Title

Cooperative Asynchronous Non-Orthogonal Multiple Access With Power Minimization Under QoS Constraints.

Authors

Zou, Xun
Ganji, Mehdi
Jafarkhani, Hamid

Publication Date

2020

DOI

10.1109/TWC.2019.2953902

Peer reviewed

Cooperative Asynchronous Non-Orthogonal Multiple Access with Power Minimization Under QoS Constraints

Xun Zou, *Student Member, IEEE*, Mehdi Ganji, *Student Member, IEEE*, and Hamid Jafarkhani, *Fellow, IEEE*

Abstract

Recent studies have demonstrated the superiority of non-orthogonal multiple access (NOMA) over orthogonal multiple access (OMA) in cooperative communication networks. In this paper, we propose a novel half-duplex cooperative asynchronous NOMA (C-ANOMA) framework with user relaying, where a timing mismatch is intentionally added in the broadcast signal. We derive the expressions for the throughputs of the strong user (acts as relay) which employs the block-wise successive interference cancellation (SIC) and the weak user which combines the symbol-asynchronous signal with the interference-free signal. We analytically prove that in the C-ANOMA systems with a sufficiently large block length, the strong user attains the same throughput to decode its own message while both users can achieve a higher throughput to decode the weak user's message compared with those in the cooperative NOMA (C-NOMA) systems. Besides, we obtain the optimal timing mismatch when the block length goes to infinity. Furthermore, to exploit the trade-off between the power consumption of the base station and that of the relay user, we solve a weighted sum power minimization problem under quality of services (QoS) constraints. Numerical results show that the C-ANOMA system can consume less power compared with the C-NOMA system to satisfy the same QoS requirements.

Index Terms

Non-orthogonal multiple access, asynchronous transmission, cooperative communication, interference cancellation, power control.

This work was supported in part by the NSF Award CCF-1526780. The authors are with the Center for Pervasive Communications and Computing, Department of Electrical Engineering and Computer Science, University of California, Irvine, CA, 92697 USA (email: {xzou4, mganji, hamidj}@uci.edu).

I. Introduction

Non-orthogonal multiple access (NOMA) has been regarded as one of the key technologies for the next generation wireless communications [1]. Compared with the conventional orthogonal multiple access (OMA), NOMA can provide massive connectivity and high spectral efficiency [2]. The key rationale behind NOMA is to allow users to share non-orthogonal wireless resources, e.g., frequency, time, and code. For multiuser detection, superposition coding and successive interference cancellation (SIC) are employed at the transmitter and receiver, respectively.

Cooperative communication is an effective approach to exploit spatial diversity available through cooperating terminals' relaying signals for one another [3–5]. Cooperative relaying network with NOMA has been extensively studied in the literature, e.g., [6–8]. It has been shown that the cooperative NOMA (C-NOMA) systems outperform the cooperative OMA systems in terms of the spectral efficiency [6] and the outage probability [7]. Instead of dedicated relay nodes, users can also be adopted as relays in a cooperative network. A key feature of NOMA is that users with better channel conditions have prior information about the messages of other users. Ding et al. [9] proposed a C-NOMA scheme to fully exploit the prior knowledge at the strong user, where the users could cooperate with each other via short-range communication channels. Yue et al. [10] compared different operation modes of the relay user in a C-NOMA system. The half-duplex relay user receives and transmits in separate time slots while the full-duplex relay user receives and transmits simultaneously. In [10], the outage probability, the ergodic rate, and the energy efficiency were analyzed in a NOMA user relaying system where the near user could switch between full-duplex and half-duplex modes to relay messages to the far user. Zhang et al. [11] studied an adaptive multiple access scheme to further improve the outage performance, which dynamically switched among the C-NOMA with user relaying, conventional NOMA, and OMA schemes, according to the level of residual self-interference and the quality of links. Wei et al. [12] solved the energy efficiency maximization problem of a full-duplex C-NOMA system under the constraint of successful SIC operation.

A. Motivations and Related Works

By intentionally introducing symbol asynchrony in the transmitted signal, asynchronous NOMA (ANOMA) systems can achieve a better throughput performance compared with the conventional (synchronous) NOMA systems [13–16]. In ANOMA systems, the receiver utilizes the oversampling technique [17] to achieve the sampling diversity gain. It has been revealed that the

cooperative communication systems can also benefit from the symbol-asynchronous transmission. Sodagari et al. [18] studied an asynchronous cognitive radio framework, where the primary user and the secondary user were not aligned in their timing. They conclude that not only can asynchronous cognitive radio reduce the interference to the primary user, but it also saves power at the secondary user compared with synchronous cognitive radio systems. An asynchronous network coding (ANC) transmission strategy for multiuser cooperative networks was investigated in [19, 20], where the received signals from multiple sources were asynchronous to each other. The proposed scheme achieves full diversity and outperforms the complex field network coding in terms of decoding complexity and bit error rate (BER).

In this paper, we consider a half-duplex cooperative ANOMA (C-ANOMA) system with user relaying, including a base station (BS), a strong user (also acting as a relay), and a weak user. The half-duplex C-ANOMA system employs a transmission scheme similar to that of the conventional half-duplex C-NOMA system [9, 10, 21]: The BS transmits the superimposed signals to two users simultaneously in the first time block and then the relay user transmits the signal to the weak user at the second time block. Different from the conventional C-NOMA systems, a symbol asynchrony is intentionally added to the downlink superposed signal in the broadcast phase of C-ANOMA systems. The weak user receives two blocks of signals via the broadcast link and the relay link separately. The questions then arise: How to realize SIC based on the symbol-asynchronous signal and then evaluate the performance of the strong user in the C-ANOMA systems? How to evaluate the performance of the weak user which combines a symbol-asynchronous signal from the broadcast link with an interference-free signal from the relay link? Moreover, compared with the cooperative systems with dedicated relay nodes, the power control strategy plays a more critical role in the cooperative systems with user relaying because the power consumption of the relay user affects the lifetime of the cooperative network. We assume that the channel information is available at transmitters [12, 21] and the system works in the delay-tolerant transmission mode [10], such that the transmitters can dynamically adjust their transmit powers according to the channel states to avoid outage and save energy. On one hand, the relay user with very limited battery capacity is more sensitive to the power consumption compared with the BS. On the other hand, the relay user can transmit signals to the weak user more efficiently because the relay user is usually closer to the weak user. As a result, an effective power control strategy is of practical interest to make a trade-off between the transmit power of the BS and that of the relay user while satisfying the quality of service

(QoS) constraints in the C-ANOMA/C-NOMA systems with user relaying. To reduce the energy consumption, the power minimization problem has been investigated in several systems, e.g., the downlink NOMA systems [22], the multicell NOMA systems [23], and the cooperative beamforming networks [24]. Besides, Liu et al. [21] and Chen et al. [25] studied the power allocation problem for half-duplex and full-duplex C-NOMA systems, respectively, to maximize the minimum achievable user rate in a NOMA user pair. To the best of our knowledge, the power minimization problem has never been studied in the C-NOMA or C-ANOMA systems with user relaying.

While this paper focuses on only the physical layer cooperation, an additional gain can be achieved by applying the cooperation at the network layer. In contrast to the physical layer cooperation which focuses on the achievable rate or throughput, the network layer cooperation is proposed to optimize the stable throughput region which captures the bursty nature of source traffic [26]. For example, the half duplex limitation can be combated by introducing the cognitive radio technology through the network level cooperation [27, 28]. The cognitive cooperation scheme investigated in [28] exploits a dedicated relay node which possesses the sensing capability to detect the idle channel. Relaying operation is done when the channel is idle, which consumes no extra channel resources and causes no bandwidth loss. Another example is the full-duplex cooperation scheme studied in [29], where the relay node works in full duplex mode relaying packets from different sources to a common destination node in a random access network. Since the network layer communication operates upon the physical layer communication, designing a more efficient physical layer cooperation scheme can enhance the performance of the network layer cooperation [30], which also motivates this work. The combination of these two approaches is also an interesting future work.

B. Contributions

In this paper, we comprehensively investigate a half-duplex C-ANOMA system with user relaying. The primary contributions of the paper are summarized as follows:

- We introduce the block-wise SIC technique into C-ANOMA systems, which is employed at the strong (relay) user. We derive the analytical expressions for throughputs achieved by the strong user to decode both users' messages and study their asymptotic performances as the block length goes to infinity. We analytically show that in the C-ANOMA systems with a sufficiently large block length, the strong user can achieve a higher throughput to

detect the weak user's message while attains the same throughput when detecting its own message compared with those in C-NOMA systems.

- We derive the expression for the combining throughput achieved by the weak user which combines the asynchronously superimposed signal from the broadcast link with the interference-free signal from the relay link. Based on the derived throughput expressions, we obtain the asymptotic throughput as the block length goes to infinity and its simple upper and lower bounds. We analytically prove that in the C-ANOMA systems with a sufficiently large block length, the combining throughput of the weak user is greater than that in the C-NOMA systems.
- We further study the optimal design of C-ANOMA systems. We analytically prove that the optimal timing mismatch to maximize the individual throughput converges to half of the symbol interval as the block length increases. Besides, we solve the weighted sum power minimization problem under the QoS constraints for C-ANOMA and C-NOMA systems. The solution is given by the explicit expressions of the powers allocated to the strong and weak users at the BS and the transmit power of the relay (strong) user. It is demonstrated that for a relatively large block length, the C-ANOMA systems consume less power compared with the C-NOMA systems in order to satisfy the same QoS requirements. In other words, under the same transmit power limits, the C-ANOMA systems can provide a higher QoS for users compared with the C-NOMA systems.

C. Organization and Notation

The remainder of the paper is organized as follows. The C-ANOMA system model is presented in Section II. The throughput performance of the C-ANOMA system is analyzed in Section III. We discuss the optimal design of the C-ANOMA system in Section IV where we investigate the optimal timing mismatch and solve the weighted power minimization problem under QoS constraints. Numerical results are presented in Section V. Finally, we draw the conclusions in Section VI.

Notations: $(\cdot)^H$ denotes the Hermitian transpose, $(\cdot)^T$ denotes the transpose, $(\cdot)^{-1}$ denotes the inverse operation, $\otimes$ denotes the Kronecker product, $|x|$ denotes the absolute value of x, $\bar{x}$ denotes the complex conjugate of x, $\mathbb{E}[\cdot]$ denotes the expectation operation, $\mathcal{CN}(0,1)$ denotes the complex normal distribution with zero mean and unit variance. $\mathrm{diag}(\mathbf{x})$ stands for a diagonal matrix whose k-th diagonal element is equal to the k-th entry of vector $\mathbf{x}$.

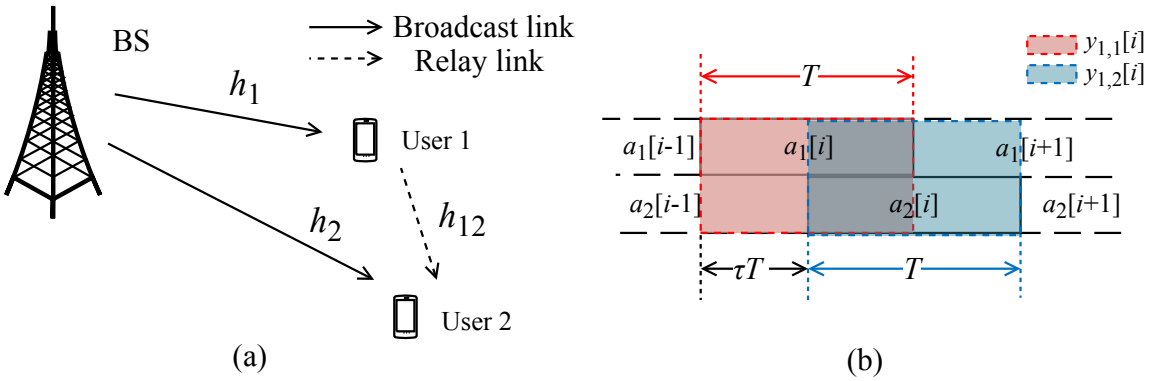

Fig. 1: (a) Illustration of a three-node C-ANOMA/C-NOMA system with user relaying. (b) Illustration of the sampling for the broadcast phase in C-ANOMA systems.

II. System Model

In this paper, as shown in Fig. 1 (a), we consider a downlink half-duplex C-ANOMA system which includes a single BS and two users equipped with a single antenna. User 1 (strong user) acts as a relay for User 2 (weak user) and adopts the decode-and-forward (DF) protocol, i.e., decodes and forwards the message to User 2 via the relay link. The downlink transmission is done in blocks, including two phases, i.e., the broadcast phase and the relay phase. In the broadcast phase, the BS broadcasts one block of superposed signal to two users simultaneously while User 1 is silent. In the relay phase, User 1 transmits the block of decoded signal to User 2 while the BS keeps silent. We assume that the channel is static within each block [31] and all the channel information is perfectly known at the BS, Users 1 and 2 [12, 21]. The channel coefficient between the BS and User i is denoted as h_i $(i = 1, 2)$ and the channel coefficient between Users 1 and 2 is denoted as h_{12}. In what follows, we present our analysis in the broadcast phase and the relay phase separately.

A. Broadcast Phase

1) C-ANOMA: In the C-ANOMA systems, a symbol mismatch is intentionally introduced in the downlink signal. As shown in Fig. 1 (b), the intended timing mismatch between the symbols for Users 1 and 2 is denoted by τT, where T is the symbol interval and τ, $0 \leq \tau < 1$, is the normalized timing mismatch. We assume that τ can be perfectly known at users. The timing mismatch information can be transmitted as part of the downlink control information through

the downlink control channel, such as the physical downlink control channel (PDCCH) in the long term evolution (LTE) system. The downlink control channel is designed to be robust in order to ensure the successful reception of the control information, e.g., by applying a low-rate coding scheme. Note that the C-ANOMA system becomes a synchronous C-NOMA system when $\tau = 0$.

Let $a_1[i] = \sqrt{P_1}s_1[i]$ and $a_2[i] = \sqrt{P_2}s_2[i]$, where $s_j[i]$ denotes the ith symbol sent to User j, $j = 1, 2$, P_j stands for the power allocated to User j. The transmitted signal at the BS is given by

$$s(t) = \sum_{i=1}^{N} a_1[i]p(t - iT) + \sum_{i=1}^{N} a_2[i]p(t - iT - \tau T). \tag{1}$$

where N denotes the number of symbols in a block, i.e., the block length, $p(\cdot)$ denotes the pulse-shaping filter. Without loss of generality, the rectangular pulse shape is adopted, i.e., $p(t) = 1/\sqrt{T}$ when $t \in [0, T]$ and $p(t) = 0$ otherwise.

The received signal at User 1 is given by

$$y_1(t) = h_1 s(t) + n_1(t) = h_1 \left(\sum_{i=1}^{N} a_1[i]p(t - iT) + \sum_{i=1}^{N} a_2[i]p(t - iT - \tau T) \right) + n_1(t), \tag{2}$$

where $n_1(t) \sim \mathcal{CN}(0, 1)$ denotes the normalized additive white Gaussian noise (AWGN).

The oversampling technique [32–34], depicted in Fig. 1 (b), is employed at the receiver to take advantage of sampling diversity in asynchronous systems. As shown in Fig. 1 (b), the receiver uses the matched filter, sampling at iT and $(i+\tau)T$, $i = 1, \cdots, N$, to obtain two sample vectors, denoted by $[y_{1,1}[1], \cdots, y_{1,1}[N]]^T$ and $[y_{1,2}[1], \cdots, y_{1,2}[N]]^T$. Specifically, the ith element in the first sample vector is given by

$$\begin{aligned} y_{1,1}[i] &= \int_0^\infty y_1(t)p(t - iT)dt \\ &= \int_0^\infty h_1 a_1[i]p(t - iT)p(t - iT)dt \\ &\quad + \int_0^\infty \{h_1 a_2[i-1]p(t - (i + 1 + \tau)T) + h_1 a_2[i]p(t - (i + \tau)T)\} p(t - iT)dt + n_{1,1}[i] \\ &= h_1 a_1[i] + \tau h_1 a_2[i-1] + (1 - \tau)h_1 a_2[i] + n_{1,1}[i], \end{aligned} \tag{3}$$

where $n_{1,1}[i] = \int_0^\infty n_1(t)p(t - iT)dt$ denotes the additive noise. The ith element in the second sample vector is given by

$$y_{1,2}[i] = \int_0^\infty y_1(t)p(t - iT - \tau T)dt = h_1 a_2[i] + \tau h_1 a_1[i+1] + (1 - \tau)h_1 a_1[i] + n_{1,2}[i], \tag{4}$$

where $n_{1,2}[i] = \int_0^\infty n_1(t)p(t - iT - \tau T)dt$ denotes the additive noise.

We can write the outputs of the two matched filters at User 1 in a matrix form as

$$\mathbf{Y}_1 = h_1\sqrt{P_1}\mathbf{R}\mathbf{G_1}\mathbf{S_1} + h_1\sqrt{P_2}\mathbf{R}\mathbf{G_2}\mathbf{S_2} + \mathbf{N}_1 \tag{5}$$

where $\mathbf{Y}_1 = [y_{1,1}[1]\ y_{1,2}[1]\ \cdots\ y_{1,1}[N]\ y_{1,2}[N]]^T$, $\mathbf{G}_1$ and $\mathbf{G}_2$ are $2N$-by-N matrices given by $\mathbf{G_1} = \mathbf{I}_N \otimes [1\ 0]^T$ and $\mathbf{G_2} = \mathbf{I}_N \otimes [0\ 1]^T$, $\mathbf{S_i} = [s_i[1]\ \cdots\ s_i[N]]^T$ $(i = 1, 2)$, $\mathbf{N}_1 = [n_{1,1}[1]\ n_{1,2}[1]\ \cdots\ n_{1,1}[N]\ n_{1,2}[N]]^T$, and

$$\mathbf{R} = \begin{bmatrix} 1 & 1-\tau & 0 & \cdots & \cdots & 0 \\ 1-\tau & 1 & \tau & 0 & \cdots & 0 \\ 0 & \tau & 1 & 1-\tau & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \tau & 1 & 1-\tau \\ 0 & \cdots & \cdots & 0 & 1-\tau & 1 \end{bmatrix}. \tag{6}$$

Note that multiplying $\mathbf{R}$ by $\mathbf{G_i}$ outputs a $2N$-by-N matrix whose columns are equal to the odd (if $i = 1$) or even (if $i = 2$) columns of $\mathbf{R}$.

We assume that the transmitted symbols are normalized and independent to each other, such that $\mathbb{E}\left[\mathbf{S_i}\mathbf{S_i}^H\right] = \mathbf{I}$ $(i = 1, 2)$. Note that the noise terms in (3) and (4) are colored due to the oversampling, and we have

$$\mathbb{E}\left\{n_{1,1}[i]n_{1,2}^H[i]\right\} = \int_0^\infty \int_0^\infty \mathbb{E}\left\{n_1(t)n_1^H(s)\right\} p\left(t - iT\right) p\left(s - iT - \tau T\right) dtds = 1 - \tau. \tag{7}$$

Thus, the covariance matrix of $\mathbf{N}_1$ in (5) is given by

$$\mathbf{R}_{\mathbf{N}_1} = \mathbb{E}\left\{\mathbf{N}_1\mathbf{N}_1^H\right\} = \mathbf{R}. \tag{8}$$

Similarly, the received samples at User 2 in the broadcast phase can be written as

$$\mathbf{Y}_2 = h_2\sqrt{P_1}\mathbf{R}\mathbf{G}_1\mathbf{S_1} + h_2\sqrt{P_2}\mathbf{R}\mathbf{G}_2\mathbf{S_2} + \mathbf{N}_2, \tag{9}$$

where the covariance matrix $\mathbf{R}_{\mathbf{N}_2} = \mathbb{E}\left\{\mathbf{N}_2\mathbf{N}_2^H\right\} = \mathbf{R}$.

2) C-NOMA: By setting $\tau = 0$, the C-ANOMA system becomes the C-NOMA system. For the C-NOMA systems, users do not use the oversampling technique. The ith sample at Users 1 and 2 in the broadcast phase will be

$$y_1[i] = h_1\sqrt{P_1}s_1[i] + h_1\sqrt{P_2}s_2[i] + n_1[i], \tag{10}$$

$$y_2[i] = h_2\sqrt{P_1}s_1[i] + h_2\sqrt{P_2}s_2[i] + n_2[i], \tag{11}$$

where $n_j[i] = \int_0^\infty n_j(t)p(t - iT)dt$, $j = 1, 2$. Note that (10) and (11) can also be derived from (5) and (9), respectively, by letting $\tau = 0$.

B. Relay Phase

In the relay phase, User 2 receives another copy of the desired signal from User 1. The ith sample received at User 2 in the relay phase is given by

$$y_{12}[i] = h_{12}\sqrt{P_r}s_2[i] + n_{12}[i], \tag{12}$$

where P_r is the transmit power of User 1 and $n_{12}[i] = \int_0^\infty n_{12}(t)p(t-iT)dt$ is the additive noise. Note that the C-NOMA and C-ANOMA systems coincide in the relay phase.

For ease of the following analysis, we rewrite the received samples from the relay link in (12) into the matrix format, i.e.,

$$\mathbf{Y}_{12} = h_{12}\sqrt{P_r}\mathbf{S}_2 + \mathbf{N}_{12}, \tag{13}$$

where $\mathbf{Y}_{12} = [y_{12}[1], y_{12}[2], \cdots, y_{12}[N]]^T$, $\mathbf{N}_{12} = [n_{12}[1], n_{12}[2], \cdots, n_{12}[N]]^T$, and the covariance matrix $\mathbf{R}_{\mathbf{N}_{12}} = \mathbb{E}\left\{\mathbf{N}_{12}\mathbf{N}_{12}^H\right\} = \mathbf{I}_N$.

Combining all the received samples of User 2 in C-ANOMA systems, i.e., $\mathbf{Y}_2$ in (9) and $\mathbf{Y}_{12}$ in (13), we have

$$\tilde{\mathbf{Y}}_2 = \left[\begin{smallmatrix}\mathbf{Y}_2\\ \mathbf{Y}_{12}\end{smallmatrix}\right] = \underbrace{\left[\begin{smallmatrix}h_2\sqrt{P_1}\mathbf{R}\mathbf{G}_1\\ \mathbf{0}_N\end{smallmatrix}\right]}_{\mathbf{W}_1}\mathbf{S}_1 + \underbrace{\left[\begin{smallmatrix}h_2\sqrt{P_2}\mathbf{R}\mathbf{G}_2\\ h_{12}\sqrt{P_r}\mathbf{I}_N\end{smallmatrix}\right]}_{\mathbf{W}_2}\mathbf{S}_2 + \underbrace{\left[\begin{smallmatrix}\mathbf{N}_2\\ \mathbf{N}_{12}\end{smallmatrix}\right]}_{\mathbf{N}}. \tag{14}$$

Applying $\mathbb{E}\left\{\mathbf{N}_2\mathbf{N}_2^H\right\} = \mathbf{R}$ and $\mathbb{E}\left\{\mathbf{N}_{12}\mathbf{N}_{12}^H\right\} = \mathbf{I}_N$, the covariance matrix of the concatenated noise vector $\mathbf{N}$ is given by

$$\mathbf{R}_{\mathbf{N}} = \mathbb{E}\left\{\mathbf{N}\mathbf{N}^H\right\} = \begin{bmatrix}\mathbb{E}\left\{\mathbf{N}_2\mathbf{N}_2^H\right\} & \mathbb{E}\left\{\mathbf{N}_2\mathbf{N}_{12}^H\right\}\\ \mathbb{E}\left\{\mathbf{N}_{12}\mathbf{N}_2^H\right\} & \mathbb{E}\left\{\mathbf{N}_{12}\mathbf{N}_{12}^H\right\}\end{bmatrix} = \left[\begin{smallmatrix}\mathbf{R} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{smallmatrix}\right]. \tag{15}$$

III. Performance Analysis of C-ANOMA Systems

In this section, we analyze the individual throughput of users in the C-ANOMA and C-NOMA systems, including the strong and weak users.

A. Strong User

1) C-ANOMA: In C-ANOMA systems, the block-wise SIC is adopted at User 1, i.e., it first decodes the block of symbols intended for User 2, subtracts it from the received signal, and then decodes its intended symbols. Note that the BS transmits one block of symbols via two time blocks in the half-duplex mode [10]. Besides, an extra τT time is utilized to create the sampling diversity in the symbol-asynchronous transmission. Hence, in the half-duplex C-ANOMA systems, a block of N symbols are transmitted via $2N+\tau$ channel uses to Users 1 and

2. By considering (5) as a virtual multiple-input multiple-output (MIMO) system and treating the symbols for User 1 as noise, the throughput of User 1 to detect User 2's message is given by

$$R_{2\to1}^{\mathrm{ANOMA}} = \frac{1}{2N+\tau}\log\det\left[\mathbf{I}_{2N}+\left(\mathbf{R}_{\mathbf{N_1}}+P_1|h_1|^2\mathbf{R}\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}^H\right)^{-1}P_2|h_1|^2\mathbf{R}\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}^H\right]$$
$$\overset{(a)}{=} \frac{1}{2N+\tau}\log\det\left[\mathbf{I}_{2N}+\left(\mathbf{I}_{2N}+P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1}P_2|h_1|^2\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}\right], \tag{16}$$

where (a) is derived by applying $\mathbf{R}_{\mathbf{N_1}}=\mathbf{R}$ and $\mathbf{R}^H=\mathbf{R}$.

Under the assumption of perfect SIC, by subtracting User 2's message from the superposed signal in (5), the throughput of User 1 to detect its own message is calculated as

$$R_1^{\mathrm{ANOMA}} = \frac{1}{2N+\tau}\log\det\left(\mathbf{I}_{2N}+P_1|h_1|^2\mathbf{R}_{\mathbf{N}_1}^{-1}\mathbf{R}\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}^H\right)$$
$$= \frac{1}{2N+\tau}\log\det\left(\mathbf{I}_{2N}+P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right). \tag{17}$$

After matrix calculations, we can rewrite the throughput expressions for User 1 in (16) and (17) as functions of the receive signal-to-noise ratios (SNRs), i.e., μ_1 and μ_2, the normalized timing mismatch, τ, and the block length, N, i.e.,

$$R_{2\to1}^{\mathrm{ANOMA}} = \frac{1}{2N+\tau}\log\frac{\left(r_1^{N+1}-r_2^{N+1}\right)+\tau^2\left(r_1^N-r_2^N\right)}{r_1-r_2}+\frac{N}{2N+\tau}\log\left(\frac{\mu_1\mu_2}{1+\mu_1}\right), \tag{18}$$

$$R_1^{\mathrm{ANOMA}} = \frac{N}{2N+\tau}\log\left(1+\mu_1\right), \tag{19}$$

where

$$\mu_1 = P_1|h_1|^2, \mu_2 = P_2|h_1|^2, Q = 2\tau(1-\tau), \tag{20}$$

$$r_1 = \frac{\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}+Q+\sqrt{\left(\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}+Q\right)^2-Q^2}}{2}, \tag{21}$$

$$r_2 = \frac{\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}+Q-\sqrt{\left(\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}+Q\right)^2-Q^2}}{2}. \tag{22}$$

The detailed derivation of (18) and (19) is presented in Appendix A.

2) C-NOMA: In conventional (synchronous) NOMA systems, with perfect SIC, the throughputs of User 1 are given by [6, 10]

$$R_{2\to1}^{\mathrm{NOMA}} = \frac{1}{2}\log\left(1+\frac{\mu_2}{1+\mu_1}\right), \tag{23}$$

$$R_1^{\mathrm{NOMA}} = \frac{1}{2}\log(1+\mu_1). \tag{24}$$

We note that by setting $\tau=0$, we obtain $Q=0$, $r_2=0$, and $r_1=\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}$. Thus, $R_{2\to1}^{\mathrm{ANOMA}}|_{\tau=0}=R_{2\to1}^{\mathrm{NOMA}}$ and $R_1^{\mathrm{ANOMA}}|_{\tau=0}=R_1^{\mathrm{NOMA}}$.

3) Comparison between C-ANOMA and C-NOMA: To study the throughput performance in the systems with a relatively large block length, we consider the asymptotic case of $N \to \infty$. According to (19), the throughput of User 1 to decode its own message if $N \to \infty$ is given by

$$R_{1,\text{asymp}}^{\text{ANOMA}} \triangleq \lim_{N\to\infty} R_1^{\text{ANOMA}} = \frac{1}{2}\log\left(1+\mu_1\right) = R_1^{\text{NOMA}}. \tag{25}$$

We note from (25) that User 1 in C-ANOMA and C-NOMA systems can achieve the same throughput to detect its own message for a sufficiently large block length. It is because with perfect SIC, the throughput of User 1 to detect its own message is not affected by the symbol asynchrony of the signal for User 2. Furthermore, we derive the following theorem to compare the throughputs of User 1 to detect User 2's message in the C-ANOMA and C-NOMA systems.

Theorem 1: The throughputs of User 1 to detect User 2's message in the C-NOMA and C-ANOMA systems satisfy the following inequalities

$$\begin{aligned} R_{2\to1}^{\text{NOMA}} &\le R_{2\to1,L}^{\text{ANOMA}} \triangleq \frac{1}{2}\log\left(1+\frac{\mu_2+\frac{1}{2}\mu_1\mu_2 Q}{1+\mu_1}\right) \\ &\le R_{2\to1,\text{asymp}}^{\text{ANOMA}} \triangleq \lim_{N\to\infty} R_{2\to1}^{\text{ANOMA}} = \frac{1}{2}\log\left(\frac{\mu_1\mu_2 r_1}{1+\mu_1}\right) \\ &\le R_{2\to1,U}^{\text{ANOMA}} \triangleq \frac{1}{2}\log\left(1+\frac{\mu_2+\mu_1\mu_2 Q}{1+\mu_1}\right), \end{aligned} \tag{26}$$

where $Q = 2\tau(1-\tau)$, all the equal signs are achieved if and only if $\tau = 0$.

Proof: See Appendix B. ■

We note from Theorem 1 that for a relatively large block length, User 1 in C-ANOMA systems can achieve a higher throughput to decode User 2's message compared with that in C-NOMA systems. Besides, comparing the expressions for $R_{2\to1,L}^{\text{ANOMA}}$ and $R_{2\to1}^{\text{NOMA}}$, we find that the gain of C-ANOMA systems is related to the term $\mu_1\mu_2 Q$ which increases as the channel qualities improve.

In practice, the block length, N, is determined by several factors, such as the channel coherence time, the modulation, the sampling rate, etc., which are beyond the scope of this paper. We assume that the block length N is a predetermined parameter in this paper. We will show in Section V that the asymptotic throughput approximates the accurate one for not-so-large values of N, e.g, $N > 50$.

B. Weak User

In the half-duplex cooperative relaying scenario, the weak user, User 2, receives two blocks of symbols, one from the BS with the superposed signal through the broadcast link and the other one from User 1 with only the intended signal through the relay link.

1) C-ANOMA: Treating (14) as a virtual MIMO system and considering User 1's message as noise, the combining throughput of User 2 can be calculated as

$$R_2^{\mathrm{ANOMA}} = \frac{1}{2N+\tau} \log\det\left[\mathbf{I}_{3N} + \left(\mathbf{R}_{\mathbf{N}} + \mathbf{W}_1\mathbf{W}_1^H\right)^{-1}\mathbf{W}_2\mathbf{W}_2^H\right]. \tag{27}$$

The combining throughput of User 2 can be written as a function of the transmit powers, the channel gains, the normalized timing mismatch, and the block length in the following theorem.

Theorem 2: In the half-duplex C-ANOMA systems, the combining throughput of User 2 is given by

$$R_2^{\mathrm{ANOMA}} = \frac{1}{2N+\tau}\log\frac{\left(z_1^{N+1}-z_2^{N+1}\right)+\tau^2\left(z_1^N-z_2^N\right)}{z_1-z_2} + \frac{N}{2N+\tau}\log\left(\frac{P_1P_2|h_2|^4}{1+P_1|h_2|^2}\right), \tag{28}$$

where

$$\nu_1 = P_1|h_2|^2, \nu_2 = \frac{P_2|h_2|^2}{1+P_r|h_{12}|^2}, Q = 2\tau(1-\tau) \tag{29}$$

$$z_1 = \frac{\nu_1^{-1}+\nu_2^{-1}+\nu_1^{-1}\nu_2^{-1}+Q+\sqrt{\left[\nu_1^{-1}+\nu_2^{-1}+\nu_1^{-1}\nu_2^{-1}+Q\right]^2-Q^2}}{2}, \tag{30}$$

$$z_2 = \frac{\nu_1^{-1}+\nu_2^{-1}+\nu_1^{-1}\nu_2^{-1}+Q-\sqrt{\left[\nu_1^{-1}+\nu_2^{-1}+\nu_1^{-1}\nu_2^{-1}+Q\right]^2-Q^2}}{2}. \tag{31}$$

Proof: See Appendix C. ■

2) C-NOMA: In C-NOMA systems, User 2 adopts the maximal ratio combining (MRC) to combine the signals from the direct and relay links [10, 11]. Then, the combining throughput of User 2 is given by

$$R_2^{\mathrm{NOMA}} = \frac{1}{2}\log\left(1+P_r|h_{12}|^2+\frac{P_2|h_2|^2}{P_1|h_2|^2+1}\right). \tag{32}$$

Note that by setting $\tau = 0$, we have $Q = 0$, $z_2 = 0$, and $z_1 = \nu_1^{-1}+\nu_2^{-1}+\nu_1^{-1}\nu_2^{-1}$. Thus, the expression for the combining throughput of User 2 in C-ANOMA systems coincides with that in C-NOMA systems, i.e., $R_2^{\mathrm{ANOMA}}|_{\tau=0} = R_2^{\mathrm{NOMA}}$.

3) Comparison between C-ANOMA and C-NOMA: We derive the following theorem which compares the throughputs of the C-ANOMA and C-NOMA systems for $N \to \infty$.

Theorem 3: In C-ANOMA systems, the combining throughput of User 2 for the asymptotic case of $N \to \infty$ is given by

$$\begin{aligned} R_{2,\text{asymp}}^{\text{ANOMA}} &\triangleq \lim_{N\to\infty} R_2^{\text{ANOMA}} \\ &= \frac{1}{2}\log\Bigg[\frac{1+P_r|h_{12}|^2}{2} + \frac{P_2|h_2|^2 + P_1P_2|h_2|^4Q}{2(1+P_1|h_2|^2)} \\ &\quad + \frac{1}{2}\sqrt{\left(1+P_r|h_{12}|^2 + \frac{P_2|h_2|^2+P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\right)^2 - \left(\frac{P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\right)^2}\Bigg], \end{aligned} \tag{33}$$

where $Q = 2\tau(1-\tau)$. The combining throughputs of User 2 for the C-NOMA and C-ANOMA systems satisfy the following inequalities

$$\begin{aligned} R_2^{\text{NOMA}} &\le R_{2,L}^{\text{ANOMA}} \triangleq \frac{1}{2}\log\left(1+P_r|h_{12}|^2 + \frac{P_2|h_2|^2 + \frac{1}{2}P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\right) \\ &\le R_{2,\text{asymp}}^{\text{ANOMA}} \\ &\le R_{2,U}^{\text{ANOMA}} \triangleq \frac{1}{2}\log\left(1+P_r|h_{12}|^2 + \frac{P_2|h_2|^2 + P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\right), \end{aligned} \tag{34}$$

where the equal signs are achieved if and only if $\tau = 0$.

Proof: See Appendix D. ■

We note from (34) that the gain of C-ANOMA over C-NOMA depends on the term $P_1P_2|h_2|^4Q$, thus, a better direct channel between User 2 and the BS results in a greater performance improvement of C-ANOMA systems compared with C-NOMA systems. Moreover, according to (25), Theorems 1 and 3, it is shown that for $N \to \infty$, the throughputs of both users to detect the weak user's message in the C-ANOMA systems are larger than those in the C-NOMA systems while the throughput of the strong user to detect its own message is identical for the C-ANOMA and C-NOMA systems. In Section V, we will show by numerical results that the C-ANOMA systems outperform the C-NOMA systems in terms of the throughput to decode the weak user's message with a relatively small value of N, e.g., $N > 20$.

Furthermore, in both C-ANOMA and C-NOMA systems, the actual throughput of User 2 is affected by both the throughput of User 1 to detect User 2's message, $R_{2\to1}$, and the combining throughput of User 2, R_2. Since User 2's message needs to be detected by both Users 1 and 2, the minimum of $R_{2\to1}$ and R_2 is the bottleneck of the actual throughput of User 2, i.e.,

$R_{\mathrm{act},2} = \min\{R_{2\to1}, R_2\}$ [21]. According to Theorems 1 and 3, it is trivial to derive that for a sufficiently large N and $\tau \neq 0$, $R_{\mathrm{act},2}^{\mathrm{ANOMA}} > R_{\mathrm{act},2}^{\mathrm{NOMA}}$.

IV. C-ANOMA System Design

In this section, we study the optimal design of the C-ANOMA systems, including the optimal timing mismatch and the power control strategy.

A. Optimal Timing mismatch

We first investigate the optimal normalized timing mismatch, τ^*. Although the optimal normalized timing mismatch to maximize $R_{2\to1}^{\mathrm{ANOMA}}$ and R_2^{ANOMA} is analytically intractable for a general finite block length N, we can numerically obtain τ^* for a given finite N by simply searching in the range of $0 \leq \tau < 1$ as done in Section V. To derive the optimal τ for a large N, we study the asymptotic case of $N \to \infty$. According to (25), the throughput of User 1 to detect its own message is independent of τ. According to (26) and (33), it is easy to show that $R_{2\to1,\mathrm{asymp}}^{\mathrm{ANOMA}}$ and $R_{2,\mathrm{asymp}}^{\mathrm{ANOMA}}$ are increasing functions of Q which is given by $2\tau(1-\tau)$. Thus, maximizing $R_{2\to1,\mathrm{asymp}}^{\mathrm{ANOMA}}$ and $R_{2,\mathrm{asymp}}^{\mathrm{ANOMA}}$ is equivalent to maximizing the term $\tau(1-\tau)$. Therefore, the optimal τ to maximize the throughputs of both users to detect User 2's message converges to 0.5, i.e.,

$$\tau^* \triangleq \arg\max_{\tau} R_{2\to1,\mathrm{asymp}}^{\mathrm{ANOMA}} = \arg\max_{\tau} R_{2,\mathrm{asymp}}^{\mathrm{ANOMA}} = 0.5. \tag{35}$$

In practice, in order to reduce the resource consumption of the control information, the normalized timing mismatch can be fixed to the default value of 0.5 at both the BS and the users.

B. Power Minimization

In this paper, we consider the delay-tolerant mode where the BS and the relay user can dynamically adjust their transmit powers according to the channel states in order to avoid outage and satisfy the minimum rate requirements [10]. Our objective is to minimize the weighted sum transmit power of the BS and the relay user under the minimum rate (i.e., QoS) requirements and the individual power constraints. Then, the power minimization problem can be formulated as

$$\min_{P_1,P_2,P_r} \quad \omega_s(P_1+P_2) + \omega_r P_r, \tag{36a}$$

$$\text{s.t.} \qquad R_{2\to1}^{\text{ANOMA}} \geq R_2^*, R_1^{\text{ANOMA}} \geq R_1^*, R_2^{\text{ANOMA}} \geq R_2^*, \tag{36b}$$

$$P_1 + P_2 < P_{s,\max}, P_r < P_{r,\max}, \tag{36c}$$

where ω_s and ω_r are the non-negative weights for the transmit powers of the BS and User 1, respectively, such that $\omega_s + \omega_r = 1$. $P_{s,\max}$ and $P_{r,\max}$ stand for the maximum available powers of the BS and User 1, respectively. R_1^* and R_2^* are the target rates of Users 1 and 2's messages. Note that the choice of ω_s and ω_r provides a trade-off between the power consumption of the BS and that of the relay user. For instance, if one wants to further restrict the power consumption of the relay user due to its limited battery capacity, ω_r should be chosen greater than ω_s.

The exact expressions of $R_{2\to1}^{\text{ANOMA}}$ and R_2^{ANOMA} in (18) and (28) make the optimization problem (36) analytically intractable. To simplify the optimization problem, we replace $R_{2\to1}^{\text{ANOMA}}$ and R_2^{ANOMA} in (36b) with their *asymptotic lower bounds*, which can provide a suboptimal solution for the original optimization problem (36), i.e.,

$$\min_{P_1,P_2,P_r} \quad \omega_s(P_1 + P_2) + \omega_r P_r, \tag{37a}$$

$$\text{s.t.} \qquad R_{2\to1,L}^{\text{ANOMA}} \geq R_2^*, \tag{37b}$$

$$R_1^{\text{ANOMA}} \geq R_1^*, \tag{37c}$$

$$R_{2,L}^{\text{ANOMA}} \geq R_2^*, \tag{37d}$$

$$P_1 + P_2 < P_{s,\max}, P_r < P_{r,\max}. \tag{37e}$$

For sufficiently large values of N, Eqs. (37b) and (37d) are stronger constraints for $R_{2\to1}^{\text{ANOMA}}$ and R_2^{ANOMA} compared with those in (36b), which means that the solution of (36) can do at least as good as that of (37). In what follows, we explain that (37) can also provide a suboptimal solution of (36) for a finite N. By definition, as N increases, the exact throughputs can be arbitrarily close to the asymptotic ones. We assume that $R_{2\to1}^{\text{ANOMA}} \geq R_{2\to1,L}^{\text{ANOMA}}$ for any $N \geq N_1$ and $R_2^{\text{ANOMA}} \geq R_{2,L}^{\text{ANOMA}}$ for any $N \geq N_2$. By choosing a proper N^*, for example, $N^* = \max\{N_1, N_2\}$, we can ensure that $R_{2\to1}^{\text{ANOMA}} \geq R_{2\to1,L}^{\text{ANOMA}}$ and $R_2^{\text{ANOMA}} \geq R_{2,L}^{\text{ANOMA}}$ for the given N^*. We will show that N^* can be a reasonable value (e.g., $N^* = 100$) in the numerical results section. In practice, the actual block length is usually greater than 100. For example, in global system for mobile communications (GSM), there are approximately 156 symbols in a normal burst (a physical channel carrying information on traffic and control channels) [35]. As

a result, the optimization problem (37) can provide a suboptimal solution for the problem (36) with the block length used in practical communication systems.

By simplifying (37b), (37c), and (37d), we obtain

$$P_2 \geq \frac{\gamma_2}{|h_1|^2}\frac{1+P_1|h_1|^2}{1+\frac{1}{2}QP_1|h_1|^2}, \tag{38}$$

$$P_1 \geq \frac{\gamma_1+\epsilon}{|h_1|^2}, \tag{39}$$

$$P_r \geq \frac{\gamma_2}{|h_{12}|^2} - \frac{P_2|h_2|^2}{|h_{12}|^2}\frac{1+\frac{1}{2}QP_1|h_2|^2}{1+P_1|h_2|^2}. \tag{40}$$

where $\gamma_i = 2^{2R_i^*}-1$, $i=1,2$, is the target signal-to-interference-plus-noise ratio (SINR) to detect User i's message and $\epsilon = 2^{2R_1^*}(2^{\frac{\tau}{N}R_1^*}-1)$. The value of ϵ can be made arbitrary small with increasing N. For a sufficiently large N, i.e., $N > N^*$, we have $\epsilon < \epsilon^* \triangleq 2^{2R_1^*}(2^{\frac{\tau}{N^*}R_1^*}-1)$, hence, we can substitute (39) with a stronger constraint, i.e.,

$$P_1 \geq \frac{\gamma_1+\epsilon^*}{|h_1|^2}. \tag{41}$$

Then, by replacing the constraints with (38), (40), and (41), the optimization problem (37) becomes

$$\min_{P_1,P_2,P_r} \quad \omega_s(P_1+P_2)+\omega_r P_r, \tag{42a}$$

$$\text{s.t.} \quad \frac{\gamma_1+\epsilon^*}{|h_1|^2} \leq P_1 \leq P_{s,\max}, \tag{42b}$$

$$\frac{\gamma_2}{|h_1|^2}\frac{1+P_1|h_1|^2}{1+\frac{1}{2}QP_1|h_1|^2} \leq P_2 \leq P_{s,\max}-P_1, \tag{42c}$$

$$\zeta_r \triangleq \max\left\{0, \frac{\gamma_2}{|h_{12}|^2} - \frac{P_2|h_2|^2}{|h_{12}|^2}\frac{1+\frac{1}{2}QP_1|h_2|^2}{1+P_1|h_2|^2}\right\} \leq P_r \leq P_{r,\max}. \tag{42d}$$

Note that (42d) indicates that the feasible domain of P_r depends on P_1 and P_2 while the constraints of P_1 and P_2 in (42b) and (42c) do not rely on P_r. For any given P_1 and P_2, the weighted sum power is minimized when P_r is equal to the least possible value, i.e., $P_r = \zeta_r$. Besides, we note that increasing P_1 improves R_1^{ANOMA} while worsens $R_{2,L}^{\text{ANOMA}}$ and $R_{2\to1,L}^{\text{ANOMA}}$ due to the increased interference from User 1's message. Then, the powers P_2 and P_r have to increase to counteract the interference of User 1's message. As a result, P_1 should also be chosen as the least possible value within the feasible domain (42b) to minimize the weighted sum power, i.e., $P_1 = \frac{\gamma_1+\epsilon^*}{|h_1|^2}$. By substituting the values of P_r and P_1, the optimization problem (42) becomes

$$\min_{P_2} \quad \omega_s P_2 + \omega_r \max\left\{0, \frac{\gamma_2}{|h_{12}|^2} - \frac{P_2|h_2|^2}{|h_{12}|^2}\frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}\right\}, \tag{43a}$$

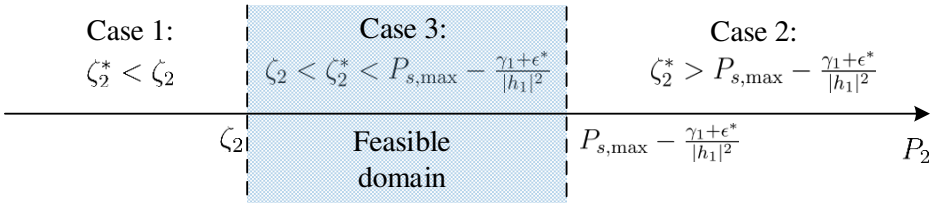

Fig. 2: Illustration of the relationship among ζ_2^*, ζ_2, and $P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$.

$$\text{s.t.} \quad \zeta_2 \le P_2 \le P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}, \tag{43b}$$

where

$$\zeta_2 = \max\left\{ \frac{\gamma_2}{|h_1|^2}\frac{1+\gamma_1+\epsilon^*}{1+\frac{1}{2}Q(\gamma_1+\epsilon^*)}, \left(\frac{\gamma_2}{|h_{12}|^2} - P_{r,\max}\right)\frac{|h_{12}|^2}{|h_2|^2}\frac{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2} \right\}, \tag{44}$$

and the rightmost term in (44) is derived by setting $\zeta_r < P_{r,\max}$.

We note from (42b) and (43b) that if $\frac{\gamma_1+\epsilon^*}{|h_1|^2} > P_{s,\max}$ or $\zeta_2 > P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$, there is no valid solution for the power minimization problem, i.e., QoS requirements cannot be satisfied with the limited transmit powers of the BS and the relay user. The following analysis is under the assumption that there are valid solutions for the power minimization problem.

By setting $\frac{\gamma_2}{|h_{12}|^2} - \frac{P_2|h_2|^2}{|h_{12}|^2}\frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2} = 0$, we obtain

$$P_2 = \zeta_2^* \triangleq \frac{\gamma_2}{|h_2|^2}\frac{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}. \tag{45}$$

Then, the objective function (43a) becomes $\omega_s P_2 + \omega_r\left(\frac{\gamma_2}{|h_{12}|^2} - \frac{P_2|h_2|^2}{|h_{12}|^2}\frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}\right)$ if $P_2 < \zeta_2^*$ and $\omega_s P_2$ otherwise. As shown in Fig. 2, we separate the following analysis into three cases according to the relationship among ζ_2^*, ζ_2, and $P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$.

1) Case 1: If $\zeta_2^* < \zeta_2$, the optimization problem (43) becomes

$$\min_{P_2} P_2, \quad \text{s.t. } \zeta_2 \le P_2 \le P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}, \tag{46}$$

In this case, it is easy to obtain that the optimal transmit powers are $P_1^* = \frac{\gamma_1+\epsilon^*}{|h_1|^2}$, $P_2^* = \zeta_2$, and $P_r^* = 0$. Intuitively, this case indicates that the channel between the BS and User 2 is strong enough such that no relay transmission is needed to satisfy the QoS requirements at User 2.

2) Case 2: If $\zeta_2^* > P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$, the optimization problem (43) becomes

$$\min_{P_2} \quad \omega_s P_2 + \omega_r\left(\frac{\gamma_2}{|h_{12}|^2} - \frac{P_2|h_2|^2}{|h_{12}|^2}\frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}\right), \tag{47a}$$

$$\text{s.t.} \quad \zeta_2 \le P_2 \le P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}. \tag{47b}$$

By omitting the constant terms, the objective function (47a) becomes

$$\min_{P_2} \left(\omega_s - \omega_r \frac{|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2 + \frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2 + (\gamma_1+\epsilon^*)|h_2|^2}\right) P_2. \tag{48}$$

Note that the solution of (47) depends on the values of ω_s and ω_r. The solutions can be given as follows: If $\omega_s = \omega_r \frac{|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}$, the optimal transmit powers are

$$P_1^* = \frac{\gamma_1+\epsilon^*}{|h_1|^2}, P_2^* \in \left[\zeta_2, P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}\right], P_r^* = \frac{\gamma_2}{|h_{12}|^2} - \frac{P_2^*|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}, \tag{49}$$

where P_2^* can be any value in the given range. We provide the intuitive explanation for the solution as follows: One can observe from (49) that P_r^* decreases with P_2^*. Under certain conditions on ω_r and ω_s, i.e., $\omega_s = \omega_r \frac{|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}$, the weighted sum power will be constant as the decrease in P_r^* is equal to the increase in P_2^*.

If $\omega_s > \omega_r \frac{|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}$, P_2^* should be chosen the least possible value. The optimal transmit powers are given by

$$P_1^* = \frac{\gamma_1+\epsilon^*}{|h_1|^2}, P_2^* = \zeta_2, P_r^* = \frac{\gamma_2}{|h_{12}|^2} - \frac{P_2^*|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2 + \frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}. \tag{50}$$

If $\omega_s < \omega_r \frac{|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}$, P_2^* should choose the largest possible value. The optimal transmit powers are given by

$$P_1^* = \frac{\gamma_1+\epsilon^*}{|h_1|^2}, P_2^* = P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}, P_r^* = \frac{\gamma_2}{|h_{12}|^2} - \frac{P_2^*|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2 + \frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}. \tag{51}$$

3) Case 3: If $\zeta_2 < \zeta_2^* < P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$, the optimization problem (43) becomes two sub-problems, i.e.,

$$\min_{P_2} P_2, \quad \text{s.t. } \zeta_2^* \le P_2 \le P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}, \tag{52}$$

and

$$\min_{P_2} P_2 \left(\omega_s - \omega_r \frac{P_2|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2 + \frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}\right), \quad \text{s.t. } \zeta_2 \le P_2 \le \zeta_2^*. \tag{53}$$

By following the derivation of Case 1, one can solve the problem (52). Similarly, by following the steps of Case 2, one can solve the problem (53). We assume that the optimal transmit powers for (52) and (53) are $[\tilde{P}_1,\ \tilde{P}_2,\ \tilde{P}_r]$ and $[\bar{P}_1,\ \bar{P}_2,\ \bar{P}_r]$, respectively. Then, the solution for Case 3 is given by

$$[P_1^*, P_2^*, P_r^*] = \underset{[P_1,P_2,P_r] \in \left\{[\tilde{P}_1,\tilde{P}_2,\tilde{P}_r],[\bar{P}_1,\bar{P}_2,\bar{P}_r]\right\}}{\arg\min} \omega_s(P_1+P_2) + \omega_r P_r. \tag{54}$$

To summarize the solutions, we provide Algorithm 1 to solve the problem (42). Note that Algorithm 1 is implemented by using only the conditional statements. All the expressions used in Algorithm 1 are closed-form. Hence, Algorithm 1 can run in constant time.

Algorithm 1 Algorithm to find the optimal powers under QoS constraints

function SOLVE_CASE_1(L, U)
  return $P_2^* = L$, $P_r^* = 0$.
end function
function SOLVE_CASE_2(L, U)
  if $\omega_s = \omega_r \frac{|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}$, **then** $P_2^* = \text{random}\left(\left[\zeta_2, P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}\right]\right)$.
  else if $\omega_s > \omega_r \frac{|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}$, **then** $P_2^* = L$.
  else $P_2^* = U$.
  return P_2^*, $P_r^* = \frac{\gamma_2}{|h_{12}|^2} - \frac{P_2^*|h_2|^2}{|h_{12}|^2} \frac{|h_1|^2+\frac{1}{2}Q(\gamma_1+\epsilon^*)|h_2|^2}{|h_1|^2+(\gamma_1+\epsilon^*)|h_2|^2}$.
end function
if $\frac{\gamma_1+\epsilon^*}{|h_1|^2} > P_{s,\max}$ or $\zeta_2 > P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$, **then** there is no solution, break.
$P_1^* = \frac{\gamma_1+\epsilon^*}{|h_1|^2}$
if $\zeta_2^* < \zeta_2$, **then** P_2^*, P_r^* = SOLVE_CASE_1(ζ_2, $P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$).
else if $\zeta_2^* > P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$, **then** P_2^*, P_r^* = SOLVE_CASE_2(ζ_2, $P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2}$).
else
  $\tilde{P}_2^*, \tilde{P}_r^* = \text{SOLVE\_CASE\_1}(\zeta_2^*, P_{s,\max} - \frac{\gamma_1+\epsilon^*}{|h_1|^2})$.
  $\hat{P}_2^*, \hat{P}_r^* = \text{SOLVE\_CASE\_2}(\zeta_2, \zeta_2^*)$.
  $P_2^*, P_r^* = \arg\min_{[P_2,P_r]\in\{[\tilde{P}_2^*,\tilde{P}_r^*],[\hat{P}_2^*,\hat{P}_r^*]\}} \omega_s P_2 + \omega_r P_r$.
return P_1^*, P_2^*, P_r^*.

C. Comparison with C-NOMA

According to (25), (26), Theorems 1 and 3, the expressions for the throughputs in C-ANOMA systems, $R_{2\to1,L}^{\text{ANOMA}}$, R_1^{ANOMA}, and $R_{2,L}^{\text{ANOMA}}$, become those in C-NOMA systems, $R_{2\to1}^{\text{NOMA}}$, R_1^{NOMA}, and R_2^{NOMA}, by setting $\tau = 0$. Therefore, the solutions derived in the previous subsection can be applied to the C-NOMA systems simply by setting $\tau = 0$ which then results in $\epsilon^* = 0$ and $Q = 0$. For the C-NOMA systems, the power minimization problem (42) becomes

$$\min_{P_1,P_2,P_r} \quad \omega_s(P_1+P_2) + \omega_r P_r, \tag{55a}$$

$$\text{s.t.} \quad \frac{\gamma_1}{|h_1|^2} \le P_1 \le P_{s,\max}, \tag{55b}$$

$$\frac{\gamma_2\left(1+P_1|h_1|^2\right)}{|h_1|^2} \le P_2 \le P_{s,\max} - P_1, \tag{55c}$$

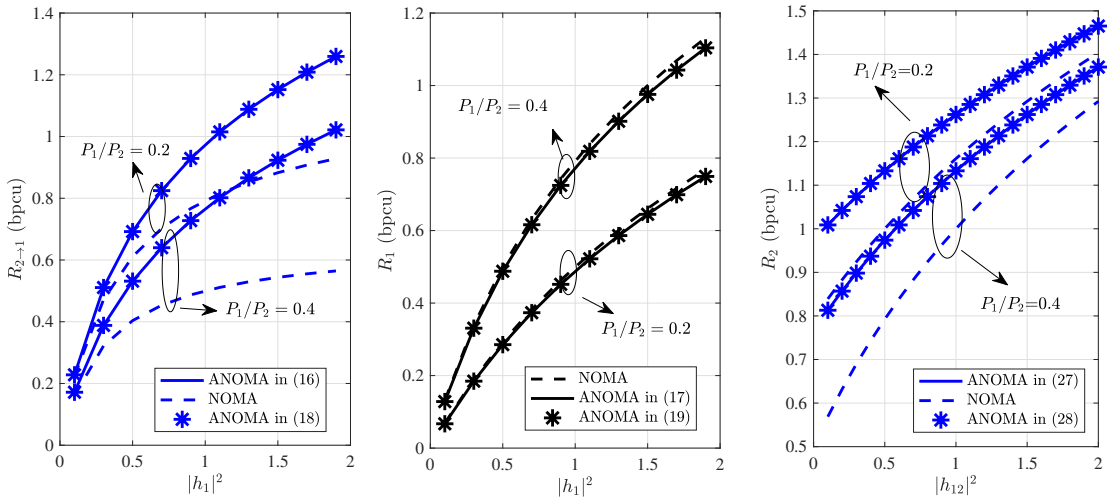

Fig. 3: The throughputs $R_{2\to1}$, R_1, and R_2 as functions of the channel gain $|h_1|^2$ or $|h_{12}|^2$ for C-ANOMA and C-NOMA systems when $N = 10$, $\tau = 0.5$, $P_1 + P_2 = 5$, $P_r = 2$, $|h_2|^2 = 1$.

$$\max\left\{0, \frac{\gamma_2}{|h_{12}|^2} - \frac{P_2|h_2|^2}{|h_{12}|^2\left(1 + P_1|h_2|^2\right)}\right\} \leq P_r \leq P_{r,\max}. \tag{55d}$$

Note that the feasible domains of P_2 and P_r in (55c) and (55d) are the subsets of those in (42c) and (42d), respectively. As a result, for a sufficiently large N, the minimization problem (42) for the C-ANOMA systems is a relaxation of the minimization problem (55) for the C-NOMA systems [36]. That is, the problem (42) provides a solution to minimize the weighted sum power within a wider feasible domain compared with (55). In other words, if $[P^*_{1,\text{ANOMA}}, P^*_{2,\text{ANOMA}}, P^*_{r,\text{ANOMA}}]$ and $[P^*_{1,\text{NOMA}}, P^*_{2,\text{NOMA}}, P^*_{r,\text{NOMA}}]$ are the optimal solutions for (42) and (55), respectively, we have

$$\omega_s(P^*_{1,\text{ANOMA}} + P^*_{2,\text{ANOMA}}) + \omega_r P^*_{r,\text{ANOMA}} \leq \omega_s(P^*_{1,\text{NOMA}} + P^*_{2,\text{NOMA}}) + \omega_r P^*_{r,\text{NOMA}}. \tag{56}$$

We note from (56) that for a sufficiently large block length, the C-ANOMA systems can consume less power compared with the C-NOMA systems in order to guarantee the same QoS. We will illustrate this phenomenon with numerical results in Section V.

V. Numerical Results

In this section, we present numerical results to compare the throughputs and power consumptions of C-NOMA and C-ANOMA systems.

First, we compare the throughputs of Users 1 and 2 in the C-NOMA and C-ANOMA systems with different ratios of P_1 to P_2 in Fig. 3. The curves of "ANOMA in (16)/(18)/(17)/(19)/(27)/(28)"

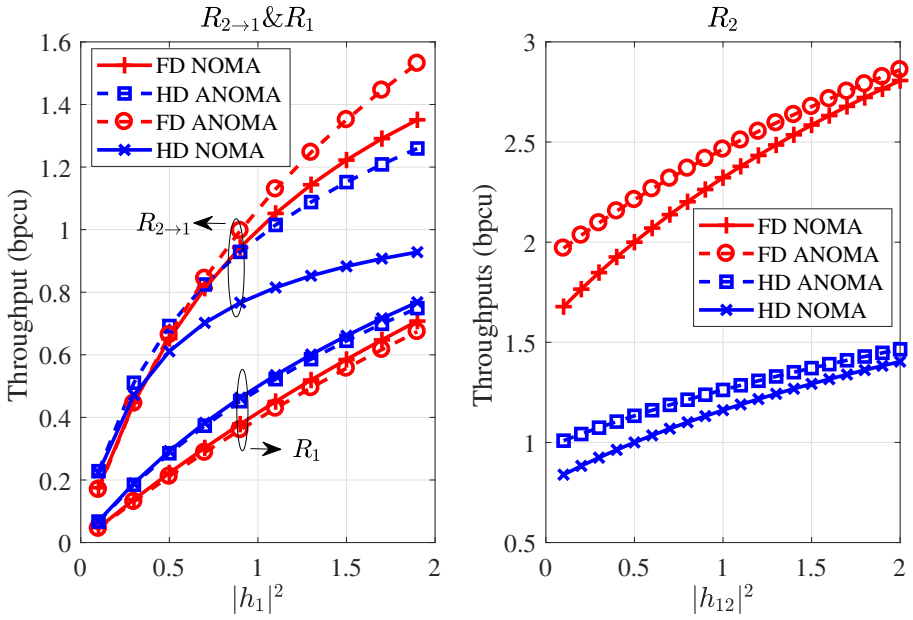

Fig. 4: The throughputs as functions of the channel gain for the full-duplex or half-duplex C-NOMA and C-ANOMA systems when $N = 10$, $\tau = 0.5$, $P_1 = 1$, $P_2 = 4$, $P_r = 2$, $|h_2|^2 = 1$, $|h_{\rm LI}|^2 = 1$.

are derived directly from the expressions in (16)/(18)/(17)/(19)/(27)/(28). First, in Fig. 3, it is shown that the throughputs calculated by (18), (19), and (28) completely align with the results of (16), (17), and (27), respectively, which verifies the correctness of (18), (19), and (28). Second, it is demonstrated that the throughputs $R_{2\to1}$ and R_2 in the C-ANOMA systems are higher than those in the C-NOMA systems. And $R_1^{\rm ANOMA}$ is less than but very close to $R_1^{\rm NOMA}$ even for a relatively small block length $N = 10$, especially when $|h_1|^2$ is small. Third, Fig. 3 shows that the throughputs in both C-ANOMA and C-NOMA systems increase with the channel gain. More specifically, the gaps of $R_{2\to1}$ between the C-ANOMA and C-NOMA systems grow wider as $|h_1|^2$ increases. In contrast, the gaps of R_2 between the C-ANOMA and C-NOMA systems shrink as $|h_{12}|^2$ increases. Note that R_2 depends on both the broadcast link from the BS and the relay link from User 1. The sampling diversity can only be obtained through the asynchronous transmission from the broadcast link. As $|h_{12}|^2$ increases, the quality of the relay link becomes more and more dominant in calculating R_2. Accordingly, the throughput gain from the sampling diversity becomes less and less noticeable as $|h_{12}|^2$ increases while $|h_2|^2$ is constant. Finally, it is evident that the actual throughput of User 2 for C-ANOMA is better than that of C-NOMA since $R_2^{\rm ANOMA} > R_2^{\rm NOMA}$ and $R_{2\to1}^{\rm ANOMA} > R_{2\to1}^{\rm NOMA}$.

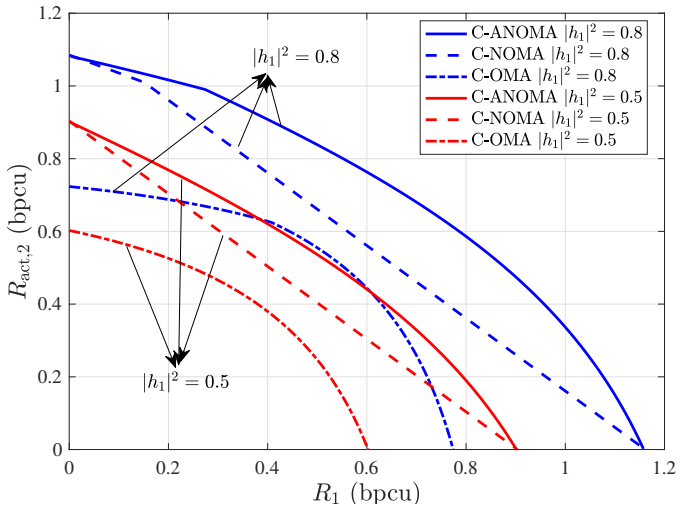

Fig. 5: The throughputs R_1 vs. $R_{\text{act},2}$ for C-ANOMA, C-NOMA, and C-OMA systems when $|h_{12}|^2 = 1$, $|h_2|^2 = 0.5$, $N = 100$, $\tau = 0.5$, $P_1 + P_2 = 5$, $P_r = 1$.

We compare the throughput performances of C-ANOMA and C-NOMA systems in the full-duplex or half-duplex mode in Fig. 4. In our simulation, we calculate the users' throughputs in the full-duplex C-NOMA systems based on the SINR expressions derived in the existing literature, for example [10]. Using the notation in [10], $|h_{\text{LI}}|^2$ stands for the level of the residual loop self-interference at the relay user caused by the full-duplex operation. For the full-duplex C-ANOMA and C-NOMA, the throughput of the relay user is calculated by treating the self-interference as noise. In Fig. 4, it is shown that in both half-duplex and full-duplex modes, $R_{2\to1}$ and R_2 in C-ANOMA outperform those in C-NOMA while R_1 in C-ANOMA is very close to that in C-NOMA. These results align with the results in Fig. 3. Besides, in some cases, the performance of full-duplex systems can be worse than that of the half-duplex systems due to the residual loop self-interference. For example, R_1 in the full-duplex C-ANOMA or C-NOMA system is worse than that in half-duplex C-ANOMA or C-NOMA system. As studied in the existing literature, the self-interference cancellation plays a crucial role in the full-duplex systems [37]. Moreover, it is also shown that R_2 in the full-duplex C-ANOMA or C-NOMA system is better than that in the half-duplex C-ANOMA or C-NOMA system because the relay user in the full-duplex mode can relay the signal to the weak user without consuming an additional time block. Note that the full-duplex operation and the realization of the self-interference require higher hardware complexity and power consumption at the relay user compared with the half-duplex operation.

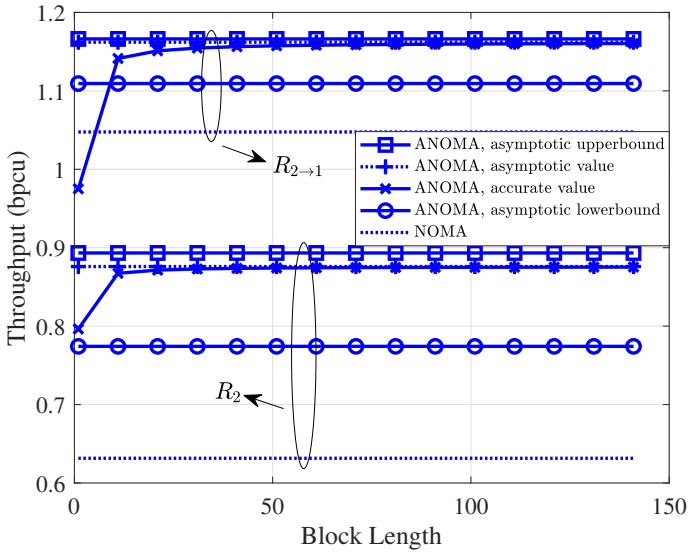

Fig. 6: The throughputs R_2 and $R_{2\to1}$ as functions of the block length N for C-ANOMA and C-NOMA systems when $\tau = 0.5$, $P_1 = 1.5$, $P_2 = 3.5$, $P_r = 2$, $|h_1|^2 = 1$, $|h_2|^2 = 0.8$, $|h_{12}|^2 = 1$.

We compare the performances of C-ANOMA, C-NOMA, and C-OMA in Fig. 5. Following the comparison between C-NOMA and C-OMA in the existing literature [9–11, 21], we adopt the time division multiple access (TDMA) C-OMA as an example of C-OMA systems. Specifically, the transmission in C-OMA occupies three time slots. In the first time slot, the BS transmits User 1's messages to User 1. In the second time slot, the BS broadcasts User 2's messages to Users 1 and 2. In the last time slot, User 1 relays User 2's messages to User 2. In Fig. 5, it is shown that C-NOMA achieves a better performance compared with C-OMA. Besides, it is also demonstrated that C-ANOMA outperforms C-NOMA because the oversampling technique provides extra sampling diversity [32–34].

In Fig. 6, we show how the throughputs in C-ANOMA systems change with the block length N. Since the expression for R_1^{ANOMA} in (19) is simple, the curves of R_1^{ANOMA} are omitted in Fig. 6. It is shown that as the block length increases, the accurate throughputs $R_{2\to1}^{\mathrm{ANOMA}}$ and R_2^{ANOMA} converge to the asymptotic ones calculated by (26) and (33), respectively. We note that the asymptotic throughputs, $R_{2\to1,\mathrm{asymp}}^{\mathrm{ANOMA}}$ and $R_{2,\mathrm{asymp}}^{\mathrm{ANOMA}}$, perfectly approximate the accurate throughputs, $R_{2\to1}^{\mathrm{ANOMA}}$ and R_2^{ANOMA}, when $N > 50$. And for both $R_{2\to1}^{\mathrm{ANOMA}}$ and R_2^{ANOMA}, the accurate throughputs exceed their asymptotic lower bounds when $N > 20$. As a result, for $N > 20$, it is reasonable to use the lower bounds of the asymptotic throughputs as the constraints

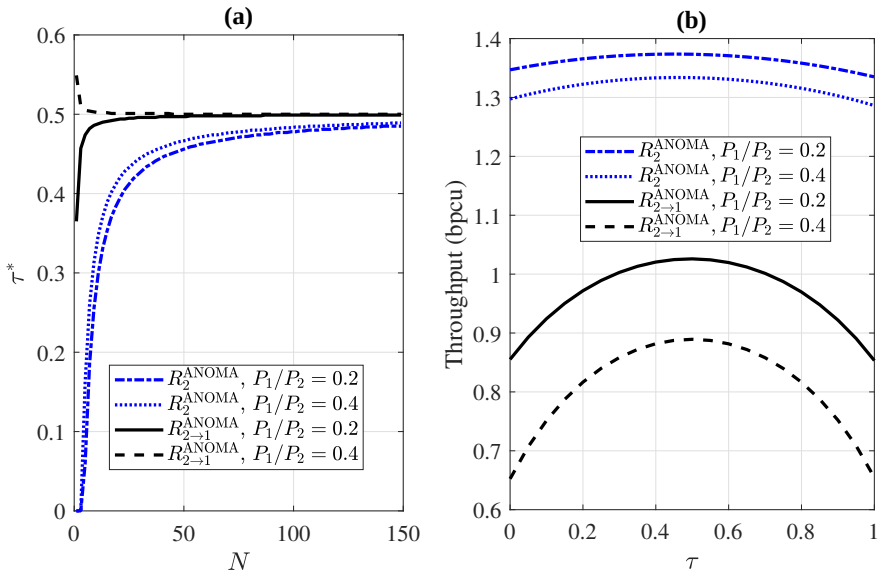

Fig. 7: (a) The optimal normalized timing mismatch τ^* to maximize the throughputs R_2^{ANOMA} and $R_{2\to1}^{\mathrm{ANOMA}}$ as a function of the block length N when $P_1+P_2=5$, $P_r=2$, $|h_1|^2=1$, $|h_2|^2=0.5$, $|h_{12}|^2=2$. (b) The throughputs R_2 and $R_{2\to1}$ as functions of the normalized timing mismatch τ when $N=50$, $P_1+P_2=5$, $P_r=2$, $|h_1|^2=1$, $|h_2|^2=0.5$, $|h_{12}|^2=2$.

(37b) and (37d) in order to simplify the optimization problem. Besides, Fig. 6 verifies Theorems 1 and 3 in addition to showing that the C-ANOMA systems outperform the C-NOMA systems for relatively small values of N.

Based on the results in Fig. 6, we discuss the time delay of the message delivery. In the non-cooperative NOMA or full-duplex C-NOMA systems, one message block is delivered to two users simultaneously via one time block. In contrast, in the half-duplex C-NOMA or C-ANOMA systems, one additional time block is needed to transmit the same message block. In order to reduce the delay, the block length is expected to be small. Besides, as shown in Fig. 6, the accurate throughputs in C-ANOMA systems increase with the block length. It is worth mentioning that when the block length is large, e.g., $N > 50$, a greater block length only results in a very subtle throughput improvement. Therefore, considering the delay and the throughput, a modest block length is desired.

We also study the optimal design of C-ANOMA systems. Fig. 7 (a) shows the optimal normalized timing mismatch τ^* to maximize R_2^{ANOMA} or $R_{2\to1}^{\mathrm{ANOMA}}$ as a function of the block length N. In our simulation, τ^* is found by exhaustive search. Although τ^* varies a lot when N

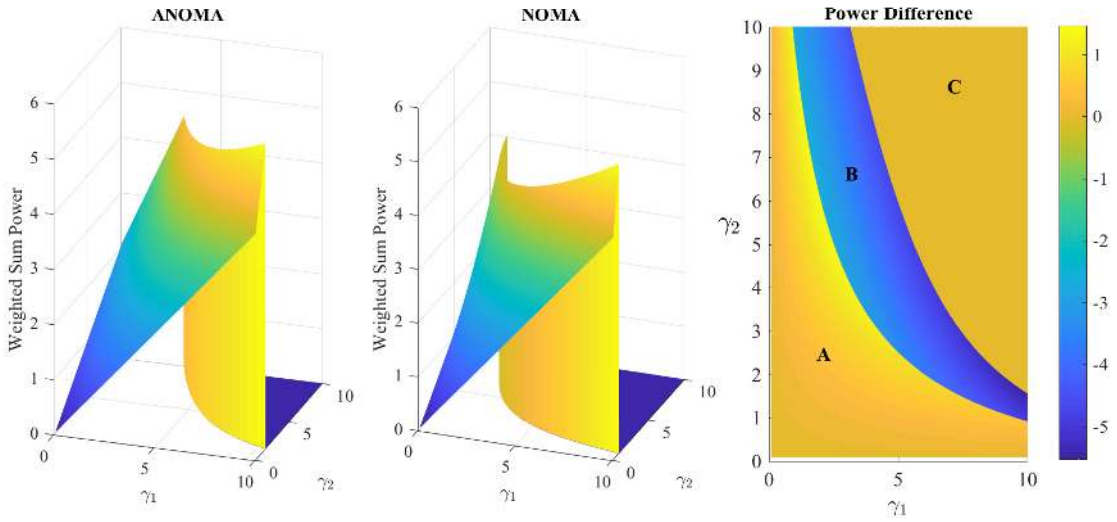

Fig. 8: The minimized weighted sum power under the QoS constraints as a function of the target SINRs, γ_1 and γ_2, for the C-NOMA and C-ANOMA systems when $\tau = 0.5$, $P_{s,\max} = 20$, $P_{r,\max} = 5$, $\omega_s = 0.2$, $\omega_r = 0.8$, $|h_1|^2 = 1$, $|h_2|^2 = 0.5$, $|h_{12}|^2 = 2$, $N = 100$. The right-most figure illustrates the difference between the minimized weighted sum power in C-NOMA systems and that in C-ANOMA systems. In that figure, A stands for the area where QoS constraints can be satisfied for both ANOMA and NOMA and $P_{\text{sum}}^{\text{ANOMA}} < P_{\text{sum}}^{\text{NOMA}}$. B stands for the area where QoS constraints can be satisfied for ANOMA, but not for NOMA. C stands for the area where QoS constraints cannot be satisfied for either ANOMA or NOMA.

is relatively small, τ^* converges to 0.5 steadily as N increases for both R_2^{ANOMA} and $R_{2\to1}^{\text{ANOMA}}$ with different ratios of P_1 to P_2, as predicted by our analytical results. This is because the timing mismatch only exists in the asynchronous transmission in the broadcast phase and will affect $R_{2\to1}^{\text{ANOMA}}$ and R_2^{ANOMA} in the same way. Moreover, Fig. 7 (b) presents how the throughputs change with the normalized timing mismatch. It is demonstrated that for both R_2^{ANOMA} and $R_{2\to1}^{\text{ANOMA}}$, the throughputs are maximized when $\tau \approx 0.5$, which verifies the results shown in Fig. 7 (a). Compared with $R_{2\to1}^{\text{ANOMA}}$, the choice of τ has a greater impact on R_1^{ANOMA}. It is because the relay link dominates the performance of User 2 if the channel of the relay link is good. Besides, it is shown that for $\tau \in [0.4, 0.6]$, the choice of τ only has a subtle effect on $R_{2\to1}^{\text{ANOMA}}$ and R_1^{ANOMA}.

Moreover, we show the minimized weighted sum power under the QoS constraints as a function of target SINRs, γ_1 and γ_2, for C-NOMA and C-ANOMA systems in Fig. 8. We set ω_s and ω_r as 0.2 and 0.8, respectively, because the power consumption of the relay user with limited battery capacity has a higher priority in the power minimization problem. In Fig. 8, the weighted sum

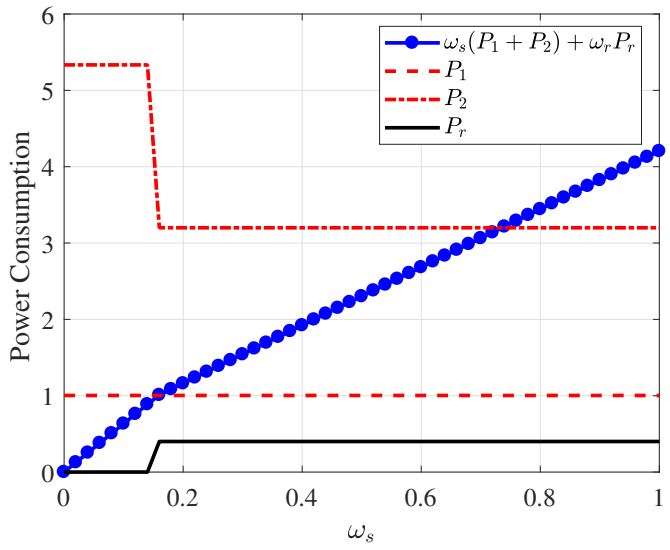

Fig. 9: The power consumptions as functions of the weight allocated to the transmit power of the BS, i.e., ω_s, for the C-ANOMA systems when $\tau = 0.5$, $P_{s,\max} = 20$, $P_{r,\max} = 5$, $|h_1|^2 = 1$, $|h_2|^2 = 0.5$, $|h_{12}|^2 = 2$, $N = 100$.

power is calculated by solving the power optimization problem (42) for the C-NOMA (setting $\tau = 0$) and C-ANOMA (setting $\tau = 0.5$) systems. In our simulation, we assume that the BS and the relay user will stop transmission (i.e., $P_1 = P_2 = P_r = 0$) if the QoS constraints cannot be satisfied. For both C-NOMA and C-ANOMA systems, it is shown in Fig. 8 that the weighted sum power increases with the target SINRs until the BS and the relay user reach their power limits and stop transmission. To further compare the power consumptions, we calculate the difference of the weighted sum powers between the C-NOMA and C-ANOMA systems and provide the results in Fig. 8. As shown in Fig. 8, C-ANOMA systems can consume less power compared with C-NOMA systems to guarantee the same QoS in the area A. In the area B, it is shown that C-ANOMA systems can still satisfy the QoS with limited transmit powers while C-NOMA systems cannot. When both γ_1 and γ_2 are large, i.e., in the area C, neither C-NOMA nor ANOMA systems can satisfy the QoS with the limited transmit powers.

Finally, we show how the power consumptions change with the weight ω_s in the power minimization problem in Fig. 9. In our simulation, we set $\omega_r = 1 - \omega_s$. In Fig. 9, the power allocated to User 1 does not change with ω_s as long as the BS has enough transmit power to support the QoS of User 1. If ω_s is large, the BS can save a large amount of power (decreases by about 2) under the help of the relay user (transmit power increases by about 0.5) because the

channel of the relay link is better than that of the broadcast link between the BS and User 2. When ω_s is small (ω_r is large), the relay user keeps silent to reduce energy consumption. When ω_s is large (ω_r is small), the BS communicates with User 2 under the help of the relay user, which takes advantage of the relay link to complement the large path loss between the BS and User 2. Hence, Fig. 9 shows that one can make a trade-off between the power consumption of the BS and that of the relay user by adjusting the weight ω_s.

VI. Conclusion

In this paper, we study the half-duplex C-ANOMA systems with user relaying. We analytically prove that for a sufficiently large block length, the strong user in C-ANOMA systems can achieve the same throughput as that in C-NOMA systems while the weak user in C-ANOMA systems benefits from the symbol-asynchronous transmission. Moreover, we analyze the optimal design of the C-ANOMA systems. As the block length increases, the optimal timing mismatch converges to half of the symbol interval. Besides, we solve a weighted sum power minimization problem under QoS constraints. Numerical results demonstrate that C-ANOMA systems can consume less power to satisfy the same QoS requirements compared with C-NOMA systems.

There are several directions worth studying in the future. For example, since the analysis of this paper is under the assumption of perfect channel estimation, the impact of imperfect channel estimation on the system performance is an interesting future work.

Appendix A
Derivation of (18) and (19)

Substituting $\mathbf{G}_1$ and $\mathbf{R}$ by their expressions, the matrix determinant term in (17) becomes

$$\det\left(\mathbf{I}_{2N}+P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)$$

$$= \det\begin{bmatrix} 1+P_1|h_1|^2 & P_1|h_1|^2(1-\tau) & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & P_1|h_1|^2\tau & 1+P_1|h_1|^2 & P_1|h_1|^2(1-\tau) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 & 0 \\ 0 & \cdots & 0 & P_1|h_1|^2\tau & 1+P_1|h_1|^2 & P_1|h_1|^2(1-\tau) \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}_{2N\times 2N}$$

$$\overset{(a)}{=} \det\begin{bmatrix} 1+P_1|h_1|^2 & P_1|h_1|^2(1-\tau) & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & P_1|h_1|^2\tau & 1+P_1|h_1|^2 & P_1|h_1|^2(1-\tau) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & P_1|h_1|^2\tau & 1+P_1|h_1|^2 \end{bmatrix}_{(2N-1)\times(2N-1)}$$

$$\overset{(b)}{=} \left(1+P_1|h_1|^2\right) \det \begin{bmatrix} 1+P_1|h_1|^2 & P_1|h_1|^2(1-\tau) & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & P_1|h_1|^2\tau & 1+P_1|h_1|^2 & P_1|h_1|^2(1-\tau) & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & P_1|h_1|^2\tau & 1+P_1|h_1|^2 & P_1|h_1|^2(1-\tau) \\ 0 & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix}_{(2N-2)\times(2N-2)}$$

$$= \cdots \overset{(c)}{=} (1+P_1|h_1|^2)^N, \tag{57}$$

where (a) and (b) are derived by applying the cofactor expansion [38], (c) is derived by applying the cofactor expansion iteratively. Thus, Eq. (19) is obtained.

According to (16), we have

$$\begin{aligned} & R_{2\to 1}^{\text{ANOMA}} \\ &= \frac{1}{2N+\tau} \log\det\left[\mathbf{I}_{2N} + \left(\mathbf{I}_{2N} + P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1} P_2|h_1|^2\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}\right] \\ &= \frac{1}{2N+\tau} \log\det\left[\left(\mathbf{I}_{2N} + P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1} \left(\mathbf{I}_{2N} + P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R} + P_2|h_1|^2\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}\right)\right] \\ &= \frac{1}{2N+\tau} \log\det\left(\mathbf{I}_{2N} + P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R} + P_2|h_1|^2\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}\right) \\ &\quad - \frac{1}{2N+\tau} \log\det\left(\mathbf{I}_{2N} + P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right) \\ &\overset{(a)}{=} \frac{1}{2N+\tau} \log\det\left(\mathbf{I}_{2N} + \mathbf{H}\mathbf{R}\right) - \frac{1}{2N+\tau} \log\det\left(\mathbf{I}_{2N} + P_1|h_1|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right), \end{aligned} \tag{58}$$

where (a) is derived because $\mathbf{G}_i\mathbf{G}_i^H$ is a $2N$-by-$2N$ matrix whose odd (if $i=1$) or even (if $i=2$) diagonal elements are 1 and all the others are 0, and $\mathbf{H} = |h_1|^2 \cdot \text{diag}\left([P_1, P_2, \cdots, P_1, P_2]\right)$.

According to Theorem 1 in [13], the term $\log\det\left(\mathbf{I}_{2N} + \mathbf{H}\mathbf{R}\right)$ in (58) can be written as

$$\log\det\left(\mathbf{I}_{2N} + \mathbf{H}\mathbf{R}\right) = N\log\left(\mu_1\mu_2\right) + \log\frac{\left(r_1^{N+1} - r_2^{N+1}\right) + \tau^2\left(r_1^N - r_2^N\right)}{r_1 - r_2}, \tag{59}$$

where $\mu_1 = P_1|h_1|^2, \mu_2 = P_2|h_1|^2, Q = 2\tau(1-\tau)$,

$$r_1 = \frac{\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}+Q + \sqrt{\left[\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}+Q\right]^2 - Q^2}}{2}, \tag{60}$$

$$r_2 = \frac{\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}+Q - \sqrt{\left[\mu_1^{-1}+\mu_2^{-1}+\mu_1^{-1}\mu_2^{-1}+Q\right]^2 - Q^2}}{2}. \tag{61}$$

Thus, Eq. (18) can be easily derived according to (57) and (59).

APPENDIX B
PROOF OF THEOREM 1

Proof: According to Corollary 1 in [13], we have

$$\lim_{N\to\infty} \frac{1}{N+\tau} \log\frac{\left(r_1^{N+1} - r_2^{N+1}\right) + \tau^2\left(r_1^N - r_2^N\right)}{r_1 - r_2} = \log r_1. \tag{62}$$

As a result, the combining throughput of User 2 for $N \to \infty$ is calculated as

$$R_{2,\text{asymp}}^{\text{ANOMA}} = \frac{1}{2}\log\left(\frac{\mu_1\mu_2 r_1}{1+\mu_1}\right) = \frac{1}{2}\log\left(\frac{1+\mu_1+\mu_2+\mu_1\mu_2 Q+\sqrt{(1+\mu_1+\mu_2)^2+2(1+\mu_1+\mu_2)\mu_1\mu_2 Q}}{2(1+\mu_1)}\right),$$

where $\mu_1 > 0$, $\mu_2 > 0$, $\tau \in [0,1)$, and $Q = 2\tau(1-\tau) > 0$. One can easily derive

$$\begin{aligned}1+\mu_1+\mu_2 &\le \sqrt{(1+\mu_1+\mu_2)^2+2(1+\mu_1+\mu_2)\mu_1\mu_2 Q}\\ &= \sqrt{(1+\mu_1+\mu_2+\mu_1\mu_2 Q)^2-(\mu_1\mu_2 Q)^2} \le 1+\mu_1+\mu_2+\mu_1\mu_2 Q,\end{aligned}$$

and the equal sign is achieved if and only if $\tau = 0$. As a result,

$$\frac{1}{2}\log\left(\frac{1+\mu_1+\mu_2+0.5\mu_1\mu_2 Q}{1+\mu_1}\right) \le R_{2,\text{asymp}}^{\text{ANOMA}} \le \frac{1}{2}\log\left(\frac{1+\mu_1+\mu_2+\mu_1\mu_2 Q}{1+\mu_1}\right).$$

Note that,

$$\frac{1}{2}\log\left(\frac{1+\mu_1+\mu_2+0.5\mu_1\mu_2 Q}{1+\mu_1}\right) \ge \frac{1}{2}\log\left(\frac{1+\mu_1+\mu_2}{1+\mu_1}\right) = R_2^{\text{NOMA}}, \tag{63}$$

where the equal sign is achieved if and only if $\tau = 0$. The proof is complete. ■

APPENDIX C
PROOF OF THEOREM 2

Proof: According to (27), the combining throughput of User 2 is given by

$$\begin{aligned}R_2^{\text{ANOMA}} &= \frac{1}{2N+\tau}\log\det\left[\mathbf{I}_{3N}+\left(\mathbf{R_N}+\mathbf{W}_1\mathbf{W}_1^H\right)^{-1}\mathbf{W}_2\mathbf{W}_2^H\right]\\
&= \frac{1}{2N+\tau}\log\det\left(\mathbf{I}_{3N}+\begin{bmatrix}\mathbf{R}+P_1|h_2|^2\mathbf{R}\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}^H & \mathbf{0}\\ \mathbf{0} & \mathbf{I}_N\end{bmatrix}^{-1}\begin{bmatrix}P_2|h_2|^2\mathbf{R}\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}^H & \sqrt{P_2P_r}h_2\bar{h}_{12}\mathbf{R}\mathbf{G}_2\\ \sqrt{P_2P_r}h_{12}\bar{h}_2\mathbf{G}_2^H\mathbf{R}^H & P_r|h_{12}|^2\mathbf{I}_N\end{bmatrix}\right)\\
&= \frac{1}{2N+\tau}\log\det\begin{bmatrix}\mathbf{I}_{2N}+P_2|h_2|^2\left(\mathbf{R}+P_1|h_2|^2\mathbf{R}\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1}\mathbf{R}\mathbf{G}_2\mathbf{G}_2^H\mathbf{R} & \sqrt{P_2P_r}h_2\bar{h}_{12}\left(\mathbf{R}+P_1|h_2|^2\mathbf{R}\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1}\mathbf{R}\mathbf{G}_2\\ \sqrt{P_2P_r}h_{12}\bar{h}_2\mathbf{G}_2^H\mathbf{R} & \mathbf{I}_N+P_r|h_{12}|^2\mathbf{I}_N\end{bmatrix}\\
&\overset{(a)}{=} \frac{1}{2N+\tau}\log\Big\{\det\left[(1+P_r|h_{12}|^2)\mathbf{I}_N\right]\det\Big[\mathbf{I}_{2N}+P_2|h_2|^2\left(\mathbf{I}_{2N}+P_1|h_2|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1}\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}\\
&\quad -\frac{P_2P_r|h_2|^2|h_{12}|^2}{1+P_r|h_{12}|^2}\left(\mathbf{I}_{2N}+P_1|h_2|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1}\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}\Big]\Big\}\\
&= \frac{N}{2N+\tau}\log\left(1+P_r|h_{12}|^2\right)+\frac{1}{2N+\tau}\log\det\Big\{\left(\mathbf{I}_{2N}+P_1|h_2|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1}\\
&\quad\cdot\left[\mathbf{I}_{2N}+P_1|h_2|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}+\left(P_2|h_2|^2-\frac{P_2P_r|h_2|^2|h_{12}|^2}{1+P_r|h_{12}|^2}\right)\left(\mathbf{I}_{2N}+P_1|h_2|^2\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}\right)\right]\Big\}\\
&= \frac{N}{2N+\tau}\log\left(1+P_r|h_{12}|^2\right)+\frac{1}{2N+\tau}\log\det\left(\mathbf{I}_{2N}+P_1|h_2|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)^{-1}\\
&\quad+\frac{1}{2N+\tau}\log\det\left(\mathbf{I}_{2N}+P_1|h_2|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}+\frac{P_2|h_2|^2}{1+P_r|h_{12}|^2}\mathbf{G}_2\mathbf{G}_2^H\mathbf{R}\right)\\
&= \frac{N}{2N+\tau}\log\left(1+P_r|h_{12}|^2\right)-\frac{\log\det\left(\mathbf{I}_{2N}+P_1|h_2|^2\mathbf{G}_1\mathbf{G}_1^H\mathbf{R}\right)}{2N+\tau}+\frac{\log\det(\mathbf{I}_{2N}+\tilde{\mathbf{H}}\mathbf{R})}{2N+\tau},\end{aligned} \tag{64}$$

where (a) is derived by applying the determinant of the block matrix, i.e., if $\mathbf{D}$ is invertible,

$$\det\left(\begin{smallmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{smallmatrix}\right) = \det(\mathbf{D})\det(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}) \tag{65}$$

and $\tilde{\mathbf{H}} = \mathrm{diag}\left(\left[P_1|h_2|^2, \frac{P_2|h_2|^2}{1+P_r|h_{12}|^2}, \cdots P_1|h_2|^2, \frac{P_2|h_2|^2}{1+P_r|h_{12}|^2}\right]\right)$.

Applying (57) and (59), Eq. (64) can be rewritten as (28). The proof is complete. ■

APPENDIX D
PROOF OF THEOREM 3

Proof: Applying (62), the combining throughput of User 2 for $N \to \infty$ is computed as

$$\begin{aligned} R_{2,\mathrm{asymp}}^{\mathrm{ANOMA}} =& \frac{1}{2}\log\left(\frac{P_1P_2|h_2|^4}{1+P_1|h_2|^2}z_1\right) \\ \overset{(a)}{=}& \frac{1}{2}\log\Bigg[\frac{1+P_r|h_{12}|^2}{2} + \frac{P_2|h_2|^2 + P_1P_2|h_2|^4Q}{2(1+P_1|h_2|^2)} \\ &+\frac{1}{2}\sqrt{\left(1+P_r|h_{12}|^2 + \frac{P_2|h_2|^2+P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\right)^2 - \left(\frac{P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\right)^2}\Bigg], \end{aligned}$$

where (a) is derived by replacing z_1 with its expression in (30). Since $Q \geq 0$,

$$\begin{aligned} &1+P_r|h_{12}|^2 + \frac{P_2|h_2|^2+P_1P_2|h_2|^4Q}{1+P_1|h_2|^2} \\ &\quad\geq \sqrt{\left(1+P_r|h_{12}|^2 + \frac{P_2|h_2|^2+P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\right)^2 - \left(\frac{P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\right)^2} \\ &\quad= \sqrt{\left(1+P_r|h_{12}|^2+\frac{P_2|h_2|^2}{1+P_1|h_2|^2}\right)^2 + \frac{2P_1P_2|h_2|^4Q}{1+P_1|h_2|^2}\left(1+P_r|h_{12}|^2+\frac{P_2|h_2|^2}{1+P_1|h_2|^2}\right)} \\ &\quad\geq 1+P_r|h_{12}|^2+\frac{P_2|h_2|^2}{1+P_1|h_2|^2}, \end{aligned}$$

where the equal signs are achieved if and only if $Q = 0$ which results in $\tau = 0$. The proof is complete. ■

REFERENCES

[1] *Study on Non-Orthogonal Multiple Access (NOMA) for NR.* TR 38.812 3GPP, Dec. 2018.

[2] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[3] J. N. Laneman, D. N. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[4] E. Koyuncu and H. Jafarkhani, "Distributed beamforming in wireless multiuser relay-interference networks with quantized feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4538–4576, Jul. 2012.

[5] Y. Jing and H. Jafarkhani, "Single and multiple relay selection schemes and their achievable diversity orders," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1414–1423, Mar. 2009.

[6] J.-B. Kim and I.-H. Lee, "Capacity analysis of cooperative relaying systems using non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1949–1952, Nov. 2015.

[7] J. Men and J. Ge, "Non-orthogonal multiple access for multiple-antenna relaying networks," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1686–1689, Oct. 2015.

[8] X. Liang, Y. Wu, D. W. K. Ng, Y. Zuo, S. Jin, and H. Zhu, "Outage performance for cooperative NOMA transmission with an AF relay," *IEEE Commun. Lett.*, vol. 21, no. 11, pp. 2428–2431, Nov. 2017.

[9] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[10] X. Yue, Y. Liu, S. Kang, A. Nallanathan, and Z. Ding, "Exploiting full/half-duplex user relaying in NOMA systems," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 560–575, Feb. 2018.

[11] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Full-duplex device-to-device-aided cooperative nonorthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4467–4471, May 2017.

[12] Z. Wei, X. Zhu, S. Sun, J. Wang, and L. Hanzo, "Energy efficient full-duplex cooperative non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 10 123–10 128, Oct. 2018.

[13] X. Zou, B. He, and H. Jafarkhani, "An analysis of two-user uplink asynchronous non-orthogonal multiple access systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1404–1418, Feb. 2019.

[14] J. Cui, G. Dong, S. Zhang, H. Li, and G. Feng, "Asynchronous NOMA for downlink transmissions," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 402–405, Oct. 2017.

[15] M. Ganji and H. Jafarkhani, "Time asynchronous NOMA for downlink transmission," in *Proc. IEEE WCNC*, Marrakech, Morocco, Apr. 2019, pp. 1–6.

[16] ——, "Improving NOMA multi-carrier systems with intentional frequency offsets," *to appear in IEEE Wireless Commun. Lett. [Early Access]*, Mar. 2019.

[17] S. Poorkasmaei and H. Jafarkhani, "Asynchronous orthogonal differential decoding for multiple access channels," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 481–493, Jan. 2015.

[18] S. Sodagari and H. Jafarkhani, "Enhanced spectrum sharing and cognitive radio using asynchronous primary and secondary users," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 832–835, Apr. 2018.

[19] X. Zhang, M. Ganji, and H. Jafarkhani, "Exploiting asynchronous signaling for multiuser cooperative networks with analog network coding," in *Proc. IEEE WCNC*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[20] X. Zhang and H. Jafarkhani, "Asynchronous network coding for multiuser cooperative communications," *IEEE Trans. Commun.*, vol. 16, no. 12, pp. 8250–8260, Dec. 2017.

[21] G. Liu, X. Chen, Z. Ding, Z. Ma, and F. R. Yu, "Hybrid half-duplex/full-duplex cooperative non-orthogonal multiple access with transmit power adaptation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 506–519, Jan. 2018.

[22] L. Lei, D. Yuan, and P. Värbrand, "On power minimization for non-orthogonal multiple access (NOMA)," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2458–2461, Dec. 2016.

[23] Y. Fu, Y. Chen, and C. W. Sung, "Distributed power control for the downlink of multi-cell NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6207–6220, Sep. 2017.

[24] Y. Jing and H. Jafarkhani, "Network beamforming using relays with perfect channel information," *IEEE Trans.*

Inf. Theory, vol. 55, no. 6, pp. 2499–2517, Jun. 2009.

[25] X. Chen, G. Liu, Z. Ma, F. R. Yu, and Z. Ding, "Power allocation for cooperative non-orthogonal multiple access systems," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1–6.

[26] N. Pappas, J. Jeon, D. Yuan, A. Traganitis, and A. Ephremides, "Wireless network-level partial relay cooperation: A stable throughput analysis," *Journal of Communications and Networks*, vol. 20, no. 1, pp. 93–101, Feb. 2018.

[27] Z. Ding, I. Krikidis, B. Rong, J. S. Thompson, C. Wang, and S. Yang, "On combating the half-duplex constraint in modern cooperative networks: Protocols and techniques," *IEEE Wireless Commun.*, vol. 19, no. 6, pp. 20–27, Dec. 2012.

[28] A. K. Sadek, K. R. Liu, and A. Ephremides, "Cognitive multiple access via cooperation: Protocol design and performance analysis," *IEEE Trans. Inf. Theory*, vol. 53, no. 10, pp. 3677–3696, Oct. 2007.

[29] N. Pappas, M. Kountouris, A. Ephremides, and A. Traganitis, "Relay-assisted multiple access with full-duplex multi-packet reception," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3544–3558, Jul. 2015.

[30] B. Rong, I. Krikidis, and A. Ephremides, "Network-level cooperation with enhancements based on the physical layer," in *Proc. IEEE ITW (ITW 2010, Cairo)*, Cairo, Egypt, Jul. 2010, pp. 1–5.

[31] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[32] X. Zou and H. Jafarkhani, "Asynchronous channel training in massive MIMO systems," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[33] X. Zou, B. He, and H. Jafarkhani, "On uplink asynchronous non-orthogonal multiple access systems with timing error," in *Proc. IEEE ICC*, Kansas City, MO, USA, May 2018, pp. 1–6.

[34] M. Ganji and H. Jafarkhani, "Interference mitigation using asynchronous transmission and sampling diversity," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[35] J. D. Gibson, *Mobile communications handbook*. Boca Raton, FL, USA: CRC press, 2012.

[36] A. M. Geoffrion, "Duality in nonlinear programming: a simplified applications-oriented development," *SIAM review*, vol. 13, no. 1, pp. 1–37, Jan. 1971.

[37] A. Masmoudi and T. Le-Ngoc, "Self-interference cancellation limits in full-duplex communication systems," in *Proc. IEEE GLOBECOM*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[38] D. Poole, *Linear algebra: A modern introduction*. Boston, MA, USA: Cengage Learning, 2014.