Hindawi
Wireless Communications and Mobile Computing
Volume 2018, Article ID 8978207, 21 pages
https://doi.org/10.1155/2018/8978207

Research Article

Rate-Adaptive Multiple Access for Uplink Grant-Free Transmission

Neng Ye,[1] **Aihua Wang**,[1] **Xiangming Li**,[1] **Wenjia Liu**,[2] **Xiaolin Hou**,[2] **and Hanxiao Yu**[1]

[1] *School of Information and Electronics, Beijing Institute of Technology, Beijing, China*
[2] *DOCOMO Beijing Communications Laboratories Co., Ltd., Beijing, China*

Correspondence should be addressed to Xiangming Li; xmli@bit.edu.cn

Received 30 November 2017; Accepted 11 January 2018; Published 3 July 2018

Academic Editor: Michel Kadoch

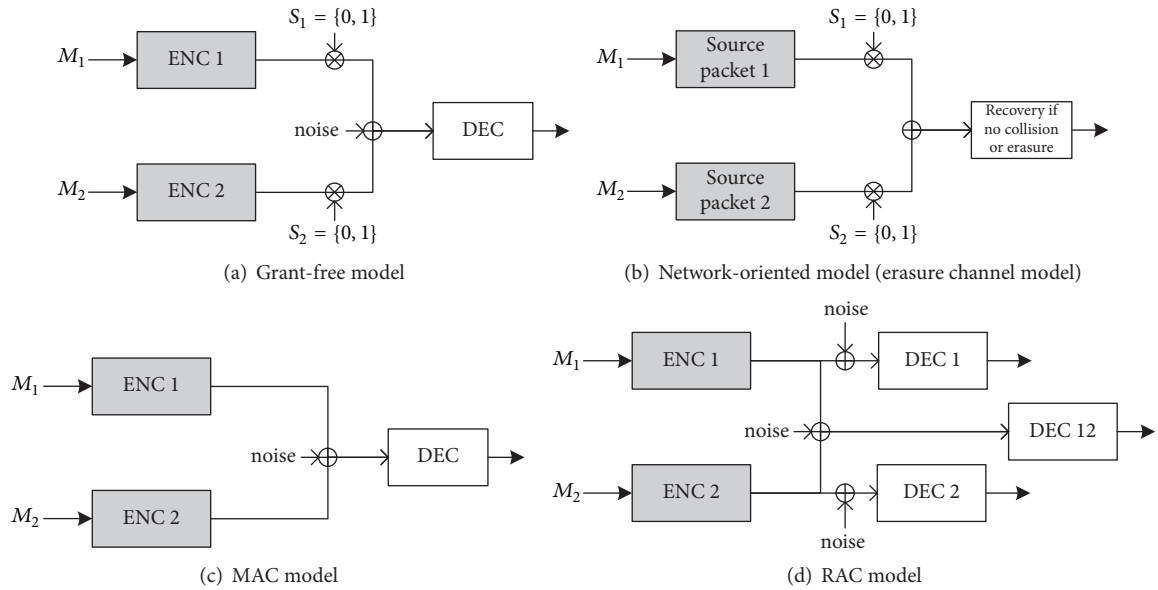
Grant-free transmission, which simplifies the signaling procedure via uplink instant transmission, has been recognized as a promising multiple access protocol to address the massive connectivity and low latency requirements for future machine type communications. The major drawback of grant-free transmission is that the contaminations among uncoordinated transmissions can reduce the data throughput and deteriorate the outage performance. In this paper, we propose a rate-adaptive multiple access (RAMA) scheme to tackle the collision problems caused by the grant-free transmission. Different from the conventional grant-free (conv-GF) scheme which transmits a single signal layer, RAMA transmits the signals with a multilayered structure, where different layers exhibit unequal protection property. At the receiver, the intra- and interuser successive interference cancellation (SIC) receiving algorithm is employed to detect multiple data streams. In RAMA, the users can achieve rate adaptation without the prior knowledge of the channel conditions, since the layers with high protection property can be successfully recovered when the interference is severe, while other layers can take advantage of the channel when the interference is less significant. Besides, RAMA also facilitates the SIC receiving since the multiple layers in the transmission signals can provide more opportunities for interference cancellation. To evaluate the system performance, we analyze the exact expressions of the throughout and the outage probability of both conv-GF and RAMA. Finally, theoretical analysis and simulation results validate that the proposed RAMA scheme can simultaneously achieve higher average throughput and lower outage performance than conv-GF. Meanwhile, RAMA shows its robustness with large user activation probability, where the collisions among users are severe.

1. Introduction

The next generation wireless communication network (5G) is expected to support various diversified usage scenarios with different performance requirements. Specifically, the most important usage scenarios for radio access are categorized into three families by Third-Generation Partnership Project (3GPP): enhanced mobile broadband (eMBB), ultrareliable and low latency communication (uRLLC), and massive machine type communication (mMTC) [1]. While eMBB aims at offering higher peak data rates and higher system throughput in mobile hotspots, the remaining two scenarios focus on the machine type communications, where mMTC is about serving massive devices with small and sporadic packets, and URLLC addresses the applications with very rigorous requirements on latency and reliability. Accordingly, the machine type communications are the extended use cases for 5G, compared with the current 4G system.

Currently, a scheduling request and grant-based access mechanism is employed in uplink data transmission. However, as shown in [2], the grant-based access mechanism expends tens of milliseconds on the signaling intersection, and the signaling overhead ratio approaches nearly a half with small packets (e.g., packets which contain dozens of bytes). Therefore, the latency and overhead requirements cannot be satisfied with grant-based access mechanism.

Recently, the grant-free access mechanism has attracted much attention from both industry and academia [3, 4], where the signaling procedure is greatly simplified. In uplink grant-free access, once the users have data in buffer, they



(a) Grant-free model (b) Network-oriented model (erasure channel model)

(c) MAC model (d) RAC model

Figure 1: Illustrations of MAC, grant-free transmission, and RAC models. M_1 and M_2 are the messages to be transmitted.

instantly transmit their signals on the preconfigured physical resources instead of waiting for grant signaling. Thus, the grant-free access mechanism is especially suitable for uRLLC and mMTC since it reduces the latency and signaling overhead by avoiding complicated signaling intersections between users and the base station (BS). Nevertheless, due to the decentralized access, unexpected collisions may occur in grant-free access and may deteriorate the system performance. As a result, the prospects and challenges of grant-free access have motivated the researchers to make further investigations on the grant-free transceivers to fulfill the requirements of machine type communications in 5G.

1.1. Related Work and Motivation of Our Research. During the past several decades, there have already been two research directions towards the problem of multiuser access in wireless networks, i.e., network-oriented research and Shannon theory oriented research.

In the former research direction, packet transition channel model is assumed, and the researchers mainly focus on the protocol design of media access control layer. One typical example is the slotted ALOHA (SA) protocol proposed for network communications [5]. After that, the listening and backoff based SA is adopted in WIFI system [6]. In more recent years, a class of graphical based SA schemes have been proposed [7], where the random packet transmission is described by a Tanner graph.

In the latter research direction, multiuser information theory is exploited to design capacity achieving multiple access technologies. In the 1970s, the capacity region of uplink multiple access channel (MAC) is derived, known as the Cover-Wyner region [8]. To approach the corner points or the hypotenuse of the Cover-Wyner region, successive interference cancellation (SIC) detection or maximum likelihood (ML) detection should be deployed at the receiver. However, these advanced receivers are not widely accepted by the industry and the academia, due to the concern of high computational complexity, until half a century of Moore's law has made the complexity less noticeable. In the most recent decade, nonorthogonal multiple access (NOMA) technologies, which are promising to achieve the entire Cover-Wyner region, have attracted much attention. By multiplexing different users' signals on the same physical resource and employing advanced detector at the receiver, NOMA possesses higher spectral efficiency and higher overloading compared to the conventional orthogonal multiple access (OMA).

However, none of the above researches can fully describe the grant-free transmission, as pointed out in [9, 10]. We compare the grant-free access model with the access models dealt with in the above two research directions in Figure 1. First of all, a two-user grant-free access model is presented in Figure 1(a), where each user is activated according to a random variable S_i, $i = 1, 2$, and the received signal is polluted by the noise. In network-oriented research, the grant-free access channel is modeled with erasure channel model, as shown in Figure 1(b), where the discontinuous packet arrival can be described, but the underlaid physical layer procedures are ignored; i.e., the noise and the potential near-far effect between two users cannot be modeled. Similarly, the grant-free access cannot be modeled by the MAC model either, as shown in Figure 1(c), since the MAC model assumes full-buffer traffic and neglects the random packet arrivals.

Efforts have been made to provide a universal description of grant-free access by bridging the gap between these two research directions. A two-user random access channel (RAC) model is proposed in [10], as presented in Figure 1(d), where three auxiliary receivers are introduced to represent three different user activation conditions; i.e., user-1 is activated, user-2 is activated, and both users are activated.

The RAC model roots in the Shannon information theory, while it also reflects the burst transmissions of grant-free access. Since the RAC model involves multiple transmitters and multiple receivers, it is similar to the interference channel (IC) model proposed in multiuser information theory [11]. The capacity region of RAC is defined by taking the closure of the unions of achievable rate tuples at user-1 and user-2 with respect to the above receivers, which is a classical information theoretic approach. Obviously, different from in the MAC model, where the NOMA technologies can approach the entire capacity region of MAC, they are not capacity achieving in RAC. As illustrated in [10], rate-splitting technique is required to approach the Shannon limit, at least in two-user RAC.

However, the existing literatures mainly focus on very theoretical cases and do not provide practical designs to incorporate the rate splitting in grant-free transmission. Also they usually assume a grant-free access system with up to two users, which is far from the requirements of 5G. And the advantages are not clarified when the number of users is more than two. Nevertheless, they do provide the insightful hint, that is, to use rate splitting for grant-free users. In this paper, we aim to design a practical multiple access scheme to address the unpredictable interference in grant-free access and to fully utilize the underlaid physical channel. In the meantime, the performance gain of the proposed scheme over conventional grant-free access (conv-GF) is also clarified.

1.2. Contributions

1.2.1. Novel Grant-Free Access Scheme. To combat the unpredictable interference in grant-free access, we propose a novel multiple access scheme, namely, rate-adaptive multiple access (RAMA). The main idea of RAMA is to incorporate rate splitting at the transmitter and employ SIC receiving at the receiver. With rate splitting, the total transmission power is unequally split into several layers and each layer is assigned with an independent codeword, where the multiple layers of a single user hold unequal protection property (UEP). The layers with UEP can be successfully recovered under different levels of interference; i.e., when the collision is high, the layers with high priority can be recovered, while when collision is low, the layers with low priority can take the advantage of the channel. Besides, SIC receiver is employed to mitigate the interference among the layers and the users. In this way, RAMA can enhance the system throughput, while reducing the outage probability simultaneously. Although the grant-free users cannot know the channel conditions in advance, the actual achievable transmission rates can still adapt to the real-time conditions of channel. Thus RAMA actually achieves the rate adaptation, which is similar to the link adaptation in the grant-based transmission. Also, introducing more layers at the transmitter can provide more opportunities for interference cancellation which makes RAMA more robust than conv-GF with high user activation probability.

1.2.2. Clarification on Performance Gain. Another contribution of our work is that we analytically clarify the performance gain of RAMA over conv-GF. The existing literatures have shown that incorporating rate splitting in grant-free access with two users in an information theoretical approach can achieve performance gain. However, it is hard to illustrate the gain with multiple users, since the information theoretical model of being grant-free is rather complicated with more than two transmitters. In this paper, the throughput performance and the outage probability are analyzed with statistical methods, where users are randomly deployed in the cell. We then formulate the exact expressions of the outage probability and throughput of RAMA and conv-GF. Analysis and simulation results reveal that RAMA can achieve higher sum throughput as well as lower outage probability which illustrates that the fairness among users is also improved.

1.2.3. Constellation Design and Parameter Optimization. To facilitate the proposed RAMA scheme, we propose two RAMA amenable constellation design methods, namely, overlapping method and bundling method, respectively. In particular, the constellations, designed by the overlapping method, are composed of several base constellations, where the parameter optimization methods are discussed, including the optimizations of power coefficients and relative rotation angles among the base constellations.

The rest of this paper is organized as follows. Section 2 describes the system and channel model of grant-free transmission, followed by the description of the proposed RAMA scheme in Section 3. Some implementation issues are also discussed in Section 3 as well. Section 4 conducts the theoretical analysis on conv-GF and RAMA. Section 5 provides two design methods on RAMA amenable constellations. The simulation results are presented in Section 6. Finally, the conclusions and the future works are illustrated in Section 7.

2. System Model

Consider a grant-free access network as shown in Figure 2. Suppose that a total of M users are randomly distributed in the cell with the maximum distance R_1 and the minimum distance R_2, and the BS is deployed in the center of the cell. We model packet arrivals at the user side as Poisson distribution and define Poisson arrival rate as γ. In the network, the users are always in inactive mode to save energy if their buffer is empty. Once there are packets in the buffer, the users transfer to active mode and instantly transmit the data packets without grant signaling from the BS. Without loss of generality, we suppose that the BS has full knowledge of user activation information, e.g., via user-specific preamble or simplified random access procedure [10, 12], while the users have no knowledge about other activated users. Note that in this paper we focus on the one-shot grant-free transmission [13] and do not consider the retransmission or hybrid automatic repeat request (HARQ).

At each time index-t, one physical resource block is defined for grant-free access. Define $\mathbf{b}_m^t \in \{0,1\}^k$ as the information bit sequence of an active user-m ($1 \leq m \leq M$) with length-k at time index-t. In conv-GF, $\mathbf{b}_m^t$ is first encoded into a single coded bit sequence $\mathbf{c}_m^t \in \{0,1\}^n$

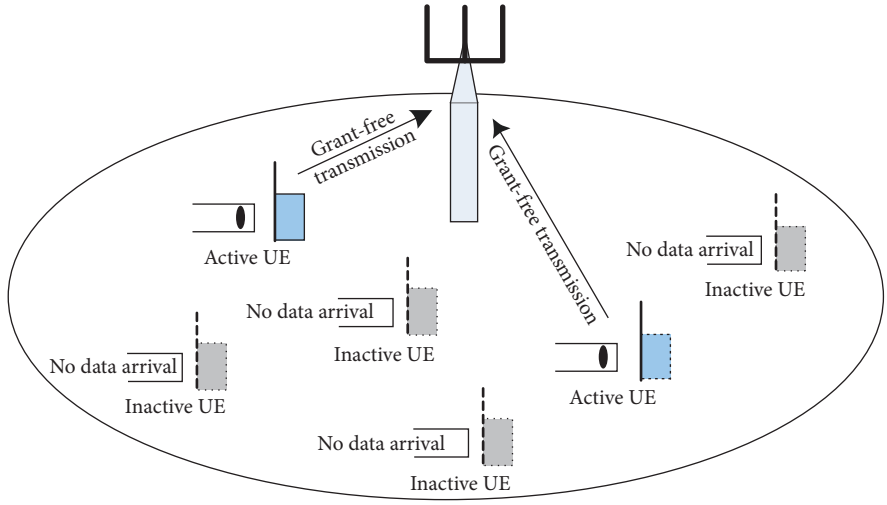


Figure 2: Grant-free access system model.

with coding rate r_m^{conv} and then modulated into a complex symbol sequence $\mathbf{x}_m^t \in \{\mathcal{X}\}^{n/\log_2(|\mathcal{X}|)}$, where $\mathcal{X}$ is the constellation with cardinality $|\mathcal{X}|$. Each symbol sequence occupies the entire physical resource block. Furthermore, we assume Rayleigh block fading channel model, where the Rayleigh distributed small-scale fading coefficient remains a constant within each block, and the fading is independent and identically distributed (i.i.d.) among different blocks or users. Therefore the channel coefficient between the BS and user-m at time index-t is given by $h_m^t = g_m^t/\sqrt{d_m^\alpha}$ [14], where g_m^t is Rayleigh fading coefficient, d_m is the distance between the BS and the user-m, and α is the decay exponent. Without loss of generality, we assume $\alpha = 2$ [15] and assume the transmission power of each user is P. Therefore, the received signal at the BS at time index-t is formulated as follows:

$$\mathbf{y}^t = \sum_{m=1}^{M} I_m^t h_m^t \sqrt{P}\mathbf{x}_m^t + n^t, \quad I_m^t = \begin{cases} 1, & \text{active} \\ 0, & \text{inactive,} \end{cases} \tag{1}$$

where I_m^t indicates the user activation, and n^t is the additive white Gaussian noise with zero mean and variance σ^2.

At the receiver, advanced multiuser detector (MUD) is usually employed to mitigate the mutual interference among the users and to distinguish different users' data streams. For simplification, we define T^t as the normalized throughput of the network which equals the number of successfully transmitted packets at time index-t, and the average normalized throughput is given by $T = E_t\{T^t\}$. The outage is defined as the event that a packet is not successfully decoded in given time index.

In this paper, we assume that the BS deploys the SIC receiver, which is a typical MUD in NOMA [4], to accomplish a good tradeoff between the detection accuracy and the computational complexity. The main idea of SIC is to firstly recover and cancel the data streams with high priority while regarding the other signals as noise and then take advantages of the residual signals. In the conventional grant-based NOMA, the users are scheduled by the BS to deliberately and cautiously multiplex together to ensure low detection error probability. However, as one may expect, the random superposition of signals in grant-free data transmission may not facilitate SIC receiving. Therefore, more elaborate designs should be made at the transmitters to enhance the total throughput and reduce the outage probability of grant-free access system.

3. The Proposed Rate-Adaptive Multiple Access

In grant-based access, each user transmits a codeword in a slot with a certain coding rate derived by channel estimation. However, in grant-free access, the users cannot anticipate the real-time traffic load and may experience unexpected interference from other active users. We illustrate the achievable data rates versus the interference with different grant-free access schemes in Figure 3. Conv-GF adopts the same transmission strategy as in grant-based access, which may lead to performance loss compared to the theoretical limit, as shown by the red arrows in Figure 3(a). When the interference is lower than the threshold, the user cannot fully utilize the potential of channel, and when the interference is higher than the threshold, the data cannot even be successfully recovered. One may expect an ideal grant-free access scheme where the achievable data rate at receiver can automatically adapt to the interference and thus follow the theoretical limits, as shown in Figure 3(b). However, this is not realistic. In this section, we propose a rate-adaptive multiple access (RAMA) scheme for grant-free data transmission, which is based on the rate-splitting technique and can be regarded as the an approximation to the ideal grant-free access as illustrated in Figure 3(c). With this aim, we firstly introduce the rate-splitting technique before presenting the proposed RAMA scheme.

3.1. Rate Splitting. Rate splitting (RS) was originally introduced in the multiuser information theory as a technique

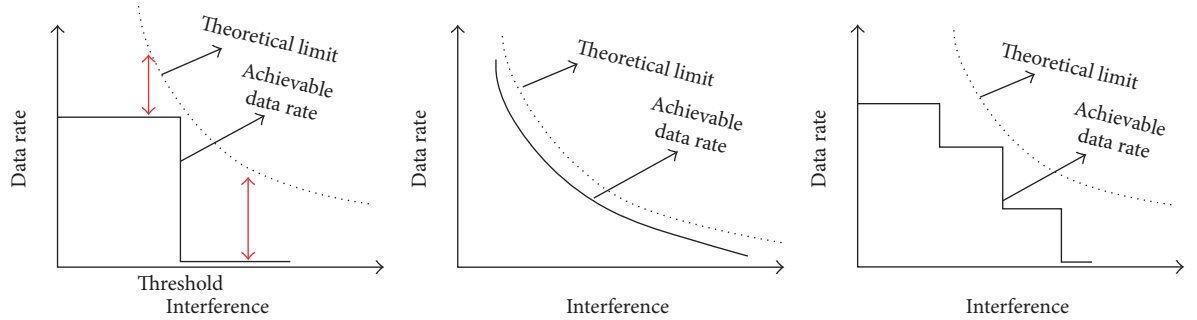


Figure 3: Achievable data rate versus interference, with different grant-free access schemes.

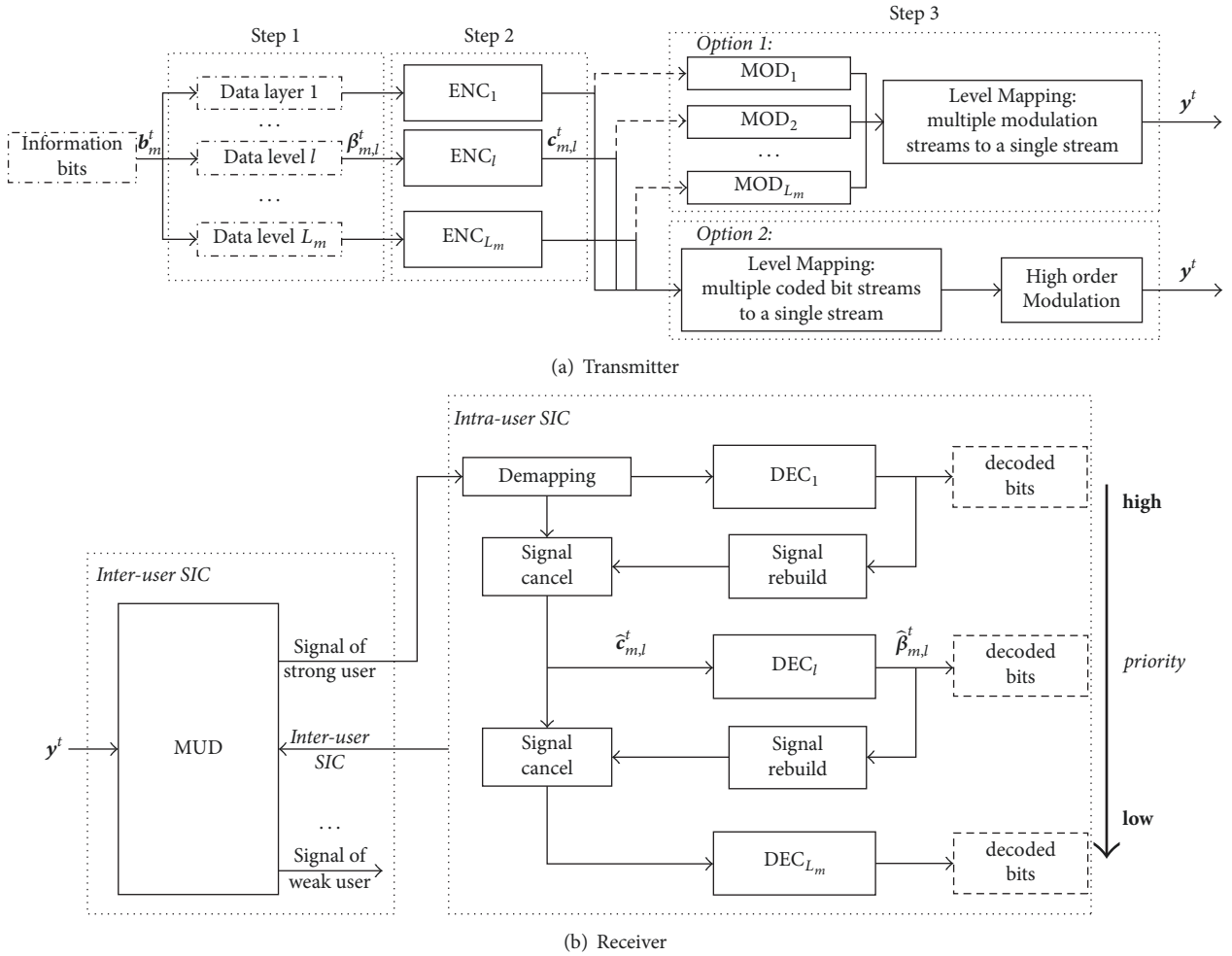


Figure 4: RAMA transmitter and receiver structure.

to prove the capacity bounds of broadcasting channel (BC), multiple access channel (MAC), and interference channel (IC) [16]. The core idea of RS is to split the original message into two or several independent layers and transmit them simultaneously. During these years, RS has attracted the attention of researchers for its potentials to reach every point in MAC [17], to enhance the fairness among the users in the network [18], and to promote the security in MIMO network [19], etc.

3.2. RAMA for Grant-Free Transmission. The proposed RAMA scheme is demonstrated in Figure 4, where RS technique and SIC are adopted at the transmitters and the receiver, respectively. When each user has data in buffer, it

Require: the received signal $\mathbf{y}^t$.
Ensure: the estimated information bits.
(1) Transverse all possible SIC orders and conduct SIC receiving for each SIC order.
(2) Find the optimal SIC order which achieves the largest throughput and output the estimated information bits.

Algorithm 1: Optimal SIC-Based Detection Algorithm.

instantly transmits signals according to the RAMA scheme, as shown in Figure 4(a), with three steps.

Step 1 (data reorganization). At the active user-m, information bit sequence $\mathbf{b}_m^t \in \{0,1\}^n$ is partitioned and reorganized by a bijection B

$$\mathrm{B} : \mathbf{b}_m^t \longmapsto \left(\boldsymbol{\beta}_{m,1}^t, \boldsymbol{\beta}_{m,2}^t, \ldots, \boldsymbol{\beta}_{m,L_m}^t\right), \tag{2}$$

where L_m is the number of layers and $\boldsymbol{\beta}_{m,2}^t$ $(1 \le l \le L_m)$ is the information bit sequence for lth layer. We assign different priority levels to data layers, where layers with higher priority will experience greater protection.

Step 2 (single-layer channel coding). Each subsequence $\boldsymbol{\beta}_{m,l}^t$ is encoded with a channel encoder ENC_l with rate $r_{m,l}^{\mathrm{RAMA}}$

$$\mathrm{ENC}_l : \boldsymbol{\beta}_{m,l}^t \longmapsto \mathbf{c}_{m,l}^t, \tag{3}$$

where $\boldsymbol{c}_{m,l}^t$ is the coded bit sequence. Generally, high priority layers are encoded with low coding rate.

Step 3 (layer-aggregation). Two options can be used to aggregate different layers into one symbol sequence.

In option 1, each coding layer $\boldsymbol{c}_{m,l}^t$ is independently modulated with modulator MOD_l:

$$\mathrm{MOD}_l : \mathbf{c}_{m,l}^t \longmapsto \mathbf{x}_{m,l}^t, \tag{4}$$

where $\mathbf{x}_{m,l}^t$ is the modulation symbol sequence with each element drawn from the constellation $\mathcal{X}_{m,l}$. All layers $\mathbf{x}_{m,l}^t$ $(1 \le l \le L_m)$ are then superimposed together to get the composite constellation symbol sequence $\mathbf{x}_m^t$ with certain power coefficient $\boldsymbol{\lambda}_m^t = [\lambda_{m,1}^t, \lambda_{m,2}^t, \ldots, \lambda_{m,L_m}^t]$ and phase rotation angle $\boldsymbol{\vartheta}_m^t = [\vartheta_{m,1}^t, \vartheta_{m,2}^t, \ldots, \vartheta_{m,L_m}^t]$, where $\mathbf{x}_m^t$ is given by

$$\begin{aligned} \mathbf{x}_m^t &= \sum_{l=1}^{L_m} \lambda_{m,l}^t \mathbf{x}_{m,l}^t e^{j\vartheta_{m,l}^t}, \\ \mathbf{x}_m^t &\in \left\{\mathcal{X}_{(\boldsymbol{\lambda}_m^t, \boldsymbol{\vartheta}_m^t)}\right\}^{n/\log_2(|\mathcal{X}_{(\boldsymbol{\lambda}_m^t, \boldsymbol{\vartheta}_m^t)}|)}, \end{aligned} \tag{5}$$

and $\mathcal{X}_{(\boldsymbol{\lambda}_m^t, \boldsymbol{\vartheta}_m^t)}$ is the composite constellation defined by $\boldsymbol{\lambda}_m^t$ and $\boldsymbol{\vartheta}_m^t$. The layers with higher priority are assigned with larger power coefficients.

In option 2, multiple coding layers are firstly mapped to a single bit stream and then modulated with high-order constellation, similar to the coded modulation, as $(\mathbf{c}_{m,1}^t, \mathbf{c}_{m,2}^t, \ldots, \mathbf{c}_{m,L_m}^t) \longmapsto \mathbf{x}_m^t$, where $\mathbf{x}_m^t \in \{\mathcal{X}\}^{n/\log_2(|\mathcal{X}|)}$. The mapping ensures that the coded bits of higher priority layers hold larger minimum Euclidean distances.

At the receiver, multiuser detection algorithm should be employed to distinguish different users' signals, since multiple users may collide due to the uncoordinated transmission. We show an optimal SIC-based detection algorithm in Algorithm 1, which requires traversing all possible SIC orders. To reduce the computational complexity, we propose a simplified detection algorithm as demonstrated in Algorithm 2. Note that, by employing Algorithm 2, we can simplify the analysis of outage performance in Section 5.

As shown in Figure 3(c), the advantage of RAMA can be intuitively illustrated as follows. In RAMA, the data of each user is transmitted with multiple signal layers, where all the layers have different reliability. As for a certain user, when the external interference is significant, only the signal layer with higher reliability can be solved. Otherwise, when the external interference is not significant, the signal layers with lower reliability can also be successfully recovered. Besides, the layered structure can also mitigate the mutual interference, since the recovered signal layers can be reconstructed and cancelled via SIC receiving. As a result, even if the grant-free user cannot foresee the channel occupancy, each user's actual transmission rate can still adapt to the real-time conditions of channel.

3.3. Implementation Issues

3.3.1. Priority Setting. As mentioned above, the multiple layers in the transmission signals of RAMA exhibit UEP, and the data with disparate priority shall be mapped to corresponding layers. Therefore, the order of the importance of data sets shall be decided in practice. The data sets can be randomly assigned with different priority. Moreover, in some usage scenarios, different data sets naturally have different levels of importance. For example, in mMTC, data sets may contain user identity and application data. The data set containing the user identity is regarded as having high priority. Once the BS knows the user identities, the BS may schedule these users with grant-based transmission to mitigate the collision [20]. Another example happens when grant-free uplink transmission collides with the uplink control information (UCI) on the same resources [21]. In this case, the user may piggyback the UCI report into grant-free data transmission, and the UCI report and grant-free data have different priority; e.g., if the data is for URLLC and the UCI is for eMBB, the former has higher priority.

3.3.2. Frame Structure. The proposed RAMA scheme can be incorporated into existing frame structure designed for grant-free transmission [22]. However, the transmission

Require: the received signal $\mathbf{y}^t$, the number of active user M_{ac}, maximum SIC iteration number S_{max}, maximum layer number L_{max}, and flag f.
Ensure: the estimated information bits.
Set $f = 1$.
(2) **while** $f = 1$ & $1 \le s \le S_{max}$ **do**
Set $f = 0$.
(4) **for** $1 \le l \le L_{max}$ **do**
for $1 \le m \le M_{ac}$ **do**
(6) *Step 1.* Detect the lth signal layer of the user with mth strongest channel gain while regarding interference as noise.
if this signal layer is successfully recovered **then**
(8) *Step 2.* Output the estimated information bits of this signal layer.
Step 3. Reconstruct and cancel this signal layer from $\mathbf{y}^t$.
(10) *Step 4.* Set flag $f = 1$.
end if
(12) **end for**
end for
(14) **end while**

Algorithm 2: The Proposed SIC-Based Detection Algorithm.

block sizes (TBSs) defined for LTE may not satisfy the need of RAMA, since RAMA contains more than one data block in each transmission and some data blocks may only have much less amount of bits. Thus more TBSs should be defined for RAMA.

3.3.3. Retransmission. Due to the UEP of RAMA, some signal layers may not have enough signal to interference and noise ratio (SINR) to be recovered, and retransmission can be employed to make use of the received signals. For the retransmission, either grant-based or grant-free transmission is available depending on specific reliability or latency requirements.

4. Performance Analysis of Conv-GF and RAMA

In this section, we analyze and compare the outage and throughput performance of both conv-GF and RAMA and show the advantages of RAMA.

4.1. Outage Performance Analysis of Grant-Free Access. The performance of conv-GF and RAMA is analyzed in incremental steps. First of all, we study the channel statistics in Lemma 1.

Lemma 1. *For each active user, the probability density function (PDF) of channel gains at time index-t, i.e., $|h^t|^2$, is given by*

$$f_{|h^t|^2}(z) = -\frac{1}{(R_1^2 - R_2^2) z^2}\left(\left(e^{-R_1^2 z/2}\left(R_1^2 z + 2\right)\right) - \left(e^{-R_2^2 z/2}\left(R_2^2 z + 2\right)\right)\right), \quad (6)$$

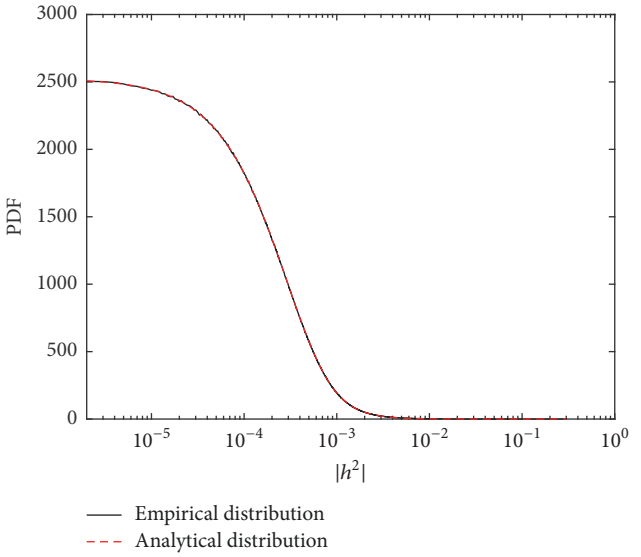


Figure 5: Comparison between empirical and analytical distributions of $|h^t|^2$, with $R_1 = 100$, $R_2 = 10$, and 10^8 samples.

and the cumulative density function (CDF) of $|h_m^t|^2$ is

$$F_{|h^t|^2}(z) = \int_{x=0}^{z} f_{|h_m^t|^2}(x)\, dx = 1 + \frac{2e^{-(R_1^2 z)/2} - 2e^{-(R_2^2 z)/2}}{z\left(R_1^2 - R_2^2\right)}. \quad (7)$$

Proof. Please refer to Appendix A. □

The empirical and analytical distributions of $|h^t|^2$ are compared in Figure 5, which shows a perfect match. In the following, we omit the index t for simplicity.

Denote the event that M_{ac} users become active at time index-t as $K_{M_{\text{ac}}}$, and its probability can be given by

$$P\left(K_{M_{\text{ac}}}\right) = \frac{(\lambda M)^{M_{\text{ac}}} e^{-\lambda M}}{M_{\text{ac}}!}. \tag{8}$$

Further, define $G_{\{|h_1|^2,\ldots,|h_{M_{\text{ac}}}|^2\}}$ as the event where the channel gains of the M_{ac} active users form the set $\boldsymbol{H} = \{|h_1|^2,\ldots,|h_{M_{\text{ac}}}|^2\}$ at time index-t. Note that $G_{\{\ldots,|h_i|^2,\ldots,|h_j|^2,\ldots\}}$ and $G_{\{\ldots,|h_j|^2,\ldots,|h_i|^2,\ldots\}}$ are exactly the same event.

Proposition 2. *The PDF of $G_{\{|h_1|^2,\ldots,|h_{M_{ac}}|^2\}}$ is given by*

$$P\left(G_{\{|h_1|^2,\ldots,|h_{M_{ac}}|^2\}}\right) = M_{ac}! \prod_{m=1}^{M_{ac}} f_{|h|^2}\left(|h_m|^2\right). \tag{9}$$

Proof. For an active user, the PDF of $|h_m|^2$ is $f_{|h|^2}(|h_m|^2)$. Since the channel coefficients of different users are independent, the joint probability density of channel gain vector $(|h_1|^2, |h_2|^2, \ldots, |h_m|^2)$ is $\prod_{m=1}^{M_{\text{ac}}} f_{|h|^2}(|h_m|^2)$. We note that sweeping the elements in $\boldsymbol{H}$ does not make it a distinct event, and there are a total number of $M_{\text{ac}}!$ permutations of $(|h_1|^2, |h_2|^2, \ldots, |h_m|^2)$, which corresponds to the same event $G_{\{|h_1|^2,\ldots,|h_{M_{\text{ac}}}|^2\}}$. Thus the PDF of $G_{\{|h_1|^2,\ldots,|h_{M_{\text{ac}}}|^2\}}$ is given by multiplexing $M_{\text{ac}}!$ with $\prod_{m=1}^{M_{\text{ac}}} f_{|h|^2}(|h_m|^2)$. □

Up to now, the channel gains of M_{ac} users are unordered, and thus analysis with SIC receiver is hard. However, since permuting the elements in $\boldsymbol{H}$ does not change the set itself, we can always assume that the set $\{|h_1|^2,\ldots,|h_{M_{\text{ac}}}|^2\}$ is sorted with $|h_i| < |h_j|$ if $i > j$. The normalization condition of P_G holds as follows:

$$\int_{z_1 \geq \cdots \geq z_{M_{\text{ac}}} \geq 0} P\left(G_{\{z_1,\ldots,z_{M_{\text{ac}}}\}}\right) dz_1 \cdots z_{M_{\text{ac}}} = 1. \tag{10}$$

Few literatures have considered the outage probability of uplink NOMA. The work [14] studied the outage probability of NOMA, where the outage events of the users in the uplink NOMA system are considered to be mutually independent, and the user which is successfully recovered is considered as having no correlation with the remaining users. However, this argument is incomplete, since the outage event of the previous users may indicate the channel conditions of the remaining users. In the following example, we show the argument in [14] is incomplete.

Example 1. Consider a special two-user uplink NOMA system with unit transmission power and unit-variance Gaussian noise. The ordered channel coefficients, namely, h_1 and h_2 ($h_1 > h_2$), may choose values from the set $\{2, 3\}$ with equal probability. Define the outage event of user-i as E_i, $i = 1, 2$, and E_i^c as the complementary set of E_i. Assume that the target data rates of both users are $r_1 = \log_2(1 + 1) = 1$ and $r_2 = \log_2(1 + 3) = 2$, respectively. Then we find that event E_1^c happens only when $h_1 = 3$ and $h_2 = 2$, and, in this case, E_2 must happen. Otherwise, when E_1 happens, we have $h_1 = 3$ and $h_2 = 3$,or $h_1 = 2$ and $h_2 = 2$, where E_2 happens with half probability. As a result, E_1 is correlated to E_2. In this example, we see that the outage events of previous users actually constrain the probability spaces of the channel gains of the rest of the users and thus influence the outage events.

As illustrated in *Example 1*, to get the outage probability of mth user, it is not appropriate to decompose the outage event into several independent events. Instead, we directly deal with the outage event of the active users by applying high dimensional integration in the following derivations. With this aim, we define the following outage events to analyze the outage probability of conv-GF.

Without loss of generality, we assume that all users adopt the same transmission rate; i.e., $r_m^{\text{conv}} = r^{\text{conv}}, \forall m$ [22]. We use $E_{m,M_{\text{ac}}}^{\text{conv}}$ to represent the outage event where the signals of 1st to $(m-1)$th users are successfully recovered, and the signals of mth to M_{ac}th users cannot be recovered. In the following, we assume capacity achieving channel coding and modulation, if not specified. Thus $E_{m,M_{\text{ac}}}^{\text{conv}}$ is given by

$$E_{m,M_{\text{ac}}}^{\text{conv}} \triangleq \left\{\hat{r}_j^{\text{conv}} \geq r_j^{\text{conv}},\ 1 \leq j < m,\ \hat{r}_m^{\text{conv}} < r_m^{\text{conv}}\right\}, \tag{11}$$

where $\hat{r}_j^{\text{conv}} = \log_2(1 + \text{SINR}_j^{\text{conv}})$ and $\text{SINR}_j^{\text{conv}}$ is the received SINR of user-m. According to (11), we readily have the following proposition.

Proposition 3. *The conditional probability of event $E_{m,M_{ac}}^{conv}$ given $G_{\{|h_1|^2,\ldots,|h_{M_{ac}}|^2\}}$ is derived as*

$$P\left(E_{m,M_{ac}}^{conv} \mid G_{\{|h_1|^2,\ldots,|h_m|^2,\ldots,|h_{M_{ac}}|^2\}}\right) = \begin{cases} 1, & \mathscr{C}_{m,M_{ac}}^{conv} \\ 0, & otherwise, \end{cases} \tag{12}$$

$$1 \leq m \leq M_{ac},$$

where $\phi = 2^{r^{conv}} - 1$, $1 \geq i \geq M_{ac}$, and it is given by

$$\mathscr{C}_{m,M_{ac}}^{conv} = \left\{\left(|h_1|^2,\ldots,\left|h_{M_{ac}}\right|^2\right) \mid \frac{\left|h_j\right|^2 P}{\sum_{i=j+1}^{M_{ac}} \left|h_i\right|^2 P + \sigma^2} \geq \phi,\ 1 \leq j < m,\ \frac{\left|h_m\right|^2 P}{\sum_{i=m+1}^{M_{ac}} \left|h_i\right|^2 P + \sigma^2} < \phi\right\}. \tag{13}$$

Averaging (12) over the entire probability space of $G_{\{z_1,\ldots,z_{M_{ac}}\}}^t$, we have

$$P\left(E_m^{conv}\right) = \int_{z_1 \geq \cdots \geq z_{M_{ac}} \geq 0} P\left(E_m^{conv} \mid G_{\{z_1,\ldots,z_{M_{ac}}\}}\right) \times P\left(G_{\{z_1,\ldots,z_{M_{ac}}\}}^t\right) dz_1 \cdots z_{M_{ac}}, \tag{14}$$

and the exact expression of $P(E_{m,M_{ac}}^{conv})$ is given by (15).

Due to the noncontinuous max operations in the integral regions, it is generally difficult to integrate (15) with either numerical intergeneration or approximation. Therefore, the exact expressions of (15) which do not contain the max

operations should be derived. When $\phi < 1$, the exact expressions of $P(E^{conv}_{m,M_{\mathrm{ac}}})$ can be recursively derived as shown in Appendix B, which does not involve the max operations. When $\phi \geq 1$, the exact expression of $P(E^{conv}_{m,M_{\mathrm{ac}}})$ is given by (16). Without loss of generality, we assume $\phi \geq 1$ in the following analysis:

$$\begin{aligned} P\left(E^{conv}_{m,M_{\mathrm{ac}}}\right) &= M_{\mathrm{ac}}! \int_0^{+\infty} f_{|h|^2}\left(z_{M_{\mathrm{ac}}}\right) \\ &\cdot \int_{z_{M_{\mathrm{ac}}-1}=z_{M_{\mathrm{ac}}}}^{+\infty} f_{|h|^2}\left(z_{M_{\mathrm{ac}}-1}\right) \cdots \int_{z_{m+1}=z_{m+2}}^{+\infty} f_{|h|^2}\left(z_{m+1}\right) \\ &\cdot \int_{z_m=z_{m+1}}^{\phi\left(\sum_{j=m+1}^{M_{\mathrm{ac}}} z_j+\sigma^2/P\right)} f_{|h|^2}\left(z_m\right) \\ &\times \int_{z_{m-1}=\max\left(z_m,\phi\left(\sum_{j=m}^{M_{\mathrm{ac}}} z_j+\sigma^2/P\right)\right)}^{+\infty} f_{|h|^2}\left(z_{m-1}\right) \\ &\cdots \int_{z_1=\max\left(z_2,\phi\left(\sum_{j=2}^{M_{\mathrm{ac}}} z_j+\sigma^2/P\right)\right)}^{+\infty} f_{|h|^2}\left(z_1\right) dz_1 \cdots z_{M_{\mathrm{ac}}} \end{aligned} \tag{15}$$

$$\begin{aligned} P\left(E^{conv}_{m,M_{\mathrm{ac}}}\right) &= M_{\mathrm{ac}}! \int_0^{+\infty} f_{|h|^2}\left(z_{M_{\mathrm{ac}}}\right) \\ &\cdot \int_{z_{M_{\mathrm{ac}}-1}=z_{M_{\mathrm{ac}}}}^{+\infty} f_{|h|^2}\left(z_{M_{\mathrm{ac}}-1}\right) \cdots \int_{z_{m+1}=z_{m+2}}^{+\infty} f_{|h|^2}\left(z_{m+1}\right) \\ &\times \int_{z_m=z_{m+1}}^{\phi\left(\sum_{j=m+1}^{M_{\mathrm{ac}}} z_j+\sigma^2/P\right)} f_{|h|^2}\left(z_m\right) \\ &\cdot \int_{z_{m-1}=\phi\left(\sum_{j=m}^{M_{\mathrm{ac}}} z_j+\sigma^2/P\right)}^{+\infty} f_{|h|^2}\left(z_{m-1}\right) \\ &\cdots \int_{z_1=\phi\left(\sum_{j=2}^{M_{\mathrm{ac}}} z_j+\sigma^2/P\right)}^{+\infty} f_{|h|^2}\left(z_1\right) dz_1 \cdots dz_m \cdots z_{M_{\mathrm{ac}}} \end{aligned} \tag{16}$$

However, it is still nontrivial to derive a general and closed-form expression of (16). However, according to the requirements in [5G traffic model], the average number of new packets in each time index is at the level of 10^0, where 2 is a typical value. Therefore, in Appendix C, we derive the compact outage expressions of the outage probabilities for some special cases where active user number is smaller than or equal to 3, i.e., $M_{\mathrm{ac}} \leq 3$, which may constitute the mainstreams of grant-free transmission in the practice. Define the outage event of an active user with conv-GF as E^{RAMA}. Now we are ready to give the expressions of the outage and throughput performance of conv-GF.

Theorem 4. *The average outage probability and the throughput of conv-GF are given by*

$$P\left(E^{conv}\right) = \frac{\sum_{M_{ac}=1}^{\infty} P\left(K_{M_{ac}}\right)\left(\sum_{m=1}^{M_{ac}} P\left(E^{conv}_{m,M_{ac}}\right)\left(M_{ac}-m+1\right)\right)}{\sum_{M_{ac}=1}^{\infty} P\left(K_{M_{ac}}\right) M_{ac}}, \tag{17}$$

$$T^{RAMA} = \sum_{M_{ac}=1}^{\infty}\left(P\left(K_{M_{ac}}\right)\sum_{m=1}^{M_{ac}}\left(P\left(E^{conv}_{m,M_{ac}}\right)\left(mr^{conv}\right)\right)\right), \tag{18}$$

respectively.

Proof. When M_{ac} users are active in a time index, the average amount of the outage users is $\sum_{m=1}^{M_{\mathrm{ac}}} P(E^{\mathrm{conv}}_{m,M_{\mathrm{ac}}})(M_{\mathrm{ac}}-m+1)$. Averaging the above value over (8), we get (17). Similarly, T^{RAMA} is given by multiplexing the transmission rate of users with the successfully recovered users, as derived in (18). □

With the similar approach, the outage probability of RAMA can also be derived as follows:

$$\begin{aligned} \mathscr{C}^{\mathrm{RAMA},(1)}_{\{m_1,m_2\},M_{\mathrm{ac}}} = \Bigg\{ & \left(|h_1|^2,\ldots,\left|h_{M_{\mathrm{ac}}}\right|^2\right) | \\ & \frac{|h_i|^2 \alpha P}{\sum_{j=i+1}^{M_{\mathrm{ac}}}\left|h_j\right|^2 \alpha P+\sum_{j=1}^{M_{\mathrm{ac}}}\left|h_j\right|^2(1-\alpha)P+\sigma^2} \geq \varphi_1, \\ & \frac{\left|h_{m_1}\right|^2 \alpha P}{\sum_{j=m_1+1}^{M_{\mathrm{ac}}}\left|h_j\right|^2 \alpha P+\sum_{j=1}^{M_{\mathrm{ac}}}\left|h_j\right|^2(1-\alpha)P+\sigma^2} < \varphi_1, \\ & \frac{|h_k|^2(1-\alpha)P}{\sum_{j=m_1}^{M_{\mathrm{ac}}}\left|h_j\right|^2 \alpha P+\sum_{j=k+1}^{M_{\mathrm{ac}}}\left|h_j\right|^2(1-\alpha)P+\sigma^2} \geq \varphi_2, \\ & \frac{\left|h_{m_2}\right|^2(1-\alpha)P}{\sum_{j=m_2+1}^{M_{\mathrm{ac}}}\left|h_j\right|^2 \alpha P+\sum_{j=m_1}^{M_{\mathrm{ac}}}\left|h_j\right|^2(1-\alpha)P+\sigma^2} \\ & < \varphi_2,\ 1 \leq i < m_1,\ 1 \leq k < m_2 \Bigg\} \end{aligned} \tag{19}$$

$$\begin{aligned} & \mathscr{C}^{\mathrm{RAMA},(2)}_{\{m_1,m_2\},M_{\mathrm{ac}}} \\ & = \bigcup_{1\leq m_{1,1}<m_1,1\leq m_{2,1}\leq m_2} \Bigg\{ \left(|h_1|^2,\ldots,\left|h_{M_{\mathrm{ac}}}\right|^2\right) | \\ & \mathscr{C}^{\mathrm{RAMA},(1)}_{\{m_{1,1},m_{2,1}\},M_{\mathrm{ac}}}, \\ & \frac{|h_i|^2 \alpha P}{\sum_{j=i+1}^{M_{\mathrm{ac}}}\left|h_j\right|^2 \alpha P+\sum_{j=m_{1,1}}^{M_{\mathrm{ac}}}\left|h_j\right|^2(1-\alpha)P+\sigma^2} \geq \varphi_1, \\ & \frac{\left|h_{m_1}\right|^2 \alpha P}{\sum_{j=m_1+1}^{M_{\mathrm{ac}}}\left|h_j\right|^2 \alpha P+\sum_{j=m_{1,1}}^{M_{\mathrm{ac}}}\left|h_j\right|^2(1-\alpha)P+\sigma^2} \\ & < \varphi_1, \frac{\left(m_2-m_{2,1}\right)|h_k|^2(1-\alpha)P}{\sum_{j=m_1}^{M_{\mathrm{ac}}}\left|h_j\right|^2 \alpha P+\sum_{j=k+1}^{M_{\mathrm{ac}}}\left|h_j\right|^2(1-\alpha)P+\sigma^2} \\ & \geq \left(m_2-m_{2,1}\right)\varphi_2, \end{aligned}$$

$$\frac{\left|h_{m_2}\right|^2(1-\alpha)P}{\sum_{j=m_2+1}^{M_{\mathrm{ac}}}\left|h_j\right|^2\alpha P+\sum_{j=m_1}^{M_{\mathrm{ac}}}\left|h_j\right|^2(1-\alpha)P+\sigma^2} < \varphi_2,\ m_{1,1}\le i<m_1,\ m_{2,1}\le k\le m_2\Bigg\}. \tag{20}$$

4.2. Outage Performance Analysis of RAMA. In this subsection, we analyze the outage performance of RAMA. Without loss of generality, we assume that each user splits its signal into two layers with RAMA, since introducing more layers may lead to severe error propagation [23]. Besides, we assume all users adopt the same transmission procedure with the same coefficients and denote r_1^{RAMA} and r_2^{RAMA} as the transmission rate of the layer-1 and layer-2 at each user, respectively. The power splitting ratio is defined as α for each user; i.e., the transmission power of layer-1 and layer-2 is αP and $(1-\alpha)P$, respectively.

Similar to conv-GF, we denote $E^{\mathrm{RAMA}}_{\{m_1,m_2\},M_{\mathrm{ac}}}$ as the outage event where the 1st to (m_1-1)th user's first layers and the 1st to (m_2-1)th user's second layers are successfully recovered, respectively, while the rest of the layers cannot be recovered. Furthermore, we assume that layer-1 exhibits higher protection than layer-2; i.e., layer-1 can always be successfully detected once layer-2 can be successfully detected. Therefore, $E^{\mathrm{RAMA}}_{\{m_1,m_2\},M_{\mathrm{ac}}}$ is given by

$$E^{\mathrm{RAMA}}_{\{m_1,m_2\},M_{\mathrm{ac}}} \triangleq \Big\{\widehat{r}^{\mathrm{RAMA}}_{j,1}\ge r_1^{\mathrm{RAMA}},\ 1\le j<m_1,\ \widehat{r}^{\mathrm{RAMA}}_{k,2} \ge r_2^{\mathrm{RAMA}},\ 1\le k<m_2,\ \widehat{r}^{\mathrm{RAMA}}_{m_1,1}<r_1^{\mathrm{RAMA}},\ \widehat{r}^{\mathrm{RAMA}}_{m_2,2} < r_2^{\mathrm{RAMA}}\Big\}, \tag{21}$$

where $\widehat{r}^{\mathrm{RAMA}}_{j,l}=\log_2(1+\mathrm{SINR}^{\mathrm{RAMA}}_{j,l})$ and $\mathrm{SINR}^{\mathrm{RAMA}}_{j,l}$ is the received SINR of the lth layer of the jth user. Due to the fact that the 1st layer exhibits higher protection than the 2nd layer, we always have $m_1\ge m_2$. Besides, we define the outage event of active users, namely, E_m^{RAMA}, which happens when at least 1 layer of the 1st to $(m-1)$th users is recovered, and none of the layers of mth to M_{ac}th users is successfully decoded, and E_m^{RAMA} can be readily defined as

$$E_m^{\mathrm{RAMA}} \triangleq \bigcup E^{\mathrm{RAMA}}_{\{m,m_2\},M_{\mathrm{ac}}},\quad 1\le m_2\le m. \tag{22}$$

As mentioned before, Algorithm 1 is the optimal SIC-based multiuser detection algorithm for RAMA. However, since Algorithm 1 involves a traversing operation, which does not have an exact mathematical expression, we study the performance of RAMA by assuming Algorithm 2. Before that, we first show the optimality of the proposed Algorithm 2 by Lemma 5.

Lemma 5. *Assume that all users split and encode their signals with the same power coefficients and the same transmission rates, respectively; the proposed Algorithm 1 is the optimal SIC-based multiuser detection algorithm of RAMA; i.e., Algorithm 1 achieves the same outage and throughput performance as Algorithm 2.*

Proof. First of all, we note that Algorithm 1 traverses all possible successive cancellation orders, and therefore it is the optimal multiuser detection algorithm based on SIC. Next, we show the optimality of Algorithm 2 by contradiction. Assume that the lth layer of the mth user happens to be decoded by Algorithm 1 but not by Algorithm 2. According to the assumptions of this paper, the 1st to lth layers of the 1st to $(m-1)$th users and the 1st to $(l-1)$th layers of the mth user can be successfully recovered by both Algorithms 1 and 2. After cancelling the aforementioned layers, the lth layer of mth users is the most reliable layer among the remaining signal layers. Therefore, Algorithm 1 will recover this layer while regarding other layers as noise, according to the assumption. However, Algorithm 2 can also decode this layer just as Algorithm 1, which contradicts the assumption. After all, Algorithm 2 is optimal. □

To model the effect of SIC receiving, we define $E^{\mathrm{RAMA},(s)}_{\{m_1,m_2\},M_{\mathrm{ac}}}$ as the outage event that, after sth iteration in Algorithm 2, m_1th user's first layer and m_2th user's second layer cannot be successfully decoded, with the 1st to (m_1-1)th user's first layers and the 1st to (m_2-1)th user's second layers, when M_{ac} users are active. In this case, $m_1\ge m_2$.

Proposition 6. *The conditional probability of event $E^{RAMA,1}_{\{m_1,m_2\},M_{ac}}$ given $G_{\{|h_1|^2,\ldots,|h_{M_{ac}}|^2\}}$ is derived as*

$$P\left(E^{RAMA,(1)}_{\{m_1,m_2\},M_{ac}} \mid G_{\{|h_1|^2,\ldots,|h_{M_{ac}}|^2\}}\right) = \begin{cases}1, & \left\{\left(\left|h_1\right|^2,\ldots,\left|h_{M_{ac}}\right|^2\right)\in\mathscr{C}^{RAMA,(1)}_{\{m_1,m_2\},M_{ac}}\right\},\\ 0, & otherwise,\end{cases} \tag{23}$$
$$1\le m_1\le M_{ac},\ 1\le m_2\le M_{ac},$$

and $\mathscr{C}^{RAMA,(1)}_{\{m_1,m_2\},M_{ac}}$ is a region given by (19), where $\varphi_i=2^{r_i^{RAMA}}-1$, $1\le i\le 2$.

Similarly, $P(E^{RAMA,(2)}_{\{m_1,m_2\},M_{ac}} \mid G_{\{|h_1|^2,\ldots,|h_m|^2,\ldots,|h_{M_{ac}}|^2\}})$ is given by

$$P\left(E^{RAMA,(2)}_{\{m_1,m_2\},M_{ac}} \mid G_{\{|h_1|^2,\ldots,|h_m|^2,\ldots,|h_{M_{ac}}|^2\}}\right) = \begin{cases}1, & \left\{\left(\left|h_1\right|^2,\ldots,\left|h_{M_{ac}}\right|^2\right)\in\mathscr{C}^{RAMA,(2)}_{\{m_1,m_2\},M_{ac}}\right\},\\ 0, & otherwise,\end{cases} \tag{24}$$
$$1\le m_1\le M_{ac},\ 1\le m_2\le M_{ac},$$

where, $\mathscr{C}^{RAMA,(2)}_{\{m_1,m_2\},M_{ac}}$ is a region given by (20).

Using mathematical induction, the general expression of $P(E^{RAMA,(s+1)}_{\{m_1,m_2\},M_{ac}} \mid G_{\{|h_1|^2,\ldots,|h_m|^2,\ldots,|h_{M_{ac}}|^2\}})$ *is given by*

$$P\left(E^{RAMA,(s+1)}_{\{m_1,m_2\},M_{ac}} \mid G_{\{|h_1|^2,\ldots,|h_m|^2,\ldots,|h_{M_{ac}}|^2\}}\right) = \begin{cases} 1, & \left\{\left(|h_1|^2,\ldots,\left|h_{M_{ac}}\right|^2\right) \in \mathscr{C}^{RAMA,(s+1)}_{\{m_1,m_2\},M_{ac}}\right\}, \\ 0, & otherwise, \end{cases} \quad (25)$$

$$1 \le m_1 \le M_{ac},\ 1 \le m_2 \le M_{ac},$$

and $\mathscr{C}^{RAMA,(s+1)}_{\{m_1,m_2\},M_{ac}}$ *is a region given by*

$$\mathscr{C}^{RAMA,(s+1)}_{\{m_1,m_2\},M_{ac}} = \bigcup_{1\le m_{1,s}<m_1, 1\le m_{2,s}\le m_2} \left\{\left(|h_1|^2,\ldots,\left|h_{M_{ac}}\right|^2\right) \mid \mathscr{C}^{RAMA,(s)}_{\{m_{1,s},m_{2,s}\},M_{ac}},\right.$$

$$\frac{|h_i|^2 \alpha P}{\sum_{j=i+1}^{M_{ac}} \left|h_j\right|^2 \alpha P + \sum_{j=m_{1,s}}^{M_{ac}} \left|h_j\right|^2 (1-\alpha) P + \sigma^2} \ge \varphi_1,$$

$$\frac{\left|h_{m_1}\right|^2 \alpha P}{\sum_{j=m_1+1}^{M_{ac}} \left|h_j\right|^2 \alpha P + \sum_{j=m_{1,s}}^{M_{ac}} \left|h_j\right|^2 (1-\alpha) P + \sigma^2}$$

$$< \varphi_1, \frac{(m_2 - m_{2,s}) |h_k|^2 (1-\alpha) P}{\sum_{j=m_1}^{M_{ac}} \left|h_j\right|^2 \alpha P + \sum_{j=k+1}^{M_{ac}} \left|h_j\right|^2 (1-\alpha) P + \sigma^2}$$

$$\ge (m_2 - m_{2,s}) \varphi_2,$$

$$\left.\frac{\left|h_{m_2}\right|^2 (1-\alpha) P}{\sum_{j=m_2+1}^{M_{ac}} \left|h_j\right|^2 \alpha P + \sum_{j=m_1}^{M_{ac}} \left|h_j\right|^2 (1-\alpha) P + \sigma^2} < \varphi_2,\ m_{1,s} \le i < m_1,\ m_{2,s} \le k \le m_2 \right\}. \quad (26)$$

$P(E^{RAMA,(s+1)}_{\{m_1,m_2\},M_{ac}})$ *is given by averaging (25) over (9) as follows:*

$$P\left(E^{RAMA,(s+1)}_{\{m_1,m_2\},M_{ac}}\right) = M_{ac}! \sum_{M_{ac}=0}^{\infty} \int_{z_1>\cdots>z_l>0} P\left(E^{RAMA,(s+1)}_{\{m_1,m_2\},M_{ac}} \mid G_{\{z_1,\ldots,z_{M_{ac}}\}}\right) \cdot P\left(G_{\{z_1,\ldots,z_{M_{ac}}\}}\right) dz_1 \cdots z_{M_{ac}}. \quad (27)$$

We define the outage event of an active user, i.e., E^{RAMA}, to be event where none of its signal layers are successfully decoded by the BS. The outage performance of RAMA is shown in Theorem 7.

Theorem 7. *The exact expressions of the average outage probability and the average throughput of RAMA after sth SIC are given by (28) and (29), respectively.*

$$P\left(E^{RAMA}\right) = \frac{\sum_{M_{ac}=1}^{\infty} \left(P\left(K_{M_{ac}}\right) \sum_{m_1=1}^{M_{ac}} \left(\left(\sum_{m_2=1}^{M_{ac}} P\left(E^{RAMA,(s)}_{\{m_1,m_2\},M_{ac}}\right)\right)(M_{ac}-m_1+1)\right)\right)}{\sum_{M_{ac}=1}^{\infty} P\left(K_{M_{ac}}\right) M_{ac}}, \quad (28)$$

$$T^{RAMA} = \sum_{M_{ac}=1}^{\infty} \left(P\left(K_{M_{ac}}\right) \sum_{m_1=1}^{M_{ac}} \left(\sum_{m_2=1}^{M_{ac}} \left(P\left(E^{RAMA,(s)}_{\{m_1,m_2\},M_{ac}}\right)\left(m_1 r_1^{RAMA} + m_2 r_2^{RAMA}\right)\right)\right)\right). \quad (29)$$

Proof. The proof is similar to the proof of Theorem 4. □

As mentioned before, RAMA introduces multiple layers at the transmitters; therefore the power and transmission rates allocation among different layers can act as an additional degree of freedom to match the statistical characteristics of interference and to further enhance the network throughput of RAMA. The optimal power and transmission rates of RAMA are given by

$$\max_{\alpha, r_1^{\mathrm{RAMA}}, r_2^{\mathrm{RAMA}}} \quad T^{\mathrm{RAMA}}, \quad \text{s.t.} \quad r_1^{\mathrm{RAMA}} + r_2^{\mathrm{RAMA}} = r^{\mathrm{conv}}, \quad 0 \le \alpha \le 1. \quad (30)$$

Note that conv-GF only considers a single signal layer, and thus it is not as adjustable as RAMA.

4.3. Comparisons. In this subsection, we compare the outage performance achieved by conv-GF and RAMA. First of all, we note that, by setting $\alpha = 1$, $\varphi_1 = \phi$ and $\varphi_2 = 0$ in (30), the outage and throughput performance of RAMA are exactly the same as conv-GF. Therefore, it is straightforward that RAMA outperforms conv-GF with sophisticatedly designed parameters.

To visually illustrate the advantage of RAMA with respect to the outage performance, we compare the *complementary outage regions* of conv-GF and RAMA, i.e., $(\bigcup_{1\le m\le M_{\mathrm{ac}}} \mathscr{C}^{\mathrm{conv}}_{m,M_{\mathrm{ac}}})^c$ and $(\bigcup_{1\le m_1,m_2\le M_{\mathrm{ac}}} \mathscr{C}^{\mathrm{RAMA},(s)}_{\{m_1,m_2\},M_{\mathrm{ac}}})^c$, with

$M_{\text{ac}} = 2$ and 3, as shown in Figures 6 and 7, respectively. In the simulation, conv-GF and RAMA are set with the same transmission rates which takes value in $\{0.8, 1, 1.2\}$. The transmission rates and the power coefficients of RAMA are optimized according to (30). In Figure 6, the black cycles represent all possible realizations of $\boldsymbol{H}$ with $M_{\text{ac}} = 2$, and the red crosses represent the realizations of $\boldsymbol{H}$ such that the signals of the two users are successfully recovered. We found that, with the increase of total transmission rate, i.e., ϕ, the outage performance of conv-GF becomes worse, while RAMA can achieve successful transmissions with almost all channel realizations. In Figure 7, the black dots represent all possible realizations of $\boldsymbol{H}$ with $M_{\text{ac}} = 3$, and the red cycled points represent the realizations of $\boldsymbol{H}$ such that the three users can be successfully recovered. The advantage of RAMA over conv-GF with respect to the outage performance with $M_{\text{ac}} = 3$ is more significant than $M_{\text{ac}} = 2$, which validate the robustness of RAMA. We also observe that when the interference among users is severe, i.e., the areas selected by cycles in Figures 6(b), 7(b), 6(c), and 7(c), conv-GF cannot ensure successful transmissions, while RAMA still achieves high throughput. To sum up, RAMA achieves high data rate in low interference region, while the robustness can also be assured in high interference region.

5. RAMA Amenable Constellations

In the above analysis, we have considered the ideal situation where Gaussian-distributed continuous alphabet is assumed at the transmitter. However, only finite alphabets can be deployed in practice. Therefore, we focus on the design and optimization of RAMA amenable constellations in this section.

The RAMA amenable constellations are composed of several subconstellations, where each subconstellation corresponds to a signal layer at transmitter. We call the equivalent channel experienced by each signal layer as a *subchannel*, as mentioned in Section 3.2. To facilitate RAMA, our aim is to construct the subchannels such that they have UEP. Corresponding to the two options in Step 3 of the RAMA scheme in Section 3.2, we propose two methods to design RAMA amenable constellations as well as the subchannels in the following.

5.1. Overlapping Method. Corresponding to option 1 in Step 3 of RAMA, we propose the overlapping method, where the composite constellation is constructed by overlapping several base constellations, i.e., $\{\mathcal{X}_{(\boldsymbol{\lambda}_m^t, \boldsymbol{\vartheta}_m^t)}\}^{n/\log_2(|\mathcal{X}_{(\boldsymbol{\lambda}_m^t, \boldsymbol{\vartheta}_m^t)}|)}$, as shown in (6). In Figure 8, we show two examples of the RAMA amenable constellations, where BPSK and QPSK are employed as the basic building blocks. When the power coefficients, i.e., λ_1 and λ_2, of different base constellations are different, the bits in the composite constellation normally have different constellation constrained capacity. Therefore, we regard each bit (or several bits) as a subchannel where one signal layer can be transmitted, as shown in Figures 8(a) and 8(b), respectively. We note that, with the proposed constellations, even all signal layers are encoded with the same coding rate (to save the hardware resources), and they still exhibit UEP property, which facilitates the SIC receiving of RAMA.

Furthermore, the power coefficients, i.e., $\boldsymbol{\lambda}_m^t$, and rotation angles, i.e., $\boldsymbol{\vartheta}_m^t$ shall be optimized to adapt to RAMA. In this paper, we sequentially optimize these coefficients. The optimization of $\boldsymbol{\lambda}_m^t$ includes the following three steps.

Step 1. Define different reliability levels for different subchannels, e.g., different BLER targets for the codewords transmitted in different subchannels.

Step 2. Map the reliability levels to the capacity of different subchannels.

Step 3. Adjust $\boldsymbol{\lambda}_m^t$ to meet the capacity requirements of different subchannels.

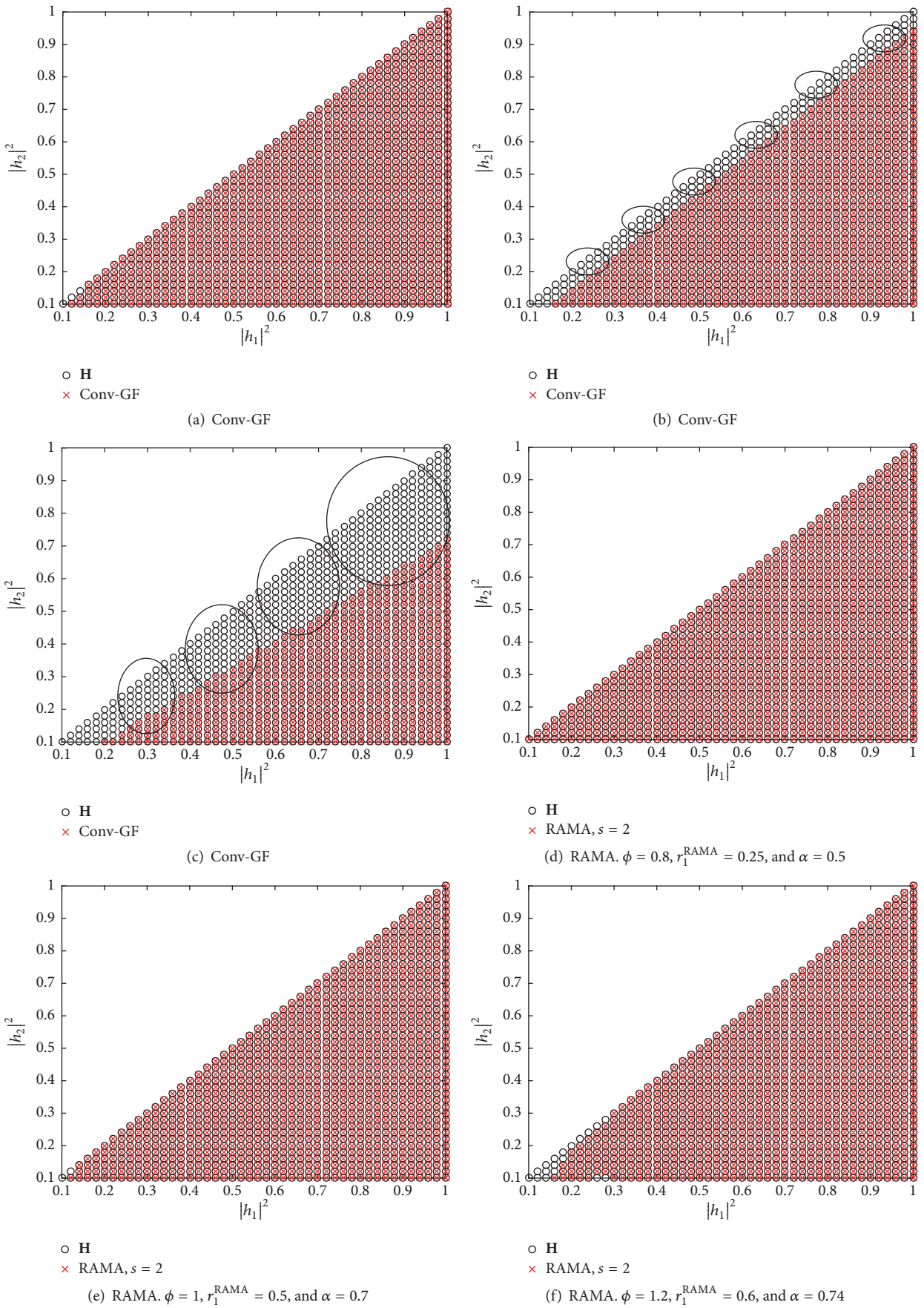
With the fixed $\boldsymbol{\lambda}_m^t$, $\boldsymbol{\vartheta}_m^t$ should be optimized to achieve optimal constellation constrained capacity. When the overlapped layers at transmitter are set to 2, optimal rotation angles can be derived by

$$\max_{\boldsymbol{\vartheta}_m} \quad m\left(\boldsymbol{\vartheta}_m^t\right), \tag{31}$$

where $m(\boldsymbol{\vartheta}_m)$ is given by [24]

$$m\left(\boldsymbol{\vartheta}_m\right) = \sum_{x_1 \in \mathcal{X}_m^t(\boldsymbol{\lambda}_m, \boldsymbol{\vartheta}_m^t)} \log\left(\sum_{x_2 \in \mathcal{X}_m^t(\boldsymbol{\lambda}_m, \boldsymbol{\vartheta}_m^t)} \frac{1}{\left|\mathcal{X}_m^t\left(\boldsymbol{\lambda}_m^t, \boldsymbol{\vartheta}_m^t\right)\right|^2} \exp\left(-\frac{1}{4\sigma^2}\left\|x_1 - x_2\right\|^2\right)\right) \tag{32}$$

5.2. Bundling Method. When high-order constellations are applied in RAMA, as illustrated in option 2 of Step 3 in Section 3.2, we propose the bundling method to construct subchannels, where different numbers of bits are bundled for different subchannels. We show an example in Figure 9, where a 16-QAM constellation is employed as the composite constellation of RAMA. We use the first bit as subchannel-1 and the remaining three bits as subchannel-2. Suppose that high priority data stream and low priority data stream are transmitted in subchannel-1 and subchannel-2, respectively. To ensure that high priority data stream is better protected, low-rate channel coding should be adopted. After cancelling the signal transmitted in subchannel-1, the residual constellation is shown at the right-hand side of Figure 9.



(a) Conv-GF

(b) Conv-GF

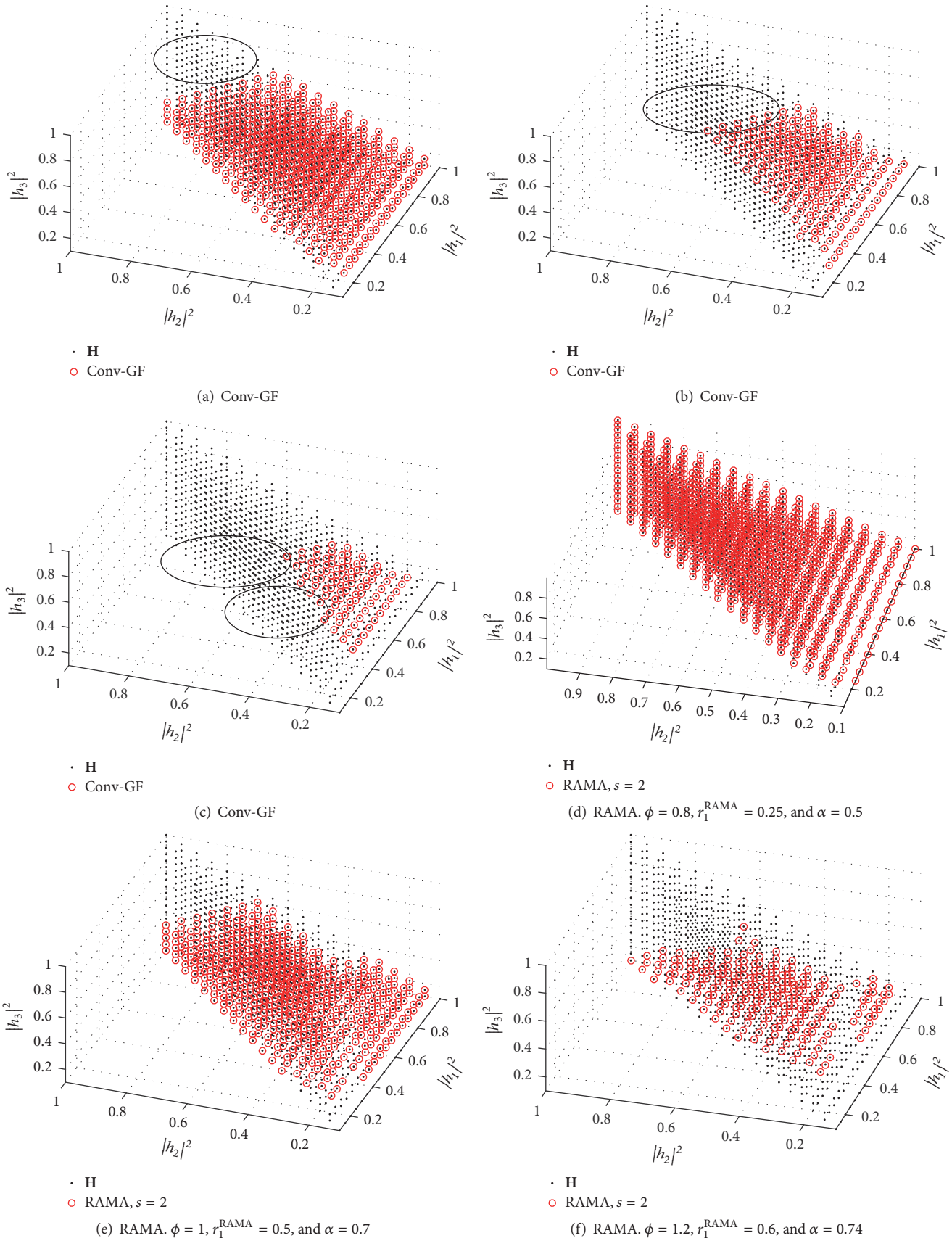
(c) Conv-GF

(d) RAMA. $\phi = 0.8$, $r_1^{\text{RAMA}} = 0.25$, and $\alpha = 0.5$

(e) RAMA. $\phi = 1$, $r_1^{\text{RAMA}} = 0.5$, and $\alpha = 0.7$

(f) RAMA. $\phi = 1.2$, $r_1^{\text{RAMA}} = 0.6$, and $\alpha = 0.74$

Figure 6: Comparison on the complementary outage regions of conv-GF and RAMA with $M_{\text{ac}} = 2$. The SNR of each user is 10 dB.



(a) Conv-GF

(b) Conv-GF

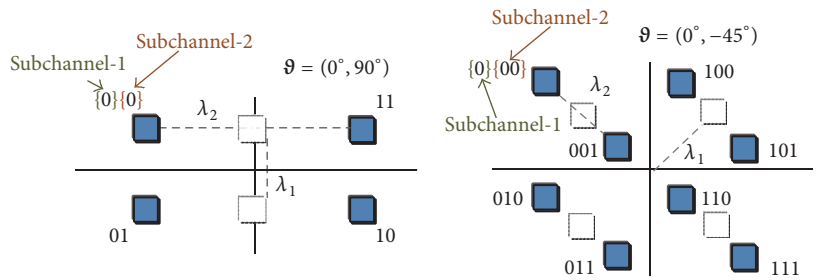
(c) Conv-GF

(d) RAMA. $\phi = 0.8$, $r_1^{\mathrm{RAMA}} = 0.25$, and $\alpha = 0.5$

(e) RAMA. $\phi = 1$, $r_1^{\mathrm{RAMA}} = 0.5$, and $\alpha = 0.7$

(f) RAMA. $\phi = 1.2$, $r_1^{\mathrm{RAMA}} = 0.6$, and $\alpha = 0.74$

Figure 7: Comparison on the complementary outage regions of conv-GF and RAMA with $M_{\mathrm{ac}} = 3$. The SNR of each user is 10 dB.



(a) 2 bits per symbol (b) 3 bits per symbol

FIGURE 8: Illustrations of RAMA amenable constellations, which are constructed by the overlapping method. λ_1 and λ_2 are the power coefficients for each of the base constellations, and ϑ is the rotation angle vector.

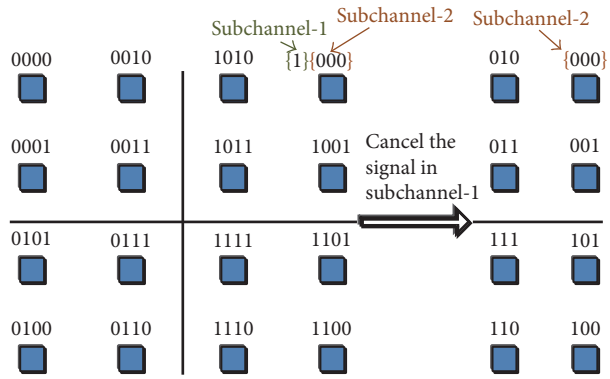


FIGURE 9: An illustration of RAMA amenable constellation with 4 bits per symbol, which are constructed by the bundling method.

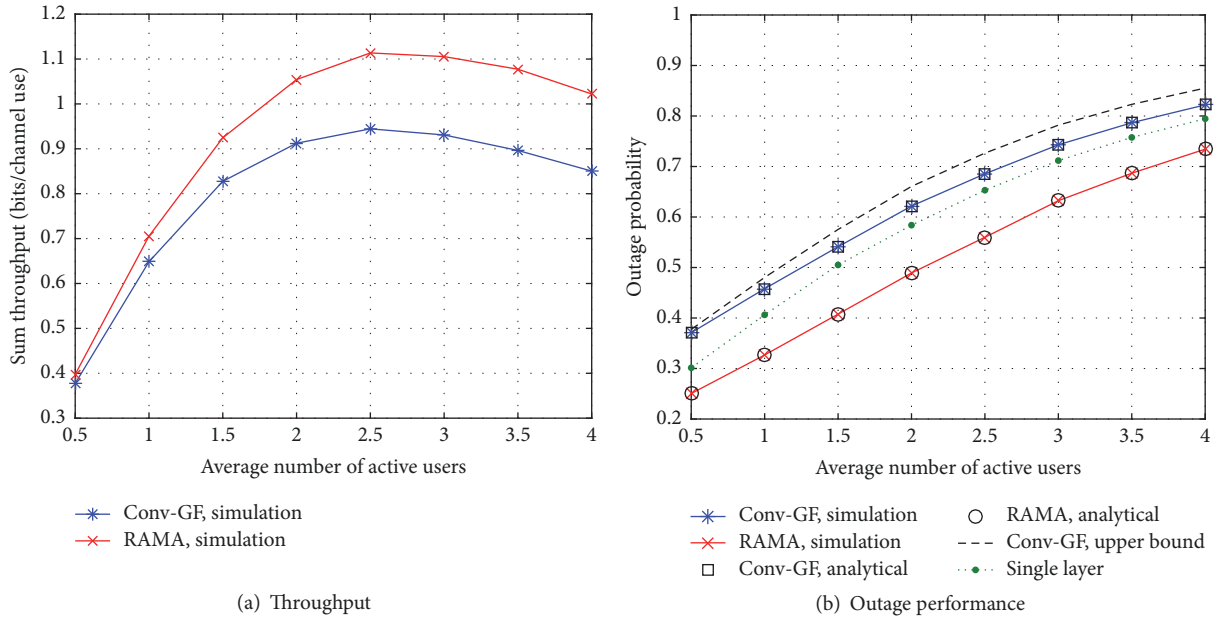
We note that the distances between the constellation points in the composite constellation cannot be arbitrarily adjusted. Therefore the UEP property is mainly offered by using different coding rates at different signal layers, which may lead to heavy burden on hardware resources. This fact makes the bundling method less flexible compared with the overlapping method.

6. Simulation Results

In this section, we evaluate the performance of the proposed RAMA scheme. First of all, we assume that Gaussian-distributed continuous alphabet and ideal SIC receiving are applied at the transmitter and the receiver, respectively. Based on these settings and the theoretical analysis in Section 4, we compare the outage and throughput performance of RAMA with that of the conv-GF [22]. Next, we validate the advantages of RAMA in practical situations, where the RAMA amenable constellations designed in Section 5 and realistic MMSE-SIC receiver are assumed.

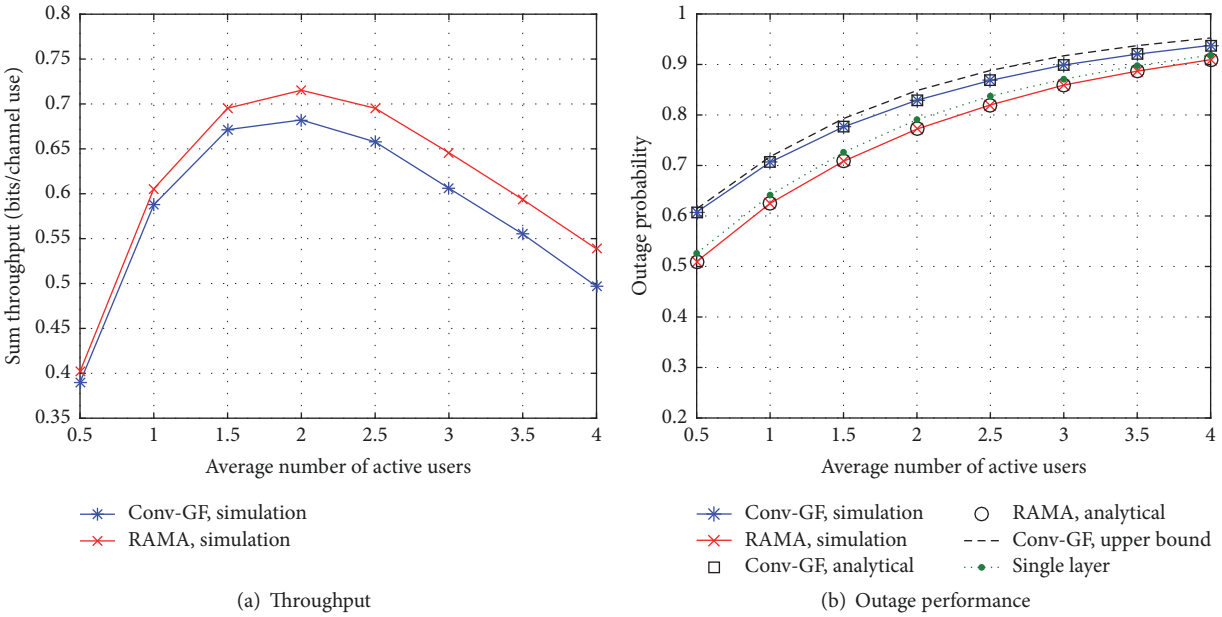
6.1. Ideal Settings. As mentioned before, we assume Gaussian-distributed continuous alphabet and ideal SIC receiving. Figures 10 and 11 compare the analytical and simulation results of conv-GF and RAMA, with $M = 100$, $R_1 = 100$, and $R_2 = 10$. The average SNR of each user is assumed as 10 dB. The analytical results of conv-GF and RAMA which are, respectively, shown by "□" and "◯" is derived by (17) and (28), respectively, via Monte-Carlo sampling method. The upper bounds of the outage performance of conv-GF, which are shown by dashed lines, are derived by integrating the results in Appendix C into (17) and setting $P(E_{1,M_{ac}}^{conv}) = 1$, $M_{ac} > 3$. The upper bound is more tight when ϕ becomes larger, since the outage probability raises sharply when $M_{ac} > 3$. The simulation results show that RAMA can simultaneously achieve higher throughput and lower outage probability than conv-GF.

In the simulation, we consider a special case, namely, "single-layer," where only layer-1 of each user is transmitted in RAMA and layer-2 is assumed as noise. Figures 10(b) and 11(b) show that the outage performance can be enhanced compared with conv-GF; however, the gains are much smaller than RAMA. These results validate that the outage performance gain of RAMA comes from two aspects: first of all, layer-1 of each user has high protection property due to its low coding rate and high power ratio; secondly, the layered structure facilitates the interference cancellation and further enhances the outage performance, since the users with small channel gains can benefit from the cancellation of highly protected layers of the users with large channel gains, and this is different from conv-GF where the signals of the users should be entirely recovered before cancellation.



(a) Throughput (b) Outage performance

FIGURE 10: Comparisons of the throughput and outage performance between conv-GF and RAMA, with $\phi = 1.2$, $r_1^{\text{RAMA}} = 0.6$, $r_2^{\text{RAMA}} = 0.6$ and $\alpha = 0.74$.



(a) Throughput (b) Outage performance

FIGURE 11: Comparisons of the throughput and outage performance between conv-GF and RAMA, with $\phi = 2$, $r_1^{\text{RAMA}} = 1$, $r_2^{\text{RAMA}} = 1$, and $\alpha = 0.75$.

6.2. Realistic Settings. In this subsection, we conduct a link level simulation of conv-GF and RAMA with realistic settings. We assume a single-cell OFDM-based uplink system with single antenna at both the BS and the user. The number of users is 100, and the activation probability varies from 0.002 to 0.06. The average received SNR of each user is assumed to take value from $[4, 20]$ dB uniformly [25]. We also assume that the small-scale channel coefficients follows Rayleigh fading, and the correlation coefficients among the small-scale channels of the symbols within a transmission block are set as 0.2, 0.5, and 0.8, where the larger the correlation coefficient, the flatter the wireless channel.

For conv-GF, we apply 1/3 rate Turbo coding and QPSK modulation, while, for RAMA, we assume $L_m = 2$ with 1/3 rate Turbo coding for both layers and apply the constellation provided in Figure 8(a) with $\lambda_1/\lambda_2 = 2, 4, 6$. The length of information bits is assumed as 1024. At the receiver, we apply Algorithm 2 to separate different users signals. Specifically,

Table 1: Simulation parameters.

Parameters	Values or Assumptions
User number	100
Activation probability	0.002 : 0.002 : 0.06
Waveform	OFDM
Average received SNR	Uniformly distributed in [4, 20] dB
Small scale channel model	Rayleigh fading, with correlation coefficients, equals 0.2, 0.5, and 0.8
Transmission mode	TM 1 (SISO)
Length of information bits	1024
Channel coding	1/3 rate Turbo
Modulation of conv-GF	QPSK
Modulation of RAMA	Use Figure 8(a) with $\lambda_1/\lambda_2 = 2, 4, 6$
Channel estimation	Ideal
Receiver	MMSE-SIC

minimum mean square error (MMSE) detection is employed in Step 1 of Algorithm 2, which is given by [26]

$$\hat{\mathbf{x}}_{m,l} = \left(h_m^t\right)^H \left(\sum_{I_j^t=1, j\neq m} h_m^t \left(h_m^t\right)^H + \sigma I^2 \right) \mathbf{y}, \quad (33)$$

where $\hat{\mathbf{x}}_{m,l}$ is the estimated signal of mth user. Note that another advanced demodulation technique, e.g., message passing algorithm (MPA), is not precluded in RAMA. The detailed simulation settings can be found in Table 1.

Figures 12 and 13 compare the throughput and the outage performance of conv-GF and RAMA. Generally, the system throughput enhances with the increase of the channel correlation coefficients, i.e., γ. With elaborately selected power coefficients of constellation, RAMA achieves larger throughput, as well as lower outage probability than conv-GF. This result also reflects that the fairness among grant-free users is improved with RAMA. Moreover, the performance gains of RAMA are more significant when the underlaying physical channel is not flat. Besides, we also observe that RAMA achieves high robustness when the activation probability goes larger (e.g., $\gamma = 0.06$), while conv-GF experiences significant drop in total throughput.

7. Conclusion and Future Work

In this paper, we have proposed the RAMA scheme for uplink grant-free data transmission. By employing layered signal structure at the transmitter and intra- and interuser SIC at the receiver, the proposed RAMA scheme achieves significant throughput and outage performance gain over conv-GF, which have been validated by analysis and simulations. The actual transmission data rate can adapt to the actual channel conditions of active users, which cannot be foreseen. RAMA also achieves high robustness when the activation probability of the users is large. Despite all this improvement, we discuss some open issues of RAMA that are worth further study in the following.

Joint Design with Spreading-Based NOMA. In this paper, we have mainly focused on the grant-free transmission based on the power domain NOMA, where symbol level spreading is not included. RAMA can also codeploy with other state-of-the-art NOMA schemes, where the main idea is to incorporate multiple independent signal layers at the transmitter.

Location-Based Access. Although the grant-free users cannot acquire the accurate channel information, they may estimate their large scale channel coefficients, e.g., via reference signal receiver power (RSRP) at the downlink. This side information may serve as an important factor based on which the grant-free users choose suitable power and transmission data rates for different layers in RAMA.

Error Propagation. Since multiple layers are introduced in RAMA, a natural problem is the error propagation among different layers during SIC receiving. This issue may be addressed by deploying the joint-detection based receiver, which also requires further study.

Appendix

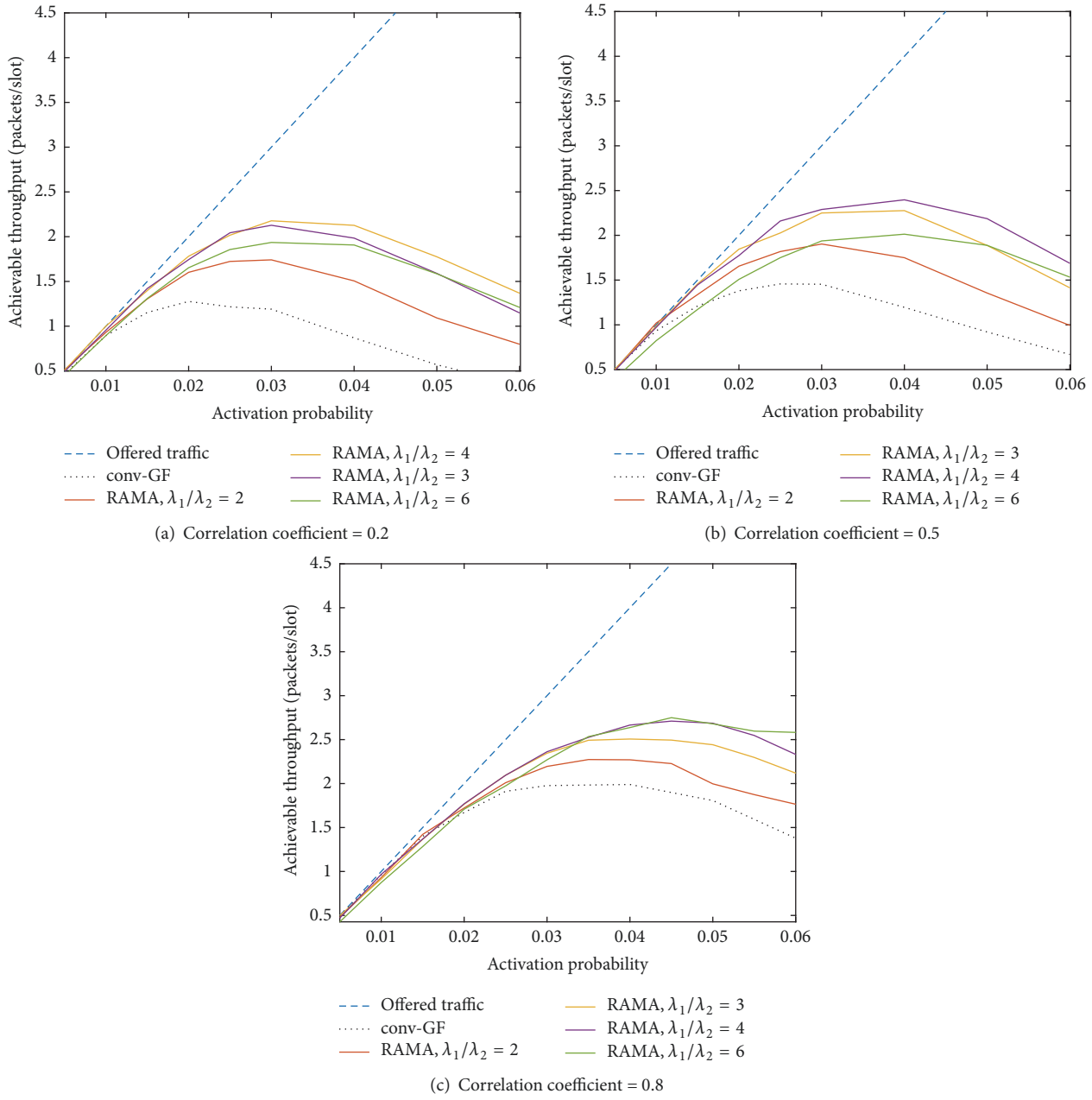
A. Proof of Lemma 1

As assumed, the users are uniformly distributed in the cell. Therefore the PDF of the distance between the users and the BS is given by

$$f_r(v) = \frac{2v}{R_1^2 - R_2^2}, \quad R_2 \le v \le R_1. \quad (A.1)$$

The PDF of the square of the distance, i.e., $x = v^2$, is

$$\begin{aligned} f_{r^2}(y) &= f_r\left(y^{-1/2}\right)\left(y^{-1/2}\right)'_y = \frac{2x^{1/2}}{R_1^2 - R_2^2}\left(\frac{1}{2}x^{-1/2}\right) \\ &= \frac{1}{R_1^2 - R_2^2}, \quad R_2^2 \le y \le R_1^2. \end{aligned} \quad (A.2)$$



(a) Correlation coefficient = 0.2

(b) Correlation coefficient = 0.5

(c) Correlation coefficient = 0.8

Figure 12: Comparison on the throughput performance between conv-GF and RAMA.

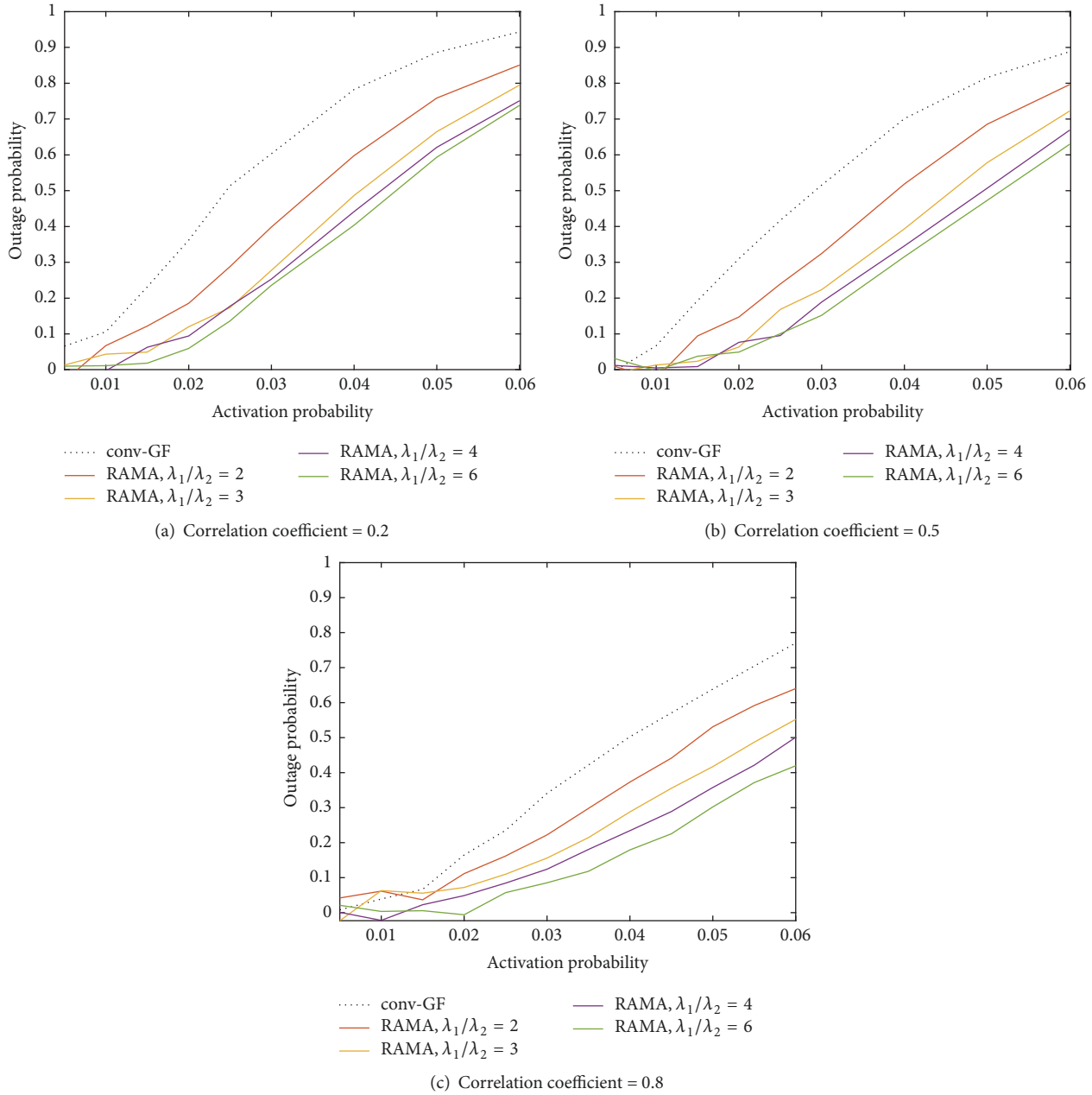
The square of the magnitude of the small-scale fading coefficient follows the $\chi^2(v)$ distribution with two degrees of freedom, i.e., $v = 2$, and is given by

$$f_{|g|^2}(x) = \frac{1}{2}e^{-x/2}. \tag{A.3}$$

Then, by integrating (A.2) and (A.3), we have

$$\begin{aligned} f_{|h|^2}(z) &= \int_0^\infty f_{|g|^2}(zy)\, f_{r^2}(y)\, y\, dy \\ &= \int_{d_2^2}^{d_1^2} \frac{1}{2}e^{-zy/2}\frac{1}{R_1^2 - R_2^2}\, y\, dy = \frac{2}{z^2\left(R_1^2 - R_2^2\right)} \\ &\quad \cdot \int_{d_2^2}^{d_1^2} e^{-zy/2}\left(\frac{zy}{2}\right) d\left(\frac{zy}{2}\right) \\ &= \frac{1}{z^2\left(R_1^2 - R_2^2\right)}\left(e^{-R_2^2 z/2}\left(R_2^2 z + 2\right)\right. \\ &\quad \left. - e^{-R_1^2 z/2}\left(R_1^2 z + 2\right)\right), \end{aligned} \tag{A.4}$$

where the last equation is due to $\int e^{-t} t\, dt = -te^{-t} - e^{-t} + C$.



(a) Correlation coefficient = 0.2

(b) Correlation coefficient = 0.5

(c) Correlation coefficient = 0.8

Figure 13: Comparison on the outage probability between conv-GF and RAMA.

The CDF of $f_{|h|^2}(z)$ is calculated as follows. We note that the indefinite integration of $f_{|h|^2}(z)$ is given by

$$h(z) = \int f_{|h|^2}(z)\,dz = \frac{2e^{-(R_1^2 z)/2} - 2e^{-(R_2^2 z)/2}}{z\left(R_1^2 - R_2^2\right)}, \tag{A.5}$$

where $h(z) \longrightarrow -1,\ z \longrightarrow 0^+$. The CDF of $|h|^2$ is given as

$$F_{|h|^2}(z) = h(z) - h(0) = 1 + h(z), \tag{A.6}$$

$$P\left(E_{m,M_{\mathrm{ac}}}^{\mathrm{conv},t}\right) = M_{\mathrm{ac}}! \times \int_0^{+\infty} f_{|h^t|^2}\left(z_{M_{\mathrm{ac}}}\right) \int_{z_{M_{\mathrm{ac}}-1}=z_{M_{\mathrm{ac}}}}^{+\infty} f_{|h^t|^2}\left(z_{M_{\mathrm{ac}}-1}\right) \times \cdots \times \int_{z_{m+1}=z_{m+2}}^{+\infty} f_{|h^t|^2}\left(z_{m+1}\right)$$

$$\times \int_{z_m=z_{m+1}}^{\phi(\sum_{j=m+1}^{M_{\mathrm{ac}}} z_j+\sigma^2/P)} f_{|h^t|^2}\left(z_m\right) \times \int_{z_{m-1}=\max(z_m,\phi(\sum_{j=m}^{M_{\mathrm{ac}}} z_j+\sigma^2/P))}^{+\infty} f_{|h^t|^2}\left(z_{m-1}\right) \times \cdots$$

$$\times \Bigg[\underbrace{\int_{z_3=\max(z_4,\phi(\sum_{j=4}^{M_{\mathrm{ac}}} z_j+\sigma^2/P))}^{+\infty} f_{|h^t|^2}(z_3) \int_{z_2=\max(z_3,(\phi/(1-\phi))(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P))}^{+\infty} f_{|h^t|^2}(z_2) \int_{z_1=z_2}^{+\infty} f_{|h^t|^2}(z_1)}_{\mathrm{I}_1}$$

$$+ \underbrace{\int_{z_3=\max(z_4,\phi(\sum_{j=4}^{M_{\mathrm{ac}}} z_j+\sigma^2/P))}^{+\infty} f_{|h^t|^2}(z_3) \int_{z_2=\max(z_3,\phi(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P))}^{(\phi/(1-\phi))(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)} f_{|h^t|^2}(z_2) \int_{z_1=\phi(\sum_{j=2}^{M_{\mathrm{ac}}} z_j+\sigma^2/P)}^{+\infty} f_{|h^t|^2}(z_1)}_{\mathrm{I}_2} \Bigg] dz_1 \cdots z_{M_{\mathrm{ac}}} \tag{A.7}$$

$$I_1 = \int_{z_3=\max(z_4,\phi(\sum_{j=4}^{M_{\mathrm{ac}}} z_j+\sigma^2/P))}^{+\infty} f_{|h^t|^2}(z_3) \int_{z_2=(\phi/(1-\phi))(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)}^{+\infty} f_{|h^t|^2}(z_2) \int_{z_1=z_2}^{+\infty} f_{|h^t|^2}(z_1), \tag{A.8}$$

$$I_2 = \int_{z_3=z_4}^{+\infty} f_{|h^t|^2}(z_3) \int_{z_2=z_3}^{(\phi/(1-\phi))(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)} f_{|h^t|^2}(z_2) \int_{z_1=\phi(\sum_{j=2}^{M_{\mathrm{ac}}} z_j+\sigma^2/P)}^{+\infty} f_{|h^t|^2}(z_1) + \int_{z_3=\max(z_4,\phi(\sum_{j=4}^{M_{\mathrm{ac}}} z_j+\sigma^2/P))}^{+\infty} f_{|h^t|^2}(z_3)$$
$$\cdot \int_{z_2=\phi(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)}^{(\phi/(1-\phi))(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)} f_{|h^t|^2}(z_2) \int_{z_1=\phi(\sum_{j=2}^{M_{\mathrm{ac}}} z_j+\sigma^2/P)}^{+\infty} f_{|h^t|^2}(z_1). \tag{A.9}$$

B. Recursive Expression of (15)

We assume that $\phi \geq 1/2$ in the following derivation, where the derivation with $\phi < 1/2$ follows the same approach.

To eliminate the max operations and derive the exact expressions of (15), we consider two cases, i.e., $z_2 \geq \phi(\sum_{j=2}^{M_{\mathrm{ac}}}+\sigma^2/P)$ and $z_2 < \phi(\sum_{j=2}^{M_{\mathrm{ac}}}+\sigma^2/P)$, where $z_2 \geq (\phi/(1-\phi))(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)$ and $z_2 \geq (\phi/(1-\phi))(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)$, respectively. Therefore, (15) is given by (A.7), where I_1 and I_2 can be further simplified in the following.

Observe that $1 > \phi \geq 1/2$, I_1 can be simplified as derived in (A.8). Besides, to eliminate the max operation in I_2, we consider two cases, i.e., $z_3 \geq \phi(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)$ and $z_2 < \phi(\sum_{j=3}^{M_{\mathrm{ac}}}+\sigma^2/P)$, where $z_3 \geq (\phi/(1-\phi))(\sum_{j=4}^{M_{\mathrm{ac}}}+\sigma^2/P)$ and $z_3 \geq (\phi/(1-\phi))(\sum_{j=4}^{M_{\mathrm{ac}}}+\sigma^2/P)$, respectively. And I_2 is given by (A.9). Till now, the max operations related to z_1 and z_2 are completely eliminated. By recursively conducting the procedures between (A.7) and (A.9), the exact expression of (15) can be derived.

C. Proof of Proposition 6

Assuming $M_{\mathrm{ac}} = 1$, $P(E_{1,1}^{\mathrm{conv}})$ is given by

$$P\left(E_{1,1}^{\mathrm{conv}}\right) = \int_{z_1=0}^{\phi\sigma^2/P} f_{|h^t|^2}(z_1)\, dz_1 = F_{|h^t|^2}\left(\frac{\phi\sigma^2}{P}\right) = 1 + \frac{2e^{-(R_1^2\phi\sigma^2/P)/2} - 2e^{-(R_2^2\phi\sigma^2/P)/2}}{\phi\sigma^2\left(R_1^2 - R_2^2\right)/P}. \tag{C.1}$$

Furthermore, if $M_{\mathrm{ac}} = 2$, $P(E_{1,2}^{\mathrm{conv}})$ and $P(E_{2,2}^{\mathrm{conv}})$ are given by

$$P\left(E_{1,2}^{\mathrm{conv}}\right) = 2! \int_{z_2=0}^{+\infty} f_{|h^t|^2}(z_2) \cdot \int_{z_1=z_2}^{\phi(z_2+\sigma^2/P)} f_{|h^t|^2}(z_2)\, dz_1 dz_2 = 2\int_{z_2=0}^{+\infty} f_{|h^t|^2}(z_2) \cdot \left(F_{|h^t|^2}\left(\phi\left(z_2+\sigma^2/P\right)\right) - F_{|h^t|^2}(z_2)\right) dz_2$$
$$= 2\left(\mathscr{F}(+\infty) - \mathscr{F}(0)\right) - 1, \tag{C.2}$$

$$P\left(E_{2,2}^{\mathrm{conv}}\right) = 2! \int_{z_2=0}^{\phi\sigma^2/P} f_{|h^t|^2}(z_2) \cdot \int_{z_1=\phi(z_2+\sigma^2/P)}^{+\infty} f_{|h^t|^2}(z_1)\, dz_1\, dz_2$$
$$= 2\int_{z_2=0}^{\phi\sigma^2/P} f_{|h^t|^2}(z_2) \cdot \left(F_{|h^t|^2}(+\infty) - F_{|h^t|^2}\left(\phi\left(z_2+\frac{\sigma^2}{P}\right)\right)\right) dz_2$$
$$= 2\left(F_{|h^t|^2}\left(\phi\left(\frac{\sigma^2}{P}\right)\right) - \mathscr{F}\left(\phi\left(\frac{\sigma^2}{P}\right)\right) + \mathscr{F}(0)\right). \tag{C.3}$$

respectively, where

$$\mathscr{F}(z) = \int f_{|h^t|^2}(z)\, F_{|h^t|^2}\left(\phi\left(z + \frac{\sigma^2}{P}\right)\right) dz. \quad \text{(C.4)}$$

Assume high SNR and $\phi = 1$, and (C.4) can be calculated as $\mathscr{F}(z) = (1/2)(F(z))^2$.

When $M_{ac} = 3$, $P(E_{1,3}^{conv})$, $P(E_{2,3}^{conv})$, and $P(E_{3,3}^{conv})$ can be derived with the similar approaches, which is omitted due to the space limitation.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant 61771051.

References

[1] "TR 38.913: Study on scenarios and requirements for next generation access technologies, 3GPP, May 2015. (Release 14)".

[2] K. Au, L. Zhang, H. Nikopour et al., "Uplink contention based SCMA for 5G radio access," in *Proceedings of the 2014 IEEE Globecom Workshops, GC Wkshps 2014*, pp. 900–905, USA, December 2014.

[3] C. Bockelmann, N. Pratas, H. Nikopour et al., "Massive machine-type communications in 5g: Physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.

[4] Y. Yuan, Z. Yuan, G. Yu et al., "Non-orthogonal transmission technology in LTE evolution," *IEEE Communications Magazine*, vol. 54, no. 7, pp. 68–74, 2016.

[5] Y.-J. Choi and K. G. Shin, "Joint collision resolution and transmit-power adjustment for Aloha-type random access," *Wireless Communications and Mobile Computing*, vol. 13, no. 2, pp. 184–197, 2013.

[6] C. Zhu, L. Shu, T. Hara, L. Wang, S. Nishio, and L. T. Yang, "A survey on communication and data management issues in mobile sensor networks," *Wireless Communications and Mobile Computing*, vol. 14, no. 1, pp. 19–36, 2014.

[7] E. Paolini, G. Liva, and M. Chiani, "Coded slotted ALOHA: a graph-based method for uncoordinated multiple access," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, vol. 61, no. 12, pp. 6815–6832, 2015.

[8] A. D. Wyner, "Recent Results in the Shannon Theory," *IEEE Transactions on Information Theory*, vol. 20, no. 1, pp. 2–10, 1974.

[9] M. Médard, J. Huang, A. J. Goldsmith, S. P. Meyn, and T. P. Coleman, "Capacity of Time-Slotted ALOHA Packetized Multiple-Access Systems Over the AWGN Channel," *IEEE Transactions on Wireless Communications*, vol. 3, no. 2, pp. 486–499, 2004.

[10] P. Minero, M. Franceschetti, and D. N. C. Tse, "Random access: An information-theoretic perspective," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 909–930, 2012.

[11] R. H. Etkin, D. N. C. Tse, and H. Wang, "Gaussian interference channel capacity to within one bit," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5534–5562, 2008.

[12] "ZTE, R1-1701608: Performance evaluation of nonorthogonal multiple access in 2-step random access procedure, 3GPP, 2017".

[13] Y. Wu and J. Fang, *Large-Scale Antenna-Assisted Grant-free Non-Orthogonal Multiple Access via Compressed Sensing*, 2016, https://arxiv.org/abs/1609.00452.

[14] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink Nonorthogonal Multiple Access in 5G Systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, 2016.

[15] H. Zheng, X. Li, and N. Ye, "Random Subchannel Selection of Store-Carry and Forward Transmissions in Traffic Hotspots," *IEEE Communications Letters*, vol. 21, no. 9, pp. 2073–2076, 2017.

[16] A. El Gamal and Y.-H. Kim, *Network information theory*, Cambridge University Press, Cambridge, UK, 2011.

[17] B. Rimoldi, R. Urbanke, and B. H. Rimoldi, "A Rate-Splitting Approach to the Gaussian Multiple-Access Channel," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 364–375, 1996.

[18] H. Joudeh and B. Clerckx, "Rate-Splitting for Max-Min Fair Multigroup Multicast Beamforming in Overloaded Systems," *IEEE Transactions on Wireless Communications*, 2017.

[19] B. Clerckx, H. Joudeh, C. Hao, M. Dai, and B. Rassouli, "Rate splitting for MIMO wireless networks: A promising PHY-layer strategy for LTE evolution," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 98–105, 2016.

[20] *NTT DOCOMO, INC., R1-1713952: UL data transmission without UL grant, 3GPP*, 2017.

[21] "Intel, R1-1712592: UL data transmission without grant, 3GPP, 2017".

[22] *NTT DOCOMO, INC., R1-167392: Discussion on multiple access for UL mMTC, 3GPP*, 2016.

[23] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, "Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access," in *Proceedings of the 2013 21st International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2013*, pp. 770–774, Japan, November 2013.

[24] N. Ye, A. Wang, X. Li, W. Liu, X. Hou, and H. Yu, "On Constellation Rotation of NOMA with SIC Receiver," *IEEE Communications Letters*, pp. 1-1.

[25] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-user shared access for internet of things," in *Proceedings of the 83rd IEEE Vehicular Technology Conference, VTC Spring 2016*, China, May 2016.

[26] K. Higuchi and A. Benjebbour, "Non-Orthogonal Multiple Access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. E98B, no. 3, pp. 403–414, 2015.