On HARQ-IR for Downlink NOMA Systems

Jinho Choi

***Abstract*— Non-orthogonal multiple access (NOMA) can exploit the power difference between the users to achieve a higher spectral efficiency. Thus, the power allocation plays a crucial role in NOMA. In this paper, we study the power allocation for hybrid automatic repeat request (HARQ) in NOMA with two users. For the power allocation, we consider the error exponents of the outage probabilities in HARQ with incremental redundancy (IR) and derive them based on large deviations. While a closed-form expression for the error exponent (or rate function) without interference is available, there is no closed-form expression for the error exponent with interference. Thus, we focus on the derivation of a lower bound on the error exponent in this paper. Based on the error exponents, we formulate a power allocation problem for HARQ-IR in NOMA to guarantee a certain low outage probability for a given maximum number of retransmissions. From the simulation results, we can confirm that it is possible to guarantee a certain outage probability by the proposed power allocation method.**

***Index Terms*— Non-orthogonal multiple access, HARQ, power allocation, error exponent.**

I. Introduction

RECENTLY, non-orthogonal multiple access (NOMA) has been studied for cellular systems in order to improve the spectral efficiency [1]–[4]. As opposed to orthogonal multiple access (OMA), in NOMA, multiple users of different channel gains within a cell can be supported in the same frequency band and time slot (or radio resource block) simultaneously by exploiting the power domain. For example, if there are one user of high channel gain (for convenience, this user is referred to as user 1) and the other user of low channel gain (this user is referred to as user 2), a base station (BS) can transmit two signals to both users simultaneously using superposition coding [5] or multiuser superposition transmission (MUST) schemes [6]. In this case, the BS usually allocates a high transmission power to user 2 and a low transmission power to user 1. From this, at user 1, the signal to user 2 might be decodable when superposition coding is employed. Thus, for successive interference cancellation (SIC) at user 1, user 1 decodes the signal to user 2 first and then decodes his/her signal after subtracting the decoded signal to user 2. At user 2, the signal to user 2 can be decoded without significant interference from the signal to user 1, which is weak. NOMA can be extended for downlink coordinated two-point systems [7] and multiple input multiple output (MIMO) systems [4].

Manuscript received March 31, 2016; revised May 20, 2016; accepted June 23, 2016. Date of publication June 28, 2016; date of current version August 12, 2016. This work was supported by the GIST Research Institute (GRI) in 2016. The associate editor coordinating the review of this paper and approving it for publication was Z. Ding.

The author is with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 500 712, South Korea (e-mail: jchoi0114@gist.ac.kr).

Digital Object Identifier 10.1109/TCOMM.2016.2585651

While SIC plays a crucial role in NOMA, in [8], it is also shown that the gain of NOMA over OMA can be achieved within a constant gap without using SIC at a receiver. In [9], the power allocation is investigated with actual achievable rates for a practical MUST scheme.

For reliable transmissions, hybrid automatic repeat request (HARQ) protocols can be employed [10], [11]. Among various HARQ protocols, an HARQ protocol with incremental redundancy (HARQ-IR) is known to achieve the capacity [12]. In [13], the performance of HARQ-IR is studied in block-fading channels. In [14], asymptotic performances of HARQ-IR are considered using the notion of large deviations [15], [16]. Recently, HARQ protocols are applied to NOMA in [17].

In this paper, we study HARQ-IR for downlink NOMA, which is referred to as NOMA-HARQ-IR for convenience. For NOMA-HARQ-IR, we consider superposition coding for two users who share a common radio resource block. Each user sends binary feedback signals for acknowledgment (ACK) or negative-acknowledgment (NACK). In NOMA-HARQ-IR, it is assumed that the BS does not know instantaneous CSI, but statistical CSI. Thus, it is important to decide the transmission rates and powers based on statistical CSI for reasonable performances. In order to decide the transmission rates and powers, we consider the outage probabilities for a given maximum number of retransmissions.

While the outage probability without interference is relatively well studied in HARQ-IR [13], [14], that with interference, which is related to the case of decoding of the signal to user 2 in NOMA-HARQ-IR, is not investigated. Thus, we focus on the outage probabilities of the signal to user 2 at users 1 and 2. Due to the interfering signal, which is the signal to user 1, it is not easy to derive a closed-form expression for the outage probability. Thus, we consider an asymptotic case based on large deviations and derive a lower-bound on the exponent of the outage probability. With the derived exponents of the outage probabilities, we consider a power allocation problem to maximize the minimum of the exponents. While the resulting power allocation approach is robust in terms of minimizing the outage probability, it does not provide an optimal power allocation result to maximize the throughput (note that the maximization of the throughput requires closed-form expressions for the outage probabilities, which are not available). Furthermore, since lower-bounds on the exponents of the outage probabilities of the signal to user 2 are considered, the resulting power allocation would be biased to the signal to user 2 (i.e., the more power would be allocated to the signal to user 2 than needed). However, due to lower-bounds on the exponents, the proposed approach can guarantee a certain low outage probability for a given maximum number

of retransmissions, which might be important for applications requiring reliable transmissions.

Note that NOMA-HARQ-IR in this paper differs from HARQ for NOMA in [17]. In this paper, we consider power allocation with statistical CSI as mentioned earlier, while in [17], power allocation is carried out with CSI feedback. CSI feedback may be used in NOMA for a better performance as in [18]. However, in this paper, we only assume binary feedback (ACK or NACK) as usual in HARQ protocols and binary feedback is not used for power allocation (i.e., power allocation is fixed during retransmissions). In addition, in [17], Chase combining (CC) is considered, while IR[1] is used for HARQ in this paper.

The rest of the paper is organized as follow. In Section II, we present the system model of NOMA with HARQ-IR. In Section III, using large deviations, we derive the error exponents of the outage probabilities. To guarantee a certain outage probability in NOMA-HARQ-IR, we consider a power allocation scheme using the error exponents in Section IV. We discuss simulation results in Section V. The paper is concluded in Section VI with some remarks.

Notation: Matrices and vectors are denoted by upper- and lower-case boldface letters, respectively. $\mathbb{E}[\cdot]$ and $\mathrm{Var}(\cdot)$ denote the statistical expectation and variance, respectively. $\mathcal{CN}(\mathbf{a}, \mathbf{R})$ represents the distribution of circularly symmetric complex Gaussian (CSCG) random vector with mean vector $\mathbf{a}$ and covariance matrix $\mathbf{R}$. The indicator function is denoted by $\mathbb{1}(\mathcal{A})$, which is 1 if $\mathcal{A}$ is true and 0 otherwise.

II. System Model

In this section, we present the system model for a NOMA system with two users and apply HARQ-IR to NOMA.

A. NOMA System With Two Users

Suppose that there are two users for downlink NOMA. For convenience, we assume that user 1 is close to the BS and user 2 is far away from the BS. We assume block fading channels, where h_t and g_t denote the channel coefficients from the BS to user1 and user 2, respectively, during time slot t. It is expected that $\mathbb{E}[|h_t|^2] > \mathbb{E}[|g_t|^2]$ as user 1 is closer to the BS than user 2. We assume that during each time slot, a coded signal block is to be transmitted to each user. Let $\mathbf{s}_{k,t}$ denote the coded signal block of length N, i.e., $\mathbf{s}_{k,t} \in \mathbb{C}^N$, to user k during time slot t. Then, the received signal at user 1 and user 2 during time slot t, denoted by $\mathbf{r}_{(1),t}$ and $\mathbf{r}_{(2),t}$, respectively, are given by

$$\begin{aligned} \mathbf{r}_{(1),t} &= h_t(\mathbf{s}_{1,t} + \mathbf{s}_{2,t}) + \mathbf{n}_t \\ \mathbf{r}_{(2),t} &= g_t(\mathbf{s}_{1,t} + \mathbf{s}_{2,t}) + \tilde{\mathbf{n}}_t, \end{aligned} \tag{1}$$

where $\mathbf{n}_t \in \mathbb{C}^N$ and $\tilde{\mathbf{n}}_t \in \mathbb{C}^N$ are the background noise vectors at user 1 and user 2, respectively, and assumed to be independent and $\mathbf{n}_t, \tilde{\mathbf{n}}_t \sim \mathcal{CN}(0, \mathbf{I})$.

Throughout the paper, we assume that $\mathbf{s}_{k,t} \sim \mathcal{CN}(0, P_k\mathbf{I})$ for tractable[2] analysis, where P_k is the transmission power to user k. Furthermore, we assume capacity achieving codes for $\mathbf{s}_{k,t}$. As in [19], the power allocation is important for NOMA. For coded systems, if each signal block is encoded independently with code rate R_k for user k, the power allocation is to be satisfied the following inequalities:

$$\begin{aligned} \log_2\left(1 + \frac{\alpha_t P_2}{\alpha_t P_1 + 1}\right) &\geq R_2 \\ \log_2(1 + \alpha_t P_1) &\geq R_1, \end{aligned} \tag{2}$$

and

$$\log_2\left(1 + \frac{\beta_t P_2}{\beta_t P_1 + 1}\right) \geq R_2, \tag{3}$$

where $\alpha_t = |h_t|^2$ and $\beta_t = |g_t|^2$. If the conditions in (2) hold, user 1 can decode $\mathbf{s}_{2,t}$ first. Then, after subtracting $\mathbf{s}_{2,t}$ from $\mathbf{r}_{(1),t}$, user 1 can decode $\mathbf{s}_{1,t}$ without any interference of $\mathbf{s}_{2,t}$. At user 2, the signal to user 1, i.e., $\mathbf{s}_{1,t}$, is assumed to be interference and decoding is carried out with it. Thus, the condition in (3) is necessary for successful decoding.

As shown in (2) and (3), for the power allocation, the BS needs to know the channel coefficients, which may require CSI feedback from users. To avoid this, we may consider a different approach such as HARQ protocols that only require binary feedback (ACK or NACK) for reliable transmissions.

B. HARQ-IR for NOMA

In this section, we apply HARQ-IR to NOMA for reliable transmissions with users' binary feedback signals, which is referred to as NOMA-HARQ-IR throughout the paper.

For convenience, define

$$\begin{aligned} V_{(1),t} &= \log_2\left(1 + \frac{\alpha_t P_2}{\alpha_t P_1 + 1}\right) \\ V_{(2),t} &= \log_2\left(1 + \frac{\beta_t P_2}{\beta_t P_1 + 1}\right) \\ W_t &= \log_2(1 + \alpha_t P_1). \end{aligned} \tag{4}$$

In HARQ-IR, R_k becomes the initial rate [13], as the effective code rate decreases with retransmissions [12]. Denote by $T_{(k),m}$ the the number of retransmissions to decode the signal to user m at user k. Then, we have

$$T_{(1),2} = \min_T \left\{ T \,\middle|\, \sum_{t=1}^{T} V_{(1),t} \geq R_2 \right\} \tag{5}$$

and

$$T_{(2),2} = \min_T \left\{ T \,\middle|\, \sum_{t=1}^{T} V_{(2),t} \geq R_2 \right\}. \tag{6}$$

If there is any user who cannot decode the signal to user 2, the BS has to retransmit. From this, the number of retransmissions of the signal to user 2 in NOMA becomes

$$T_2 = \max\{T_{(1),2}, T_{(2),2}\}. \tag{7}$$

At user 1, since the signal to user 2 is to be decoded first, the number of retransmissions of the signal to user 1 is given by

$$T_1 = \max\left\{ \min_T \left\{ T \,\middle|\, \sum_{t=1}^{T} W_t \geq R_1 \right\}, T_2 \right\}. \tag{8}$$

[1]Note that IR can provide a higher spectral efficiency than repetition time diversity (RTD) for CC [12].

[2]That is, we consider a Gaussian codebook of codewords of length N. This code can achieve the capacity and simplify analysis.

In fact, T_1 is an upper-bound as the BS can retransmit coded blocks to user 1 with a total power of $P_1 + P_2$ without any signal to user 2 once both the users send the ACK feedback of the signal to user 2. However, for tractable analysis, we assume that P_1 remains unchanged. As a result, in NOMA-HARQ-IR, the total number of retransmissions becomes T_1 in (8). From (8), we can see that it is desirable that the numbers of retransmissions are the same, i.e.,

$$T_{(1),2} = T_{(2),2} = T_1 \tag{9}$$

to minimize the number of retransmissions. Unfortunately, since the channel coefficients are random and unknown to the transmitter, we cannot always guarantee (9). In addition, in practice, it may be desirable to have a smaller number of retransmissions for the signal to user 2 as $T_1 \geq T_2$ according to (8).

In this paper, we focus on the rate determination and power allocation for NOMA-HARQ-IR based on the *statistical CSI* of user 1 and channel 2. We will consider the outage probabilities to decide the rates and powers, which are decided prior to transmissions in NOMA-HARQ-IR. Note that it might be possible to adapt the powers during retransmissions based on CSI feedback as in [18], which may provide a better performance. But, we do not consider this in this paper. Furthermore, we consider the case that the channel coefficients are independent and identically distributed (iid) (i.e., randomly varying from one block to another) as will be assumed in (12), causal CSI feedback does not help to adaptively decide the power for the next block. For a given maximum (target) number of retransmissions, T, the rates and powers can be decided to have sufficiently low outage probabilities of the signals as follows:

$$\Pr\left(\sum_{i=1}^{T} W_t \leq R_1\right) \leq \delta, \tag{10}$$

$$\Pr\left(\sum_{i=1}^{T} V_{(m),t} \leq R_2\right) \leq \delta, \ m = 1, 2, \tag{11}$$

where δ is a sufficiently small constant. In this case, successful transmissions can be achieved within T retransmissions with a high probability (say $\geq 1 - \delta$). It is noteworthy that this approach does not necessarily maximize the throughput.

III. Error Exponents of Outage Probabilities

In this section, we focus on the error exponents of the outage probabilities in (11) and (10) under the following assumption:

A) α_t and β_t are independent and follow the following distributions:

$$\alpha_t \sim \frac{1}{\bar{\alpha}} \exp\left(-\frac{\alpha_t}{\bar{\alpha}}\right), \ \alpha_t \geq 0;$$
$$\beta_t \sim \frac{1}{\bar{\beta}} \exp\left(-\frac{\beta_t}{\bar{\beta}}\right), \ \beta_t \geq 0, \tag{12}$$

where $\bar{\alpha} = \mathbb{E}[\alpha_t]$ and $\bar{\beta} = \mathbb{E}[\beta_t]$. That is, we consider independent Rayleigh fading channels for users 1 and 2. In addition, we assume that

$$\bar{\alpha} \geq \bar{\beta}. \tag{13}$$

A. Outage Probability of Signals to User 1

In this subsection, we consider an asymptotic analysis of the outage probability of the signal to user 1 at user 1 when the signal to user 2 is removed, i.e., that in (10).

Using large deviations [15], [16], under **A)**, we can find a closed-form expression for an upper-bound on the outage probability in (10). For iid $\{\alpha_t\}$ or $\{W_t\}$, using the Chernoff bound [5], we have

$$\Pr\left(\sum_{t=1}^{T} W_t \leq T\mu_1\right) \leq \min_{s\geq 0} e^{sT\mu_1} \left(\mathbb{E}[e^{-sW_t}]\right)^T,$$

where $\mu_1 < \mathbb{E}[W_t]$ is a parameter[3] to be decided for R_1. Note that for a given T, we have $R_1 = \mu_1 T$ from (10). According to [15] and [16], it can be further shown that

$$\Pr\left(\sum_{t=1}^{T} W_t \leq T\mu_1\right) \doteq \exp\left(-T\ell_1(\mu_1, P_1)\right), \tag{14}$$

where $\ell_1(\mu_1, P_1)$ is the rate function and $\doteq$ represent the asymptotic equality that is given by

$$f(T) \doteq g(T) \Rightarrow \lim_{T\to\infty} \frac{1}{T} \ln \frac{f(T)}{g(T)} = 0.$$

In [14], the rate function in (14) is derived, which is given by

$$\begin{aligned}\ell_1(\mu_1, P_1) &= \max_{s\geq 0} \{s\mu_1 - \lambda(s)\} \\ &= \max_{s\geq 0} \Big\{s\mu_1 - M \ln(\bar{\alpha} P_1) \\ &\qquad + \ln \psi\left(1, 2 - \frac{s}{\log 2}, \frac{1}{\bar{\alpha} P_1}\right)\Big\},\end{aligned} \tag{15}$$

where $\lambda(s) = \ln \mathbb{E}[e^{-sW_t}]$ and $\psi(a, b, z)$ is the confluent hypergeometric function of the second kind [20, p. 344]. For convenience, $\ell_1(\mu_1, P_1)$ is referred to as the (negative) exponent of the outage probability in (10).

B. Outage Probabilities of Signals to User 2

In this subsection, we study the outage probabilities of the signal to user 2 at both users 1 and 2 for a given T, which are in (11). Due to the presence of interference in this case, the analysis is not easy. Thus, we consider bounds in this subsection for tractable analysis.

Consider the following lower-bound on the signal-to-interference-plus-noise ratio (SINR) at user 1 when decoding the signal to user 2:

$$\begin{aligned}\frac{P_2}{P_1 + \frac{1}{\alpha_t}} &\geq D_{(1),2}(\epsilon) \\ &= \frac{P_2}{P_1 + 1/\epsilon} \mathbb{1}(\alpha_t \geq \epsilon),\end{aligned} \tag{16}$$

where $\epsilon > 0$ is a parameter to be decided. Since

$$V_{(1),t} \geq \log_2\left(1 + D_{(1),2}\right),$$

[3]The value of μ_1 has to be decided to keep the outage probability low. In (36), a feasible range for μ_1 is derived for a given total transmission power.

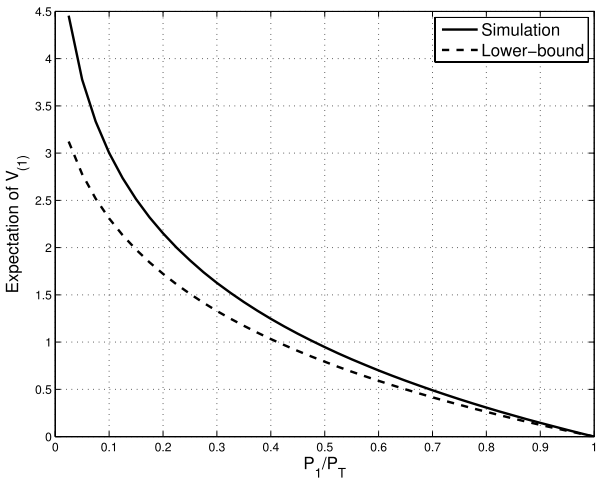


Fig. 1. Comparisons of $\mathbb{E}[V_{(1),t}]$ and lower-bound, $\mathbb{E}[\tilde{V}_{(1),t}]$, when $P_T = 20$ dB and $\bar{\alpha} = 1$.

we can decide ϵ to make $\log_2\left(1 + D_{(1),2}\right)$ a better lower-bound on $V_{(1),t}$ in terms of the mean as follows:

$$\hat{\epsilon}_1 = \underset{\epsilon \geq 0}{\operatorname{argmax}}\, \mathbb{E}\left[\log_2\left(1 + D_{(1),2}\right)\right]. \tag{17}$$

Using (16), we have a binary random variable to approximate $V_{(1),t}$ as follows:

$$\tilde{V}_{(1),t} = \begin{cases} \log_2\left(1 + \dfrac{\hat{\epsilon}_1 P_2}{\hat{\epsilon}_1 P_1 + 1}\right), & \text{if } \alpha_t \geq \hat{\epsilon}_1; \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Fig. 1 shows $\mathbb{E}[V_{(1),t}]$ and $\mathbb{E}[\tilde{V}_{(1),t}]$ for different values of P_1 when $P_T = 20$ dB and $\bar{\alpha} = 1$, where P_T represents the total transmission power (i.e., $P_T = P_1 + P_2$). We can see that $\tilde{V}_{(1),t}$ can be a good lower-bound when P_1 approaches P_T.

Similarly, we can also define a lower bound on $V_{(2),t}$ at user 2 as

$$\tilde{V}_{(2),t} = \begin{cases} \log_2\left(1 + \frac{\hat{\epsilon}_2 P_2}{\hat{\epsilon}_2 P_1 + 1}\right), & \text{if } \beta_t \geq \hat{\epsilon}_2; \\ 0, & \text{otherwise,} \end{cases} \tag{19}$$

where

$$\hat{\epsilon}_2 = \underset{\epsilon \geq 0}{\operatorname{argmax}}\, \mathbb{E}\left[\log_2\left(1 + \frac{P_2}{P_1 + 1/\epsilon}\mathbb{1}(\beta_t \geq \epsilon)\right)\right]. \tag{20}$$

Property 1: Under **A)**, there are unique solutions to (17) and (20). In addition,

$$\hat{\epsilon}_1 \geq \hat{\epsilon}_2. \tag{21}$$

Proof: See Appendix A. ∎

Since $\tilde{V}_{(m),t} \leq V_{(m),t}$, we can have upper-bounds on the following outage probabilities:

$$\Pr\left(\sum_{t=1}^{T} V_{(m),t} \leq T\mu_2\right) \leq \Pr\left(\sum_{t=1}^{T} \tilde{V}_{(m),t} \leq T\mu_2\right), \ m = 1, 2, \tag{22}$$

where μ_2 is a parameter to be decided for R_2. Since we have binary random variables in (22), it is easy to find the upper-bound on the outage probability using large deviations.

For convenience, define

$$\omega_{(m)} = \log_2\left(1 + \frac{\hat{\epsilon}_m P_2}{\hat{\epsilon}_m P_1 + 1}\right), \ m = 1, 2. \tag{23}$$

Property 2: Under **A)**, let

$$\begin{aligned} p_{(1),2} &= \Pr(\alpha_t \geq \hat{\epsilon}_1) = e^{-\frac{\hat{\epsilon}_1}{\bar{\alpha}}} \\ p_{(2),2} &= \Pr(\beta_t \geq \hat{\epsilon}_2) = e^{-\frac{\hat{\epsilon}_2}{\bar{\beta}}} \end{aligned} \tag{24}$$

and assume that

$$\mu_2 < \min\{\omega_{(1)} p_{(1),2}, \omega_{(2)} p_{(2),2}\}. \tag{25}$$

Then, when T is a sufficiently large, for binary random variables, $\tilde{V}_{(m),t}$, we have

$$\Pr\left(\sum_{t=1}^{T} \tilde{V}_{(m),t} \leq T\mu_2\right) \doteq \exp\left(-T\ell_2\left(\frac{\mu_2}{\omega_{(m)}}, p_{(m),2}\right)\right), \tag{26}$$

where $\ell_2(q, p)$ is the rate function which is given by

$$\ell_2(q, p) = -q \ln \frac{q}{p} - (1 - q) \ln \frac{1 - q}{1 - p}. \tag{27}$$

Proof: Since $\tilde{V}_{(m),t}$ is a binary random variable, $\sum_{t=1}^{T} \tilde{V}_{(m),t}$ becomes a binomial random variable. Using the Chernoff bound for the binomial random variable, we can obtain (26). For details, see [15]. ∎

Let the (negative) exponent of the outage probability in (11) be

$$\begin{aligned} &\bar{\ell}_2(\mu_2, P_1, P_2) \\ &\quad = \min\left\{\ell_2\left(\frac{\mu_2}{\omega_{(1)}}, p_{(1),2}\right), \ell_2\left(\frac{\mu_2}{\omega_{(2)}}, p_{(2),2}\right)\right\}. \end{aligned} \tag{28}$$

According to (28), for a large T, we can claim that

$$\begin{aligned} \Pr\left(\sum_{t=1}^{T} V_{(m),t} \leq T\mu_2\right) &\leq \Pr\left(\sum_{t=1}^{T} \tilde{V}_{(m),t} \leq T\mu_2\right) \\ &\doteq e^{-T\bar{\ell}_2(\mu_2, P_1, P_2)}, \ m = 1, 2. \end{aligned} \tag{29}$$

From (29), for a given T and (P_1, P_2), we are able to decide μ_2 or $R_2 = T\mu_2$ with a sufficiently low outage probability.

IV. Rate Determination and Power Allocation for NOMA-HARQ-IR

In this section, we consider the rate determination and power allocation for NOMA-HARQ-IR with the exponents of the outage probabilities.

In order to have sufficiently low outage probabilities in HARQ-IR for a given T, from (10) and (11), we need to have

$$\mathbb{E}[W_t] > \frac{R_1}{T} = \mu_1 \tag{30}$$

$$\min\{\mathbb{E}[V_{(1),t}], \mathbb{E}[V_{(2),t}]\} > \frac{R_2}{T} = \mu_2. \tag{31}$$

Property 3: Under **A)**, we have

$$\mathbb{E}[V_{(2),t}] \leq \mathbb{E}[V_{(1),t}]. \tag{32}$$

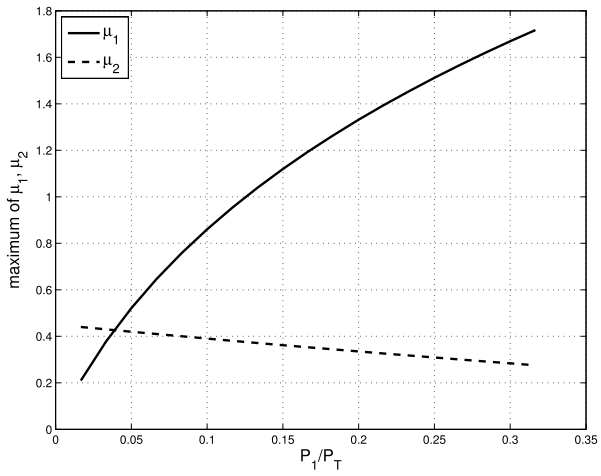


Fig. 2. The maximum values of μ_1 and μ_2 when $P_T = 10$ dB and $(\bar{\alpha}, \bar{\beta}) = (1, 1/8)$.

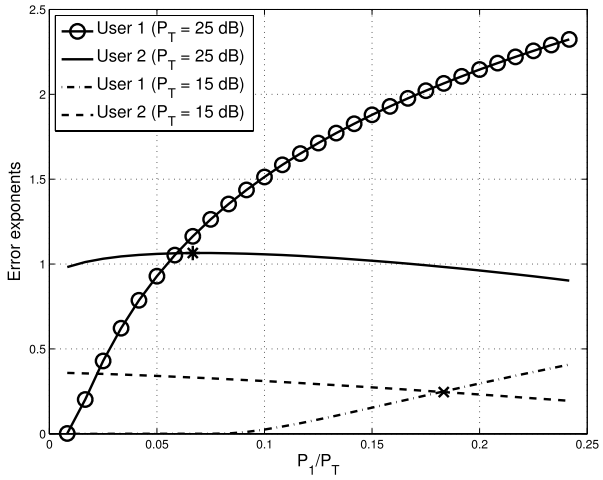


Fig. 3. The error exponents of the outage probabilities of the signals to user 1 and user 2 for different values of $\varphi = P_1/P_T$ with $P_2 = P_T - P_1$.

Proof: See Appendix B. ∎

From (32), (31) becomes

$$\mu_2 < \mathbb{E}[V_{(2),t}]. \tag{33}$$

For the power allocation, we consider the following power constraint:

$$P_1 + P_2 \leq P_T. \tag{34}$$

For convenience, let

$$\begin{aligned} \nu_1(P_1) &= \mathbb{E}[W_t] \\ \nu_2(P_1, P_2) &= \mathbb{E}[\tilde{V}_{(2),t}] \leq \mathbb{E}[V_{(2),t}]. \end{aligned} \tag{35}$$

It is easy to see that $\nu_1(P_1)$ is an increasing function of P_1 and $\nu_2(P_1, P_2)$ is an increasing function of P_2 and a decreasing function of P_1. Thus, from (34), for a given P_T, μ_1 and μ_2 should be decided to satisfy the following inequalities:

$$\begin{aligned} \mu_1 &< \bar{\mu}_1(P_T) = \nu_1(P_T) \\ \mu_2 &< \bar{\mu}_2(P_T) = \nu_2(0, P_T). \end{aligned} \tag{36}$$

From (36), we can find feasible (normalized) rates, μ_1 and μ_2, for a given P_T.

In Fig. 2, we show the upper bounds on μ_m, $m = 1, 2$, when $P_T = 10$ dB and $(\bar{\alpha}, \bar{\beta}) = (1, 1/8)$. It is shown that $\bar{\mu}_1(P_T)$ is larger than $\bar{\mu}_2(P_T)$ unless P_1 is much smaller than P_T (i.e., $P_1 \ll P_T$). Thus, μ_1 would be usually larger than μ_2. That is, the transmission rate to the user near to the BS (i.e., user 1) is higher than that to the user far away from the BS (i.e., user 2).

For a given P_T, suppose that we have decided feasible rates as $\mu_m < \bar{\mu}_m(P_T)$, $m = 1, 2$. Then, for given μ_1 and μ_2, we could decide P_1 and P_2 such that the exponents of the outage probabilities are the same. That is, as mentioned earlier, since (9) cannot be guaranteed, we consider low outage probabilities for a given T as in (10) and (11). To this end, the power allocation can be carried out as follows:

$$\begin{aligned} &\max_{P_1, P_2} \min\{\ell_1(\mu_1, P_1), \bar{\ell}_2(\mu_2, P_1, P_2)\} \\ &\text{subject to } P_1 + P_2 \leq P_T. \end{aligned} \tag{37}$$

An important feature of this power allocation is that it can be performed without the maximum number of transmissions, T, as the exponents are not functions of T. Thus, the resulting power allocation can be used for any maximum number of transmissions, T.

In addition, as in (36), the determination of the feasible rates is independent of T. Consequently, the rate determination and power allocation to provide low outage probabilities in this paper can be carried out without T.

Property 4: Suppose that μ_2 satisfies the inequality in (36). Let P_1^* and P_2^* denote the solution powers of (37). Then, we have

$$P_1^* + P_2^* = P_T. \tag{38}$$

Proof: See Appendix C. ∎

The result in Property 4 can help simplify the optimization as we only need one-dimensional search as (37) reduces to

$$\varphi^* = \underset{0 \leq \varphi \leq 1}{\operatorname{argmax}} \min\{\ell_1(\mu_1, \varphi P_T), \bar{\ell}_2(\mu_2, \varphi P_T, (1-\varphi)P_T)\}, \tag{39}$$

where $\varphi = \frac{P_1}{P_T}$. The power allocation in (39) is referred to as the error exponent based power allocation (EE-PA). Note that this power allocation is biased to user 2 as the error exponent of the outage probability of user 2 is obtained from an upper-bound on the outage probability of user 2 as in (29). Thus, the EE-PA scheme may allocate more power to the signal to user 2 than needed. However, since successful decoding of the signal to user 2 is required for SIC prior to decoding of the signal to user 1 in NOMA-HARQ-IR as in (8), this biased power allocation would not be problematic.

For convenience, we denote by ℓ^* the error exponent that corresponds to φ^*, i.e.,

$$\ell^* = \min\{\ell_1(\mu_1, \varphi^* P_T), \bar{\ell}_2(\mu_2, \varphi^* P_T, (1-\varphi^*)P_T)\}. \tag{40}$$

Clearly, the outage probabilities become less than or equal to $e^{-T\ell^*}$. Thus, the maximum number of retransmissions, T, can be decided if ℓ^* is known with a guaranteed outage probability.

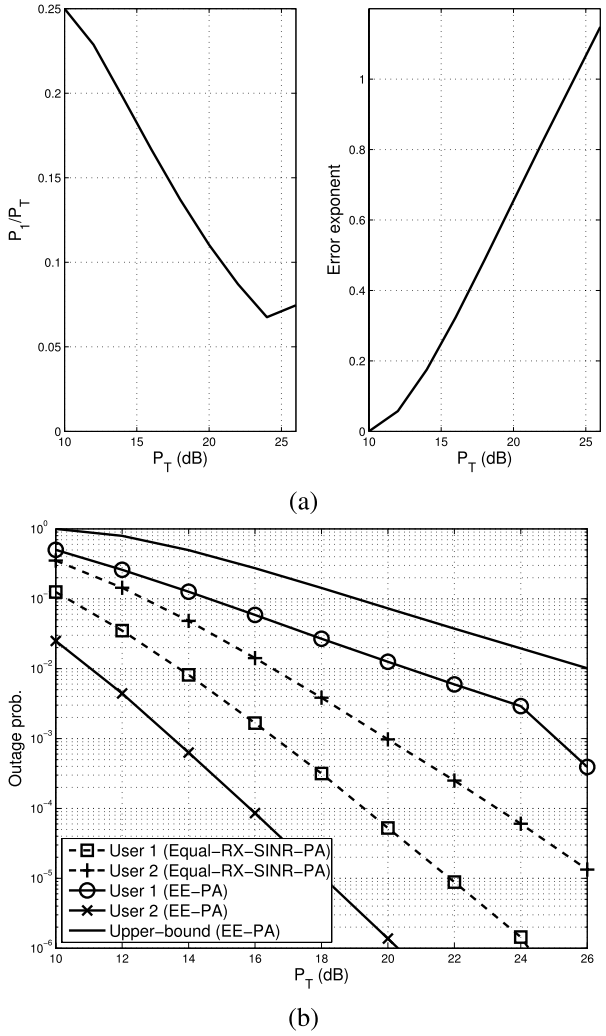


Fig. 4. Performances of NOMA-HARQ-IR for different values of P_T when $(\mu_1, \mu_2) = (1.5, 0.5)$ and $d = 1.5$: (a) the power allocation and error exponents (i.e., φ^* and ℓ^*); (b) outage probabilities with $T = 4$.

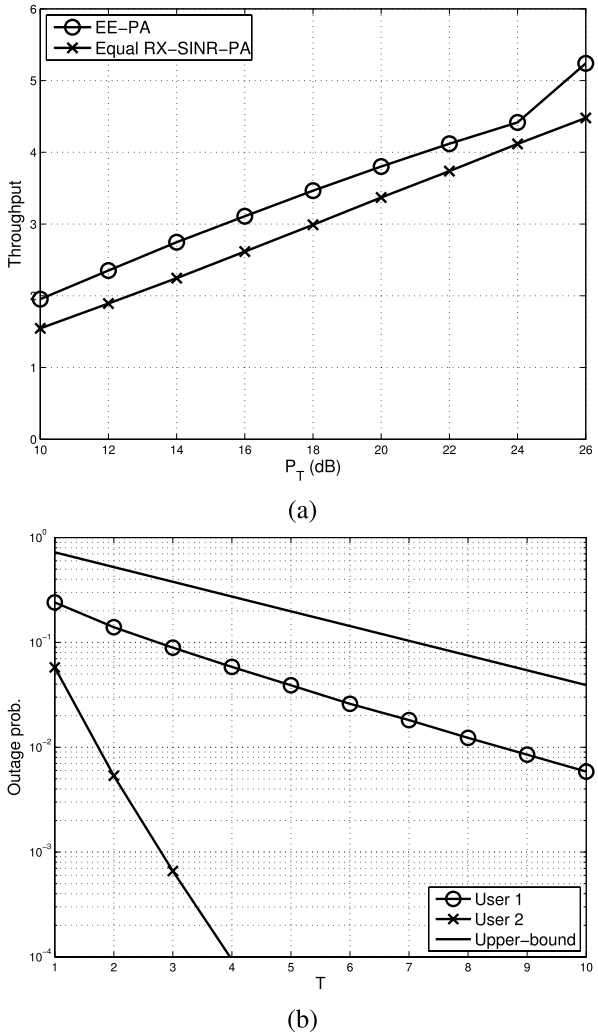


Fig. 5. Performances of NOMA-HARQ-IR with two different power allocation schemes when $(\mu_1, \mu_2) = (1.5, 0.5)$ and $d = 1.5$: (a) the throughputs versus P_T (with $T = 20$); (b) the outage probabilities versus the maximum (target) number of retransmissions T (with $P_T = 16$ dB).

Alternatively, if δ and T are given in (11) and (10), we can find P_T to guarantee $\delta \leq e^{-T\ell^*}$ using the EE-PA scheme.

For illustration purposes, we consider an example with $(\bar{\alpha}, \bar{\beta}) = (1, \frac{1}{d^\eta})$, where d is the normalized distance between user 2 and the BS (we assume that the distance between user 1 and the BS is normalized to be 1) and η is the path loss exponent. We assume that $\eta = 3$, $(\mu_1, \mu_2) = (1.5, 0.5)$, and $d = 1.5$. With $P_T = 15$ and 25 dB, the error exponents are shown in Fig. 3 for different values of $\varphi = P_1/P_T$ with $P_2 = P_T - P_1 = (1-\varphi)P_T$. The optimal value of φ is the point where the error exponents of the two users are the same when $P_T = 15$ dB, which is shown by "$\times$" mark. On the other hand, when $P_T = 25$ dB, the optimal value of φ is the maximum point of $\bar{\ell}_2(\mu_2, \varphi P_T, (1-\varphi)P_T)$, which is shown by "$*$" mark. From Fig. 3, we can see that ℓ^* with $P_T = 25$ dB is higher than that with $P_T = 15$. That is, a higher total power results in a larger error exponent. In addition, φ^* with $P_T = 25$ dB is larger than that with $P_T = 15$ dB. This implies that the power ratio to user 1, $\frac{P_1}{P_T}$, decreases as the total power P_T increases, while $\frac{P_2}{P_T}$ increases.

V. Simulation Results

In this section, we present simulation results. For comparison purposes, we consider a simple power allocation scheme that satisfies the following equality:

$$\bar{\alpha} P_1 = \frac{\bar{\beta} P_2}{\bar{\beta} P_1 + 1}, \tag{41}$$

with $P_T = P_1 + P_2$. From this power allocation, the received signal-to-noise ratio (SNR) (of the signal to user 1) at user 1 would be the same as the receive-SINR (of the signal to user 2) at user 2, which is referred to as the equal receive-SINR power allocation (equal RX-SINR-PA). For simulations, we assume that $\bar{\alpha} = 1$ and $\bar{\beta} = 1/d^\eta$ with $\eta = 3$ under **A**).

Fig. 4 (a) shows the power allocation result from the EE-PA in (39) and the corresponding error exponent, ℓ^*, for different values of P_T when $(\mu_1, \mu_2) = (1.5, 0.5)$ and $d = 1.5$. It is interesting to see that $\varphi^* = \frac{P_1}{P_T}$ decreases with P_T and then increases. As in Fig.3, when P_T is low, $\varphi^* = \frac{P_1^*}{P_T}$, is the point that makes the error exponents of the two users equal,

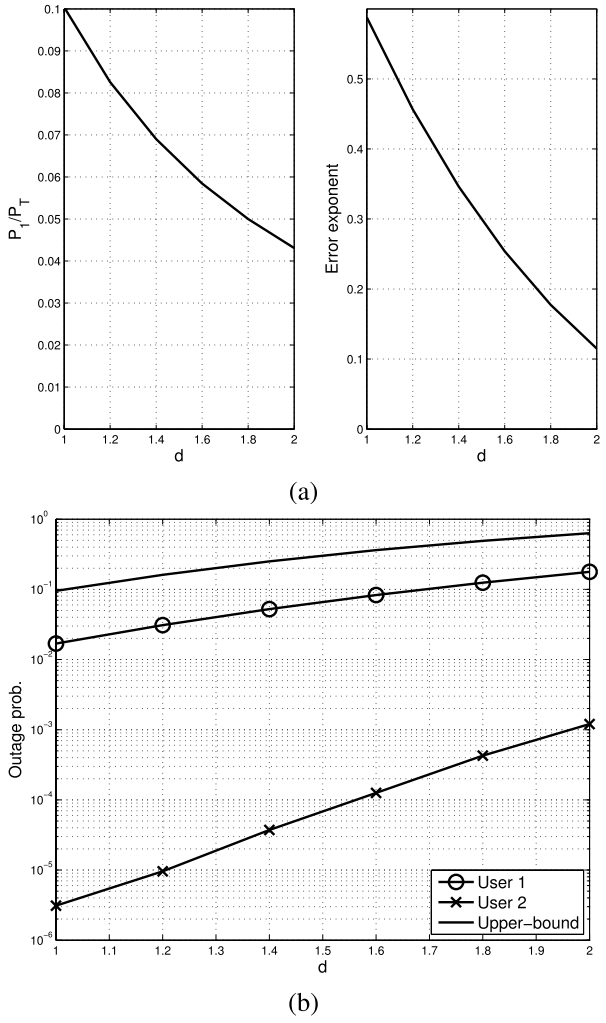


Fig. 6. Performances of NOMA-HARQ-IR for different values of d when $P_T = 20$ dB and $(\mu_1, \mu_2) = (1.5, 1.0)$: (a) φ^* and ℓ^*; (b) the outage probabilities (with $T = 4$).

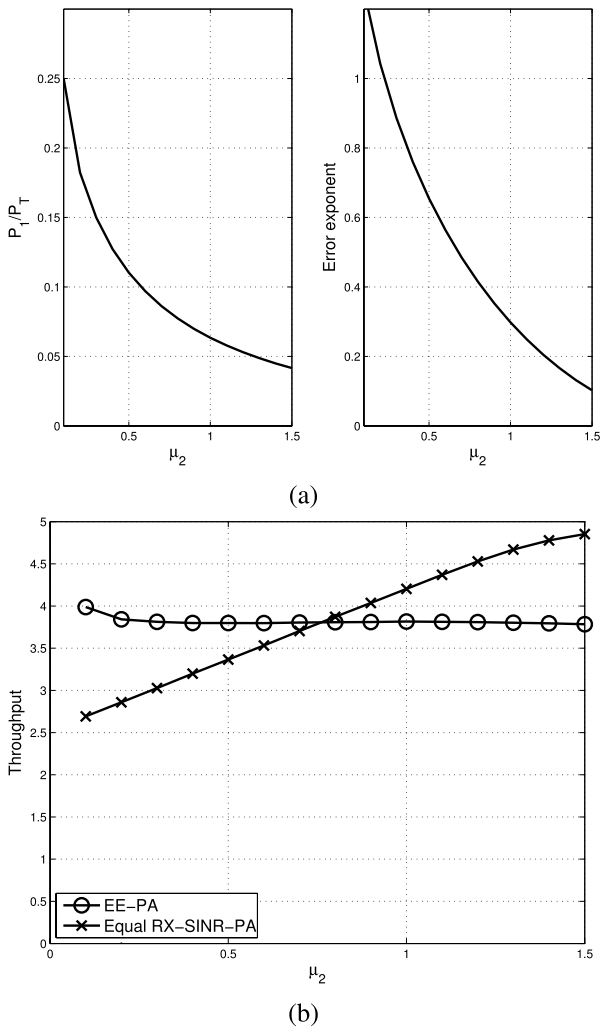


Fig. 7. Performances of NOMA-HARQ-IR for different values of μ_2 when $\mu_1 = 1.5$, $P_T = 20$ dB, and $d = 1.5$: (a) φ^* and ℓ^*; (b) the throughputs of EE-PA and equal RX-SINR-PA (with $T = 20$).

which decreases with P_T. However, if P_T is sufficiently large, φ^*, becomes the maximum point of $\bar{\ell}_2(\mu_2, \varphi P_T, (1-\varphi)P_T)$, which increases with P_T.

In Fig. 4 (b), we show the outage probabilities with $T = 4$ for the EE-PA and equal RX-SINR-PA schemes. The upper-bound on the outage probability is given by $e^{-T\ell^*}$ in the EE-PA scheme. We note that the outage probability at user 2 is much lower than the upper-bound. As in (29), in order to find the outage probability of the signal to user 2 in the EE-PA scheme, we consider an upper-bound. Fig. 4 (b), we can see that this bound is not tight. As a result, the outage behavior of the EE-PA scheme is not better than that of the equal RX-SINR-PA scheme (on the other hand, the EE-PA scheme provides a better throughput than the equal RX-SINR-PA scheme as will be shown in Fig. 5 (a)). Thus, we may need to derive a better bound as a further research topic.

In Fig. 5 (a), we show the throughputs of NOMA-HARQ-IR for different values of P_T with the two different power allocation methods: (i) the EE-PA; (ii) the equal RX-SINR-PA. We assume that $T = 20$, $(\mu_1, \mu_2) = (1.5, 0.5)$, and $d = 1.5$. We can see that the EE-PA scheme can provide a higher throughput than the equal RX-SINR-PA scheme. Fig. 5 (b) shows the outage probabilities at user 1 and user 2 for different values of T. We can see that the exponent in (40) by the EE-PA scheme is similar to the exponent of the outage probability of user 1, while it is larger than that of the outage probability of user 2. As mentioned earlier, this is due to the upper-bound in (29).

Fig. 6 shows the performances of NOMA-HARQ-IR for different values of d when $P_T = 20$ dB and $(\mu_1, \mu_2) = (1.5, 1.0)$. As d increases, it is expected to allocate more power to the signal to user 2 as shown in Fig. 6 (a). We can also see that the error exponent decreases with d. Accordingly, as shown in Fig. 6 (b), the outage probabilities increase with d. From Figs. 4 (b), 5 (b), and 6 (b), we can observe that all the outage probabilities are lower than the upper-bound, $e^{-T\ell^*}$. Clearly, this confirms that the EE-PA scheme is an effective means to guarantee a certain outage probability for a given maximum number of retransmissions, T.

Fig. 7 shows the performances of NOMA-HARQ-IR for different values of μ_2 when $\mu_1 = 1.5$, $P_T = 20$ dB, and $d = 1.5$. For a fixed μ_1, as μ_2 increases, we need to allocate a higher

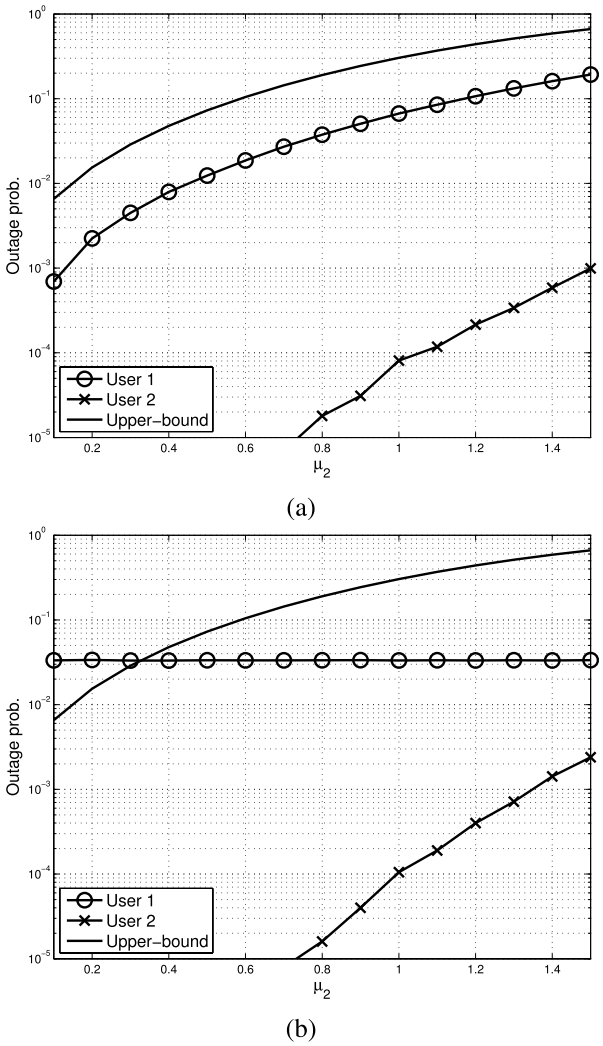


Fig. 8. Outage probabilities (with $T = 4$) versus μ_2 ($\mu_1 = 1.5$, $P_T = 20$ dB, and $d = 1.5$): (a) the EE-PA scheme; (b) the equal-RX-SINR-PA scheme.

power to the signal to user 2. Thus, as shown in Fig. 7 (a), the power to the signal to user 1 decreases with μ_2, which also results in the decrease of the error exponent. In Fig. 7 (b), the throughputs of the EE-PA and equal-RX-SINR-PA schemes are shown. As μ_2 approaches $\mu_1 = 1.5$, we can see that the equal-RX-SINR-PA scheme can provide a higher throughput than the EE-PA scheme. As mentioned earlier, since the EE-PA scheme is not to maximize the throughput, it can provide a lower throughput than the equal-RX-SINR-PA scheme.

Fig. 8 shows the outage probabilities of EE-PA and equal RX-SINR-PA with $T = 4$ for different values of μ_2 when $\mu_1 = 1.5$, $P_T = 20$ dB, and $d = 1.5$. Clearly, we can see that the EE-PA scheme can guarantee a certain outage probability through the upper-bound. Note that since the equal RX-SINR-PA scheme does not take into account outage probability, it is not necessary to guarantee any outage probability.

VI. Concluding Remarks

In this paper, we studied HARQ-IR for NOMA with two users. Unlike conventional HARQ-IR for single-user systems, NOMA-HARQ-IR needed to take into account the interference in decoding the signal to the user far away from the BS (i.e., user 2), where the interference is the signal to the user close to the BS (i.e., user 1). For the rate determination and power allocation, we considered the outage probabilities for a given maximum number of retransmissions. In particular, the error exponents have been taken into account to allocate the power to maximize the minimum of the error exponents for given rates based on large deviations. From this, we could guarantee a certain outage probability for a given total power and a maximum number of retransmissions. To this end, we derived a closed-form expression for a lower-bound on the exponent of the outage probability of the signal to user 2. Based on simulation results, we confirmed that it is possible to guarantee a certain outage probability by the proposed power allocation method, i.e., the EE-PA scheme. However, as the EE-PA scheme does not maximize the throughput, we may need a further study to find a power allocation scheme to maximize the throughput for NOMA-HARQ-IR.

Appendix A
Proof of Property 1

We first consider (17). Under **A)**, it can be shown that

$$\frac{d\mathbb{E}[\ln(1+D_{(1),2})]}{d\epsilon} = e^{-\frac{\epsilon}{\bar{\alpha}}}\left(\frac{P}{1+P\epsilon}-\frac{P_1}{1+P_1\epsilon}\right) - \frac{e^{-\frac{\epsilon}{\bar{\alpha}}}}{\bar{\alpha}}\left(\ln(1+P\epsilon)-\ln(1+P_1\epsilon)\right). \tag{42}$$

To find the maximum, the derivative is set to 0, which results in

$$\bar{\alpha}\underbrace{\frac{P-P_1}{(1+P\epsilon)(1+P_1\epsilon)}}_{=f(\epsilon)} = \underbrace{\ln\frac{1+P\epsilon}{1+P_1\epsilon}}_{=g(\epsilon)}. \tag{43}$$

Since $P > P_1$, $f(\epsilon)$ decreases with ϵ, while $g(\epsilon)$ increases with ϵ. In addition, $f(0) = (P-P_1) > g(0) = 0$ and $f(\infty) = 0 < g(\infty) = \ln\frac{P}{P_1}$. Consequently, there must be a unique solution ϵ that satisfies the equality in (43), i.e., $\hat{\epsilon}_1$ is the unique solution that satisfies

$$\bar{\alpha} f(\hat{\epsilon}_1) = g(\hat{\epsilon}_1).$$

In addition, since $\bar{\alpha} > \bar{\beta}$, we can have $\hat{\epsilon}_1 > \hat{\epsilon}_2$, $\hat{\epsilon}_2$ is the unique solution that satisfies

$$\bar{\beta} f(\hat{\epsilon}_2) = g(\hat{\epsilon}_2).$$

Appendix B
Proof of Property 3

Define the following function:

$$G(x) = \ln\left(1+\frac{P_2 x}{P_1 x+1}\right).$$

Clearly, $V_{(1),t} = \frac{1}{\ln 2}G(\alpha_t)$ and $V_{(2),t} = \frac{1}{\ln 2}G(\beta_t)$. It can be easily shown that $G(x)$ is an increasing function of x. In addition, under **A)**, it follows

$$\mathbb{E}[G(\beta_t)] = \mathbb{E}\left[G\left(\frac{\bar{\beta}}{\bar{\alpha}}\alpha_t\right)\right].$$

Since $\bar{\beta} \leq \bar{\alpha}$ and $G(x)$ is an increasing function of x, it can be shown that

$$\mathbb{E}[G(\beta_t)] = \mathbb{E}\left[G\left(\frac{\bar{\beta}}{\bar{\alpha}}\alpha_t\right)\right] \leq \mathbb{E}[G(\alpha_t)].$$

This completes the proof.

Appendix C
Proof of Property 4

From (23), we can see that $\omega_{(2)}$ increases with P_2 when P_1 is fixed. Thus, for a fixed P_1, $\frac{\mu_2}{\omega_{(2)}}$ in the effective exponent in (28) decreases with P_2. Furthermore, since $\ell_2(q, p)$ is the relative entropy [5], we can see that $\ell_2(q, p)$ decreases with q when $q < p$ and p is fixed. From this, if $P_2 < P_2'$, we have

$$\bar{\ell}_2(\mu_2, P_1, P_2) < \bar{\ell}_2(\mu_2, P_1, P_2'). \tag{44}$$

Since the maximum of P_2 for a given P_1 is $P_T - P_1$, the maximum exponent becomes $\bar{\ell}_2(\mu_2, P_1, P_T - P_1)$. Thus, for a given P_1, it can be shown that

$$\begin{aligned}&\min\{\ell_1(\mu_1, P_1), \bar{\ell}_2(\mu_2, P_1, P_2)\}\\ &\quad \leq \min\{\ell_1(\mu_1, P_1), \bar{\ell}_2(\mu_2, P_1, P_T - P_1)\}\end{aligned}$$

if $P_2 < P_T - P_1$. This confirms that if P_1^* is the solution to (37), $P_2^* = P_T - P_1^*$, which completes the proof.

References

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. 77th IEEE VTC-Spring*, Jun. 2013, pp. 1–5.

[2] B. Kim *et al.*, "Non-orthogonal multiple access in a downlink multiuser beamforming system," in *Proc. IEEE Military Commun. Conf. (MILCOM)*, Nov. 2013, pp. 1278–1283.

[3] A. G. Perotti and B. M. Popovic. (Oct. 2014). "Non-orthogonal multiple access for degraded broadcast channels: RA-CEMA." [Online]. Available: http://arxiv.org/abs/1410.5579

[4] Z. Ding, F. Adachi, and H. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[6] 3GPP. "TP for classification of must schemes," 3GPP, Sophia Antipolis, France, Tech Rep. R1-154999, 2015.

[7] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 313–316, Feb. 2014.

[8] S.-L. Shieh and Y.-C. Huang, "A simple scheme for realizing the promised gains of downlink nonorthogonal multiple access," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1624–1635, Apr. 2016.

[9] J. Choi, "On the power allocation for a practical multiuser superposition scheme in NOMA systems," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 438–441, Mar. 2016.

[10] S. B. Wicker, *Error Control Systems for Digital Communication and Storage*. Upper Saddle River, NJ, USA: Prentice-Hall, 1995.

[11] S. Lin and D. J. Costello, Jr., *Error Control Coding: Fundamentals and Applications*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1983.

[12] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inf. Theory*, vol. 47, no. 5, pp. 1971–1988, Jul. 2001.

[13] P. Wu and N. Jindal, "Performance of hybrid-ARQ in block-fading channels: A fixed outage probability analysis," *IEEE Trans. Commun.*, vol. 58, no. 4, pp. 1129–1141, Apr. 2010.

[14] J. Choi, "On large deviations of HARQ with incremental redundancy over fading channels," *IEEE Commun. Lett.*, vol. 16, no. 6, pp. 913–916, Jun. 2012.

[15] A. Weiss, "An introduction to large deviations for communication networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 939–952, Aug. 1995.

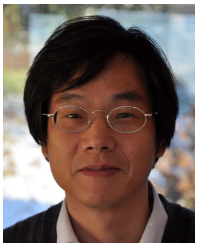
[16] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[17] A. Li, A. Benjebbour, X. Chen, H. Jiang, and H. Kayama, "Investigation on hybrid automatic repeat request (HARQ) design for NOMA with SU-MIMO," in *Proc. IEEE 26th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Aug./Sep. 2015, pp. 590–594.

[18] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[19] 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 Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[20] I. S. Gradshteyn and M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed. San Diego, CA, USA: Academic, 2000.



Jinho Choi (SM'02) was born in Seoul, South Korea. He received the B.E. (*magna cum laude*) degree in electronics engineering from Sogang University, Seoul, in 1989, and the M.S.E. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology, Daejeon, in 1991 and 1994, respectively. He has been with the Gwangju Institute of Science and Technology (GIST) as a Professor since 2013. He was with the College of Engineering, Swansea University, U.K., as a Professor and the Chair of Wireless. His research interests include wireless communications and array/statistical signal processing. He has authored two books published by Cambridge University Press in 2006 and 2010, respectively. He received the 1999 Best Paper Award for Signal Processing from EURASIP, and the 2009 Best Paper Award from the WPMC Conference. He is currently an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS. He had served as an Associate Editor or an Editor of other journals including the IEEE COMMUNICATIONS LETTERS, the *Journal of Communications and Networks*, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the *ETRI Journal*.