Cutoff Rate of Sparse Code Multiple Access in Downlink Broadcast Channels

Li Li, Zheng Ma, *Member, IEEE*, Li Wang, *Member, IEEE*, Ping Zhi Fan, *Fellow, IEEE* and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—For the sake of supporting massive connectivity in future 5G networks, non-orthogonal multiple access (NOMA) techniques are advocated. As a promising NOMA technique, in recent years sparse code multiple access (SCMA) has attracted substantial attention. However, there is a paucity of studies on the theoretical analysis of its error-freely achievable data rate, especially in the downlink context. Hence, we derive the cutoff rate of SCMA in downlink broadcast channels, which indicates the lower-bound of a system's error-freely achievable rate. However, we will demonstrate that when considering the conventional categorization of pairwise error events, the accuracy of the cutoff rate rapidly degrades in the low-SNR region owing to the fact that multi-user SCMA systems typically encounter an extremely large constellation size. Alternatively, by invoking Bergmans' concept from 1973 in the categorization of pairwise error events, we obtain a more accurate cutoff rate both in the low- and the high-SNR regions. Moreover, we provide insights into the cutoff rate derivation process, which reveals some general guidelines for designing a beneficial codebook, capable of improving SCMA with respect to its original low-density signature (LDS) based counterpart.**

I. Introduction

Code-division multiple access (CDMA) constitutes one of the most successful multi-user communication systems, allowing a certain number of users to share a common physical channel, where the users are assigned a unique user-specific spreading sequence referred to as "signature". However, owing to the dispersive channel conditions as well as to the potential correlation amongst the spreading sequences employed, CDMA encounters both inter-symbol interference (ISI) and multi-user interference (MUI). Hence, overcoming these interference problems becomes a critical design issue of CDMA techniques.

This work is supported by the NSFC projects (no. 61501383 and no. 61571373), the Key International Cooperation Project of SIchuan Province (no. 2017HH0002), the NSFC China-Swedish project (no. 6161101297), the National Science and Technology Major Project (no. 2016ZX03001018-002), the Huawei HIRP Flagship Project under Grant YB201504, the 111 Project under Grant 111-2-14, and the fundamental research funds for the Central universities (no. 2682015CX064). The financial support of the European Union under the auspices of the Concerto project, and that of the European Research Council's Advanced Fellow Grant are also gratefully acknowledged.

L. Li, Z. Ma, and P. Z. Fan are with the Provincial Key Lab of Information Coding and Transmission, Southwest Jiaotong University, Chengdu, 610031, China (e-mail: ll5e08, zma, pzfan@home.swjtu.edu.cn).

L. Wang is with the R&D center of Huawei Technologies in Stockholm, Sweden.

L. Hanzo is with the School of ECS, University of Southampton, SO17 1BJ, United Kingdom (e-mail: lh@ecs.soton.ac.uk).

Apparently, constructing a set of orthogonal user-specific spreading sequences, e.g. employing the Householder and Gram-Schmidt algorithms [1], provides a straightforward solution for synchronous CDMA systems. However, it is impossible to retain orthogonality, when the number of users U exceeds the spreading factor. Hence, in order to support more than U users, we have to sacrifice the orthogonality of the signals [2].

In order to design this rank-deficient system, an efficient solution is to invoke optimum sequence design for meeting the Welch-bound-equality (WBE) [3]. After early success in this area [4]–[6], some more sophisticated sequence design regimes were developed, for example by invoking hierarchical orthogonal subsets [7], or by adding a feedback mechanism for adapting the spreading sequences [8], [9]. Another promising technique relies on the design of an optimum multiuser detector (MUD). This sort of MUD is capable of approaching the best possible error probability [10], although at the cost of imposing a high computational complexity, which severely impairs its practicality. Consequently, a lot of efforts have been dedicated to improving Verdú's work in [10] for striking a better tradeoff between the complexity imposed and the achievable error performance. In simple terms, these MUDs may be categorized into the class of linear detectors [11]–[13] and that of nonlinear detectors [14]–[16]. However, none of these MUDs achieve a satisfactory performance for rank-deficient systems.

A compelling tradeoff was struck between the complexity and robustness in rank-deficient CDMA systems by Hoshyar *et al.* [17] upon combining their novel low-density signature (LDS) scheme[1] with the message passing algorithm (MPA) of [19], where a large fraction of the chips within a user-specific signature were assigned zero values. Hence the number of users interfering with each other at a specific chip-index was reduced. As a result, the signal search space involved in the maximum-likelihood (ML) criterion based chip-level hypothesis testing process is exponentially reduced. Meanwhile, benefiting from the sparse nature of every user's signature, the signature matrix constituted by all the users' signatures becomes similar to the parity-check matrix (PCM) of low-density parity check codes (LDPC) [20] and consequently lends itself to detection by the MPA, which hence approached the near-optimum BER performance demonstrated in [17].

[1]The LDS is conceptually heralded earlier in [18].

In the original LDS scheme [17], an indicator matrix $\mathbf{F}$ consisting of all binary elements was introduced for representing the spreading pattern. In more detail, the position of 1's in the k^{th} column of $\mathbf{F}$ represents the set of chips over which the k^{th} user spreads his user data. In the MPA of LDS schemes, the indicator matrix $\mathbf{F}$ plays a similar role to the PCM of LDPC codes. Hence, arranging for all the nonzero elements in a column of $\mathbf{F}$ to have identical values of 1 facilitates employing the belief propagation process of the MPA. Regretfully however, assigning an identical value of 1 to all the nonzero elements of $\mathbf{F}$ results in a situation, where the output chip-sequence of a user, - which is generated by spreading a modulated symbol of that user over the associated column of $\mathbf{F}$ - also has identical nonzero elements. The problem is namely that arranging for all the nonzero elements of the output chip-sequence of a user to have the same value impairs the attainable spreading-induced "coding" gain. Hence arranging for the related values to be different is expected to lead to higher "coding" gain. This philosophy was invoked in [21], where the "diversity" of nonzero elements in the output chip-sequence of a user is realized by assigning carefully selected different numbers to the nonzero elements of $\mathbf{F}$.

Inspired by the above-mentioned contributions [17], [21], Nikopour *et al.* [22] suggested a novel multiple access (MA) strategy, namely the sparse code multiple access (SCMA), which inherited the sparse binary indicator matrix $\mathbf{F}$ of the original LDS [17], whilst still assigning a fixed value of unity to all the nonzero elements in $\mathbf{F}$. Then, the diversity of the nonzero elements in the output chip-sequence of a user was realized by introducing a multi-dimensional constellation technique. As a benefit of these multi-dimensional constellation techniques [23]–[25], the resultant shaping gain and coding gain provided by the carefully constructed multi-dimensional constellations resulted in the superiority of SCMA over its LDS counterparts. This improvement of SCMA over its LDS counterpart is achieved without imposing any extra computational complexity. As a result, the user-specific constellation design termed as the "codebook design" of the SCMA scheme was discussed by Taherzadeh *et al.* again in their ensuing work [26].

As a benefit of its remarkable capability of supporting rank-deficient multiuser communications, substantial efforts have been dedicated to further improving the SCMA technique [27]–[31]. For instance, Yu *et al.* attempted to further narrow the performance discrepancy between the SCMA system and the single user system, employing more advanced lattices [27], while Xiao *et al.* used more powerful iterative decoding algorithms [28]. As another important milestone, Chen *et al.* further reduced the complexity [29], while the throughput of SCMA operating in downlink broadcast channels was studied by Nikopour *et al.* [30]. However, only experimental results were shown. The constellation-constrained capacity C of SCMA system was analyzed in [31]. However, only a point-to-point AWGN channel was considered and no closed-form solution was found. On the other hand, the cutoff rate R_0 obtained by Gallager in [32] quantified the data rate for a given SNR, below which reliable communication does become possible even with the aid of randomly generated codewords, provided that the length of codewords tends to infinity. Since we have $R_0 \leq C$, the cutoff rate is also regarded as the lower-bound of a system's error-free data rate [33]. In contrast to capacity, the cutoff rate is typically formulated as a closed-form bound, but it becomes inaccurate at low SNRs owing to relying on the union bound approximation of [34].

Against this background, our major contributions are:

- By invoking the cloud concept introduced in [35] for the categorization of pairwise error events and by exploiting Hölder's inequality [36], we obtained an appropriate approximation of the cutoff rate of SCMA downlink broadcast channels for both AWGN and Rayleigh fading scenarios, which has a closed-from and remains accurate even at low SNRs.
- We provide deep insights into our cutoff rate derivation process, and conceive general guidelines for designing high-performance SCMA codebooks for improving SCMA compared to the LDS technique.

The rest of this paper is organized as follows. Our system model is introduced in Section II, where the concept of cloud partitioning of the synthetic SCMA codebook is also illustrated. The derivation of cutoff rate is provided in Section III and its accuracy is confirmed by simulations. Then, in Section IV, a range of insights into our cutoff rate derivations are offered. Finally, we conclude in Section V.

Notation: Lowercase and capital characters, both in boldface are used to represent vectors and matrices, respectively. Calligraphic characters such as $\mathscr{X}$ represent constellation alphabets.

II. System model and Cloud Partitioning of the Synthetic SCMA Codebook

A. System Model

The multiuser SCMA downlink is considered for broadcasting user-specific information from the base station (BS) to U mobile users (MUs), as depicted in Fig.1, which spans a SCMA transmission block. Observe at the BS of Fig.1 that the information bits destined for the u^{th} MU are firstly mapped to an N-dimensional constellation point $\mathbf{c}_u$, which is selected from its original complex-valued constellation alphabet $\mathscr{C}_u \subset \mathbb{C}^N$ with an equal probability. Let an Ω-dimensional vector $\mathbf{s}_u = [\mathbf{s}_{1u}, \mathbf{s}_{2u}, \ldots, \mathbf{s}_{\Omega u}]$ denote the sparse signature adopted for the u^{th} MU. Then, similar to [22], a binary mapping matrix $\mathbf{V} \in \mathbb{B}^{\Omega \times N}$ can be obtained by omitting all the zero-columns of $\mathrm{diag}(\mathbf{s}_u)$. Consequently, the spreading process invoked for the u^{th}

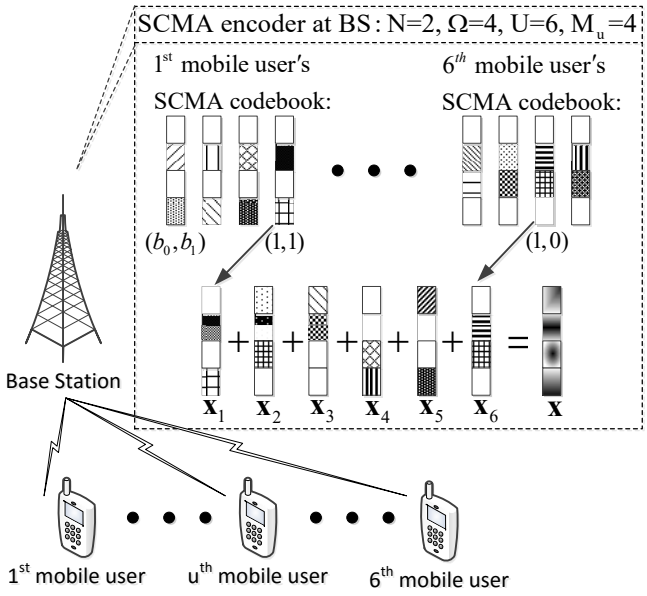

Fig. 1: System model of the SCMA downlink spanning a single transmission block, where the SCMA encoding process implemented at the BS is visualized. A configuration of $N = 2, \Omega = 4, U = 6, M_u = 4$ is adopted as an example.

MU at the BS may be formulated as[2]

$$\mathbf{x}_u = \mathbf{V}\mathbf{c}_u, \tag{1}$$

where the set of spread multidimensional constellation points $\mathbf{x}_u$ constitute the final complex-valued Ω-dimensional constellation alphabet $\mathscr{X}_u$ of the u^{th} MU, which will be referred to as the "u^{th} user-specific SCMA codebook" herein. Moreover, let $M_u = |\mathscr{X}_u|$ represent its cardinality, where it is assumed that all the user-specific SCMA codebooks have the same dimension of Ω.

Observe at the BS of Fig.1 that the spread constellation points of all MUs, i.e. $\{\mathbf{x}_u\}_{u=1}^{U}$, are further multiplexed over Ω shared orthogonal channel resources, such as TDMA time-slots or OFDM tones. Hence the final overlapped symbol spanning Ω channel resources is given by

$$\mathbf{x} = \sqrt{\frac{\sum_{u=1}^{U} P_u}{E\left(\left\|\sum_{u=1}^{U}\mathbf{x}_u\right\|^2\right)}}\sum_{u=1}^{U}\mathbf{x}_u. \tag{2}$$

Consequently, the Ω-dimensional constellation alphabet consists of all the possible values of $\mathbf{x}$ in (2), which may be represented by $\mathscr{X}$, namely by the "synthetic SCMA codebook".

In more depth, P_u denotes the power available for the u^{th} MU, while $E\left(\left\|\sum_{u=1}^{U}\mathbf{x}_u\right\|^2\right)$ represents the power averaged over all constellation points given by the sum of $\sum_{u=1}^{U}\mathbf{x}_u$. Hence, the amplification factor $f_{\text{am}} = \sqrt{\sum_{u=1}^{U} P_u / E\left(\|\cdot\|^2\right)}$ in (2) is introduced for constraining the average power of signals to the total available power of all U MUs, which will be finally transmitted by the BS. Furthermore, as stated in [17], [21], in order to guarantee having a unique output by the MPA employed in the SCMA detector, the synthetic SCMA codebook has to ensure that no pair of different permutations of $\{\mathbf{x}_u\}_{u=1}^{U}$ yields the same $\mathbf{x}$ of (2). Hence the synthetic SCMA codebook has to satisfy the following properties

$$\forall\, \mathbf{x}_u \in \mathscr{X}_u : \left(\{\mathbf{x}'_u\}_{u=1}^{U} \neq \{\mathbf{x}''_u\}_{u=1}^{U}\right) \Rightarrow \left(\sum_{u=1}^{U}\mathbf{x}'_u \neq \sum_{u=1}^{U}\mathbf{x}''_u\right);$$

$$|\mathscr{X}| = M = \prod_{u=1}^{U} M_u. \tag{3}$$

Naturally, every mobile user has to store the synthetic SCMA codebook for invoking the MPA algorithm.

According to the above properties, the received signal of the u^{th} MU in Fig.1, is given by

$$\mathbf{y}_u = \mathbf{h}_u\left(f_{\text{am}}\sum_{u=1}^{U}\mathbf{x}_u\right) + \mathbf{n}_u, \tag{4}$$

where $\mathbf{h}_u = [h_{1u}, h_{2u}, \ldots, h_{\Omega u}]^T$ represents the channel's impulse response (CIR) for the Ω channel resources of the BS-to-u^{th} MU link. Then, $\mathbf{n}_u = [n_{1u}, n_{2u}, \ldots, n_{\Omega u}]^T$ is the complex-valued AWGN noise vector imposed on the u^{th} MU, whose elements are independent and obey $n_{\omega u} \sim \mathcal{CN}(0, \sigma^2)$.

In the forthcoming description, our analyses will be carried on in a range of steps as follows:

1) The proposed novel partitioning of the synthetic SCMA codebook is illustrated in Section II-B.
2) The original formulation of the cutoff rate of our system is derived in Section III-A.
3) The approximations of the original formulas obtained in last step are discussed in Section III-B.
4) The accuracy of these approximations are verified in Section III-C.

B. Could Partitioning of the Synthetic SCMA Codebook

According to (2), (4), the signal transmitted by the BS of Fig.1 carries the joint message of all MUs. The u^{th} MU's receiver seen in Fig.1 is capable of completely recovering this joint message based on an observation of $\mathbf{y}_u$ in (4). However, recovering the joint message is not necessary, since the u^{th} MU only requires the message carried in the specific component $\mathbf{x}_u$ of $\mathbf{x}$ as its valid information. Explicitly, instead of decoding the joint message $\mathbf{x}$, it will only focus on correctly decoding $\mathbf{x}_u$. Hence, from the perspective of the u^{th} MU, the set of constellation points $\mathbf{x}$ having the same component of "$\mathbf{x}_u = \mathbf{x}_u[m]$"[3] constitutes

[2]In SCMA encoder, the information bits destined for a MU may be further divided into multiple data streams. Each of the data stream requires a specific signature and hence invokes an additional column in the SCMA signature matrix, which was interpreted as occupying a SCMA layer in [22], [30]. Without loss of generality, throughout this paper, we assume that a MU only utilizes a single SCMA layer.

[3]$\mathbf{x}_u[m]$ represents the value of $\mathbf{x}_u$ specified as the m^{th} legitimate constellation point in $\mathscr{X}_u$.

a sub-alphabet of the synthetic SCMA codebook, namely $\mathscr{X}_{|\mathbf{x}_u[m]}$, which may be formulated as

$$\mathscr{X}_{|\mathbf{x}_u[m]} = \left\{ \mathbf{x} \in \mathscr{X} : \mathbf{x} = f_{\text{am}} \left(\mathbf{x}_u[m] + \sum_{\substack{v=1 \\ v \neq u}}^{U} \mathbf{x}_v \right) \right\};$$
$$\mathbf{x}_{|\mathbf{x}_u[m]} \in \mathscr{X}_{|\mathbf{x}_u[m]};\ \left|\mathscr{X}_{|\mathbf{x}_u[m]}\right| = M_{|u} = \prod_{\substack{v=1 \\ v \neq u}}^{U} M_v, \quad (5)$$

where $\mathbf{x}_{|\mathbf{x}_u[m]}$ represents an element of $\mathscr{X}_{|\mathbf{x}_u[m]}$.

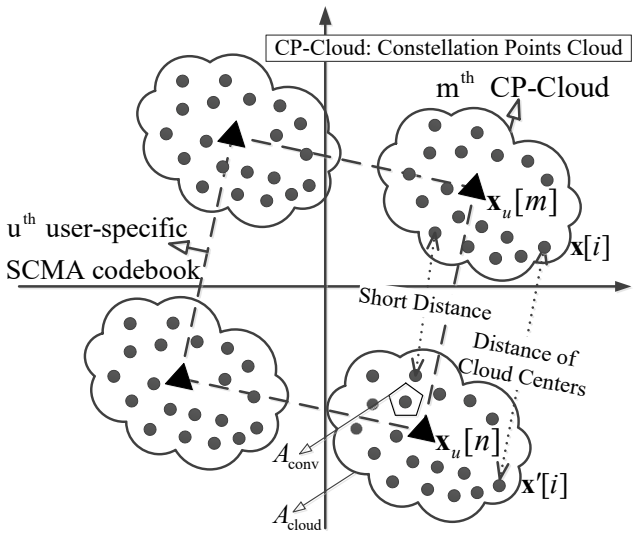

Fig. 2: Partitioning of the synthetic SCMA codebook from the perspective of the u^{th} MU, where $M_u = 4$ is adopted as an example. The synthetic SCMA codebook consists of all filled dots, while the u^{th} user-specific SCMA codebook consists of all solid triangles, which are regarded as virtual cloud centers.

The resultant group of $\{\mathscr{X}_{|\mathbf{x}_u[m]}\}_{m=1}^{M_u}$ actually realizes a partitioning of the synthetic SCMA codebook $\mathscr{X}$, which is visualized in Fig.2, where $\mathscr{X}_{|\mathbf{x}_u[m]}$ represents the "m^{th} constellation point cloud (CP-Cloud)" of the u^{th} MU. Furthermore, as shown in Fig.2, $\mathbf{x}_u[m]$ may be labelled as the virtual center of the m^{th} CP-Cloud. Hence, if we only concentrate our attention on the cloud centers of Fig.2, they happen to construct the u^{th} user-specific SCMA codebook $\mathscr{X}_u$. As a result, at u^{th} MU, the decoding process becomes equivalent to determining the specific CP-Cloud, which the received signal $\mathbf{y}_u$ belongs to.

III. Approximation of Cutoff Rate: Derivations and Verifications

A. Preliminaries

We assume that during L SCMA transmission blocks, a subset of size 2^k is selected from all the possible symbol sequences having a length of L, namely $\mathscr{X}_u^L$, for transmitting k information bits of the u^{th} MU. Its elements $\mathbf{X}_u^L = \left[\mathbf{x}_u^1, \mathbf{x}_u^2, \ldots, \mathbf{x}_u^L\right]$ are referred to as codewords, where the component of $\mathbf{x}_u^\ell$ is the spread constellation point defined in (1), which is broadcast by the BS during the ℓ^{th} SCMA transmission block. We assumed furthermore that all the codewords and their associated components are drawn independently from the probability density function (PDF) $\mathbf{P}\left(\mathbf{X}_u^L\right) = \prod_{\ell=1}^{L} \mathbf{p}\left(\mathbf{x}_u^\ell\right)$. According to the system model of Section II-A, the codeword $\mathbf{X}_u^L$ is actually embedded in $\mathbf{X}^L = \left[\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^L\right]$, which is the final symbol sequence broadcast by the BS[4]. However, as mentioned in Section II-B, instead of determining which specific instantiation of $\mathbf{X}^L$ is transmitted in the SCMA downlink of Fig.1, the u^{th} MU only attempts to decode its information-bearing codewords $\mathbf{X}_u^L$.

Hence, similar to [33], the pairwise error probability of the codewords averaged over the set of randomly generated codewords $\mathscr{X}_u^L$ observed at u^{th} MU is given by

$$\overline{\mathbf{P}_{\mathbf{X}_u^L \to \mathbf{X}_u^{L\prime}}} \leq \prod_{\ell=1}^{L} \left(\sum_{\mathbf{x}_u^\ell \in \mathscr{X}_u} \sum_{\mathbf{x}_u^{\ell\prime} \in \mathscr{X}_u} \mathbf{p}\left(\mathbf{x}_u^\ell\right) \mathbf{p}\left(\mathbf{x}_u^{\ell\prime}\right) \boldsymbol{\Delta}_{\mathbf{x}_u^\ell \to \mathbf{x}_u^{\ell\prime}}^{\lambda} \right), \quad (6)$$
$$= \left(\sum_{\mathbf{x}_u \in \mathscr{X}_u} \sum_{\mathbf{x}_u{}' \in \mathscr{X}_u} \mathbf{p}\left(\mathbf{x}_u\right) \mathbf{p}\left(\mathbf{x}_u{}'\right) \boldsymbol{\Delta}_{\mathbf{x}_u \to \mathbf{x}_u{}'}^{\lambda} \right)^{L}, \lambda > 0, \quad (7)$$

where $\mathbf{X}_u^L \to \mathbf{X}_u^{L\prime}$ represents a pairwise error event, when $\mathbf{X}_u^L$ is transmitted, but during the ML detection, it is incorrectly decoded to $\mathbf{X}_u^{L\prime}$. Then, the equivalence between (6) and (7) relies on the assumption that $\mathbf{p}\left(\mathbf{x}_u^\ell\right)$ obeys an i.i.d distribution and $\mathbf{x}_u^\ell$ has the same alphabet of $\mathscr{X}_u$, regardless of its block index ℓ. Furthermore, $\boldsymbol{\Delta}_{\mathbf{x}_u \to \mathbf{x}_u{}'}^{\lambda}$ is the Chernov parameter defined in [33, 6.8-10]. Consequently, the codeword-level error probability averaged over the ensemble of $\mathscr{X}_u^L$ for the u^{th} MU is given by

$$\overline{\mathbf{P}_e} \leq \sum_{\substack{\mathbf{X}_u^{L\prime} \in \mathscr{X}_u^L \\ \mathbf{X}_u^{L\prime} \neq \mathbf{X}_u^L}} \overline{\mathbf{P}_{\mathbf{X}_u^L \to \mathbf{X}_u^{L\prime}}} = \left(2^k - 1\right) \cdot \overline{\mathbf{P}_{\mathbf{X}_u^L \to \mathbf{X}_u^{L\prime}}}$$
$$\leq 2^{LR_c} \cdot \overline{\mathbf{P}_{\mathbf{X}_u^L \to \mathbf{X}_u^{L\prime}}} \leq 2^{-L(R_0(\mathbf{p},\lambda) - R_c)}, \quad \lambda > 0, \quad (8)$$

where we have $2^k = 2^{L \cdot \frac{k}{L}} = 2^{LR_c}$, while $R_c = \frac{k}{L}$ represents the coding rate expressed in terms of bits per multidimensional signal period. Moreover, we stipulate that $R_0(\mathbf{p}, \lambda) = -\log_2\left[E\left[\boldsymbol{\Delta}_{\mathbf{x}_u \to \mathbf{x}_u{}'}^{\lambda}\right]\right]$.

As shown in (8), provided that we have $R_c \leq R_0(\mathbf{p}, \lambda)$, $\overline{\mathbf{P}_e}$ of the u^{th} MU is reduced to an infinitesimally low value, if the codeword length L tends to ∞, which means that reliable communication is achieved. Hence the maximized $R_0(\mathbf{p}, \lambda)$ in (8) provides a reasonable bound for R_c, below which reliable communications become possible. It is also referred to as the cutoff rate and formulated as

$$R_0 = \max_{\mathbf{p}(\mathbf{x}_u)} \sup_{\lambda > 0} R_0(\mathbf{p}, \lambda). \quad (9)$$

[4] $\mathbf{x}^\ell$ is the synthetic constellation point defined in (2) and its superscript ℓ indicates the associated transmission block index.

Since symmetric channels are involved in our system model, it has been shown in [33] that the maximization of $R_0(\mathbf{p}, \lambda)$ in (9) is achieved by assigning $\lambda = \frac{1}{2}$ and employing the uniform distribution of $\mathbf{p}(\mathbf{x}_u) = \frac{1}{M_u}$. Correspondingly, by exploiting (7), (8), and (9), we arrive at[5]

$$R_0 = -\log_2 \left(\sum_{\substack{m=1, \\ \in \mathscr{X}_u}}^{M_u} \sum_{\substack{n=1, \\ \in \mathscr{X}_u}}^{M_u} \mathbf{p}(\mathbf{x}_u[m])\, \mathbf{p}(\mathbf{x}'_u[n])\, \boldsymbol{\Delta}_{m \to n} \right), \tag{10}$$

where $\boldsymbol{\Delta}_{m \to n}$ is a simplified notation of the Bhatacharyya parameter $\boldsymbol{\Delta}_{\mathbf{x}_u[m] \to \mathbf{x}'_u[n]}$ defined in [33], which is given by

$$\boldsymbol{\Delta}_{m \to n} = \int \left[\int \sqrt{\mathbf{p}(\mathbf{y}, \mathbf{h} \mid \mathbf{x}_u[m])\, \mathbf{p}(\mathbf{y}, \mathbf{h} \mid \mathbf{x}'_u[n])}\, d\mathbf{y} \right] d\mathbf{h}. \tag{11}$$

The variables "$\mathbf{y}, \mathbf{h}$" utilized in (11) have the same definitions as that stipulated in (4). Since throughout this section we always concentrate on a specific MU, the original subscript u is omitted for compactness of expression in Section III. We assume that perfect channel state estimation is available at the receiver represented by the channel impulse response (CIR) $\mathbf{h}$. Furthermore, $\mathbf{p}(\mathbf{y}, \mathbf{h} \mid \mathbf{x}_u[m])$ in (11) equals to the average probability that a certain pair of $\mathbf{y}, \mathbf{h}$ are observed at the receiver, given that an arbitrary symbol pertaining to the m^{th} CP-Cloud of Fig.2 is transmitted. According to this definition and with the aid of Bayes's theorem, the Bhatacharyya parameter may be rewritten as (12), (13).

According to the philosophy behind the Bhatacharyya parameter [33], as indicated by the under-braces in (12), the components under the square-root operation in (12) actually represent a cloud-style categorization of the pairwise error events, which is in accordance with the cloud-style codebook partitioning demonstrated in Fig.2. It implies that only the specific decision, which results in an inter-cloud-transition in Fig.2 will constitute an error-event. Hence, the number of decision regions involved in the ML detection process becomes identical to that of the CP-Clouds.

B. Approximation of Cutoff Rate for Both AWGN and Rayleigh Channels

Recall from Section III-A that in the integral of (13) a square root of the double sum of the products "$\mathbf{p}(\mathbf{y} \mid \mathbf{h}, \mathbf{x}[i])\, \mathbf{p}(\mathbf{y} \mid \mathbf{h}, \mathbf{x}'[j])$" is involved. For example, in a SCMA system supporting 6 MUs, where every MU employs a common modulation order of $M_u = 4$, the number of these products under the square-root operation in (13) is close to $M_{|u}^2 = 2^{20}$, which makes it excessively complex to directly solve the associated integral in (13). Hence, we will attempt to obtain reasonable bounds for the system's cutoff rate. Furthermore, according to the specific form of the Bhatacharyya parameter of (11), we conclude that $\forall\, m = n : \boldsymbol{\Delta}_{m \to n} = 1$. Hence we may pay more attention to the situations of $m \neq n$ during the ensuing analysis.

1) In AWGN Channels: After applying Hölder's inequality [36] to (12), we arrive at[6]

$$\sqrt{\sum_{\substack{i=1, \in \\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}} \mathbf{p}(\mathbf{y} \mid \mathbf{x}[i]) \sum_{\substack{j=1, \in \\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}} \mathbf{p}(\mathbf{y} \mid \mathbf{x}'[j])} \geq \sum_{\substack{i,j=1, \\ (i,j) \in \mathscr{P}}}^{M_{|u}} \sqrt{\mathbf{p}(\mathbf{y} \mid \mathbf{x}[i])\, \mathbf{p}(\mathbf{y} \mid \mathbf{x}'[j])}. \tag{14}$$

The right-hand side of (14) has $M_{|u}$ products. In more detail, we firstly select a pair of filled dots in Fig.2 by ensuring that one comes from the m^{th} CP-Cloud, while the other from the n^{th} CP-Cloud. This pair of synthetic constellation points may be simply labeled by (i, j) and determines one of these product items given by $\mathbf{p}(\mathbf{y} \mid \mathbf{x}[i])\, \mathbf{p}(\mathbf{y} \mid \mathbf{x}'[j])$. Then, similarly, we select the second pair of filled dots from the remaining part of the m^{th} CP-Cloud as well as from the remaining part of the n^{th} CP-Cloud, respectively, for formulating another product item. This procedure continues, until all the $M_{|u}$ pairs of (i, j) have become specified. Obviously, a set of $M_{|u}$ pairs of (i, j) obtained throughout the above-mentioned process completes a point-to-point pairing between two different CP-Clouds of Fig. 2, yielding the set of $\mathscr{P}$ in (14). The number of all possible pairing patterns is identical to that of all possible permutations of $M_{|u}$ objects, which is given by the factorial $M_{|u}!$.

Since AWGN channels are considered, both $\mathbf{h}$ and $\mathbf{p}(\mathbf{h})$ in (12) may be fixed to unity. Hence, after substituting the

[5] In (10), the complete form of $\sum_{\substack{m=1, \\ \in \mathscr{X}_u}}^{M_u} \sum_{\substack{n=1, \\ \in \mathscr{X}_u}}^{M_u}$ is actually $\sum_{\substack{m=1, \\ \mathbf{x}_u[m] \in \mathscr{X}_u}}^{M_u} \sum_{\substack{n=1, \\ \mathbf{x}'_u[n] \in \mathscr{X}_u}}^{M_u}$. The subscript of the summations was simplified here for the sake of arriving at a compact formula that fits into a single journal column. This operation will be also applied to other similar equations.

[6] We have considered a range of classical inequalities and found that Hölder's inequality provides the tightest bound in the set considered.

$$\mathbf{\Delta}_{m\to n} = \int \frac{\mathbf{p}(\mathbf{h})}{M_{|u}} \int \sqrt{\underbrace{\sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}} \mathbf{p}\left(\mathbf{y} \mid \mathbf{h}, \mathbf{x}[i]\right)}_{\text{from m}^{\text{th}}\text{ CP}-\text{Cloud}} \underbrace{\sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}} \mathbf{p}\left(\mathbf{y} \mid \mathbf{h}, \mathbf{x}'[j]\right)}_{\text{to n}^{\text{th}}\text{ CP}-\text{Cloud}}} d\mathbf{y}\, d\mathbf{h}, \tag{12}$$

$$= \int \frac{\mathbf{p}(\mathbf{h})}{M_{|u}} \int \sqrt{\sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}} \sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}} \mathbf{p}\left(\mathbf{y} \mid \mathbf{h}, \mathbf{x}[i]\right) \mathbf{p}\left(\mathbf{y} \mid \mathbf{h}, \mathbf{x}'[j]\right)} d\mathbf{y}\, d\mathbf{h}. \tag{13}$$

inequality of (14) into (12), we arrive at

$$\begin{aligned}
&\mathbf{\Delta}_{m\to n} \\
&\geq \frac{1}{M_{|u}} \sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{P}}}^{M_{|u}} \int \sqrt{\mathbf{p}\left(\mathbf{y} \mid \mathbf{x}[i]\right) \mathbf{p}\left(\mathbf{y} \mid \mathbf{x}'[j]\right)}\, d\mathbf{y} \\
&= \frac{1}{M_{|u}} \sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{P}}}^{M_{|u}} \int \sqrt{\frac{1}{\left(\pi N_0\right)^{2\Omega}} e^{-\frac{|\mathbf{y}-\mathbf{x}[i]|^2+|\mathbf{y}-\mathbf{x}'[j]|^2}{N_0}}}\, d\mathbf{y} \qquad (15) \\
&= \frac{1}{M_{|u}} \sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{P}}}^{M_{|u}} e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}} \int \frac{1}{\left(\pi N_0\right)^{\Omega}} e^{-\frac{\left|\mathbf{y}-\frac{\mathbf{x}[i]+\mathbf{x}'[j]}{2}\right|^2}{N_0}}\, d\mathbf{y} \\
&= \frac{1}{M_{|u}} \sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{P}}}^{M_{|u}} e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}, \qquad (16)
\end{aligned}$$

where (15) is based on the fact that $\mathbf{p}\left(\mathbf{y} \mid \mathbf{x}\right)$ is the joint PDF of Ω multivariate Gaussian random variables. These multivariate Gaussian random variables have been referred to in the sentences below (4) and $N_0 = \sigma^2$. Then, with the aid of (16), we are capable of obtaining a range of lower bounds of $\mathbf{\Delta}_{m\to n}$, where each of them is associated with a particular pairing pattern $\mathscr{P}$. However, in order to derive the tightest possible bound, we have to maximize the right-hand side of (16). Hence we may opt for

$$\begin{aligned}
\mathbf{\Delta}_{m\to n}\big|_{\text{AWGN}}^{\text{Lower}} &= \arg\max_{\mathscr{P}} \frac{1}{M_{|u}} \sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{P}}}^{M_{|u}} e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}} \\
&= \frac{1}{M_{|u}} \sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{P}_{m,n}}}^{M_{|u}} e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}, \qquad (17)
\end{aligned}$$

where $\mathscr{P}_{m,n}$ is the optimum point-to-point pairing pattern between the m^{th} and n^{th} CP-Clouds in Fig.2, which maximizes the right-hand side of (16) and hence results in a reasonable lower bound of $\mathbf{\Delta}_{m\to n}$. Furthermore, for a given synthetic SCMA codebook, $\mathscr{P}_{m,n}$ is significantly affected both by the particular choice of the pair of CP-Cloud indices m, n and by the actual SNR value. The method of finding $\mathscr{P}_{m,n}$ will be demonstrated later in Section IV.

Upon substituting (17) into (10), and exploiting the monotonicity of (10), as well as bearing in mind that $\forall\, m = n : \mathbf{\Delta}_{m\to n} = 1$, we obtain an upper-bound of the cutoff rate R_0 for the SCMA downlink operating in AWGN channels, which may be formulated as (18).

Then, for the sake of deriving the lower bound of the cutoff rate, we may invoke the following simple inequality

$$\begin{aligned}
&\sqrt{\sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}} \sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}} \mathbf{p}\left(\mathbf{y} \mid \mathbf{x}[i]\right) \mathbf{p}\left(\mathbf{y} \mid \mathbf{x}'[j]\right)} \\
&\leq \sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}} \sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}} \sqrt{\mathbf{p}\left(\mathbf{y} \mid \mathbf{x}[i]\right) \mathbf{p}\left(\mathbf{y} \mid \mathbf{x}'[j]\right)}. \qquad (19)
\end{aligned}$$

After substituting (19) into (13) and fixing $\mathbf{h}, \mathbf{p}\left(\mathbf{h}\right)$ to unity for AWGN channels, we arrive at

$$\begin{aligned}
&\mathbf{\Delta}_{m\to n} \\
&\leq \frac{1}{M_{|u}} \sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}} \sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}} \int \sqrt{\mathbf{p}\left(\mathbf{y} \mid \mathbf{x}[i]\right) \mathbf{p}\left(\mathbf{y} \mid \mathbf{x}'[j]\right)}\, d\mathbf{y}, \\
&= \frac{1}{M_{|u}} \sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}} \sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}} e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}, \qquad (20)
\end{aligned}$$

where the derivation of the integral invoked in (20) is the same as that in (15), (16). Similarly, after substituting (20) into (10), and exploiting its monotonicity as well as the fact of $\forall\, m = n : \mathbf{\Delta}_{m\to n} = 1$, a lower bound of the cutoff rate R_0 is obtained, which may be formulated as (21).

2) In Rayleigh Channels: Similar to the strategy exploited in Section III-B1, Hölder's inequality is applied to the generalized formula of $\mathbf{\Delta}_{m\to n}$ in (12) again. In contrast to the above AWGN scenario, the major difference is constituted by replacing the fixed value of $\mathbf{h}$ by a Rayleigh

$$R_0|^{\text{Upper}}_{\text{AWGN}} = -\log_2\left(\sum_{\substack{m=1,\\ \in\mathscr{X}_u}}^{M_u}\sum_{\substack{n=1,\neq m,\\ \in\mathscr{X}_u}}^{M_u}\frac{1}{M_u^2}\left[\frac{1}{M_{|u}}\sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{P}_{m,n}}}^{M_{|u}} e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}\right]+\frac{1}{M_u}\right). \tag{18}$$

$$R_0|^{\text{Lower}}_{\text{AWGN}} = -\log_2\left(\sum_{\substack{m=1,\\ \in\mathscr{X}_u}}^{M_u}\sum_{\substack{n=1,\neq m,\\ \in\mathscr{X}_u}}^{M_u}\frac{1}{M_u^2}\left[\frac{1}{M_{|u}}\sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}}\sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}} e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}\right]+\frac{1}{M_u}\right). \tag{21}$$

random variable. Hence we arrive at

$$\boldsymbol{\Delta}_{m\to n} \geq \frac{1}{M_{|u}}\times \sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{R}}}^{M_{|u}}\int \mathbf{p}(\mathbf{h})\left[\int\sqrt{\mathbf{p}\left(\mathbf{y}\mid\mathbf{h},\mathbf{x}[i]\right)\mathbf{p}\left(\mathbf{y}\mid\mathbf{h},\mathbf{x}'[j]\right)}\,d\mathbf{y}\right]d\mathbf{h} \tag{22}$$

$$= \frac{1}{M_{|u}}\sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{R}}}^{M_{|u}}\int_0^{\infty}\mathbf{p}(\mathbf{h})\cdot e^{-\frac{\mathbf{h}^2|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}\,d\mathbf{h} \tag{23}$$

$$= \frac{1}{M_{|u}}\sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{R}}}^{M_{|u}}\int_0^{\infty}\prod_{\omega}^{\Omega} e^{-\frac{\mathbf{h}_\omega^2|\mathbf{x}[i]_\omega-\mathbf{x}'[j]_\omega|^2}{4N_0}}\cdot\frac{\mathbf{h}_\omega}{\sigma^2}\cdot e^{-\frac{\mathbf{h}_\omega^2}{2\sigma^2}}\,d\mathbf{h}_\omega \tag{24}$$

$$= \frac{1}{M_{|u}}\sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{R}}}^{M_{|u}}\prod_{\omega}^{\Omega}\frac{1}{1+\frac{|\mathbf{x}[i]_\omega-\mathbf{x}'[j]_\omega|^2\cdot 2\sigma^2}{4N_0}}, \tag{25}$$

where the concept of point-to-point pairing between two different CP-Clouds of Fig.2 is employed again in (22). Furthermore, we have $\mathbf{p}(\mathbf{y}|\mathbf{h},\mathbf{x}[i]) = \frac{1}{(\pi N_0)^{\Omega}}\exp\left(-\frac{|\mathbf{y}-\mathbf{h}\cdot\mathbf{x}[i]|^2}{N_0}\right)$. For the sake of avoiding any ambiguity, in Rayleigh fading scenarios, a pairing pattern consisting of $M_{|u}$ pairs of (i,j) is referred to as $\mathscr{R}$, which is discriminated from $\mathscr{P}$ in (16). Then, the inner integral shown on the right-hand side of (22) can be solved with the aid of a similar derivation to that in (15) and (16), where additionally, $\mathbf{h}$ is regarded to be a constant coefficient throughout the inner integration. In (24), the subscript of ω represents the index of the associated dimension. For example, $\mathbf{x}[i]_\omega$ denotes the specific component of $\mathbf{x}[i]$ at the ω^{th} dimension. If we assume a normalized Rayleigh distribution, we arrive at $E\left(\mathbf{h}_\omega^2\right)=2\sigma^2=1$.

Again, substituting different pairing patterns $\mathscr{R}$ into (25) will result in diverse lower bounds. In the same spirit as in (17), we opt for the optimum symbol-pairing pattern of $\mathscr{R}_{m,n}$ for Rayleigh fading scenarios, which is capable of maximizing the right-hand side of (25) and yields the tightest bound as

$$\boldsymbol{\Delta}_{m\to n}|^{\text{Lower}}_{\text{Ray.}} = \frac{1}{M_{|u}}\sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{R}_{m,n}}}^{M_{|u}}\prod_{\omega}^{\Omega}\frac{1}{1+\frac{|\mathbf{x}[i]_\omega-\mathbf{x}'[j]_\omega|^2}{4N_0}}. \tag{26}$$

The method of obtaining $\mathscr{R}_{m,n}$ will be elaborated on later in Section IV.

Now, by substituting (26) into (10), as well as by exploiting the monotonicity of (10), we arrive at a plausible upper-bound of R_0 for the SCMA downlink in Rayleigh channels, which may be formulated as (27).

Then, we proceed by deriving the lower bound of R_0 for Rayleigh fading scenarios. We find that the philosophy invoked for approaching (21) may be exploited again here. To elaborate a little further, when applying the inequality shown in (19), the additional random variable $\mathbf{h}$ is invoked. Hence, for Rayleigh fading scenarios, (20) may be rewritten as

$$\boldsymbol{\Delta}_{m\to n} \leq \frac{1}{M_{|u}}\sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}}\sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}}\prod_{\omega}^{\Omega}\frac{1}{1+\frac{|\mathbf{x}[i]_\omega-\mathbf{x}'[j]_\omega|^2}{4N_0}}, \tag{28}$$

where the derivation of the integral shown in (25) is used again.

Finally, after substituting (28) into (10), as well as by exploiting its monotonicity again, a plausible lower bound of R_0 is obtained for the SCMA downlink in Rayleigh fading channels , which may be formulated as (29).

C. Verification of the Bounds Derived

We will now verify the cutoff rate bounds in the following steps:

- The straightforward relationship between the cutoff rate and the union bound of symbol error probability (SEP) is explicitly outlined.
- Verifying the lower- and upper-bounds of cutoff rate may then be readily transformed to verifying the upper- and lower-bounds of the SEP union bound, respectively.
- The associated verifications will be provided in Fig.3 and in Fig.5.
- The high accuracy of our approximation of the cutoff rate will be evidenced again more directly in Fig.7 and Fig.8.

$$R_0\big|_{\text{Ray.}}^{\text{Upper}} = -\log_2\left(\sum_{\substack{m=1,\\ \in\mathscr{X}_u}}^{M_u}\sum_{\substack{n=1,\neq m,\\ \in\mathscr{X}_u}}^{M_u}\frac{1}{M_u^2}\left[\frac{1}{M_{|u}}\sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{R}_{m,n}}}^{M_{|u}}\prod_{\omega}^{\Omega}\frac{1}{1+\frac{|\mathbf{x}[i]_\omega-\mathbf{x}'[j]_\omega|^2}{4N_0}}\right]+\frac{1}{M_u}\right). \tag{27}$$

$$R_0\big|_{\text{Ray.}}^{\text{Lower}} = -\log_2\left(\sum_{\substack{m=1,\\ \in\mathscr{X}_u}}^{M_u}\sum_{\substack{n=1,\neq m,\\ \in\mathscr{X}_u}}^{M_u}\frac{1}{M_u^2}\left[\frac{1}{M_{|u}}\sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}}\sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}}\prod_{\omega}^{\Omega}\frac{1}{1+\frac{|\mathbf{x}[i]_\omega-\mathbf{x}'[j]_\omega|^2}{4N_0}}\right]+\frac{1}{M_u}\right). \tag{29}$$

As mentioned in Section I, the cutoff rate quantifies the data rate, below which reliable transmission becomes possible. Since $R_0\big|^{\text{Lower}}$ given in (21) or (29) is not higher than R_0, this cutoff rate function remains valid after replacing R_0 by its lower bound of $R_0\big|^{\text{Lower}}$. However, a loose lower bound may impair the practicality of this prediction. Hence the discrepancy between $R_0\big|^{\text{Lower}}$ and R_0 is critical. On the other hand, replacing R_0 by its upper bound of $R_0\big|^{\text{Upper}}$ given in (18) or (27) may result in predicting an optimistic data rate. As stated in Section I, the relationship between the cutoff rate and the system's capacity C obeys $R_0 \leq C$. Hence, if the upper bound $R_0\big|^{\text{Upper}}$ we derived is sufficiently tight, it may only slightly exceed the value of R_0 and still be less than the system's capacity. Hence it may be regarded as a satisfactory approximation of R_0.

According to above analysis, let us now verify, whether the system's reliable data rate predicted by $R_0\big|^{\text{Upper}}$ is still lower than the actually achievable level. If this is true, then the discrepancy between $R_0\big|^{\text{Upper}}$ and $R_0\big|^{\text{Lower}}$ will be further investigated.

As demonstrated in (8), the concept of cutoff rate is exploited by the derivation process of our error probability bound. More explicitly, based on the CP-Cloud concept illustrated in Fig.2 and on the union bound concept [33], we may arrive at

$$\overline{\mathbf{p}_e} \leq \frac{1}{M_u}\sum_{\substack{m=1,\\ \in\mathscr{X}_u}}^{M_u}\sum_{\substack{n=1,\neq m,\\ \in\mathscr{X}_u}}^{M_u}\boldsymbol{\Delta}_{m\to n} = \text{UB}, \tag{30}$$

where $\overline{\mathbf{p}_e}$ represents the actual average multidimensional-symbol error probability (M-SEP) at a MU. Then, the right-hand side of (30) represents the union bound of $\overline{\mathbf{p}_e}$, which may be abbreviated as "UB" for simplicity.

In Section III-B, the inequalities (14), (19) were invoked for reducing or increasing the Bhatacharyya parameter of $\boldsymbol{\Delta}_{m\to n}$, as shown in (16), (20), which further result in the upper or lower bounds of the cutoff rate, respectively. Similarly, if we substitute the same lower- and upper-bounding algebraic operations of $\boldsymbol{\Delta}_{m\to n}$ into (30), we obtain the lower and upper estimates of the UB. In the case of the AWGN scenario, we have

$$\text{UB} \geq \sum_{\substack{m=1,\\ \in\mathscr{X}_u}}^{M_u}\sum_{\substack{n=1,\neq m,\\ \in\mathscr{X}_u}}^{M_u}\frac{1}{M_u}\left[\frac{1}{M_{|u}}\sum_{\substack{i,j=1;\\ (i,j)\in\mathscr{P}_{m,n}}}^{M_{|u}}e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}\right], \tag{31}$$

$$\text{UB} \leq \sum_{\substack{m=1,\\ \in\mathscr{X}_u}}^{M_u}\sum_{\substack{n=1,\neq m,\\ \in\mathscr{X}_u}}^{M_u}\frac{1}{M_u}\left[\frac{1}{M_{|u}}\sum_{\substack{i=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[m]}}}^{M_{|u}}\sum_{\substack{j=1,\in\\ \mathscr{X}_{|\mathbf{x}_u[n]}}}^{M_{|u}}e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}\right]. \tag{32}$$

The right-hand sides of (31) and (32) may be simply termed as $\text{UB}\big|_{\text{AWGN}}^{\text{Lower}}$ and $\text{UB}\big|_{\text{AWGN}}^{\text{Upper}}$, respectively.

According to the close relationship between the system's error performance and cutoff rate as demonstrated in (8), as well as by comparing (31) and (32) to (18) and (21), we may claim that verifying whether the system's reliable data rate predicted by $R_0\big|^{\text{Upper}}$ is still lower than the actually achievable level is equivalent to verifying whether the system's M-SEP predicted by $\text{UB}\big|_{\text{AWGN}}^{\text{Lower}}$ is still higher than the actual performance of $\overline{\mathbf{p}_e}$. These observations reflect the effectiveness of the approximation of $\boldsymbol{\Delta}_{m\to n}$ shown in (16) or (25).

Furthermore, $R_0\big|_{\text{AWGN}}^{\text{Upper}}$ of (18) and $R_0\big|_{\text{Ray.}}^{\text{Upper}}$ of (27) as well as $\text{UB}\big|_{\text{AWGN}}^{\text{Lower}}$ of (31) are all derived by invoking Hölder's inequality to approximate $\boldsymbol{\Delta}_{m\to n}$. Hence, in order to emphasize this primary common characteristic among them, we can readily refer to these bounds as the approximate Chernov bound (Approxi. CB) in the ensuing simulations. Similarly, all the loose bounds, e.g. $R_0\big|_{\text{AWGN}}^{\text{Lower}}$ of (21), $R_0\big|_{\text{Ray.}}^{\text{Lower}}$ of (29), as well as $\text{UB}\big|_{\text{AWGN}}^{\text{Upper}}$ of (32), may be generally termed as relaxed Chernov bounds (Relaxed CB). The system parameters employed during our ensuing Monte-Carlo simulation-based verifications are summarized in Table I. More particularly, for the sake of a fair comparison to the family of orthogonal multiple access (OMA) systems in our forthcoming simulations, the SNR observed in a specific dimension (one orthogonal CR) of a single MU in an OMA system is regarded as our SNR criterion, i.e. we have $\text{SNR} = \frac{P_u}{N_0}$. Correspondingly, owing to introducing the amplification factor f_{am} into (2), the average power of a

signal transmitted by the BS is $\sum_{u=1}^{U} P_u$. This power is distributed among a number of Ω dimensions and we assume in Section II-A that the complex-valued noise imposed on every channel resource dimension obeys $n_\omega \sim \mathcal{CN}(0, N_0)$. We may also assume that every MU has the same available power. Hence, according to the above-mentioned power allocation scheme, the actual SNR observed in a particular dimension of a SCMA MU with respect to an OMA system is given by $\frac{U \cdot P_u}{\Omega \cdot N_0} = f \cdot \mathrm{SNR}$, which is in accordance with the multiplexing and diversity capability of the SCMA system.

	Full-rank	Rank-deficient
Number of MUs	$U = 4$	$U = 6$
Selected MU	The Worst	The Worst
Orthogonal CRs	$\Omega = 4$	$\Omega = 4$
Normalized User-load	$f = 100\%$	$f = 150\%$
MA Technique	SCMA	
Channel Model	AWGN or Rayleigh	
Maximum Samples	10^7 (SCMA Transmission Blocks)	
SNR	P_u/N_0	
Size of Every $\mathscr{X}_u$	$M_u = 4$	

TABLE I: System Parameters

In the spirit of our previous analysis in the early part of Section III-C, the theoretical results of M-SEP are verified by simulations in Fig.3, where again, "Relaxed CB" represents the theoretical M-SEP predicted by $\mathrm{UB}|_{\mathrm{AWGN}}^{\mathrm{Upper}}$ and "Approxi. CB" is the one given by $\mathrm{UB}|_{\mathrm{AWGN}}^{\mathrm{Lower}}$. Moreover, "Prac. ML Det." indicates the actual performance obtained by ML detection.

Observe in Fig.3 that throughout the entire SNR region, the theoretical M-SER given by "Approxi. CB" is always higher than the associated simulation-based performance represented by "Prac. ML Det.", regardless of whether the full-rank or the rank-deficient scenario is considered. Hence (31) constitutes an effective upper bound of $\overline{\mathbf{p}_e}$, which confirms the accuracy of the approximation of $\mathbf{\Delta}_{m\to n}$ shown in (16) and implies that the reliable data rate predicted by $R_0|_{\mathrm{AWGN}}^{\mathrm{Upper}}$ of (18) may still be lower than the system's actually achieved rate. In other words, $R_0|_{\mathrm{AWGN}}^{\mathrm{Upper}}$ of (18) will most probably not violate the system's capacity. Hence $R_0|_{\mathrm{AWGN}}^{\mathrm{Upper}}$ may be regarded as a satisfactory approximation of R_0 in AWGN scenarios.

Then, it is demonstrated in Fig.3 that in the same loading scenario, the curve of "Approxi. CB" and that of "Relaxed CB" merge with each other in the high-SNR region of $\mathrm{SNR} \geq 15\,\mathrm{dB}$. But the "Approxi. CB" bound gets significantly more close to the associated practical performance in the moderate and low-SNR regions of $\mathrm{SNR} \leq 13\,\mathrm{dB}$, where it demonstrates higher accuracy. This implies that the bound of $R_0|_{\mathrm{AWGN}}^{\mathrm{Upper}}$ given in (18) will retain its accuracy both in the low- and high- SNR regions. By contrast, the accuracy of $R_0|_{\mathrm{AWGN}}^{\mathrm{Lower}}$ given in (21) erodes in the low-SNR region. This is evidenced in Fig 4, where both the approximation of R_0 given by $R_0|_{\mathrm{AWGN}}^{\mathrm{Upper}}$ and that given by $R_0|_{\mathrm{AWGN}}^{\mathrm{Lower}}$ are

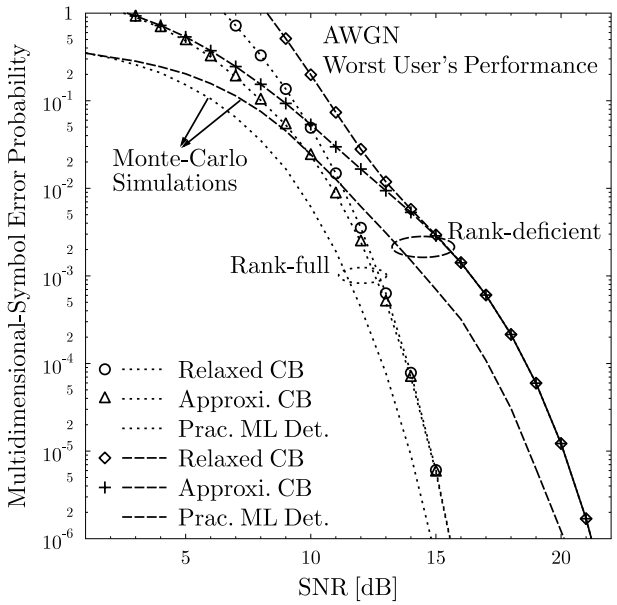

Fig. 3: Theoretical vs Simulation-based M-SEP performance, where the system parameters shown in Table I are adopted. Furthermore, all the "Approxi. CB" curves are calculated according to (31), while those of "Relaxed CB" curves are obtained according to (32).

explicit.

In more detail, for instance, observe in the context of the right-most dotted curve of Fig 4, which corresponds to $R_0|_{\mathrm{AWGN}}^{\mathrm{Lower}}$ in a rank-deficient scenario that it is capable of approaching the right-most dashed curve for $\mathrm{SNR} \geq 15\,\mathrm{dB}$. However, its accuracy rapidly decays upon reducing the SNR. It even suggests having a negative throughput for SNR values below $7\,\mathrm{dB}$, although the curve is not drawn below 0 bits/s.

We conclude that the experimental results recorded in AWGN scenarios are in line with our arguments provided at the beginning of this subsection. Hence we may claim that $R_0|^{\mathrm{Upper}}$ indeed represents a satisfactory approximation of R_0. This conclusion retains its validity also in Rayleigh fading scenarios, which may be verified by the associated simulation results shown in Fig.5 and Fig.6. Explicitly, both the comparison between the theoretical M-SEP and the practical M-SEP shown in Fig.5 and the approximations of the cutoff rate shown in Fig.6 demonstrate trends similar to those exhibited both in Fig.3 and Fig.4. Furthermore, we would like to emphasize that the same system parameters as those listed in Table I are employed again in Fig.5 and Fig.6, except that the AWGN channels are replaced by the Rayleigh fading channels. Correspondingly, the curves labelled by "Approxi. CB" and "Relaxed CB" in Fig.5 are given by the theoretical M-SEP predicted by $\mathrm{UB}|_{\mathrm{Ray.}}^{\mathrm{Lower}}$ and $\mathrm{UB}|_{\mathrm{Ray.}}^{\mathrm{Upper}}$, respectively. Similar to (31) and (32), $\mathrm{UB}|_{\mathrm{Ray.}}^{\mathrm{Lower}}$ and $\mathrm{UB}|_{\mathrm{Ray.}}^{\mathrm{Upper}}$ may be readily obtained by

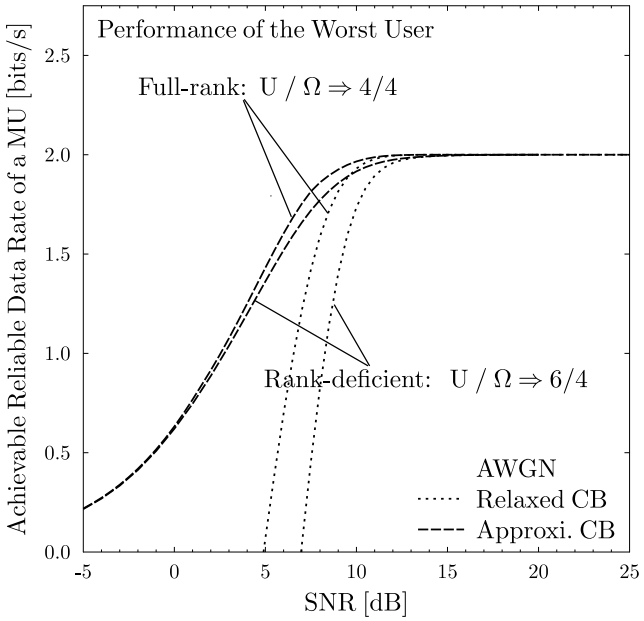

Fig. 4: Estimates of the reliable data rate of a single MU in AWGN channels. The theoretical results calculated according to $R_0\big|_{\text{AWGN}}^{\text{Upper}}$ in (18) are denoted by "Approxi. CB" and those obtained according to $R_0\big|_{\text{AWGN}}^{\text{Lower}}$ in (21) are denoted by "Relaxed CB".

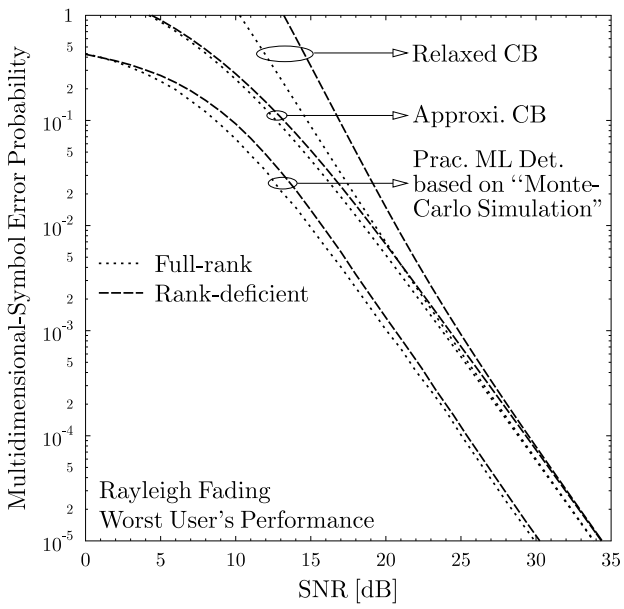

Fig. 5: Theoretical vs simulation results in terms of M-SEP, where the AWGN channels considered in Fig.3 are replaced by Rayleigh fading channels.

simply substituting (25) and (28) into (30), respectively. Therefore, their specific formulae are omitted due to space-limitations.

IV. Insights of Cutoff Rate Derivation

In this section, we firstly visualize both the advantage and disadvantage of the SCMA system by comparing its achievable error-free date rate to both that of the OMA and LDS systems. Then, a beneficial method of generating an appropriate pairing pattern $\mathscr{P}$ for the sake of tightening the upper bound of cutoff rate is provided. Furthermore, the benefit of employing our cloud-style categorization of the pairwise error events is revealed. Finally, the specific impact of the product distance distribution is demonstrated.

A. Comparison among cutoff rates of SCMA, LDS, and OMA

As mentioned in Section I, the cutoff rate constitutes a reliable metric of predicting the superiority or inferiority of different systems. It was also discussed in Section I that LDS constitutes another important NOMA technique, which may in fact be regarded as the root of the SCMA scheme. Hence, we will compare the cutoff rate of a single MU in the downlink of SCMA system both to that of the LDS system, as well as to that of the OMA system.

This comparison is shown both in Fig.7 and in Fig.8, where the system parameters listed in Table I are employed. Herein, the concept of both the regular LDS and that of the

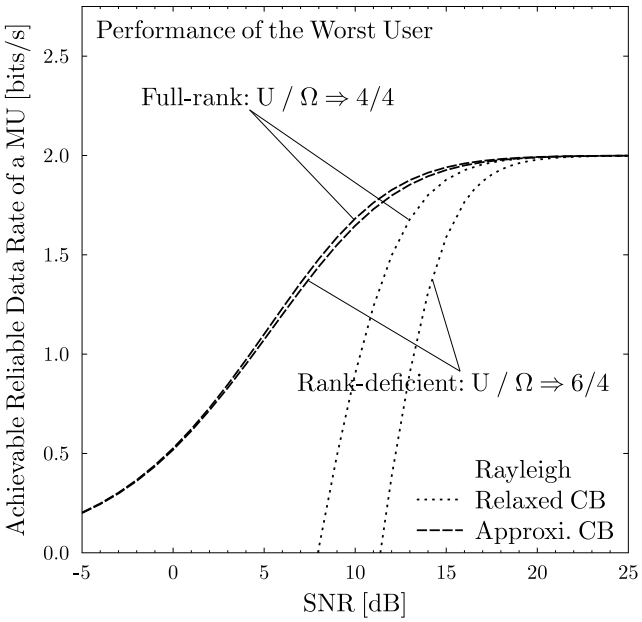

Fig. 6: Estimates of the reliable data rate of a single MU in Rayleigh fading channels. The theoretical results calculated according to $R_0\big|_{\text{Ray}}^{\text{Upper}}$ in (27) are denoted by "Approxi. CB" and those obtained according to $R_0\big|_{\text{Ray}}^{\text{Lower}}$ in (29) are denoted by "Relaxed CB".

irregular LDS introduced in [31] is employed, which are abbreviated as "Re. LDS" and "Irre. LDS", respectively. Then, the classic 4-ary QAM constellation is employed by the MU of the OMA system for AWGN channels. By contrast, in Rayleigh fading channels, the conventional 4-ary QAM is replaced by its rotated version as advocated in [25, Fig.1]. Furthermore, the reliable data rate of every multiple access regime involved in Fig.7 and Fig.8 is evaluated according to the same philosophy as demonstrated in (21).

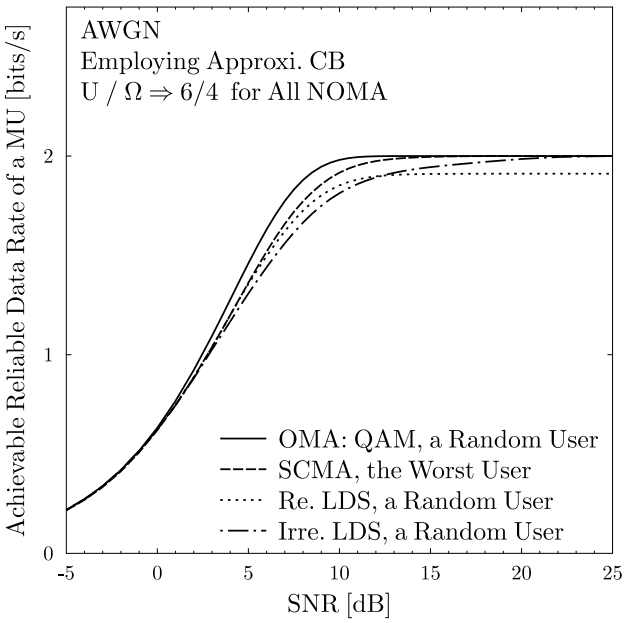

Fig. 7: Comparison among SCMA, LDS and orthogonal multiple access (OMA) in the scenario of AWGN channels, where the reliable data rate of a single MU in each system is adopted as our metric, which is calculated by exploiting (21).

It is demonstrated in Fig.7 that SCMA outperforms both the regular LDS and the irregular LDS right across the entire SNR region both in AWGN and in Rayleigh fading channels. However, none of the NOMA techniques is capable of outperforming the OMA technique in terms of a single MU's performance. For example, as shown in Fig.8, even though the performance of SCMA gets close to that of OMA in Rayleigh fading channels, it still cannot exceed its orthogonal counterpart. The fundamental reason behind this trend will be revealed in Section IV-B1.

B. Seek of an Appropriate Pairing Pattern for Tightening the Upper Bound of Cutoff Rate

As stated in Section III-B, the accuracy of our approximation of the cutoff rate is determined by the specific selection of the pairing pattern, namely of $\mathscr{P}$ or $\mathscr{R}$ for AWGN or Rayleigh fading channels, respectively. However, the search for the optimum pairing pattern defined in (17) or in (26) would impose an excessive computational

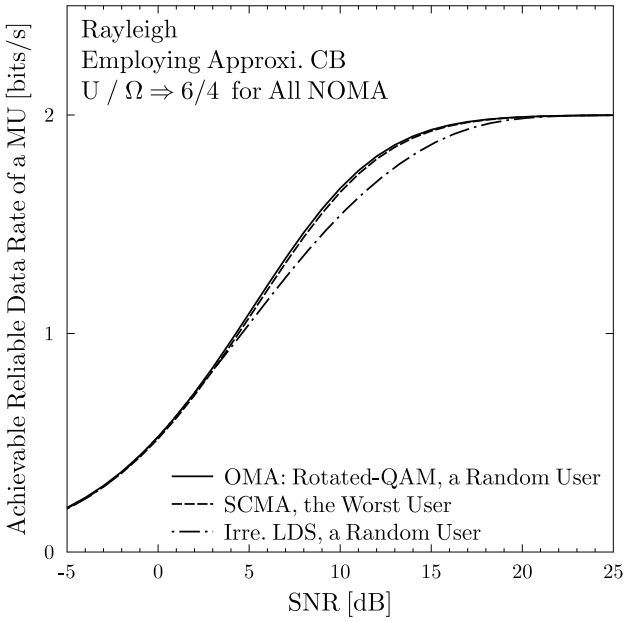

Fig. 8: Comparison among SCMA, LDS and OMA in the scenario of Rayleigh fading channels, where the reliable data rate of a single MU in each system is adopted as our metric, which is calculated by exploiting (21).

complexity. Hence meritorious near-optimum solutions will be discussed in this subsection. We will first consider the AWGN scenario and then extend the associated methods to the Rayleigh fading scenario. Furthermore, a range of conclusions related to the SCMA system and the particular choice of the pairing pattern will also be drawn.

Firstly, let us focus our attention on the AWGN scenario. As defined in (17), $\mathscr{P}_{m,n}$ represents the optimum symbol-wise pairing pattern between the m^{th} and n^{th} CP-Clouds in Fig.2, which aims for maximizing the right-hand side of (16). Choosing the appropriate pairing pattern constitutes a critical factor in terms of generating a satisfactory approximation of the cutoff rate, as evidenced in Fig.3 and Fig.4.

In order to clarify the method of determining $\mathscr{P}_{m,n}$, let us observe Fig.2 again. Recall that there are $M_{|u}$ solid dots in every CP-Cloud, which correspond to the elements of every sub-alphabet of $\mathscr{X}_{|\mathbf{x}_u[m]}, m = 1, 2, \ldots$ and are also denoted as $\mathbf{x}[i]$ or $\mathbf{x}'[j]$ from (12) onward. We may index the solid dots of every CP-Cloud in the same order, which implies that a pair of dots pertaining to different CP-Clouds but having the same index within the cloud will also have the same component of $\sum_{\substack{v=1 \\ v\neq u}}^{U} \mathbf{x}_v$ in (5). As a result, $\mathbf{x}[i]$ and $\mathbf{x}'[j]$ involved in (17) exhibits the explicit property that

$$\left|\mathbf{x}[i] - \mathbf{x}'[i]\right|^2 = f_{\text{am}}^2 \cdot \left|\mathbf{x}_u[m] - \mathbf{x}_u[n]\right|^2, \qquad (33)$$

which means that the Euclidean distance between a pair of solid dots having the same index equals to the Euclidean distance between their cloud centers. This Euclidean distance is also termed as "Distance of Cloud-Centers" and

illustrated in Fig.2.

Then, we invoke a squared matrix $\mathbf{D}_{\rm Euc} \in \mathbb{R}^{M_{|u} \times M_{|u}}$, whose elements at the i^{th} row, j^{th} column are given by

$$\mathbf{D}_{\rm Euc}(i,j) = e^{-\frac{|\mathbf{x}[i]-\mathbf{x}'[j]|^2}{4N_0}}, \tag{34}$$

where the expression at the right-hand side comes from (16). Hence the matrix $\mathbf{D}_{\rm Euc}$ has $M_{|u}$ rows and each row index is related to an ordered index of a solid dot in the m^{th} CP-Cloud. Similarly, $\mathbf{D}_{\rm Euc}$ has $M_{|u}$ columns and each column index is related to an ordered index of a solid dot in the n^{th} CP-Cloud.

According to the above stipulations in Section IV-B, the general procedure of constructing a legitimate point-to-point pairing pattern of $\mathscr{P}$, which was described below (14) and above (15), now could be detailed as follows

- An element at the i^{th} row and j^{th} column of $\mathbf{D}_{\rm Euc}$ is selected and consecutively, the entire i^{th} row, as well as j^{th} column of $\mathbf{D}_{\rm Euc}$ are deleted;
- Another element is selected from the remaining part of $\mathbf{D}_{\rm Euc}$, whose size has been reduced to $\left(M_{|u}-1\right)$ rows and $\left(M_{|u}-1\right)$ columns. Consecutively, the above-mentioned deletion is applied to $\mathbf{D}_{\rm Euc}$ again;
- Repeat the last two steps until $\mathbf{D}_{\rm Euc}$ is reduced to a matrix having only a single element.

Apparently, the set of indices of all the elements selected during the above procedure constitutes a possible realization of $\mathscr{P}$. Then, the sum of all the selected elements, namely $\sum_{(i,j)\in\mathscr{P}} \mathbf{D}_{\rm Euc}(i,j)$ is exactly the maximization-based objective function in (17).

1) A Particular Realization of $\mathscr{P}$: During the above-mentioned procedure of constructing a legitimate $\mathscr{P}$, we may always select the specific element located at the diagonal of $\mathbf{D}_{\rm Euc}$, i.e. we let $i = j$ for all selected elements. The resultant pairing pattern of $\mathscr{P}$ may then be specifically denoted by $\mathscr{P}_{\rm Diag}$, which is given by $\mathscr{P}_{\rm Diag} = \{(1,1),(2,2),\ldots,\left(M_{|u},M_{|u}\right)\}$. Therefore, according to (33), each component in the sum of $\sum_{(i,j)\in\mathscr{P}_{\rm Diag}} \mathbf{D}_{\rm Euc}(i,j)$ is actually related to the same Euclidean distance, i.e. to the "Distance of Cloud Centers" of $f_{\rm am}^2 \cdot \left|\mathbf{x}_u[m]-\mathbf{x}_u[n]\right|^2$ shown in Fig.2. Hence, after substituting $\sum_{(i,j)\in\mathscr{P}_{\rm Diag}} \mathbf{D}_{\rm Euc}(i,j)$ into the right-hand side of (31), we can obtain a lower estimate of the UB of a SCMA MU, which is given by

$$\left.{\rm UB}\right|_{\mathscr{P}_{\rm Diag}}^{lower} = \sum_{\substack{m=1,\\ \in\mathscr{X}_u}}^{M_u} \sum_{\substack{n=1,\neq m,\\ \in\mathscr{X}_u}}^{M_u} \frac{1}{M_u} e^{-\frac{f_{\rm am}^2\cdot|\mathbf{x}_u[m]-\mathbf{x}_u[n]|^2}{4N_0}}. \tag{35}$$

It is interesting to point out that the expression on the right-hand side of (35) happens to be the classical UB of a MU in the single-link direct transmission scenario, which employs the amplified constellation of $\mathscr{X}_u$. Since $\left.{\rm UB}\right|_{\mathscr{P}_{\rm Diag}}^{lower}$ is a lower-bound estimate of the practical UB observed by a SCMA MU, **it implies that a MU supported by a SCMA system cannot outperform its orthogonal counterpart while employing the same constellation, power amplification factor and channel resources**. This is also the fundamental reason for the phenomenon observed in Fig.7.

2) Suboptimum Realization of $\mathscr{P}$: It has been clarified below (14) that the number of legitimate patterns for $\mathscr{P}$ equals to $M_{|u}!$. When considering the system parameters listed in Table I, $1024!$ legitimate pairing patterns have to be considered by the procedure of finding $\mathscr{P}_{m,n}$ of (17), which is excessively complex to complete in a brute-force manner.

Alternatively, we attempt to replace $\mathscr{P}_{m,n}$ by one of its near-optimum counterparts, which may be found at the cost of an affordable computational complexity. In more detail, according to the general procedure of determining a legitimate $\mathscr{P}$ introduced earlier in Section IV-B, once an element of $\mathbf{D}_{\rm Euc}(i,j)$ has been selected, its other elements also located in the i^{th} row or j^{th} column, such as $\mathbf{D}_{\rm Euc}(i,k)$ or $\mathbf{D}_{\rm Euc}(t,j)$ have to be excluded in the ensuing selections. $\mathbf{D}_{\rm Euc}(i,j)$ may be the maximum element in the entire set of elements included in either the i^{th} row or the j^{th} column. Then, let $\mathbf{D}'_{\rm Euc}(i,j)$ represent the maximum element of the remaining part of $\mathbf{D}_{\rm Euc}$ after deleting the i^{th} row and j^{th} column. However, $\mathbf{D}_{\rm Euc}(i,k)$ and $\mathbf{D}_{\rm Euc}(t,j)$ may also be a pair of relatively large elements. Hence it is still possible that we have $\mathbf{D}_{\rm Euc}(i,j) + \mathbf{D}'_{\rm Euc}(i,j) < \mathbf{D}_{\rm Euc}(i,k) + \mathbf{D}_{\rm Euc}(t,j)$. In this case, according to the objective function of (17), which aims for maximizing the sum of $\sum_{(i,j)\in\mathscr{P}} \mathbf{D}_{\rm Euc}(i,j)$, it is better to opt for $\mathbf{D}_{\rm Euc}(i,k)$ or $\mathbf{D}_{\rm Euc}(t,j)$ instead of selecting the maximum element $\mathbf{D}_{\rm Euc}(i,j)$ in that step.

Furthermore, we found that the elements in the diagonal of $\mathbf{D}_{\rm Euc}$, which are also used in (35) have moderate values among all the elements of $\mathbf{D}_{\rm Euc}$. Hence we regard $\mathbf{D}_{\rm Euc}(i,i) + \mathbf{D}_{\rm Euc}(j,j)$ as a benchmark and compare the sum of $\mathbf{D}_{\rm Euc}(i,i) + \mathbf{D}_{\rm Euc}(j,j)$ to that of $\mathbf{D}_{\rm Euc}(i,j)+\mathbf{D}_{\rm Euc}(j,i)$. Then, the pair of elements resulting in a higher sum is selected. This is equivalent to converting the optimization of a huge set, which has $M_{|u}$ elements to the individual optimization of a range of small constituent sets, where each of these small sets only has two elements. Our algorithm that performs the search for finding a near-optimum pairing pattern is summarized by the pseudo-code of Algorithm 1.

Actually, during our simulations, the optimum pairing pattern $\mathscr{P}_{m,n}$ introduced in Section III-B1 is replaced by its suboptimum approximation of $\mathscr{P}_{\rm Subopt}$, which is obtained according to Algorithm 1. To elaborate a little further, once a pair of elements in $\mathbf{D}_{\rm Euc}$ that satisfies the condition of the 8^{th} line in Algorithm 1 is detected, a performance degradation will be imposed on our SCMA aided MU with respect to an OMA aided MU. This leads to the conclusions provided at the end of Section IV-B1.

These "inferior" elements in $\mathbf{D}_{\rm Euc}$ are due to the pairs of solid dots having a "Short Distance" between two different CP-Clouds, which have been shown in Fig.2. We attribute

Algorithm 1 Search for suboptimum pairing of $\mathscr{P}_{\rm Subopt}$

Inputs: m, n, N_0;
Initialization: $k = 0, \mathbf{D} \in \mathbb{R}^{M_{|u} \times M_{|u}}$;
$\mathbf{D} =$ **Construct-D**$_{\rm Ecu}$ (m, n, N_0);
Benchmark $= 2 \cdot \mathbf{D}(0, 0)$;
while $\mathbf{D}.\text{cols}() + \mathbf{D}.\text{rows}() > 2$ **do**
  $[i, j] =$ **max-element-index** $(\mathbf{D})$;
  if $i! = j$ **then**
    if $\mathbf{D}(i, j) + \mathbf{D}(j, i) >$ Benchmark **then**
      $\mathscr{P}_{\rm Subopt}(k) = \mathbf{D}(i, j); k++$
      $\mathscr{P}_{\rm Subopt}(k) = \mathbf{D}(j, i); k++$
    else
      $\mathscr{P}_{\rm Subopt}(k) = \mathbf{D}(i, i); k++$
      $\mathscr{P}_{\rm Subopt}(k) = \mathbf{D}(j, j); k++$
    $\mathbf{D} =$ **Del-row-col**$(\mathbf{D}, i, j)$;
  else
    $\mathscr{P}_{\rm Subopt}(k) = \mathbf{D}(i, j); k++$
    break
for $i = k$ to $M_{|u}$ **do**
  $\mathscr{P}_{\rm Subopt}(i) =$ Benchmark;
return $\mathscr{P}_{\rm Subopt}$

the existence of these "Short Distances" in Fig.2 to the specific component of $\sum_{\substack{v=1 \\ v \neq u}}^{U} \mathbf{x}_v$ in (5), which distorts a SCMA MU's original constellation of $\mathscr{X}_u$. From this perspective, a CP-Cloud seen in Fig.2 may be regarded as a dispersion of its cloud center. **Hence, intuitively, the criterion of designing a high-performance synthetic SCMA codebook for improving the performance of a specific MU is that of concentrating the dispersed solid dots of a CP-Cloud around the cloud center as close as possible in a specific MU's partitioning diagram**.

3) Extension to Rayleigh Fading Channels: The philosophies introduced in Section IV-B1 and Section IV-B2 may also be directly applied to Rayleigh fading scenarios. In contrast to the AWGN scenario, the only difference for the Rayleigh fading scenario is that a new square-shaped matrix, namely $\mathbf{D}_{\rm Prod} \in \mathbb{R}^{M_{|u} \times M_{|u}}$ is defined for replacing $\mathbf{D}_{\rm Euc}$ in (34), whose element at the i^{th} row and j^{th} column is given by

$$\mathbf{D}_{\rm Prod}(i, j) = \prod_{\omega}^{\Omega} \frac{1}{1 + \frac{|\mathbf{x}[i]_\omega - \mathbf{x}'[j]_\omega|^2}{4N_0}}. \tag{36}$$

Hence similar conclusions may be obtained, but the details are omitted owing to space-limitations.

C. Benefit of Cloud Partitioning of Synthetic SCMA Codebook

It has been evidenced in Section III-C by Fig.3 and Fig. 4 that $R_0\big|_{\rm AWGN}^{\rm Upper}$ in (18) provides a more accurate approximation of the cutoff rate than $R_0\big|_{\rm AWGN}^{\rm Lower}$ in (21), which retains an adequate accuracy even in the low-SNR region. This phenomenon actually benefits from our cloud-style partitioning of our synthetic SCMA codebook illustrated in Fig. 2. The detailed reason behind this will be revealed later in this subsection.

Let us reconsider the average M-SEP of a MU, whose union bound was previously formulated in (30). Let us assume furthermore that a constellation point $\mathbf{x}[i]$ in the synthetic SCMA codebook $\mathscr{X}$ is broadcast by the BS, which carries the u^{th} MU's information symbol of $\mathbf{x}_u[m]$. At u^{th} MU, $\mathbf{x}[i]$ is detected as $\mathbf{x}'[j]$. The u^{th} MU's information symbol in $\mathbf{x}'[j]$ may be denoted by $\mathbf{x}_u[n]$. Once we have $m \neq n$, a pairwise error event that incorrectly converts $\mathbf{x}[i]$ to $\mathbf{x}'[j]$ occurs at the u^{th} MU, which is the classic definition of pairwise error events. According to this definition of pairwise error event and by exploiting the union bound concept [33] again, we obtain the conventional union bound of the average M-SEP for a MU, which may be formulated as

$$\overline{\mathbf{p}_e} \leq \frac{1}{M} \sum_{\mathbf{x}[i] \in \mathscr{X}} \sum_{\substack{\mathbf{x}'[j] \in \mathscr{X}, \\ \mathbf{x}_u[n] \neq \mathbf{x}_u[m]}} e^{-\frac{|\mathbf{x}[i] - \mathbf{x}'[j]|^2}{4N_0}}. \tag{37}$$

It can be readily shown that this union bound is identical to that of $\text{UB}\big|_{\rm AWGN}^{\rm Upper}$ given in (32). Hence, according to the straightforward relationship between $\text{UB}\big|_{\rm AWGN}^{\rm Upper}$ and $R_0\big|_{\rm AWGN}^{\rm Lower}$, which has been discussed in Section III-C, $R_0\big|_{\rm AWGN}^{\rm Lower}$ in (21) may be alternatively defined as a result of still employing the conventional categorization of pairwise error events.

It is widely recognized that the inaccuracy of the union bound in estimating the error probability is dominated by the discrepancy between an original decision region and its extended version in a simplified binary constellation based system. Intuitively, the higher the number of the original decision regions involved in a detection algorithm, the larger the above-mentioned discrepancy becomes. In the system model considered herein, the number of decision regions required by the conventional categorization of pairwise error events is as high as $M_{|u} \cdot (M_u - 1) + 1$. Correspondingly, the actual decision region of every filled dot in Fig. 2 is very small. For example, the actual decision region of $\mathbf{x}'[j]$, namely $A_{\rm conv}$ is represented by a pentagon seen at the bottom-right of Fig. 2. Hence, when applying the union bound to approximate the pairwise error probability of $\mathbf{P}_{\mathbf{x}[i] \to \mathbf{x}'[j]}$, $A_{\rm conv}$ will be replaced by a half area of the entire coordinate plane $A_{\rm half}$. Therefore, the inaccuracy of this approximation is proportional to $|A_{\rm half} - A_{\rm conv}|$. By contrast, the number of decision regions required by the cloud style categorization of pairwise error event is only M_u, which implies that only the pairwise error events such as $\mathbf{x}_u[m] \to \mathbf{x}_u[n]$ are taken into account. Explicitly, the actual decision region of $\mathbf{x}_u[n]$ may be visualized by the cloud centered around $\mathbf{x}_u[n]$, which may be denoted by $A_{\rm cloud}$ that is significantly larger than $A_{\rm conv}$. Consequently, when applying the union bound theorem, the inaccuracy of the approximation becomes proportional to $|A_{\rm half} - A_{\rm cloud}|$.

Since $A_{\text{cloud}} \gg A_{\text{conv}}$, the inaccuracy imposed by employing the union bound theorem will be significantly mitigated. Furthermore, when employing the system parameters listed in Table I, the number of decision regions increases from 4 to 3073, if the could style categorization of the pairwise error event is replaced by the conventional categorization. This is the reason for the poor performance of $R_0\big|_{\text{AWGN}}^{\text{Lower}}$.

The above-mentioned trends and findings remain true also in the context of Rayleigh fading channels. It is also evidenced by the trends of Fig.3, where the SEP bounds approximated by "Approxi. CB" get more close to the Monte-Carlo simulation-based results than the SEP bounds approximated by "Relaxed CB". Hence, based on our discussions in this subsection, we may conclude that our approximation of the cutoff rate of SCMA system benefits substantially from the cloud-based partitioning of the synthetic SCMA codebook in both AWGN and Rayleigh fading scenarios.

D. Distribution of Product Distance in Rayleigh Fading Channels

It has been demonstrated in Fig.5 that in contrast to the full-rank configuration, the performance of the rank-deficient configuration operating in a Rayleigh fading scenario only incurs a slight degradation in the high-SNR regions of $\text{SNR} \geq 25\,\text{dB}$. Below we would like to reveal the main reason behind this phenomenon.

The product distance between two constellation points was defined in [24, (9)] and it is widely recognized that the performance of a constellation in fading scenarios is determined by the distribution of all of its legitimate product distances. Naturally, the specific number and value of the short product distances plays an important role. Hence the distribution of product distances recorded for the synthetic SCMA codebook of either the full-rank or of the rank-deficient configuration is investigated in Fig.9. Particularly, the product distances higher than ten times of the shortest product distance in the same constellation are excluded from the statistics owing to their negligible impact.

Observe in Fig.9 that although the rank-deficient configuration results in a lower minimum product distance than its full-rank counterpart, encountering this minimum product distance has a low probability. Hence, the rank-deficient and full-rank configurations exhibit a similar product distance distribution This fact suggests that the performance of the rank-deficient configuration is expected to be only marginally inferior to that of the full-rank configuration, especially in the relatively high SNR region. This is in line with the situation demonstrated in Fig.5.

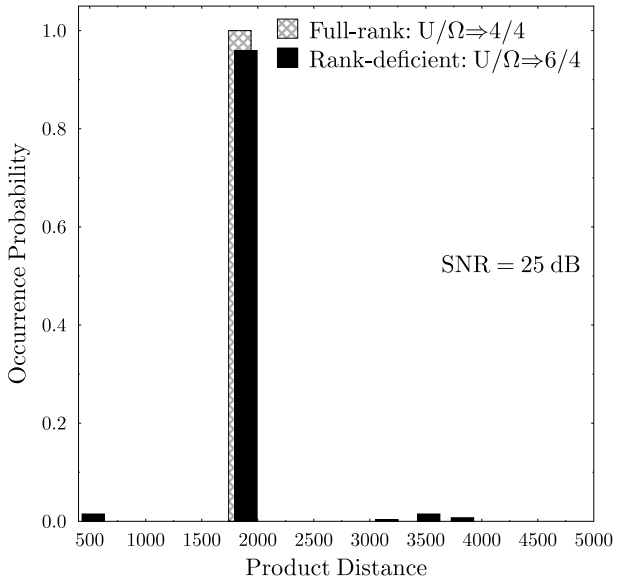

Fig. 9: Distribution of short product distances for the synthetic SCMA codebook of either the full-rank or of the rank-deficient configuration, where $\text{SNR} = 25\,\text{dB}$ is adopted.

V. Conclusions

In this paper, we derived a beneficial approximation of the cutoff rate of SCMA downlink broadcast channels, where both AWGN and Rayleigh fading scenarios have been evaluated. Its accuracy across the entire SNR region has been characterised by the associated theoretical discussions and experimental evidences. Furthermore, a range of insights into our derivations have been provided. For example, some general guidelines of designing a high-performance SCMA codebook have been provided, which were also summarized by sentences in boldface in Section IV. The advantage of SCMA compared to LDS has also been demonstrated. The drawback of applying the conventional categorization of pairwise error events to the cutoff rate derivation in the context of SCMA systems has also been revealed.

References

[1] G. Golub and C. V. Loan, *Matrix Computations*. Baltimore, MD: The John Hopkins University Press, 3rd ed., 1996.

[2] P. Botsinis, D. Alanis, Z. Babar, H. Nguyen, D. Chandra, S. X. Ng, and L. Hanzo, "Quantum-aided Multi-User Transmission in Non-Orthogonal Multiple Access Systems," *IEEE Access*, vol. PP, pp. 1–1, July 2016.

[3] L. Welch, "Lower Bounds on the Maximum Cross-Correlation of Signals," *IEEE Transactions on Information Theory*, vol. 20, pp. 397–399, 1974.

[4] M. Rupf and J. Massey, "Optimum Sequences Multisets for Synchronous Code-Division Multiple-Access Channels," *IEEE Transactions on Information Theory*, vol. 40, pp. 1261–1266, Jul. 1994.

[5] P. Viswanath and V. Anantharam, "Optimal Sequences and Sum Capacity of Synchronous CDMA Systems," *IEEE Transactions on Information Theory*, vol. 45, pp. 1984–1991, Sep. 1999.

[6] L. L. Yang and L. Hanzo, "Slow Frequency-Hopping Multicarrier DS-CDMA for Transmission over Nakagami Multipath Fading Channels," *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 7, pp. 1211–1221, 2001.

[7] F. Vanhaverbeke and M. Moeneclaey, "Optimal Signature Sets for Oversaturated Quasi-Scalable Direct-Sequence Spread-Spectrum Systems," *IEEE Transactions on Information Theory*, vol. 51, pp. 1136–1139, Mar. 2005.
[8] G. Rajappan and M. L. Honig, "Signature Sequence Adaptation for DS-CDMA with Multipath," *IEEE Journal on Selected Areas in Communications*, vol. 20, pp. 384–395, Feb. 2002.
[9] W. Santipach and M. L. Honig, "Signature Optimization for CDMA with Limited Feedback," *IEEE Transactions on Information Theory*, vol. 51, pp. 3475–3492, Oct. 2005.
[10] S. Verdú, "Minimum Probability of Error for Asynchronous Gaussian Multiple-Access Channels," *IEEE Transactions on Information Theory*, vol. 32, pp. 85–94, Jan. 1986.
[11] R. Lupas and S. Verdú, "Linear Multiuser Detectors for Synchronous Code-Division Multiple-Access Channels," *IEEE Transactions on Information Theory*, vol. 35, pp. 123–136, Jan. 1989.
[12] M. Tsatsanis, "Inverse Filtering Criteria for CDMA Systems," *IEEE Transactions on Signal Process*, vol. 45, pp. 102–112, Jan. 1997.
[13] H. V. Poor and S. Verdú, "Probability of Error in MMSE Multiuser Detection," *IEEE Transactions on Information Theory*, vol. 43, pp. 858–871, May. 1997.
[14] P. Patel and J. Holtzman, "Analysis of a Simple Successive Interference Cancellation Scheme in a DS/CDMA System," *IEEE Journal on Selected Areas in Communications*, vol. 12, pp. 796–807, May. 1994.
[15] M. Varanasi and B. Aazhang, "Near-Optimum Detection for Synchronous Code-Division Multiple-Access System," *IEEE Transactions on Communications*, vol. 39, pp. 725–736, May 1991.
[16] X. Wang and H. V. Poor, "Iterative (Turbo) Soft Interference Cancellation and Decoding for Coded CDMA," *IEEE Transactions on Communications*, vol. 47, pp. 1046–1061, Jul. 1999.
[17] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel Low-Density Signature for Synchronous CDMA Systems Over AWGN Channel," *IEEE Transactions on Signal Processing*, vol. 56, pp. 1616–1626, April 2008.
[18] J. Choi, "Low Density Spreading for Multicarrier Systems," in *Proceedings of the Eighth International Symposium on Spread-Spectrum Techniques and Applications, ISSSTA*, pp. 575–578, Aug.30-Sept.2 2004.
[19] T. J. Richardson and R. L. Urbanke, "The Capacity of Low-Density Parity-Check Codes under Message-Passing Decoding," *IEEE Transactions on Information Theory*, vol. 47, pp. 599–618, Feb. 2001.
[20] R. Gallager, "Low-Density Parity Check Codes," *IEEE Transactions on Information Theory*, vol. 8, pp. 21–28, Jan. 1962.
[21] J. van de Beek and B. M. Popovic, "Multiple Access with Low-Density Signatures," in *Proceedings of Global Telecommunications Conference, GLOBECOM*, pp. 1–6, Nov. 2009.
[22] H. Nikopour and H. Baligh, "Sparse Code Multiple Access," in *Proceedings of IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications, PIMRC*, pp. 332–336, Sept. 2013.
[23] J. H. Conway and N. J. Sloane, *Sphere Packings, Lattices and Groups*. Springer-Verlag, New York, 2nd ed., 1993.
[24] J. Boutros, E. Viterbo, C. Rastello, and J. C. Belfiore, "Good Lattice Constellations for Both Rayleigh Fading and Gaussian Channels," *IEEE Transactions on Information Theory*, vol. 42, pp. 502–518, Mar. 1996.
[25] J. Boutros and E. Viterbo, "Signal Space Diversity: A Power- and Bandwidth-Efficient Diversity Technique for the Rayleigh Fading Channel," *IEEE Transactions on Information Theory*, vol. 44, pp. 1453–1467, Jul. 1998.
[26] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA Codebook Design," in *Proceedings of IEEE 80th Vehicular Technology Conference, VTC2014-Fall*, pp. 1–5, Sept. 2014.
[27] L. Yu, X. Lei, P. Fan, and D. Chen, "An Optimized Design of SCMA Codebook based on Star-QAM Signaling Constellations," in *Proceedings of IEEE International Conference on WCSP*, pp. 1–5, Oct. 2015.
[28] B. Xiao, K. Xiao, S. Zhang, Z. Chen, B. Xia, and H. Liu, "Iterative Detection and Decoding for SCMA Systems with LDPC Codes," in *Proceedings of IEEE International Conference on WCSP*, pp. 1–6, Oct. 2015.
[29] J. Chen, Z. Zhang, S. He, J. Hu, and G. E. Sobelman, "Sparse Code Multiple Access Decoding Based on a Monte Carlo Markov Chain Method," *IEEE Signal Processing Letters*, vol. 23, pp. 639–643, May 2016.
[30] H. Nikopour, E. Yi, A. Bayesteh, K. Au, M. Hawryluck, H. Baligh, and J. Ma, "Scma for Downlink Multiple Access of 5G wireless Networks," in *Proceedings of IEEE Global Communications Conference, GLOBECOM*, pp. 3940–3945, Dec. 2014.
[31] M. Cheng, Y. Wu, and Y. Chen, "Capacity Analysis for Non-Orthogonal Overloading Transmissions under Constellation Constraints," in *Proceedings of IEEE International Conference on WCSP*, pp. 1–5, Oct. 2015.
[32] R. G. Gallager, "A Simple Derivation of the Coding Theorem and Some Applications," *IEEE Transactions on Information Theory*, vol. IT-11, pp. 3–18, Jan. 1965.
[33] J. G. Proakis, *Digital Communications*. McGraw-Hill, 5th ed., 2007.
[34] D. Divsalar and E. Biglieri, "Upper Bounds to Error Probabilities of Coded Systems Beyond the Cutoff Rate," *IEEE Transactions on Communications*, vol. 51, pp. 2011–2018, Dec. 2003.
[35] P. Bergmans, "Random Coding Theorem for Broadcast Channels with Degraded Components," *IEEE Transactions on Information Theory*, vol. 19, pp. 197–207, Mar. 1973.
[36] A. Z. Grinshpan, "Weighted Inequalities and Negative Binomials," *Advances in Applied Mathematics*, vol. 45, pp. 564–606, April. 2010.

Li Li received his Ph.D. degree from Southampton Wireless (SW) Group, School of Electronics and Computer Science, University of Southampton in October 2013. Upon the completion of his Ph.D. degree, he conducted research as a senior research assistant in School of Electronics and Computer Science at the University of Southampton from December 2013 to December 2014, where he participated in the European Union Concerto project. In January 2015, he joined the Provincial Key Lab of Information Coding and Transmission, Southwest Jiaotong University, Chengdu, China, serving as a lecturer.

His research interests include channel coding, iterative detection, non-coherent transmission technologies, cooperative communications, network coding and non-orthogonal multiple access technologies.

Zheng Ma (M'07) received the B.Sc. and Ph.D degrees in communications and information system in 2000 and 2006 respectively from Southwest Jiaotong University. Dr. Ma was a visiting scholar of University of Leeds in UK in 2003. In 2003 and 2005, he was a visiting scholar in Hong Kong University of Science and Technology. From 2008 to 2009, he was a visiting research fellow in department of communication systems, Lancaster University, UK. He is currently a professor in Southwest Jiaotong University, and servers as deputy dean of school of information science and technology. Dr Mas research interests include: information theory & coding, signal design & applications, FPGA/DSP Implementation, and professional mobile radio (PMR). He has published more than 60 research papers in high quality journals and conferences. He is currently the Editor for IEEE Communications Letters. He is also the chairman of ComSoc in IEEE Chengdu section.

Li Wang (S'09-M'10) was born in Chengdu, China, in 1982. He received the Ph.D. degree from the University of Southampton, Southampton, U.K., in 2010. From 2010 to 2012, he conducted research as a Senior Research Fellow with the School of Electronics and Computer Science, University of Southampton. During his academic period, he was involved in a number of projects, such as those from UK's EPSRC, Mobile VCE and Indian-UK Advanced Technology Centre (IU-ATC). In March 2012, he joined the R&D Center of Huawei Technologies in Stockholm, Sweden, and is currently working as a Principle Engineer in Wireless Network Algorithm Lab. He has authored over 40 research papers in IEEE/IET journals and conferences, and also coauthored one JohnWiley/IEEE Press book. He has wide research interests in both radio transmission technology (RTT) and radio resource management (RRM) areas for future wireless communication technologies and networks, including PHY layer modeling, advanced iterative receiver design, noncoherent transmission techniques, link adaptation, power control, scheduling, cross-layer cross-module system design, CoMP, massive MIMO, mmWave as well as communication system intelligentization. He has now been conducting pioneering cross-discipline researches to build next-generation communication systems with artificial intelligence (AI). Upon his significant contributions in this field, he was the recipient of Huawei Individual Contribution Award in 2015 and Furture Star Award in 2017.

Lajos Hanzo FREng, FIEEE, FIET, Fellow of EURASIP, DSc received his degree in electronics in 1976 and his doctorate in 1983. In 2009 he was awarded an honorary doctorate by the Technical University of Budapest and in 2015 by the University of Edinburgh. In 2016 he was admitted to the Hungarian Academy of Science. During his 40-year career in telecommunications he has held various research and academic posts in Hungary, Germany and the UK. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, UK, where he holds the chair in telecommunications. He has successfully supervised 111 PhD students, co-authored 20 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10 000 pages, published 1665 research contributions at IEEE Xplore, acted both as TPC and General Chair of IEEE conferences, presented keynote lectures and has been awarded a number of distinctions. Currently he is directing a 50-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council's Advanced Fellow Grant and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and he offers a range of industrial courses. He is also a Governor of the IEEE VTS. During 2008 - 2012 he was the Editor-in-Chief of the IEEE Press and a Chaired Professor also at Tsinghua University, Beijing. For further information on research in progress and associated publications please refer to http://www-mobile.ecs.soton.ac.uk. Lajos has 30000 citations and an H-index of 68.

Ping Zhi Fan received his PhD degree in Electronic Engineering from the Hull University, UK. He is currently a professor and director of the institute of mobile communications, Southwest Jiaotong University, China. He is a recipient of the UK ORS Award, the Outstanding Young Scientist Award by NSFC, and the chief scientist of a National 973 major research project. He served as general chair or TPC chair of a number of international conferences, and is the guest editor-in-chief, guest editor or editorial member of several international journals. He is the founding chair of IEEE VTS BJ Chapter and the IEEE ComSoc CD Chapter, founding chair of IEEE Chengdu Section. He also served as a board member of IEEE Region 10, IET(IEE) Council and IET Asia-Pacific Region. He has over 200 research papers published in various academic English journals (IEEE/IEE/IEICE, etc), and 8 books (incl. edited) published by John Wiley & Sons Ltd/RSP etc, and is an inventor of 22 granted PCT and Chinese patents. His current research interests include high mobility wireless communications, 5G technologies, wireless networks for big data, signal design & coding, etc. He is a fellow of IEEE, IET(IEE), CIE and CIC, and is an IEEE VTS Distinguished Lecturer.