On the Performance of HARQ Protocols With Blanking in NOMA Systems

Zeina Mheich, Wenjuan Yu, Pei Xiao, Atta ul Quddus and Amine Maaref

***Abstract*—In this paper, we investigate the throughput performance of single-packet and multi-packet hybrid-automatic repeat request (HARQ) with blanking for downlink non-orthogonal multiple access (NOMA) systems. While conventional single-packet HARQ achieves high throughput at the expense of high latency, multi-packet HARQ, where several data packets are sent in the same channel block, can achieve high throughput with low latency. Previous works have shown that multi-packet HARQ outperforms single-packet HARQ in orthogonal multiple access (OMA) systems, especially in the moderate to high signal-to-noise ratio regime. This work amalgamates multi-packet HARQ with NOMA to achieve higher throughput than the conventional single-packet HARQ and OMA, which has been adopted in the legacy mobile networks. We conduct theoretical analysis for the throughput per user and also investigate the optimization of the power and rate allocations of the packets, in order to maximize the weighted-sum throughput. It is demonstrated that the gain of multi-packet HARQ over the single-packet HARQ in NOMA systems is reduced compared to that obtained in OMA systems due to inter-user interference. It is also shown that NOMA-HARQ cannot achieve any throughput gain with respect to OMA-HARQ when the error propagation rate of the NOMA detector is above a certain threshold.**

***Index Terms*—Non-orthogonal multiple access, hybrid automatic repeat request, throughput, power allocation, rate allocation.**

I. Introduction

The rise of data-hungry applications and the new paradigm of machine-type communications, in the past few years, have motivated the research of new multiple access techniques to improve the spectral efficiency and to increase the connectivity of massive number of devices [1]. To meet these new requirements, NOMA has emerged as a promising technique for future mobile networks, due to the fact that it enables overloading which improves the connectivity and the spectral efficiency compared to conventional OMA. The proposed NOMA techniques can be mainly classified into two categories [2]: power-domain NOMA (PD-NOMA) [3] and code-domain NOMA (CD-NOMA) [4], [5], [6]. In NOMA schemes, the same resource, e.g. time/frequency/code, is used by multiple users, which is different from conventional OMA schemes, e.g. time/frequency/code-division multiple access, relying on orthogonality to avoid interference between users. The receiver of NOMA employs multi-user detection techniques, such as successive interference cancellation (SIC) in PD-NOMA and message passing algorithm in sparse code multiple access (SCMA) to separate user signals.

In recent years, NOMA has triggered new research directions such as codebook design for SCMA [7], [8], [9], performance analysis for NOMA [10]–[17], etc. Furthermore, the diversity of 5G requirements calls for rethinking of the hybrid-automatic repeat request (HARQ) design to tailor them for new technologies such as NOMA [18] and cooperative NOMA [19]. HARQ is adopted in many cellular standards such as Long Term Evolution (LTE) and also adopted in 5G [20]. HARQ protocols are used to combat the impairments of wireless channels by allowing multiple retransmissions. They combine powerful channel coding with ARQ error-control to improve the communication reliability and the system throughput. Recently, HARQ has been combined with NOMA, in order to reap the aforementioned benefits of HARQ into NOMA systems as well [21], [22]. In addition, NOMA systems are subject to higher HARQ retransmission probability compared to OMA systems, since they suffer from other source of interference due to overloading. When multiple signals are superimposed on the same resource, a decoding failure for one signal will affect other signals as well. For instance, SIC detectors suffer from error propagation where the decoding failure of one signal affects the decoding performance of the remaining signals that should be decoded subsequently. By incorporating HARQ into NOMA systems, the receiver can request for a retransmission whenever a decoding error is detected for any of the superimposed signals. This has recently led to a considerable interest in combining HARQ with NOMA [23], [24], [25] as well as designing new HARQ schemes tailored for NOMA systems [26]. References [23] and [27] show that the performance of HARQ-NOMA system is superior to that of HARQ-OMA system over fading channels.

The improvement of the system reliability by using HARQ comes at the expense of higher latency. In [28], the authors propose a novel broadcasting HARQ strategy that achieves high throughput with low latency. More specifically, the broadcast approach enables the receiver to decode with the rates that are matched to fading realizations. That is, the transmitter sends the superposition of many packets (layers), which can be interpreted as a PD-NOMA of multiple packets. The better the channel condition, the more layers can be reliably decoded at the receiver. Motivated by the throughput gain of multi-

This work was supported in part by a research grant from Huawei Technologies Canada Co. Ltd. It was also supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/P03456X/1. (Corresponding author: Zeina Mheich.)

Zeina Mheich, Wenjuan Yu, Pei Xiao and Atta ul Quddus are with the 5G Innovation Centre, Institute for Communication Systems, University of Surrey, U.K. (e-mail: {z.mheich, w.yu, p.xiao, a.quddus}@surrey.ac.uk)

Amine Maaref is with the Canada Research Center, Huawei Technologies Company Ltd., Ottawa, ON, Canada. (e-mail: Amine.Maaref@huawei.com)

layer HARQ over conventional single-layer HARQ and its capability to lower the latency, several works have been devoted to the study of new multi-layer HARQ protocols (called also, multi-packet HARQ protocols). Specifically, [29] aims to jointly encode across several data packets in incremental redundancy (IR) HARQ, in order to create redundancy packets relative to multiple messages. Results show that multiple-packet IR may offer substantial throughput gains with respect to the conventional single-packet IR approach. In [30], the authors study superposition coding and time-sharing encoding strategies to allow two different packets to share the same channel block, and optimize their parameters to maximize the throughput. A cross-packet HARQ was proposed later, where multiple packets are encoded jointly into a channel block [31], [32]. However, all the aforementioned studies are conducted for the single user case. For the multi-user case, [33] investigates HARQ schemes over multiple-antenna multi-user systems and [34] proposes an ARQ protocol to enhance the aggregate throughput efficiency in multi-user systems. When the HARQ protocol is used in a multi-user context, users can help each other to correctly decode their own messages. Indeed, the transmitter can enforce *blanking*, i.e., transmitting no information for the users that have successfully decoded their own packets, in order to decrease the interference on weak users [33]. Such blanking-based HARQ protocols have also been used in the context of SCMA systems in [35] to improve the overall throughput of the system. Other works have recently studied the performance of single-packet HARQ protocols in NOMA systems [21], [22]. However, all the aforementioned works on HARQ in NOMA systems do not consider multi-packet HARQ. Unlike OMA systems, NOMA systems suffer from interference between users. Therefore, it is questionable if multi-packet HARQ provides substantial benefits over single-packet HARQ in the moderate to high signal-to-noise ratio (SNR) regime in NOMA systems.

This work deals with the performance of single-packet and multi-packet HARQ protocols in downlink NOMA systems where the transmitter enforces blanking for users whose packets have been correctly decoded before the maximum number of allowed retransmissions is reached. The main contributions of this paper are summarized as follows.

- This paper studies the throughput performance of blanking-based single-packet and multi-packet HARQ for downlink NOMA systems with block fading channels. Multiple packets are multiplexed using PD-NOMA for each user, as in [28]. The throughput expression contains integrals which cannot be expressed in closed form. Thus, we derive an expression for the throughput involving solely single-integrals, which are easy to solve numerically, for the special case when maximum two transmissions are allowed.
- For the downlink HARQ-based NOMA system under study, a weighted-sum throughput maximization problem is formulated and then solved, where the rate and power allocation are optimized for all the packets. To the best of our knowledge, our work is the first of its kind which derives the achievable throughput regions of NOMA-HARQ systems.
- Numerical results show that the gain of multi-packet HARQ over the single-packet HARQ depends on the SNR as well as the target throughput of users in both PD-NOMA that utilizes superposition coding at the transmitter and SIC at the receiver [3] and SCMA with SIC at the receiver[1] [37], [38], [39]. In the presence of strong interference, it is shown that the gain of multi-packet HARQ decreases.
- Finally, the impact of imperfect SIC is studied. It is shown that NOMA-HARQ cannot achieve any throughput gain with respect to OMA-HARQ when the error propagation rate of the NOMA detector is above a certain value.

Table I
SUMMARY OF KEY NOTATIONS.

Notation	Meaning
I	Number of users
$\mathcal{I}$	Set of user indexes
p	Average transmission power
$p_m^{\mathcal{E}}$	Effective transmission power at the mth round where the event $\mathcal{E}$ specifies the decoded and undecoded packets
$F_{X(r)}$	CDF of the rth ordered statistic
$f_{X(r)}$	PDF of the rth ordered statistic
N_p^i	Number of packets sent simultaneously to user u_i
$P_j^{u_i}$	The jth packet of user u_i
$\alpha_j^{u_i}$	Fraction of power allocated to the jth packet of user u_i
$R_j^{u_i}$	Rate of the jth packet of user u_i
s_m^i	The channel gain of user u_i at the mth HARQ round
$\mathbf{s}^i$	The channel gain vector of user u_i
M	Maximum number of HARQ rounds
$\Theta_j^i(m)$	The event "$P_j^{u_i}$ is decoded successfully at round m but its decoding failed at round $m-1$"
$\bar{\Theta}_j^i(m)$	The event "$P_j^{u_i}$ is in outage at round m"
$P_o(.)$	Outage probability
$q(.)$	Success probability
$\eta(.)$	Throughput
ℓ_i	Decoder state of user u_i
$\mathcal{L}(.)$	Set of admissible values of user decoder states

The remainder of this paper is organized as follows. Section II introduces the system model. Section III defines the main performance metrics that will be analyzed in this work. Section IV gives the single-integral forms of the outage and success probabilities when the maximum number of transmissions is limited to two, as well as the general expressions when the maximum number of transmissions is greater than two. Numerical results are presented in Section V. Finally, Section VI concludes the paper. The key notations used in the paper are summarized in Table I.

II. System model and preliminaries

A. NOMA system model

We consider the downlink transmission scenario between a single transmitter and I users denoted by u_i, where $i \in \mathcal{I} = \{1, \cdots, I\}$. The channel fading experienced by user

[1] It is not straightforward to analyse theoretically SCMA with MPA because the closed-form expression of the signal-to-interference ratio after MPA detection is not easy to obtain [36]. Thus we approximate MPA with SIC as it is also widely adopted in the literature of SCMA [37], [38], [39].

u_i is denoted by h_i and is modeled according to a unit-variance Rayleigh fading distribution. In order to support massive connectivity, the transmitter employs a NOMA scheme to multiplex all users' data. In the following, PD-NOMA with SIC is firstly taken into consideration and theoretical analysis will be conducted. The extension to SCMA with SIC will be considered as well[2]. According to the PD-NOMA principle, the transmitted signal broadcasted to all users is the superposition of all users' signals with different power levels, given as follows:

$$x = \sum_{i=1}^{I} \alpha^{u_i} \cdot p \cdot x_i, \tag{1}$$

where p is the average transmission power at the transmitter, i.e., $\mathbb{E}[x^2] \leq p$, α^{u_i} is the percentage of power allocated to user u_i, i.e., $0 < \alpha^{u_i} < 1$ and $\sum\limits_{i=1}^{I} \alpha^{u_i} = 1$, and x_i is the transmitted signal of user u_i drawn from a capacity-achieving codebook following a zero-mean unit-variance complex Gaussian distribution[3]. Then, the received signal y_i by user u_i is given by

$$y_i = h_i \cdot x + n_i, \tag{2}$$

where n_i is the additive white Gaussian noise with zero-mean and unit-variance, i.e., $n_i \sim \mathcal{CN}(0, 1)$. Henceforth, we denote by $s_i = |h_i|^2$. Without loss of generality, we assume that the users' channels are ordered such that $s_1 \leq s_2 \leq \cdots \leq s_I$, which means that user u_i has the i^{th} weakest channel.

When the channel gains are not ordered, their probability density function (PDF) and cumulative density function (CDF) are respectively denoted by $f_S(s_i)$ and $F_S(s_i)$, $\forall i \in \mathcal{I}$. Note that the non-ordered channel gains are assumed to be statistically independent and identically distributed. However, when the users' channels are ordered such that $s_1 \leq s_2 \leq \cdots \leq s_I$, the PDF and CDF of the ordered $s_r, \forall r \in \mathcal{I}$, are given as follows, according to the order statistics theory [40]:

$$F_{S(r)}(x) = \sum_{i=r}^{I} \binom{I}{i} \Bigg[F_S(x)\Bigg]^{i} \Bigg[1 - F_S(x)\Bigg]^{I-i}, \tag{3}$$

and

$$f_{S(r)}(x) = \frac{I!}{(r-1)!(I-r)!} \Bigg[F_S(x)\Bigg]^{r-1} \Bigg[1 - F_S(x)\Bigg]^{I-r} f_S(x), \tag{4}$$

where $f_{S(r)}$ and $F_{S(r)}$ are the PDF and CDF of the rth ordered statistic s_r respectively.

In order to retrieve its own signal, each user will carry out SIC. The SIC is performed by first decoding the signal of the user with the worst channel condition u_1 considering other signals as noise, subtracting it from the superimposed signal x and then decoding the next user's signal. The SIC decoding order of all users can be written as $u_1 \rightarrow u_2 \rightarrow \cdots \rightarrow u_I$ which means that user u_i can decode user u_k's packets when $k < i$. Note that due to the ordered channels assumption, user u_i can always decode user u_k' signal when $k < i$[4]. The decoding process continues successively until each user decodes its own signal. Therefore, the maximum achievable rate by user u_k can be expressed as[5] [41]:

$$R^{u_k} = \log\left(1 + \frac{s_k \cdot \alpha^{u_k} \cdot p}{1 + s_k \cdot \sum\limits_{k < i \leq I} \alpha^{u_i} \cdot p}\right). \tag{5}$$

B. HARQ

The instantaneous channel gains s_i is known at the receiver u_i but not at the transmitter. Therefore, R^{u_i} cannot be determined at the transmitter prior to the transmission. As a result, the transmitter cannot adapt the transmission rates according to the instantaneous channel conditions. When the transmission rate is greater than the maximum achievable rate in (5), an outage occurs, meaning that the information is lost. In order to decrease the outage probability, the transmitter implements HARQ protocol to enable the receiver to request a retransmission when a decoding error occurs, via a feedback channel[6]. The transmitter can send the same information in all HARQ transmissions (repetition HARQ) or different information belonging to the same packet (incremental redundancy HARQ). In this work, we focus only on incremental-redundancy HARQ.

In the conventional single-packet HARQ protocol, the transmitter sends only one packet for each user. For example, the transmitter encodes the packet P^{u_i} for user u_i into a mother codeword using an encoder of rate R^{u_i} [nats/channel use]. In incremental redundancy HARQ, the mother codeword is divided into M blocks or sub-codewords[7]. The transmitter starts the HARQ process by sending the first block. If the receiver decodes correctly its own packet, it sends a positive acknowledgment (ACK) message to the transmitter so the latter moves to the transmission of the next packet. If the receiver fails to decode, it sends a negative acknowledgment (NACK) message to the transmitter. Upon receiving a NACK, the transmitter sends the next sub-codeword or stops the transmission of the current packet and moves to the next packet if all the M blocks have been sent to the receiver. We denote M the maximum number of HARQ transmissions or rounds.

In this work, we also consider blanking-based HARQ protocols and we assume that a new HARQ transmission begins for each user (a new packet is transmitted) if all users are ACK-ed or if the maximum number of transmissions is reached. Thus, if at least one user is NACK-ed and the number of transmissions is less than M, nothing will be sent to the remaining ACK-ed users. In the latter case, more power could

[2] In Section V, numerical results will be shown for both PD-NOMA and SCMA schemes.

[3] In SCMA, x_i denotes the codeword of user u_i.

[4] When $k < i$, user u_i can decode user u_k' signal if the condition $R^{u_k \rightarrow u_i} \geq R^{u_k}$ holds, where $R^{u_k \rightarrow u_i}$ is user u_i's data rate to decode user u_k's message, i.e., $R^{u_k \rightarrow u_i} = \log\left(1 + \frac{s_i \cdot \alpha^{u_k} \cdot p}{1 + s_i \cdot \sum\limits_{k < i' \leq I} \alpha^{u_{i'}} \cdot p}\right)$. Due to the assumption $s_k \leq s_i$, this condition always holds.

[5] Unless otherwise stated, perfect SIC is assumed, which means that the decoded signal can be perfectly removed without any residual interference.

[6] We assume that feedback channels are error-free.

[7] Throughout this paper, all HARQ rounds are assumed to have the same length, i.e., occupy the same number of channel uses.

be allocated to NACK-ed users. As a result, all users will have the same transmission number at any time slot.

To reduce the delay and increase the throughput in HARQ protocols, multi-packet HARQ protocols have been well-investigated in previous works. Multi-packet HARQ protocol is a non-orthogonal HARQ protocol where the transmitter can send more than one packet to each user in the same transmission round. The packets of each user can be multiplexed using a NOMA technique or by time sharing or can be encoded jointly [30], [31]. If the user decodes correctly all its intended packets, it returns an ACK. Otherwise, it returns a NACK with an index pointing to the last successfully packet decoded. In this work, power-domain NOMA is used for both user-multiplexing and packet-multiplexing per user. However, in general, different multiplexing techniques can be used to combine the packets of each user and the data of different users. For example, one can use SCMA for user multiplexing and PD-NOMA to combine the packets of each user. Now, assume that the transmitter can simultaneously send N_p^i packets for the user u_i, denoted by $P_1^{u_i}, \cdots, P_{N_p^i}^{u_i}$. The SIC decoding order of the packets for user u_i is given by $P_1^{u_i} \rightarrow P_2^{u_i} \cdots \rightarrow P_{N_p^i}^{u_i}$. Throughout this paper, $R_j^{u_i}$ denotes the coding rate of packet $P_j^{u_i}$ and $\alpha_j^{u_i}$ is the percentage of power allocated to $P_j^{u_i}$ at the first HARQ round, where we have $0 < \alpha_j^{u_i} < 1$ and $\sum_{i=1}^{I} \sum_{j=1}^{N_p^i} \alpha_j^{u_i} = 1$.

C. Power allocation method

We assume that if the packet $P_j^{u_i}$ is correctly decoded, its allocated power in the next round will be zero. Note that the average transmission power is equal to p in the first round. The power fraction is assumed to be constant if the packet is not yet successfully decoded, thus the "effective" average transmission power at each round can vary depending on the successfully decoded packets. Let $p_m^{\mathcal{E}}$ denotes the effective average transmission power at round m, where $m \leq M$ and the event $\mathcal{E}$ specifies the decoded and non-decoded packets before the round m. For example, let us assume a system where the transmitter communicates with one user and simultaneously sends 3 packets, i.e., $I = 1$, $N_p^1 = 3$. At $m = 1$, the allocated power percentages for packets $P_1^{u_1}, P_2^{u_1}, P_3^{u_1}$ are respectively $\alpha_1^{u_1}, \alpha_2^{u_1}, \alpha_3^{u_1}$. Suppose that the user successfully decodes packet $P_1^{u_1}$ at the first round ($m = 1$). Hence, at $m = 2$, we have $p_2^{\mathcal{E}} = \frac{p}{\alpha_2^{u_1} + \alpha_3^{u_1}}$, where $\mathcal{E} = \{\text{only } P_1^{u_1} \text{ is decoded at } m = 1\}$. This is due to the fact that in the second round, there is no power allocated to $P_1^{u_1}$ and the transmitter will send the packets $P_2^{u_1}$ and $P_3^{u_1}$ with powers $\alpha_2^{u_1} \cdot p_2^{\mathcal{E}}$ and $\alpha_3^{u_1} \cdot p_2^{\mathcal{E}}$, respectively.

In a nutshell, we aim to investigate the blanking-based multi-packet incremental-redundancy HARQ protocols for NOMA systems. An illustration of our system model is given in Fig. 1[8].

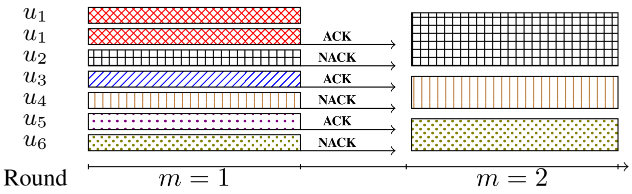

Figure 1. Multi-packet HARQ for a downlink NOMA network with six users. In the first HARQ round, the transmitter sends two packets for user 1 and one packet for the remaining users. All packets are transmitted in the same channel block. In the second HARQ round, the transmitter sends nothing to the "ACK-ed" users who have successfully decoded their own packets.

III. Outage probabilities, success probabilities and throughput

In this section, we aim to calculate the throughput for the proposed multi-packet HARQ-based NOMA system. Block-fading[9] channel model is assumed where the channel fading coefficient for each user remains constant within one HARQ round but vary independently from one round to another. This indicates that the same channel is experienced by all the signals belonging to the same round (which is equivalent to the same channel block). As a result, there is no channel diversity within the same block. The instantaneous channel state information is unknown to the transmitter, but their statistics are known. Let $\mathrm{s}^i = [s_1^i, s_2^i, \cdots, s_M^i]$ denotes the channel gain vector for user u_i, where s_m^i refers to the channel gain at round m. We assume that $s_1^i, s_2^i, \cdots, s_M^i$ are independent and identically distributed.

We start this section by defining and deriving the general expressions of the "outage" and "success" probabilities. Then, as a function of these probabilities, the general expression of the throughput is given and the weighted-sum-throughput optimization problem is formulated.

A. Outage probability: definition and calculation

Definition 1 (Outage probability): The outage probability, denoted by $P_o(u_k, \mathrm{s}^k, P_j^{u_i}, m)$, is the probability of not correctly decoding the packet $P_j^{u_i}$ by user u_k at the end of the m^{th} transmission, where $1 \leq m \leq M$ and $1 \leq j \leq N_p^j$.

We aim to calculate the outage probability $P_o(u_k, \mathrm{s}^k, P_j^{u_i}, m)$, for any user i, k, any packet j, at any HARQ round m, where $i, k \in \mathcal{I}$, $1 \leq j \leq N_p^i$ and $1 \leq m \leq M$. To proceed, let us define

$$P_j^{u_i^-} = \begin{cases} P_{j-1}^{u_i} & \text{if } j > 1, \\ P_{N_p^{i-1}}^{u_{i-1}} & \text{otherwise.} \end{cases}$$

and

$$m^- = \begin{cases} m - 1 & \text{if } m > 1, \\ \varnothing & \text{otherwise.} \end{cases}$$

Here, $P_j^{u_i^-}$ denotes the packet preceding the $P_j^{u_i}$ in the SIC order and m^- is the HARQ round preceding the round m. Moreover, let $\Theta_j^i(m)$ denote the event "the decoding of $P_j^{u_i}$

[8] If the transmitter sends multiple packets simultaneously to at least one user, we consider this case as a multi-packet HARQ scheme.

[9] The block-fading channel model is widely adopted in the performance studies of HARQ protocols.

failed at round $m-1$ but is decoded successfully at round m" and $\bar{\Theta}_j^i(m)$ denote the event "$P_j^{u_i}$ is in outage at round m".

Recall that due to SIC decoding, the user u_k can decode the packets of all the users having worse channel conditions than u_k. The SIC decoding order for user u_k is $P_1^{u_1} \rightarrow \cdots \rightarrow P_{N_p^1}^{u_1} \rightarrow \cdots \rightarrow P_1^{u_{k-1}} \rightarrow \cdots \rightarrow P_{N_p^{k-1}}^{u_{k-1}} \rightarrow P_1^{u_k} \rightarrow \cdots \rightarrow P_{N_p^k}^{u_k}$ which means that user u_k can decode user u_i's packets when $i < k$. We assume that if a packet is not successfully decoded by user u_k, then all the succeeding packets in the SIC decoding order are also lost. For example, an extreme case is that if the packet $P_1^{u_1}$ is not successfully decoded by user u_k, then all the packets are declared as lost. Hence, based on the property of SIC and Definition 1, we can write the outage probability as follows:

$$\begin{aligned}
&P_o(u_k, \mathbf{s}^k, P_j^{u_i}, m) = \\
&\Pr\Big\{ P_j^{u_i} \text{ is in outage at } m, \\
&P_1^{u_1} \cdots P_j^{u_i^-} \text{ are decoded successfully before or at } m \Big| \mathbf{s}^k \Big\} \\
&+ P_o(u_k, \mathbf{s}^k, P_j^{u_i^-}, m),
\end{aligned} \tag{6}$$

where the first term is equal to the probability that $P_j^{u_i}$ is in outage at round m while the packets preceding $P_j^{u_i}$ in the SIC decoding order, i.e., $P_1^{u_1} \cdots P_j^{u_i^-}$ are successfully decoded. The second term $P_o(u_k, \mathbf{s}^k, P_j^{u_i^-}, m)$ is the probability that $P_j^{u_i^-}$ is in outage at round m (if $P_j^{u_i^-}$ is not decoded successfully then $P_j^{u_i}$ cannot be decoded). $P_o(u_k, \mathbf{s}^k, P_j^{u_i^-}, m)$ is calculated recursively using (6).

The first term in the summation in (6) can be expressed as

$$\begin{aligned}
&\Pr\Big\{ P_j^{u_i} \text{ is in outage at } m, \\
&P_1^{u_1} \cdots P_j^{u_i^-} \text{ are decoded successfully before or at } m \Big| \mathbf{s}^k \Big\} = \\
&\sum_{m_T=1}^{m} \cdots \sum_{m_2=1}^{m_3} \sum_{m_1=1}^{m_2} \Pr\Big\{ \bar{\Theta}_j^i(m), \Theta_1^1(m_1), \Theta_2^1(m_2), \cdots, \\
&\Theta_j^{i^-}(m_T) \Big| \mathbf{s}^k \Big\},
\end{aligned} \tag{7}$$

where T is the number of packets preceding the packet $P_j^{u_i}$ in the SIC order. It can be expressed as $T = \sum\limits_{i'=1}^{i-1} N_p^{i'} + j - 1$. Note that $\Theta_j^{i^-}(m_T)$ corresponds to the event "the decoding of $P_j^{u_i^-}$ failed at round $m_T - 1$ but is decoded successfully at round m_T". In (7), the round number m_1, in which the first packet $P_1^{u_1}$ is decoded correctly, takes value between 1 and m_2 because $P_1^{u_1}$ should be decoded before the second packet $P_2^{u_1}$. Similarly, the round number m_2, in which the second packet $P_2^{u_1}$ is decoded correctly, takes value between 1 and m_3 because $P_2^{u_1}$ should be decoded before the third packet.

Let $P_{p_{\bar{m}}}^{u_{q_{\bar{m}}}}$ denote the first packet that is not decoded successfully at round $\bar{m}$ and belongs to the set $\{P_1^{u_1}, P_2^{u_1}, \cdots, P_j^{u_i^-}\}$. Clearly, this packet is the $p_{\bar{m}}$-th packet belonging to user $u_{q_{\bar{m}}}$ where $q_{\bar{m}} \leq i$. The packet $P_{p_{\bar{m}}}^{u_{q_{\bar{m}}}}$ can be determined from the round numbers $m_1, \cdots, m_T$. Hence, the probability in (7) can be further expanded as

$$\begin{aligned}
&P_{o\_m_{1\rightarrow T}}(u_k, \mathbf{s}^k, P_j^{u_i}, m) = \\
&\Pr\left\{ \mathcal{E} \triangleq \bar{\Theta}_j^i(m), \Theta_1^1(m_1), \Theta_2^1(m_2), \cdots \Theta_j^{i^-}(m_T) \middle| \mathbf{s}^k \right\} \\
&= \Pr\Bigg\{ \bigcap_{\bar{m}=1}^{m^-} \Bigg\{ \sum_{r=1}^{\bar{m}} C(s_r^k, p_r^{\mathcal{E}}, q_{\bar{m}}, p_{\bar{m}}) < MR_{p_{\bar{m}}}^{u_{q_{\bar{m}}}} \Bigg\}, \\
&\qquad \sum_{r=1}^{m} C(s_r^k, p_r^{\mathcal{E}}, i, j) < MR_j^{u_i}, \\
&\qquad \sum_{r_1=1}^{m_1} C(s_{r_1}^k, p_{r_1}^{\mathcal{E}}, 1, 1) \geq MR_1^{u_1}, \\
&\qquad \sum_{r_2=1}^{m_2} C(s_{r_2}^k, p_{r_2}^{\mathcal{E}}, 1, 2) \geq MR_2^{u_1}, \\
&\qquad \vdots \\
&\qquad \Bigg\},
\end{aligned} \tag{8}$$

where $\bigcap\limits_{i=1}^{j} \mathcal{E}_i$ denotes the joint event $(\mathcal{E}_1, \mathcal{E}_2, \cdots, \mathcal{E}_j)$ and

$$\begin{aligned}
&C(s,p,i,j) \triangleq \\
&\log\left[1 + \frac{s \cdot p \cdot \alpha_j^{u_i}}{1 + s \cdot p \cdot \big[\sum\limits_{j<j' \leq N_p^i} \alpha_{j'}^{u_i} + \sum\limits_{i'>i} \sum\limits_{1 \leq j' \leq N_p^{i'}} \alpha_{j'}^{u_{i'}} \big]}\right].
\end{aligned} \tag{9}$$

Note that (8) holds due to the well known fact that the decoding failure results from packet outage when its rate is greater than its channel capacity. The first event in (8) comes from the fact that if a packet is in outage at a certain round, then all the succeeding packets in the same round are considered to be in outage as well. The extension to SCMA with SIC is performed by replacing the function $C(s,p,i,j)$ in (9) with [39]:

$$\begin{aligned}
&C(s,p,i,j) \triangleq \\
&\log\left[1 + \frac{s \cdot p \cdot \sum\limits_{r \in \zeta(i)} \alpha_{jr}^{u_i}}{1 + s \cdot p \cdot \sum\limits_{r \in \zeta(i)} \cdot \big[\sum\limits_{j<j' \leq N_p^i} \alpha_{j'r}^{u_i} + \sum\limits_{\substack{i'>i \\ i' \in \Omega(r)}} \sum\limits_{1 \leq j' \leq N_p^{i'}} \alpha_{j'r}^{u_{i'}} \big]}\right],
\end{aligned} \tag{10}$$

where $\zeta(i)$ is the set of SCMA resources used by user u_i and $\Omega(r)$ is the set of users using resource r. We assume equal power allocation between resources for each user, i.e., $\alpha_{jr}^{u_i} = \alpha_{jr'}^{u_i}, \forall r, r' \in \zeta(i)$.

B. Success probability: definition and calculation

Definition 2 (Success probability): The success probability, denoted by $q(u_k, \mathbf{s}^k, P_j^{u_i}, m)$, is the probability of successfully decoding packet $P_j^{u_i}$ by user u_k on the m^{th} transmission while failing to decode it in all the previous transmissions.

According to this definition, the success probability can be expressed as

$$\begin{aligned} & q(u_k, \mathbf{s}^k, P_j^{u_i}, m) = \Pr\Bigg\{ P_j^{u_i} \text{ is decoded successfully at } m, \\ & P_1^{u_1} \cdots P_{j-1}^{u_i^-} \text{ are decoded successfully before or at } m \Bigg| \mathbf{s}^k \Bigg\} \\ & = \sum_{m_T=1}^{m} \cdots \sum_{m_2=1}^{m_3} \sum_{m_1=1}^{m_2} \Pr\Bigg\{ \Theta_j^i(m), \Theta_1^1(m_1), \Theta_2^1(m_2), \cdots, \\ & \Theta_j^{i^-}(m_T) \Bigg| \mathbf{s}^k \Bigg\}. \end{aligned} \tag{11}$$

Equation (11) means that $P_j^{u_i}$ can be successfully decoded at round m if all packets preceding $P_j^{u_i}$ in the SIC decoding order, i.e., $P_1^{u_1} \cdots P_{j-1}^{u_i^-}$ are successfully decoded at round m or before.

Recall that $P_{p_{\bar{m}}}^{u_{q_{\bar{m}}}}$ the first packet that is not decoded successfully at round $\bar{m}$ but belonging this time to the set $\{P_1^{u_1}, P_2^{u_1}, \cdots, P_j^{u_i}\}$, i.e., the set of packets including and preceding $P_j^{u_i}$ in the SIC order. Clearly, this packet is the $p_{\bar{m}}$th packet belonging to user $u_{q_{\bar{m}}}$ where $q_{\bar{m}} \in \mathcal{I}$ and $q_{\bar{m}} \le i$. The term inside the summations in (11) can be calculated as follows.

$$\begin{aligned} & q_{m_{1\to T}}(u_k, \mathbf{s}^k, P_j^{u_i}, m) \\ & = \Pr\Bigg\{ \mathcal{E} \triangleq \Theta_j^i(m), \Theta_1^1(m_1), \Theta_2^1(m_2), \cdots \Theta_j^{i^-}(m_T) \Bigg| \mathbf{s}^k \Bigg\} \end{aligned} \tag{12}$$

$$\begin{aligned} = \Pr\Bigg\{ & \bigcap_{\bar{m}=1}^{m^-} \Bigg\{ \sum_{r=1}^{\bar{m}} C(s_r^k, p_r^{\mathcal{E}}, q_{\bar{m}}, p_{\bar{m}}) < M R_{p_{\bar{m}}}^{u_{q_{\bar{m}}}} \Bigg\}, \\ & \sum_{r=1}^{m} C(s_r^k, p_r^{\mathcal{E}}, i, j) \ge M R_j^{u_i}, \\ & \sum_{r_1=1}^{m_1} C(s_{r_1}^k, p_{r_1}^{\mathcal{E}}, 1, 1) \ge M R_1^{u_1}, \\ & \sum_{r_2=1}^{m_2} C(s_{r_2}^k, p_{r_2}^{\mathcal{E}}, 1, 2) \ge M R_2^{u_1}, \\ & \vdots \\ \Bigg\}&. \end{aligned} \tag{13}$$

C. Throughput

The (average) throughput is an important criterion to evaluate the performance of the proposed HARQ-enabled NOMA systems. Based on the reward-renewal theorem [42], the throughput is defined as the ratio of the number of information bits received reliably by the destination to the expected number of channel uses required by the HARQ protocol to deliver the packet in up to M transmission attempts. According to [42], the throughput for each user is given by

$$\eta(u_i, \mathbf{s}^i) = \frac{\mathbb{E}\Big[\mathcal{R}(u_i, \mathbf{s}^i)\Big]}{\mathbb{E}\Big[\mathcal{D}\Big]}, \ \forall i \in \mathcal{I}, \tag{14}$$

where $\mathbb{E}\big[\mathcal{R}(u_i, \mathbf{s}^i)\big]$ and $\mathbb{E}\big[\mathcal{D}\big]$ are respectively the average reward and the expected inter-renewal delay for user u_i.

The user u_i receives a reward $\mathcal{R}_j^i = R_j^{u_i}$ if the decoding for the packet $P_j^{u_i}$ is successful. Otherwise, $\mathcal{R}_j^i = 0$. Using the definition of the outage probability, the average reward is equal to:

$$\begin{aligned} \mathbb{E}\Big[\mathcal{R}(u_i, \mathbf{s}^i)\Big] &= \sum_{j=1}^{N_p^i} \mathbb{E}\Big[\mathcal{R}_j^i\Big] \\ &= \sum_{j=1}^{N_p^i} M \cdot R_j^{u_i}\Big(1 - P_o(u_i, \mathbf{s}^i, P_j^{u_i}, M)\Big). \end{aligned} \tag{15}$$

Then, the average inter-renewal delay is the average number of HARQ rounds:

$$\begin{aligned} & \mathbb{E}\Big[\mathcal{D}\Big] = \\ & \max_{i \in \mathcal{I}} \Bigg(\sum_{m=1}^{M} m \cdot q(u_i, \mathbf{s}^i, P_{N_p^i}^{u_i}, m) + M \cdot P_o(u_i, \mathbf{s}^i, P_{N_p^i}^{u_i}, M) \Bigg). \end{aligned} \tag{16}$$

The maximization in the average delay expression (16) is due to the assumption that a new HARQ transmission begins when all users have successfully decoded their packets (blanking-based HARQ).

We aim to maximize the weighted sum throughput of all users by optimizing the power allocation ($\alpha_j^{u_i}$) and the rates ($R_j^{u_i}$) for all packets. The optimization problem is given by

$$\begin{aligned} \max_{\boldsymbol{\alpha}, \mathbf{R}} \quad & \sum_{i=1}^{I} \theta_i \cdot \eta(u_i, \mathbf{s}^i), \\ \text{s.t.} \quad & R_j^{u_i} > 0, \quad \forall i \in \mathcal{I}, j \in \{1, \cdots, N_p^i\}, \\ & 0 < \alpha_j^{u_i} < 1, \quad \forall i \in \mathcal{I}, j \in \{1, \cdots, N_p^i\}, \\ & \sum_{i=1}^{I} \sum_{j=1}^{N_p^i} \alpha_j^{u_i} = 1, \end{aligned} \tag{17}$$

where $\boldsymbol{\alpha}$ and $\mathbf{R}$ respectively denote the power percentage vector and the rate vector for all packets. Furthermore, θ_i is a weight parameter satisfying the constraints: $0 \le \theta_i \le 1$ and $\sum\limits_{i=1}^{I} \theta_i = 1$. Problem (17) should be solved for every tuple $(\theta_1, \theta_2, \cdots, \theta_I)$ in order to obtain the maximum achievable throughput region by all users.

Note that the optimization problem in (17) is non-convex, which is challenging to solve. It is not straightforward to approximate it by a convex problem because the expression of the objective function is very complicated. Moreover, using exhaustive search method to obtain the optimal solution is time-consuming, especially when the number of optimization variables increases. Some gradient-based (GB) optimization methods can be applied, but the solutions depend on the initialization vector $[\boldsymbol{\alpha}^{(0)}, \mathbf{R}^{(0)}]$. To overcome this problem, we use a GB optimization algorithm with several initialization vectors to get multiple solutions for problem (17). Then, we choose the solution that gives the maximum value of the weighted sum throughput. Note that for non-convex optimization problems, some local search optimization methods can be used as well, for e.g., hill-climbing, simulated annealing, tabu-search, etc [43]. However, all these methods cannot guarantee the convergence to the global optimum.

It is clear that in order to calculate the throughput, we need to evaluate the probabilities in (8) and (13). Unfortunately, they cannot be expressed in closed-form and can only be evaluated by Monte-Carlo simulations. However, in the following, we will show that when $M = 2$, the probabilities can be expressed in single integral forms. Note that a small value of M simplifies the theoretical analysis, which is well adopted in the literature, e.g., see [44] and [45], and the study can provide some guidance for more complicated cases. Moreover, the choice of selecting a small number of HARQ rounds, i.e., $M = 2$, is suitable for the emerging low-latency application scenarios.

D. *Performance limits*

In our model, the channel is assumed to be block-wise memoryless thus the block-wise feedback of HARQ does not change the capacity of the channel. For a memoryless channel, the rate of any capacity-achieving coding scheme is limited by the ergodic capacity. Thus, according to [46], [47], [48], the achievable throughput region when $M \to \infty$ is upper bounded by the ergodic capacity region, consisting of the convex hull of the set of all users' capacities $\{C_1, C_2, \cdots, C_I\}$ such that [13]

$$C_i = \int_0^{\infty} f_{S(i)} \cdot \log\left(1 + \frac{s^i \cdot \alpha^{u_i} \cdot p}{1 + s^i \cdot \sum\limits_{k>i} \alpha^{u_k} \cdot p}\right) \cdot ds, \qquad \forall i \in \mathcal{I}, \tag{18}$$

where $0 \leq \alpha^{u_i} \leq 1$ and $\sum\limits_{i=1}^{I} \alpha^{u_i} = 1$. By varying all the α^{u_i} between 0 and 1, the ergodic capacity region can be obtained.

IV. Throughput analysis

In this section, we derive single-integral forms for the outage and success probabilities when the maximum number of transmissions is limited to two ($M = 2$). The accuracy of the resulting expressions will be validated by Monte-Carlo simulations in Section V. To ease tracking, we first derive the single-integral forms of the outage and success probabilities in the first and second transmissions separately. Then, we generalize the expressions of these probabilities when $M > 2$.

A. *First round:* $m = 1$

According to (11) and (13), the success probability $q(u_k, \mathbf{s}, P_j^{u_i}, m = 1)$ at the first round can be written as:

$$q(u_k, \mathbf{s}^k, P_j^{u_i}, m = 1) = \Pr\left\{ \bigcap_{(v,c)} \left\{ C(s_1^k, p, v, c) \geq MR_c^{u_v} \right\} \right\}, \tag{19}$$

for all $(v, c) \in \mathcal{P}_j^i \triangleq \{1 \leq v < i, 1 \leq c \leq N_p^v\} \bigcup \{v = i, c \leq j\}$, where $\mathcal{S}_1 \bigcup \mathcal{S}_2$ denotes the union set of $\mathcal{S}_1$ and $\mathcal{S}_2$.

After simple manipulations for (19), we obtain

$$q(u_k, \mathbf{s}^k, P_j^{u_i}, m = 1) = \Pr\left\{ s_1^k \geq s_{th} \right\} = 1 - F_{S(k)}(s_{th}), \tag{20}$$

where $F_{S(k)}$ is the CDF of the ordered variables s_m^k,

$$s_{th} = \max_{(v,c) \in \mathcal{P}_j^i} s_0(v, c), \tag{21}$$

and $s_0(v, c) =$

$$\frac{e^{MR_c^{u_v}} - 1}{p\big[\alpha_c^{u_v} - (\sum\limits_{c+1 \leq j' \leq N_p^v} \alpha_{j'}^{u_v} + \sum\limits_{i' > v, 1 \leq j' \leq N_p^{i'}} \alpha_{j'}^{u_{i'}}) \cdot (e^{MR_c^{u_v}} - 1)\big]}. \tag{22}$$

The result in (20) is valid only if the denominator in (22) is positive $\forall (v, c) \in \mathcal{P}_j^i$. Otherwise, there exists at least one event in (20) which could not be satisfied, i.e., at least one packet is in outage, thus the probability $q(u_k, \mathbf{s}^k, P_j^{u_i}, m = 1)$ is equal to zero.

From (6), (7) and (8), the outage probability $P_o(u_k, \mathbf{s}^k, P_j^{u_i}, m = 1)$ can be written as

$$\begin{aligned} & p_o(u_k, \mathbf{s}^k, P_j^{u_i}, m = 1) = \\ & \Pr\left\{ C(s_1^k, p, i, j) < MR_j^{u_i}, \bigcap_{(v,c)} \left\{ C(s_1^k, p, v, c) \geq MR_c^{u_v} \right\} \right\} \\ & = \Big[F_{S(k)}(s_0(i, j)) - F_{S(k)}(\hat{s}_{th}) \Big]^+, \end{aligned} \tag{23}$$

for all $(v, c) \in \mathcal{P}_j^{i^-} = \{1 \leq v < i, 1 \leq c \leq N_p\} \bigcup \{v = i, c < j\}$, where $[\cdot]^+ = \max\{0, \cdot\}$ and

$$\hat{s}_{th} = \max_{(v,c) \in \mathcal{P}_j^{i^-}} s_0(v, c). \tag{24}$$

Again, the result in (23) holds only if the denominator in (22) is positive, $\forall (v, c) \in \mathcal{P}_j^{i^-}$. If the denominator in (22) is non-positive for some $(v, c) \in \mathcal{P}_j^{i^-}$, then the outage probability $P_o(u_k, \mathbf{s}^k, P_j^{u_i}, m = 1)$ is equal to zero. Otherwise, if the denominator in (22) is non-positive for $(v, c) = (i, j)$, then the expression of the outage probability in (23) becomes $p_o(u_k, \mathbf{s}^k, P_j^{u_i}, m = 1) = 1 - F_{S(k)}(\hat{s}_{th})$.

B. *Second round:* $m = 2$

Let $\bar{\mathcal{P}}_1$ denote the set of packets for which decoding failed at the first round by their intended users. The effective average power in the second round is assumed to be p_2, which can be calculated after knowing $\bar{\mathcal{P}}_1$, according to Section II. Let $\alpha(P_j^{u_i}) := \alpha_j^{u_i}$. We have

$$p_2 = \frac{p}{\sum\limits_{P \in \bar{\mathcal{P}}_1} \alpha(P)}. \tag{25}$$

In what follows, ℓ_i denotes the index of the last packet decoded successfully by user u_i, where $\ell_i \in \{0, 1, \cdots, N_p^i\}$. For example, if user u_3 decodes successfully its second packet $P_2^{u_3}$ but fails to decode $P_3^{u_3}$, then we have $\ell_3 = 2$. If $\ell_i = 0$, it means that user u_i has failed to decode all its intended packets. We call ℓ_i the "decoder state". According to this definition, the effective average power p_2 can be also written as

$$p_2 := p_2(\ell_1, \cdots, \ell_I) = \frac{p}{\sum\limits_{i=1}^{I} \sum\limits_{j=\ell_i+1}^{N_p^i} \alpha_j^{u_i}}. \tag{26}$$

The denominator in (26) is strictly positive since a second transmission happens only if at least one user fails to decode one of its intended packets.

The probabilities of all possible decoder states at the beginning of the second round can be calculated for each user, using the outage and success probabilities calculated in the first round. Then, after obtaining the probabilities of all decoder states, they will be used to calculate the outage and success probabilities in the second round. They can be firstly expressed as follows:

$$\begin{aligned}\Pr(\ell_i = 0) &= \Pr(\text{decoding of } P_1^{u_i} \text{ is failed at } m = 1) \\ &= p_o(u_i, \mathbf{s}^i, P_1^{u_i}, m = 1),\end{aligned} \tag{27}$$

$$\begin{aligned}\Pr(\ell_i = n) = \Pr(&\text{decoding of } P_{n+1}^{u_i} \text{ is failed at } m = 1, \\ &P_n^{u_i} \text{ is decoded successfully at } m = 1) \\ = p_o(&u_i, \mathbf{s}^i, P_{n+1}^{u_i}, m = 1) - p_o(u_i, \mathbf{s}^i, P_n^{u_i}, m = 1), \\ &\text{for } 0 < n < N_p^i,\end{aligned} \tag{28}$$

$$\begin{aligned}\Pr(\ell_i = N_p^i) = \Pr(&P_1^{u_i}, \cdots, P_{N_p^i}^{u_i} \text{ decoded successfully at} \\ &m = 1) \\ = q(&u_i, \mathbf{s}^i, P_{N_p^i}^{u_i}, m = 1).\end{aligned} \tag{29}$$

Before proceeding to the derivation of the general formulas for the outage and success probabilities in the second round, we show the steps needed to calculate the outage probability of the first packet intended for the first user, i.e. $P_1^{u_1}$, for the sake of clarity. According to (6), (7) and (8), the probability that user u_k fails to decode $P_1^{u_1}$ can be expressed as

$$\begin{aligned}&p_o(u_k, \mathbf{s}^k, P_1^{u_1}, m = 2) = \\ &\quad \sum_{(\ell_1, \cdots, \ell_I) \in \mathcal{L}_k(P_1^{u_1})} \Pr\big(\text{decoding of } P_1^{u_1} \text{ is failed at} \\ &\quad m = 2 \big| \ell_1, \cdots, \ell_I; \mathbf{s}^k\big) \cdot \Pr\big(\ell_1, \cdots, \ell_I \big| \mathcal{L}_k(P_1^{u_1})\big).\end{aligned} \tag{30}$$

We have explicitly shown in (30) the dependence of the outage probability on all user' decoder states, because each I-tuple $(\ell_1, \cdots, \ell_I)$ may result in different values of the effective average power p_2 according to (26). Thus, by knowing $(\ell_1, \cdots, \ell_I)$, we can calculate p_2 from (26). This also applies to the success probability formulas. Note that $\mathcal{L}_k(P_1^{u_1})$ denotes the set of admissible states since not all states can be allowed. For example, if $k = 1$, $P_1^{u_1}$ is in outage at $m = 2$ at user u_1 means also that the user u_1 has failed to decode all its packets at $m = 1$. Consequently, $\ell_1 = 0$. Since $P_1^{u_1}$ is not intended for user u_i where $i > 1$, we do not know the exact decoder state ℓ_i. Hence, we conclude that $\mathcal{L}_1(P_1^{u_1}) = \{\ell_1 = 0$ and $\ell_i \in \{0, 1, \cdots, N_p^i\}$ for $i > 1\}$. Let's take another example when $k = 2$. $P_1^{u_1}$ in outage at $m = 2$ for user u_2 also means that both u_1 and u_2 have failed to decode all their packets at $m = 1$. This is due to the assumption that u_1 has worse channel conditions than u_2. Moreover, due to SIC, u_2 cannot proceed to decode its own packets if the packets of user u_1 are not successfully decoded. Consequently, we have $\mathcal{L}_2(P_1^{u_1}) = \{\ell_1 = 0, \ell_2 = 0$ and $\ell_i \in \{0, 1, \cdots, N_p^i\}$ for $i > 2\}$. Then, the probabilities $\Pr\big(\ell_1, \cdots, \ell_I \big| \mathcal{L}_k(P_1^{u_1})\big)$ in (30) can be easily calculated given $\mathcal{L}_k(P_1^{u_1})$ using (27), (28) and (29).

After this illustrative example, we proceed to the derivation of the general formulas of the success and outage probabilities. Let $\mathcal{P}^i$ denote the set of packets that should be decoded by user u_i (these packets are those intended for user u_i and also could be intended for users with worse channel conditions than u_i). Define $\mathcal{P}_1^i = \mathcal{P}^i \bigcap \mathcal{P}_1$ and $\bar{\mathcal{P}}_1^i = \mathcal{P}^i \setminus \mathcal{P}_1^i$. In order to calculate the success probability, we need to first calculate the probability in (13) for a specific tuple $(\ell_1, \cdots, \ell_I)$ given the sets $\mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)$ for each $k, i \in \mathcal{I}$ and $j \in \{1, \cdots, N_p^i\}$. We observe that the admissible sets of decoder states depend on the round number at which the packets $P_1^{u_1}, \cdots, P_j^{\bar{u_i}}$ are decoded by user u_k.

$$\Pr\left\{\Theta_j^i(m), \Theta_1^1(m_1), \Theta_2^1(m_2), \cdots \middle| \ell_1, \cdots, \ell_I; \mathbf{s}^k\right\} \tag{31}$$

$$\begin{aligned}= \Pr\Bigg\{ &C(s_1^k, p, q_1, p_1) < MR_{p1}^{u_{q_1}}, \\ &\bigcap_{(v,c): P_c^{u_v} \in \bar{\mathcal{P}}_1^k} \left\{ C(s_1^k, p, v, c) + C(s_2^k, p_2, v, c) \geq MR_c^{u_v} \right\}, \\ &\bigcap_{(v,c): P_c^{u_v} \in \mathcal{P}_1^k} \left\{ C(s_1^k, p, v, c) \geq MR_c^{u_v} \right\}, \\ \Bigg\}&,\end{aligned} \tag{32}$$

for all $(v, c) \in \{1 \leq v < i, 1 \leq c \leq N_p^v\} \bigcup \{v = i, c \leq j\}$, where

$$(\ell_1, \cdots, \ell_I) \in \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)$$

and $P_{p1}^{u_{q1}}$ is the first packet whose decoding failed at the first transmission by user u_k among the packets $\{P_1^{u_1}, \cdots, P_j^{\bar{u_i}}\}$. The sets $\mathcal{P}_1^k$ and $\bar{\mathcal{P}}_1^k$ can be determined from the round numbers $m, m_1, \cdots, m_T$. It can be easily shown that the probability in (32) can be written as:

$$\Pr\left\{ s_1^k < s_{th1}, s_1^k \geq s_{th2}, s_1^k \geq s_{th3} \right\}, \tag{33}$$

where:

$$s_{th1} = s_0(q_1, p_1), \tag{34}$$

$$s_{th2} = \max_{(v,c):P_c^{u_v} \in \bar{\mathcal{P}}_1^k} \hat{s}_0(v,c;k), \tag{35}$$

$$s_{th3} = \max_{(v,c):P_c^{u_v} \in \mathcal{P}_1^k} s_0(v,c), \tag{36}$$

where $\hat{s}_0(v,c;k) =$

$$\frac{\gamma(v,c;k)}{p \cdot \alpha_c^{u_v} - \gamma(v,c;k) \cdot p \cdot \big(\sum\limits_{c+1 \leq j' \leq N_p^v} \alpha_{j'}^{u_v} + \sum\limits_{i'>v, 1 \leq j' \leq N_p^{i'}} \alpha_{j'}^{u_{i'}} \big)}, \tag{37}$$

and $\gamma(v,c;k) =$

$$\frac{e^{MR_c^{u_v}}}{1 + \frac{s_2^k p_2 \alpha_c^{u_v}}{1 + s_2^k p_2 \big(\sum\limits_{c+1 \leq j' \leq N_p^v} \alpha_{j'}^{u_v} + \sum\limits_{i'>v, 1 \leq j' \leq N_p^{i'}} \alpha_{j'}^{u_{i'}} \big)}} - 1. \tag{38}$$

Notice that γ depends on the random variable s_2^k. Finally, (33) can be written in a single-integral form as follows:

$$\begin{aligned} &\Pr\left\{ s_1^k < s_{th1}, s_1^k \geq s_{th2}, s_1^k \geq s_{th3} \right\} \\ &= \int_0^\infty ds_2^k \cdot f_{S(k)}(s_2^k) \cdot \Big[A - B \Big]^+, \end{aligned} \tag{39}$$

where

$$\begin{aligned} A &= F_{S(k)}(s_{th1}), \\ B &= F_{S(k)}\bigg(\max\{s_{th2}, s_{th3}\} \bigg). \end{aligned} \tag{40}$$

From (40), one can note that B depends on s_2^k via s_{th2}.

Remark 1: The expressions of A and B in (40) are valid only when the denominators of (37) and (22) are positive and the value of $\gamma(v,c;k)$ in (38) is also positive, $\forall (v,c) \in \bar{\mathcal{P}}_1^k$. In the following, we give the expressions of A and B if these conditions are not satisfied. If the value $\gamma(v,c;k)$ is negative for a certain packet $P_c^{u_v}$, then we set $\hat{s}_0(v,c;k) = 0$. If $\gamma(v,c;k) > 0$ but $\hat{s}_0(v,c;k) < 0$ for a certain packet $P_c^{u_v}$, then the event related to $P_c^{u_v}$ in (32) could not be satisfied; thus the probability in (31) is null ($A = B = 0$). Moreover, if $s_{th1} < 0$, then the condition $s_1^k < s_{th1}$ is always satisfied ($s_{th1} = \infty$).

In a nutshell, according to (11), (30) and (39), the success probability in the second round is equal to

$$\begin{aligned} &q(u_k, \mathbf{s}^k, P_j^{u_i}, m=2) = \\ &\sum_{m_T=1}^{2} \cdots \sum_{m_2=1}^{m_3} \sum_{m_1=1}^{m_2} \sum_{\substack{(\ell_1,\cdots,\ell_I) \in \\ \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)}} \Pr\bigg(\ell_1, \cdots, \\ &\ell_I \Big| \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T) \bigg) \cdot \int_0^\infty ds_2^k \cdot f_{S(k)}(s_2^k) \cdot \Big[A - B \Big]^+, \end{aligned} \tag{41}$$

where B depends on $\ell_1, \cdots, \ell_I$ via p_2 in (38). Moreover, both A and B depend on $m_1, \cdots, m_T$ via the sets $\mathcal{P}_1^k$ and $\bar{\mathcal{P}}_1^k$.

In order to calculate the outage probability, we need first to calculate the probability in (8) for each tuple $(\ell_1, \cdots, \ell_I) \in \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)$ and each $k, i \in \mathcal{I}$ and $j \in \{1, \cdots, N_p^i\}$.

$$\Pr\left\{ \bar{\Theta}_j^i(m), \Theta_1^1(m_1), \Theta_2^1(m_2), \cdots \middle| \ell_1, \cdots, \ell_I; \mathbf{s}^k \right\} \tag{42}$$

$$\begin{aligned} = \Pr\Bigg\{ &C(s_1^k, p, q_1, p_1) < MR_{p_1}^{u_{q_1}}, \\ &\bigcap_{(v,c):P_c^{u_v} \in \bar{\mathcal{P}}_1^k} \Big\{ C(s_1^k, p, v, c) + C(s_2^k, p_2, v, c) \geq MR_c^{u_v} \Big\}, \\ &\bigcap_{(v,c):P_c^{u_v} \in \mathcal{P}_1^k} \Big\{ C(s_1^k, p, v, c) \geq MR_c^{u_v} \Big\}, \\ &C(s_1^k, p, i, j) + C(s_2^k, p_2, i, j) < MR_j^{u_i} \\ \Bigg\}&, \end{aligned} \tag{43}$$

for all $(v,c) \in \{1 \leq v < i, 1 \leq c \leq N_p^v\} \bigcup \{v = i, c < j\}$. Equation (43) can be reformulated as

$$\begin{aligned} &\Pr\left\{ s_1^k < s_{th0}, s_1^k < s_{th1}, s_1^k \geq s_{th2}, s_1^k \geq s_{th3} \right\} \\ &= \int_0^\infty ds_2^k \cdot f_{S(k)}(s_2^k) \cdot \Big[A - B \Big]^+, \end{aligned} \tag{44}$$

where $s_{th0} = \hat{s}_0(i,j;k)$, $s_{th1}, s_{th2}, s_{th3}$ are given in (34), (35), (36) and A, B are given below.

$$\begin{aligned} A &= F_{S(k)}\bigg(\min\{s_{th0}, s_{th1}\} \bigg), \\ B &= F_{S(k)}\bigg(\max\{s_{th2}, s_{th3}\} \bigg). \end{aligned} \tag{45}$$

Remark 2: The expressions of A and B in (40) are valid only when the denominators of (37) and (22) are positive and the value of $\gamma(v,c;k)$ in (38) is also positive, $\forall (v,c) \in \bar{\mathcal{P}}_1^k$. If these conditions are not satisfied, the same analysis given in Remark 1 applies here. In addition, we have other conditions on $s_{th0} = \hat{s}_0(i,j;k)$. If $\gamma(i,j;k) < 0$, then the probability in (44) is null, because the last event in (43) could not be satisfied. If $\gamma(i,j;k) > 0$ but $s_{th0} < 0$, we set $s_{th0} = \infty$ because the last event in (43) is always satisfied in this case.

Finally, according to (6), (7), (30) and (44), the outage probability in the second round is equal to

$$\begin{aligned} &P_o(u_k, \mathbf{s}^k, P_j^{u_i}, m=2) = P_o(u_k, \mathbf{s}^k, P_j^{\bar{u_i}}, m=2) + \\ &\sum_{m_T=1}^{2} \cdots \sum_{m_2=1}^{m_3} \sum_{m_1=1}^{m_2} \sum_{\substack{(\ell_1,\cdots,\ell_I) \in \\ \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)}} \Pr\bigg(\ell_1, \cdots, \\ &\ell_I \Big| \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T) \bigg) \cdot \int_0^\infty ds_2^k \cdot f_{S(k)}(s_2^k) \cdot \Big[A - B \Big]^+. \end{aligned} \tag{46}$$

In order to calculate the outage and success probabilities for the second round, we should find the set $\mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)$ which consists of all the admissible values of the tuple $(\ell_1, \cdots, \ell_I)$ for each $k, i \in \mathcal{I}$, $j \in \{1, \cdots, N_p^i\}$ and each value of the tuple $(m_1, \cdots, m_T)$, where $m_1, \cdots, m_T$ are the transmission rounds of the T packets preceding the packet $P_j^{u_i}$ in the SIC order. After enumerating all the admissible values $(\ell_1, \cdots, \ell_I)$ for each $k, i \in \mathcal{I}$, $j \in \{1, \cdots, N_p^i\}$ and each tuple $(m_1, \cdots, m_T)$, we can demonstrate that the tuples $(\ell_1, \cdots, \ell_I)$ belonging to the set $\mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)$ satisfy the following

$$\begin{aligned} \text{If } v > k &\implies \ell_v \in \{0, 1, \cdots, N_p^v\}, \\ \text{If } i+1 \leq v \leq k &\implies \ell_v = 0, \\ \text{If } v = i \text{ and } k \neq i &\implies \ell_v \in \{0, 1, \cdots, \ell_i^{\max}\}, \\ \text{If } v = i \text{ and } k = i &\implies \ell_v = \ell_i^{\max}, \\ \text{If } v < i &\implies \ell_v \in \{0, 1, \cdots, \ell_v^{\max}\}, \end{aligned} \tag{47}$$

for each $v \in \mathcal{I}$. Here, $\ell_v^{\max}$ is the number of packets belonging to user u_v that are decoded successfully by user u_k in the first round, and $\ell_i^{\max}$ is the number of packets belonging to the set $\{P_1^{u_i}, \cdots, P_j^{u_i^-}\}$ and that are decoded successfully by user u_k in the first round. Note that $\ell_v^{\max}$ and $\ell_i^{\max}$ can be deduced from $(m_1, \cdots, m_T)$.

C. Generalization to $M > 2$

The final expressions of the outage and success probabilities for $M = 2$, given in (46) and (41), can be respectively reformulated in a general form as function of the decoder states for $m > 2$. For the user u_i, we define a vector of decoder states $\boldsymbol{\ell}_i = [\ell_i^1, \ell_i^2, \cdots, \ell_i^{M-1}]$, where ℓ_i^m is the decoder state of user u_i at the end of the m-th round. The outage probability can be written as

$$\begin{aligned} &P_o(u_k, \mathbf{s}^k, P_j^{u_i}, m) = P_o(u_k, \mathbf{s}^k, P_j^{u_i^-}, m) + \\ &\sum_{m_T=1}^{m} \cdots \sum_{m_2=1}^{m_3} \sum_{m_1=1}^{m_2} \sum_{\substack{(\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_I) \in \\ \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)}} \Pr\Bigg(\boldsymbol{\ell}_1, \cdots, \\ &\boldsymbol{\ell}_I \Big| \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)\Bigg) \cdot P_{o\_m_{1\to T}}(u_k, \mathbf{s}^k, P_j^{u_i}, m). \end{aligned} \tag{48}$$

The success probability can be written as

$$\begin{aligned} &q(u_k, \mathbf{s}^k, P_j^{u_i}, m) = \\ &\sum_{m_T=1}^{m} \cdots \sum_{m_2=1}^{m_3} \sum_{m_1=1}^{m_2} \sum_{\substack{(\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_I) \in \\ \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)}} \Pr\Bigg(\boldsymbol{\ell}_1, \cdots, \\ &\boldsymbol{\ell}_I \Big| \mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)\Bigg) \cdot q_{m_{1\to T}}(u_k, \mathbf{s}^k, P_j^{u_i}, m). \end{aligned} \tag{49}$$

Note that $\mathcal{L}_k(P_j^{u_i}, m_1, \cdots, m_T)$ is the set of all the admissible values of the tuple of decoder state vectors $(\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_I)$ for each $k, i \in \mathcal{I}$, $j \in \{1, \cdots, N_p^i\}$ and $(m_1, \cdots, m_T)$. The probabilities $P_{o\_m_{1\to T}}(u_k, \mathbf{s}^k, P_j^{u_i}, m)$ and $q_{m_{1\to T}}(u_k, \mathbf{s}^k, P_j^{u_i}, m)$ are given in (8) and (13) and can be evaluated using Monte Carlo simulations when $m > 2$.

V. Numerical results and discussions

A. Results with perfect SIC

In this section, we will present numerical results to confirm the single-integral form derived for the outage and success probabilities for $M = 2$. Simulations are also provided to evaluate the performance of multi-packet HARQ in NOMA systems. Specifically, both PD-NOMA and SCMA will be numerically studied in this section. We consider SNR values ≥ 20 dB, since using NOMA is more beneficial in this case [13].

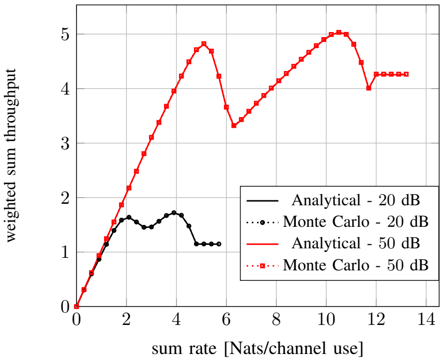

Figure 2. The weighted sum throughput as a function of the sum rate of all packets. $N_p^1 = N_p^2 = 1$, $\theta = 0.52$ and $M = 2$.

Fig. 2 shows the weighted sum throughput, i.e., $\theta \cdot \eta_1 + (1-\theta) \cdot \eta_2$, where η_i is the throughput of user u_i and $\theta \in [0, 1]$, as a function of the sum rate of all packets, i.e., $R_s = R_1^{u_1} + R_1^{u_2}$ for a NOMA system with two users, i.e., $I = 2$, using HARQ protocol. In order to obtain the optimal value of the weighted sum throughput (y-axis) for each value of R_s, we added the sum-rate constraint $R_s = R_1^{u_1} + R_1^{u_2}$ to the optimization problem given in (17). It is assumed that the transmitter sends one packet to each user, i.e., $N_p^1 = N_p^2 = 1$. The weight factor θ is equal to 0.52. The SNR in decibels (dB) is defined as SNR $= p$. This figure shows the weighted sum throughput for two SNR values, 20 dB (practical SNR regime) and 50 dB (high SNR regime), calculated in two ways, i.e., by using Monte Carlo simulations and the proposed single-integral form (named "analytical" in the legend) in Section IV. We can observe that the accuracy of the single-integral forms is confirmed. Note that this figure is plotted with optimized power allocation and packet rates. The non-smoothness of the curves is due to the non-convexity nature of the optimization problem in (17).

Figs. 3 and 4 show the weighted sum throughput (WST) as a function of the sum rate of all packets for SNR values equal to 50 dB and 20 dB respectively. The NOMA system under study consists of two users using an HARQ protocol with $N_p^1 = N_p^2 = 1$. Two values of M are considered ($M = 2$ and $M = 3$). Clearly, the maximum weighted sum throughput value increases when M becomes larger. However, we observe that when the sum rate is larger than a certain value, the weighted sum throughput becomes constant. This is because when the sum-rate increases, the optimal power allocation is given as $\alpha_1^{u_1} = 1$ and $\alpha_1^{u_2} = 0$ and the optimal rate allocation is given as $R_1^{u_1} = r$ and $R_1^{u_2} = R_s - r$, where R_s is the sum rate and r is the optimal rate value of user u_1's packet. When R_s increases, the value of r converges to a constant value, but $R_1^{u_2}$ increases. Since the throughout of user u_2 is zero, due to the power allocation $\alpha_1^{u_2} = 0$, the weighted sum throughput

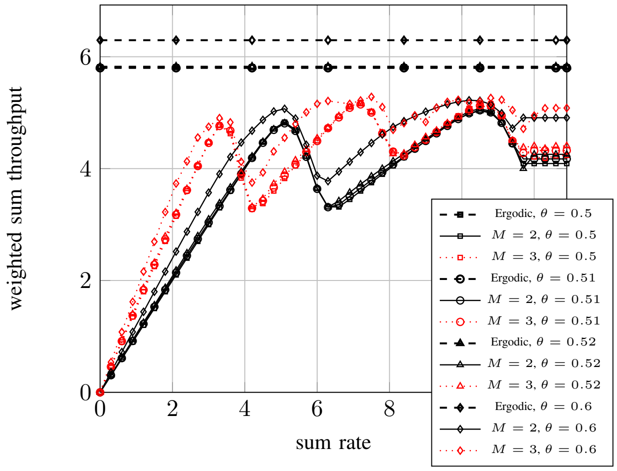

Figure 3. The WST as a function of the sum rate of all packets. $N_p^1 = N_p^2 = 1$, $I = 2$ and SNR = 50 dB.

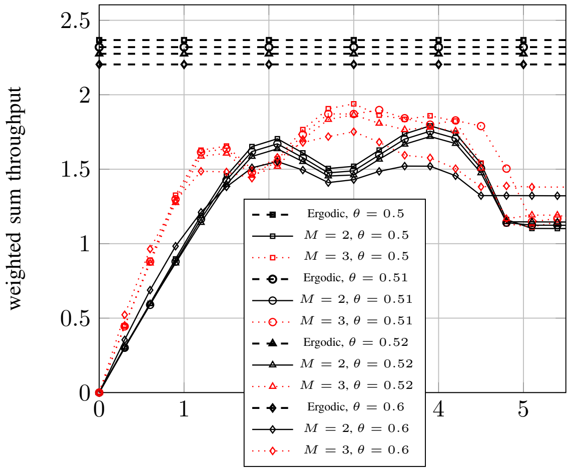

Figure 4. The WST as a function of the sum rate of all packets. $N_p^1 = N_p^2 = 1$, $I = 2$ and SNR = 20 dB.

remains constant as R_s increases.

Then, we consider a downlink system of six users, i.e., $I = 6$, where any two users are paired together to form a NOMA pair. Conventional OMA is applied for inter-NOMA-pairs' multiple access. Assume that all users are paired as follows: (u_1, u_5), (u_2, u_4) and (u_3, u_6). In the following, let r_i, where $i \in \{1, 2\}$, denote the i-th user in each NOMA pair. For example, in the first NOMA pair, we have $r_1 = u_1$ and $r_2 = u_5$. Figures 5, 6 and 7 show the maximum achievable throughput regions (η_1 versus η_2, where η_i is the throughput of user r_i, $\forall i \in \{1, 2\}$) with SNR = 50 dB and $M = 2$ for the three NOMA pairs, respectively. In each figure, both single-packet HARQ region ($N_p^1 = N_p^2 = 1$) and multi-packet HARQ regions are given. For the multi-packet HARQ, we consider three cases: 1) $N_p^1 = 2, N_p^2 = 1$, 2) $N_p^1 = 1, N_p^2 = 2$ and 3) $N_p^1 = 1, N_p^2 = 3$. Moreover, the ergodic capacity region (upper bound) is also shown, as well as the throughput regions of single-packet HARQ ($N_p = 1$) and multi-packet HARQ ($N_p = 2$), for the OMA scheme where r_1 and r_2 are orthogonal multiplexed using time sharing (lower bound). The gain of multi-packet HARQ with respect to single packet HARQ for single-user systems (equivalently, OMA systems) has been proved in [28]. The question that remains unanswered is whether the multi-packet HARQ is still more advantageous in NOMA systems. Figures 5, 6 and 7 demonstrate that multi-packet HARQ allows to enlarge the achievable throughput regions. For example, consider the pair (u_2, u_4) and a target throughput for user u_4 equal to 8 nats/channel use. By using multi-packet HARQ protocol with $N_p^1 = 1, N_p^2 = 3$ instead of single-packet HARQ, the gap to the ergodic capacity is reduced by 14.6%. However, the gain is not the same in all the achievable throughput regions. We observe that in some parts of the achievable region (especially the high-interference region) the gain is less. Since multi-packet HARQ introduces additional complexity due to the SIC decoder, in practice it is important to choose the best strategy which considers the tradeoff between performance and complexity depending on the SNR value and the target user's throughput. For example, in Figure 5, when $\eta_1 = 5$ Nats/channel use, the gain is small. In this case, using multi-packet HARQ will introduce additional complexity to the system (additional layers to be decoded) without significant gain in throughput. Thus, in this case, using single packet HARQ provides a good tradeoff between performance and complexity because it has lower complexity than multi-packet HARQ without scarifying users' throughput. The method used to calculate the SNR gain is given in [49], [50]. Note that multi-packet HARQ do not only increase the throughput, but also decrease the latency as well. Figures 5, 6 and 7 also show that noticeable throughput gains can be achieved using multi-packet HARQ-NOMA scheme instead of the conventional single-packet HARQ-OMA scheme. NOMA systems suffer from interference between users which reduces the gain obtained by using multi-packet HARQ compared with OMA systems, for the same SNR value. In our future work, we shall study if cross-packet HARQ can outperform PD-NOMA-based multi-packet HARQ for NOMA systems, especially in the high-interference region.

Figures 8, 9 and 10 show the maximum achievable throughput regions with SNR = 20 dB and $M = 2$ for the three NOMA pairs. The gain of multi-packet HARQ over single-packet HARQ is very small in this case, thus using single-packet HARQ is preferred. We can observe also that HARQ-NOMA schemes achieve significant gains over HARQ-OMA schemes.

Since SNR gains are usually more noticeable than throughput gains [49], Table II provides the maximum achievable SNR gain in dB, using multi-packet HARQ with respect to single-packet HARQ for different NOMA systems and SNR values. We can observe that the SNR gain increases with the SNR value.

To provide a comprehensive study, we also investigate the performance of multi-packet HARQ in SCMA systems. We consider an SCMA system of six users sharing 4 resources, see [5, Fig.2]. In order to focus on the performance of multi-packet HARQ and to facilitate the optimization of the weighted-sum throughput, we fix the rates and the powers of the packets allocated to users u_1, u_2, u_3, u_4. Then, we determine the achievable throughput regions of users u_5 and u_6 using single-packet HARQ and multi-packet HARQ. Fig. 11 shows

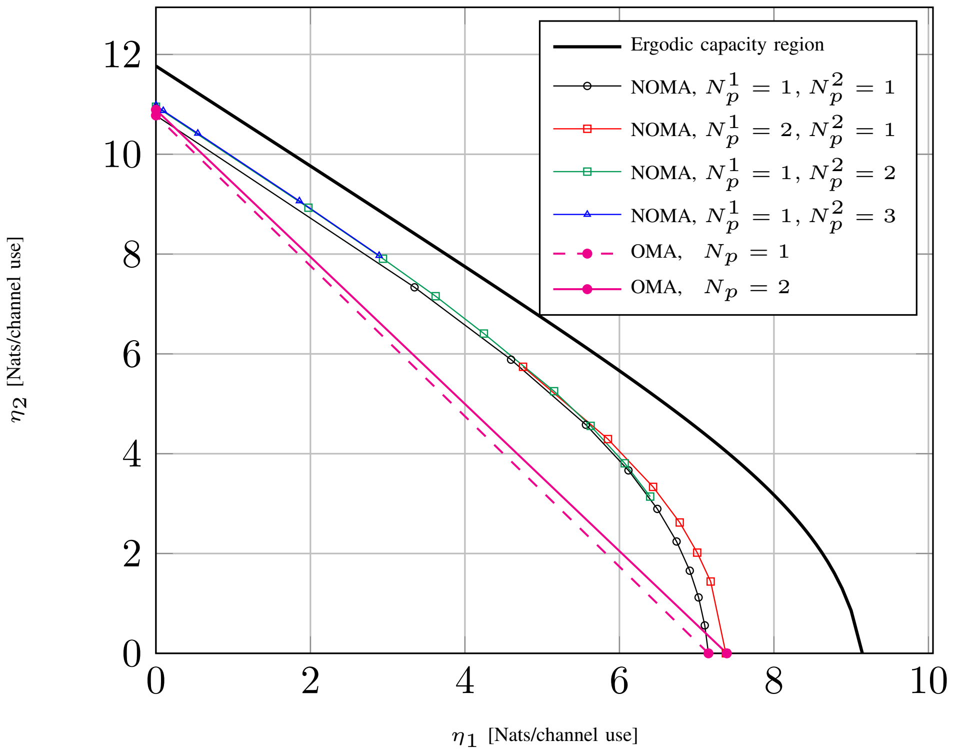

Figure 5. Achievable throughput regions. $I = 6$, $M = 2$, $(r_1, r_2) = (u_1, u_5)$, SNR = 50 dB (high SNR regime).

Table II
Maximum achievable SNR gain for multi-packet HARQ ($N_p^1 = N_p^2 = 2$) versus single-packet HARQ ($N_p^1 = N_p^2 = 1$).

SNR [dB]	20	30	40	50	50	50	50
(r_1, r_2, I)	$(1,2,2)$	$(1,2,2)$	$(1,2,2)$	$(1,2,2)$	$(3,6,6)$	$(2,4,6)$	$(1,5,6)$
Max gain [dB]	0.30	0.65	0.92	1.18	1	1.05	1.05

that multi-packet HARQ outperforms single-packet HARQ and it can clearly enlarge the achievable throughput region when SNR $= 50$ dB. On the other hand, it is observed in Fig. 12 that this gain is smaller when SNR $= 20$ dB.

B. Effect of imperfect SIC

SIC receivers are imperfect in practice and the performance can be affected by error propagation. In order to study the impact of imperfect SIC on the performance of NOMA-HARQ systems, we introduce the error propagation factor denoted by Q, where $0 \leq Q \leq 1$ [19]. Then, the performance analysis in the previous sections is extended to the case of imperfect SIC by replacing the function $C(s, p, i, j)$ in (9) with (50). Specifically, $Q = 0$ represents perfect SIC and $Q = 1$ corresponds to the worst case that SIC is totally unsuccessful. Figure 13 shows the effect of imperfect SIC on the achievable throughput regions of the NOMA pair (u_1, u_5) for $Q \in \{0.05, 0.01, 0.005, 0.001\}$. It can be observed that the achievable throughput of NOMA-HARQ systems decreases with Q. When $Q \geq 0.05$, NOMA-HARQ cannot achieve any gain with respect to OMA-HARQ. In this case, imperfect SIC is therefore a major limiting factor for NOMA systems to outperform OMA. Consequently, it is crucial to use HARQ with powerful error-correcting codes having high error detection capability in NOMA systems to cope with the issue of imperfect SIC.

$$C(s,p,i,j) \triangleq \log\left[1 + \frac{s \cdot p \cdot \alpha_j^{u_i}}{1 + s \cdot p \cdot \left[\sum\limits_{j<j'\leq N_p^i} \alpha_{j'}^{u_i} + \sum\limits_{i'>i} \sum\limits_{1\leq j'\leq N_p^{i'}} \alpha_{j'}^{u_{i'}}\right] + Q \cdot s \cdot p \cdot \left[\sum\limits_{j'<j} \alpha_{j'}^{u_i} + \sum\limits_{i'<i} \sum\limits_{1\leq j'\leq N_p^{i'}} \alpha_{j'}^{u_{i'}}\right]}\right]. \quad (50)$$

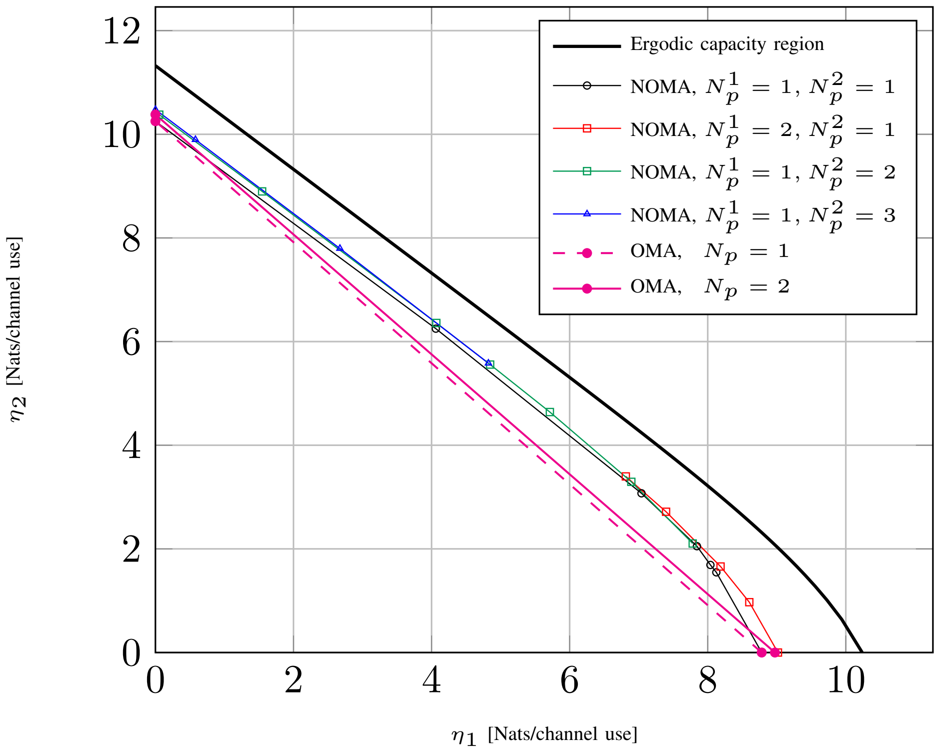

Figure 6. Achievable throughput regions. $I = 6$, $M = 2$, $(r_1, r_2) = (u_2, u_4)$, SNR = 50 dB (high SNR regime).

VI. Conclusions

In this paper, we studied the throughput performance of single-packet and multi-packet HARQ in downlink NOMA systems. We assumed that the multiple packets of each user are multiplexed on the same channel block using PD-NOMA scheme. We theoretically analyzed the throughput of blanking-based HARQ in downlink NOMA systems and provided an analytical expression of the throughput which depends on single-integral forms when the maximum number of transmissions is limited to two. The rate and power allocation were optimized for all packets in order to maximize the weighted-sum throughput. Simulation results, for both PD-NOMA and SCMA systems, have shown that the gain of multi-packet HARQ over the single-packet HARQ depends on the SNR as well as the target throughput of users. Furthermore, the SNR gain of multi-packet HARQ over single-packet HARQ increases with the SNR, and it can reach up to 1 dB in the high SNR regime (for SNR values up to $40 - 50$ dB). For practical SNR values, e.g. 20 dB, single-packet HARQ achieves almost similar performance to multi-packet HARQ. Simulation results confirmed also that HARQ-NOMA schemes achieve better throughput than HARQ-OMA schemes only when the error propagation rate of the SIC detector is lower than a certain threshold.

This work opens the door to many future directions. Specifically, the performance of cross-packet HARQ [31], layered-HARQ [32], and time-sharing HARQ [30] could be studied in NOMA systems. Moreover, future works could investigate the effect of delayed-feedback on the performance of multi-packet HARQ in NOMA systems as well as the extension to NOMA systems with multiple antennas [16].

References

[1] Huawei, “5G Scenarios and Security Designs,” White paper, November 2016. [Online]. Available: https://www-file.huawei.com/-/media/corporate/pdf/white%20paper/5g-scenarios-and-security-design.pdf

[2] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple access techniques for 5G wireless networks and beyond*. Springer, 2019.

[3] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, “Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

[4] 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, no. 4, pp. 1616–1626, 2008.

[5] H. Nikopour and H. Baligh, “Sparse code multiple access,” in *IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC)*, 2013, pp. 332–336.

[6] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, “ Pattern Division Multiple Access—A Novel Nonorthogonal Multiple Access for Fifth-Generation Radio Networks,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3185–3196, April 2017.

[7] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, “SCMA codebook design,” in *IEEE 80th Vehicular Technology Conference (VTC Fall)*, 2014, pp. 1–5.

[8] Z. Mheich, L. Wen, P. Xiao, and A. Maaref, “Unequal Error Protection SCMA Codebooks,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 4055–4058, April 2019.

[9] ——, “Design of SCMA Codebooks Based on Golden Angle Modulation,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1501–1509, Feb 2019.

[10] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, “On the Performance Gain of NOMA Over OMA in Uplink Communication Systems,” *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 536–568, Jan 2020.

[11] Q. Yang, H. Wang, D. W. K. Ng, and M. H. Lee, “NOMA in Downlink SDMA With Limited Feedback: Performance Analysis and Optimization,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2281–2294, Oct 2017.

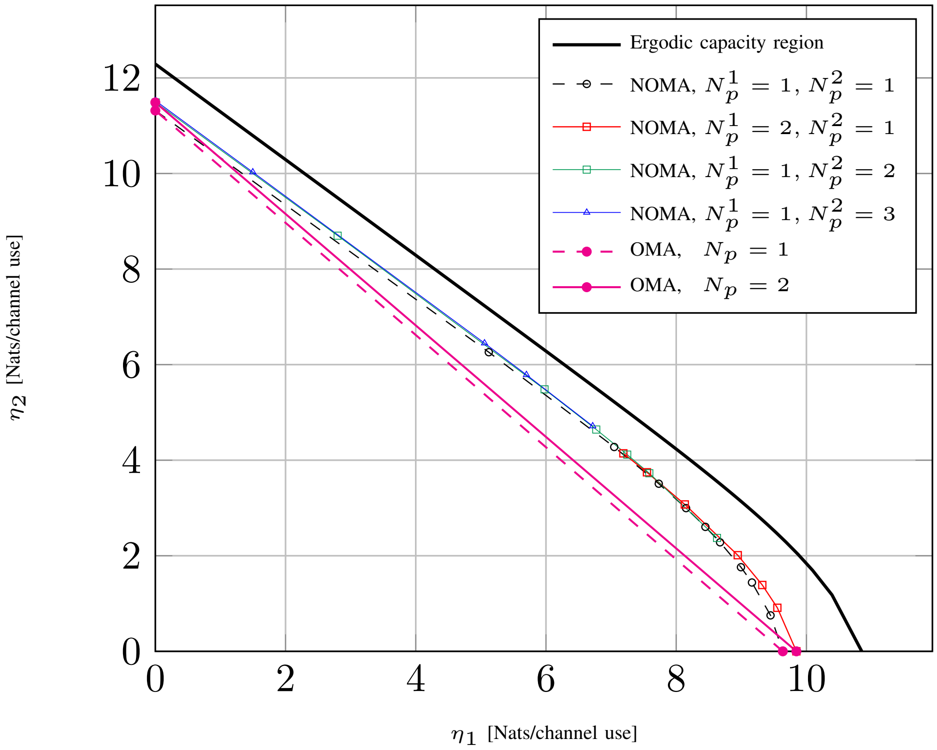

Figure 7. Achievable throughput regions. $I = 6$, $M = 2$, $(r_1, r_2) = (u_3, u_6)$, SNR = 50 dB (high SNR regime).

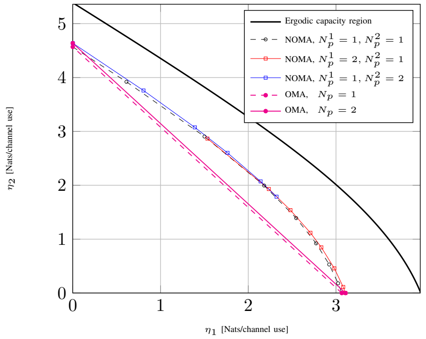

Figure 8. Achievable throughput regions. $M = 2$, $(r_1, r_2) = (u_3, u_6)$, 20 dB.

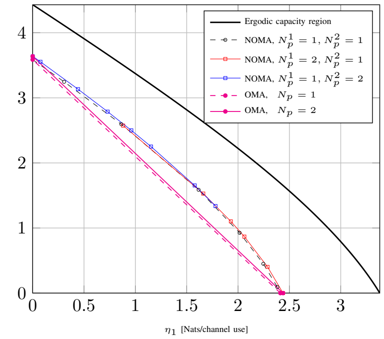

Figure 9. Achievable throughput regions. $M = 2$, $(r_1, r_2) = (u_2, u_4)$, 20 dB.

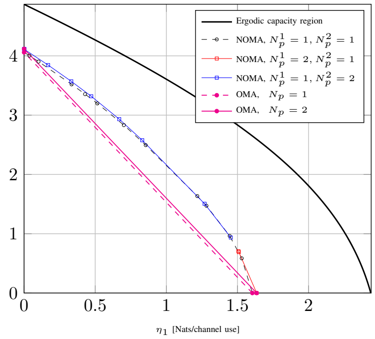

Figure 10. Achievable throughput regions. $M = 2$, $(r_1, r_2) = (u_1, u_5)$, 20 dB.

[12] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay Minimization for NOMA-MEC Offloading," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1875–1879, Dec 2018.

[13] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the Performance of Non-Orthogonal Multiple Access in 5G Systems with Randomly Deployed Users," *IEEE Signal Processing Letters*, vol. 21, no. 12, pp. 1501–1505, Dec 2014.

[14] M. Moltafet, N. M. Yamchi, M. R. Javan, and P. Azmi, "Comparison Study Between PD-NOMA and SCMA," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 2, pp. 1830–1834, Feb 2018.

[15] W. Yu, L. Musavian, and Q. Ni, "Link-Layer Capacity of NOMA Under Statistical Delay QoS Guarantees," *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4907–4922, Oct 2018.

[16] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "On the Performance of Cell-Free Massive MIMO Relying on Adaptive NOMA/OMA Mode-Switching," *IEEE Transactions on Communications*, vol. 68, no. 2, pp. 792–810, 2020.

[17] W. Yu, A. Chorti, L. Musavian, H. V. Poor, and Q. Ni, "Effective secrecy rate for a downlink NOMA network," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5673–5690, 2019.

[18] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen, "Rethink Hybrid Automatic Repeat reQuest Design for 5G: Five Configurable Enhancements," *IEEE Wireless Communications*, vol. 24, no. 6, pp. 154–160, Dec 2017.

[19] Z. Shi, S. Ma, H. ElSawy, G. Yang, and M. Alouini, "Cooperative HARQ-Assisted NOMA Scheme in Large-Scale D2D Networks," *IEEE Transactions on Communications*, vol. 66, no. 9, pp. 4286–4302, Sep. 2018.

[20] 3GPP, "TS38.321 Medium Access Control (MAC) protocol specification," 2019. [Online]. Available: http://www.3gpp.org/ftp//Specs/archive/38_series/38.321/

[21] J. Choi, "On HARQ-IR for Downlink NOMA Systems," *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3576–3584, Aug 2016.

[22] D. Cai, Z. Ding, P. Fan, and Z. Yang, "On the Performance of NOMA With Hybrid ARQ," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 10, pp. 10 033–10 038, Oct 2018.

[23] J. Choi, "On Multiple Access Using H-ARQ with SIC Techniques for Wireless Ad Hoc Networks," *Wireless Personal Communication*, vol. 69, no. 1, pp. 187–212, Mar 2013.

[24] Z. Shi, C. Zhang, Y. Fu, H. Wang, G. Yang, and S. Ma, "Achievable Diversity Order of HARQ-Aided Downlink NOMA Systems," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 471–487, Jan

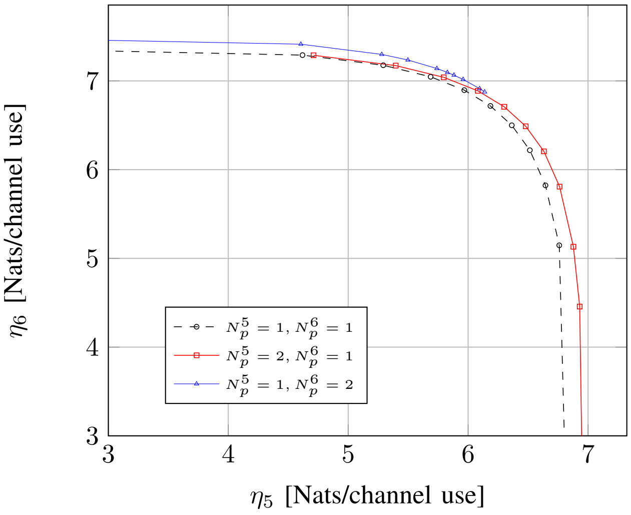

Figure 11. Achievable throughput regions of the fifth and sixth users of an SCMA of 6 users sharing 4 resources. $M = 2$, SNR = 50 dB, $\eta_1 = \eta_2 = 1.987$ and $\eta_3 = \eta_4 = 0.9985$.

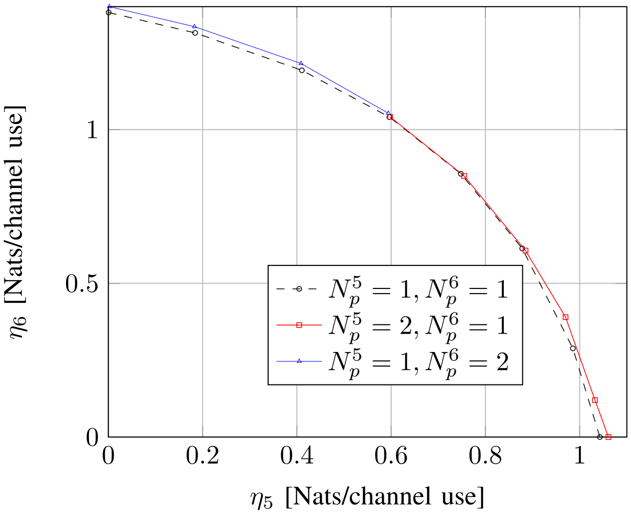

Figure 12. Achievable throughput regions of the fifth and sixth users of an SCMA of 6 users sharing 4 resources. $M = 2$, SNR = 20 dB, $\eta_1 = \eta_2 = 0.75$ and $\eta_3 = \eta_4 = 0.49$.

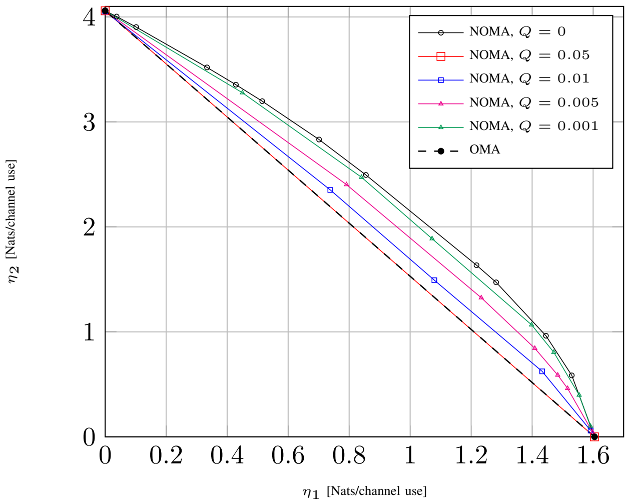

Figure 13. Achievable throughput regions with imperfect SIC. $M = 2$, $(r_1, r_2) = (u_1, u_5)$, 20 dB, $N_p^1 = 1, N_p^2 = 1$.

2020.

[25] Y. Xu, D. Cai, F. Fang, Z. Ding, C. Shen, and G. Zhu, “ HARQ-CC Enabled NOMA Designs With Outage Probability Constraints,” *arXiv:1911.01167*, 2019.

[26] B. Makki, K. Chitti, A. Behravan, and M. S. Alouini, “ A Survey of NOMA: Current Status and Open Research Challenges,” *arXiv:1912.10561*, 2019.

[27] D. Marasinghe, N. Rajatheva, and M. Latva-aho, “ Block Error Performance of NOMA with HARQ-CC in Finite Blocklength,” *arXiv:1910.13877*, 2020.

[28] A. Steiner and S. Shamai, “Multi-layer broadcasting hybrid-ARQ strategies for block fading channels,” *IEEE Transactions on Wireless Communications*, vol. 7, no. 7, pp. 2640–2650, July 2008.

[29] M. El Aoun, R. Le Bidan, X. Lagrange, and R. Pyndiah, “Multiple-packet versus single-packet incremental redundancy strategies for type-II hybrid ARQ,” in *2010 6th International Symposium on Turbo Codes Iterative Information Processing*, Sep. 2010, pp. 226–230.

[30] M. Jabi, A. E. Hamss, L. Szczecinski, and P. Piantanida, “Multipacket Hybrid ARQ: Closing Gap to the Ergodic Capacity,” *IEEE Transactions on Communications*, vol. 63, no. 12, pp. 5191–5205, Dec 2015.

[31] M. Jabi, A. Benyouss, M. Le Treust, E. Pierre-Doray, and L. Szczecinski, “Adaptive Cross-Packet HARQ,” *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2022–2035, May 2017.

[32] M. Jabi, E. Pierre-Doray, L. Szczecinski, and M. Benjillali, “How to Boost the Throughput of HARQ With Off-the-Shelf Codes,” *IEEE Transactions on Communications*, vol. 65, no. 6, pp. 2319–2331, June 2017.

[33] N. Prasad and X. Wang, “Efficient combining techniques for multi-input multi-output multi-user systems employing hybrid automatic repeat request,” *IET Communications*, vol. 5, no. 13, pp. 1785–1796, Sep. 2011.

[34] P. Larsson and N. Johansson, “Multi-User ARQ,” in *2006 IEEE 63rd Vehicular Technology Conference*, vol. 4, May 2006, pp. 2052–2057.

[35] Y. Long, Z. Chen, Z. Guo, and J. Fang, “A Novel HARQ Scheme for SCMA Systems,” *IEEE Wireless Communications Letters*, vol. 5, no. 5, pp. 452–455, Oct 2016.

[36] D. Gamal, A. H. Mehana, and K. M. F. Elsayed, “User capacity for uplink SCMA system,” *Physical Communication*, vol. 39, p. 100979, 2020.

[37] B. Di, L. Song, and Y. Li, “Radio resource allocation for uplink sparse code multiple access (SCMA) networks using matching game,” in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[38] G. Xiong and J. Sun, “An Optimal Resource Allocation Algorithm Based on Sum Rate Maximization for Uplink SCMA system,” in *2018 IEEE 18th International Conference on Communication Technology (ICCT)*, Oct 2018, pp. 805–810.

[39] T. Liu, X. Li, and L. Qiu, “Capacity for downlink massive MIMO MU-SCMA system,” in *International Conference on Wireless Communications Signal Processing (WCSP)*, Oct 2015, pp. 1–5.

[40] H. A. David and H. N. Nagaraja, *Order Statistics*. 3rd ed. John Wiley & Sons, Inc. USA, 2003.

[41] A. El Gamal and Y. H. Kim, *Network information theory*. Cambridge university press, 2011.

[42] M. Zorzi and R. R. Rao, “On the use of renewal theory in the analysis of ARQ protocols,” *IEEE Transactions on Communications*, vol. 44, no. 9, pp. 1077–1081, Sep. 1996.

[43] E.-G. Talbi, *Metaheuristics: from design to implementation*. John Wiley & Sons, 2009, vol. 74.

[44] T. V. K. Chaitanya and E. G. Larsson, “Outage-Optimal Power Allocation for Hybrid ARQ with Incremental Redundancy,” *IEEE Transactions on Wireless Communications*, vol. 10, no. 7, pp. 2069–2074, July 2011.

[45] ——, “Optimal resource allocation for IR-HARQ,” in *2011 IEEE Swedish Communication Technologies Workshop (Swe-CTW)*, Oct 2011, pp. 74–79.

[46] P. Wu and N. Jindal, “Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis,” *IEEE Transactions on Communications*, vol. 58, no. 4, pp. 1129–1141, April 2010.

[47] G. Caire and D. Tuninetti, “The throughput of hybrid-ARQ protocols for the Gaussian collision channel,” *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1971–1988, July 2001.

[48] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, "Rate Allocation and Adaptation for Incremental Redundancy Truncated HARQ," *IEEE Transactions on Communications*, vol. 61, no. 6, pp. 2580–2590, June 2013.

[49] Z. Mheich, P. Duhamel, L. Szczecinski, and M. A. Morel, "Constellation shaping for broadcast channels in practical situations," in *2011 19th European Signal Processing Conference*, Aug 2011, pp. 96–100.

[50] Z. Mheich, F. Alberge, and P. Duhamel, "Achievable rates optimization for broadcast channels using finite size constellations under transmission constraints," *EURASIP J. Wireless Comm. and Networking*, vol. 2013, p. 254, 2013.