.

[12] H. S. Jang, S. M. Kim, K. S. Ko, J. Cha, and D. K. Sung, "Spatial group based random access for M2M communications," *IEEE Commun. Lett.*, vol. 18, no. 6, pp. 961-964, 2014.

[13] C. Kalalas and J. Alonso-Zarate, "Efficient cell planning for reliable support of event-driven machine-type traffic in LTE," in *IEEE Global Commun. Conf.*, Singapore, 2017, pp. 1-7.

[14] J. B. Seo and V. C. M. Leung, "Performance characterization on random access in LTE-based two-tier small-cell networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8528-8537, Oct. 2016.

[15] A. Ilori, Z. Tang, J. He, K. Blow, and H. H. Chen, "A random channel access scheme for massive machine devices in LTE cellular networks," in *2015 IEEE Int. Conf. Commun. (ICC)*, London, 2015, pp. 2985-2990.

[16] S. D. Howard, A. R. Calderbank, and S. J. Searle, "A fast reconstruction algorithm for deterministic compressive sensing using second order Reed-Muller codes," in *42nd Annu. Conf. Inf. Sci. and Sys.*, Princeton, 2008, pp. 11-15.

[17] R. Calderbank and S. Jafarpour, "Reed Muller sensing matrices and the LASSO," in *Int. Conf. Sequences Their Applicat. (SETA)*, Paris, 2010, pp. 442-463.

[18] R. Calderbank, S. Howard and S. Jafarpour, "Sparse reconstruction via the Reed-Muller Sieve," in *IEEE Int. Sym. Inf. Theory (ISIT)*, Austin, 2010, pp. 1973-1977.

[19] L. Zhang, J. Luo and D. Guo, "Neighbor discovery for wireless networks via compressed sensing," *Perform. Eval.*, vol. 70, no. 7, pp. 457-471, 2013.

[20] H. Zhang, R. Li, J. Wang, Y. Chen and Z. Zhang, "Reed-Muller sequences for 5G grant-free massive access," in *IEEE Global Commun. Conf.*, Singapore, 2017, pp. 1-7.

[21] R. Calderbank, A. Thompson, "CHIRRUP: a practical algorithm for unsourced multiple access," *arXiv*: 1811.00879v2, 2019.

[22] J. Wang, Z. Zhang and L. Hanzo, "Joint active user detection and channel estimation in massive access systems exploiting Reed-Muller sequences," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 739-752, Jun. 2019.

[23] M. M. Nastasescu and A. R. Calderbank, "The projective Kerdock code," in *IEEE Inform. Theory Workshop*, Dublin, 2010, pp. 1-5.

[24] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*. North-Holland Elsevier, 1983.

[25] N. Sloane and E. Berlekamp, "Weight enumerator for second-order Reed-Muller codes," *IEEE Trans. Info. Theory*, vol. 16, no. 6, pp. 745-751, Nov. 1970.

Jue Wang (S'18) received the B.S. degree in communication engineering from the Department of Communication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, China, in 2016. She is currently working towards her Ph.D. degree in information and communication engineering under the supervision of Prof. Zhaoyang Zhang at Zhejiang University. Her current research interests include signal processing, massive random access, massive MIMO and machine learning.

Zhaoyang Zhang (M'02) received his Ph.D. degree from Zhejiang University, Hangzhou, China, in 1998, where he is currently a Qiushi Distinguished Professor. His current research interests are mainly focused on the fundamental aspects of wireless communications and networking, such as information theory and coding, network signal processing and distributed learning, AI-empowered communications and networking, network intelligence with synergetic sensing, computing and communication, etc. He has co-authored more than 300 peer-reviewed international journal and conference papers, and is a co-recipient of 8 best paper awards of international conferences including IEEE ICC 2019 and IEEE GlobeCom 2020. He was awarded the National Natural Science Fund for Distinguished Young Scholars by NSFC in 2017.

Dr. Zhang is serving or has served as Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS and IET COMMUNICATIONS, etc, and as General Chair, TPC Co-Chair or Symposium Co-Chair for VTC-Spring 2017 Workshop HMWC, WCSP 2013/2018, Globecom 2014 Wireless Communications Symposium, etc. He was also a keynote speaker for APCC 2018 and VTC-Fall 2017 Workshop NOMA.

Caijun Zhong (S'07-M'10-SM'14) received the B.S. degree in Information Engineering from the Xi'an Jiaotong University, Xi'an, China, in 2004, and the M.S. degree in Information Security in 2006, Ph.D. degree in Telecommunications in 2010, both from University College London, London, United Kingdom. From September 2009 to September 2011, he was a research fellow at the Institute for Electronics, Communications and Information Technologies (ECIT), Queens University Belfast, Belfast, UK. Since September 2011, he has been with Zhejiang University, Hangzhou, China, where he is currently a Professor. His current research interests include Reconfigurable intelligent surfaces assisted communications and artificial intelligence based wireless communications.

Dr. Zhong is an Editor of Science China: Information Science and China Communications. He was an editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2015-2020 and IEEE COMMUNICATIONS LETTERS from 2014 to 2018. He is the recipient of the 2013 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He and his coauthors has been awarded a Best Paper Award at the IEEE ICC 2019 and IEEE Globecom 2020.

Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk, https://en.wikipedia.org/wiki/Lajos_Hanzo) (FIEEE'04) received his Master degree and Doctorate in 1976 and 1983, respectively from the Technical University (TU) of Budapest. He was also awarded the Doctor of Sciences (DSc) degree by the University of Southampton (2004) and Honorary Doctorates by the TU of Budapest (2009) and by the University of Edinburgh (2015). He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of the IEEE Press. He has served several terms as Governor of both IEEE ComSoc and of VTS. He has published 1900+ contributions at IEEE Xplore, 19 Wiley-IEEE Press books and has helped the fast-track career of 123 PhD students. Over 40 of them are Professors at various stages of their careers in academia and many of them are leading scientists in the wireless industry. He is also a Fellow of the Royal Academy of Engineering (FREng), of the IET and of EURASIP.

L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/N004558/1, EP/P034284/1, EP/P034284/1, EP/P003990/1 (COALESCE), of the Royal Society's Global Challenges Research Fund Grant as well as of the European Research Council's Advanced Fellow Grant QuantCom.