This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMC.2018.2812722, IEEE Transactions on Mobile Computing

Optimal Power Allocation and Scheduling for Non-Orthogonal Multiple Access Relay-Assisted Networks

Yuan Wu *Senior Member IEEE*, Li Ping Qian *Senior Member IEEE*, Haowei Mao, Xiaowei Yang, Haibo Zhou *Member IEEE*, Xuemin (Sherman) Shen *Fellow IEEE*

***Abstract*—The emerging non-orthogonal multiple access (NOMA), which enables mobile users (MUs) to share same frequency channel simultaneously, has been considered as a spectrum-efficient multiple access scheme to accommodate tremendous traffic growth in future cellular networks. In this paper, we investigate the NOMA downlink relay-transmission, in which the macro base station (BS) first uses NOMA to transmit to a group of relays, and all relays then use NOMA to transmit their respectively received data to an MU. In specific, we propose an optimal power allocation problem for the BS and relays to maximize the overall throughput delivered to the MU. Despite the non-convexity of the problem, we adopt the vertical decomposition and propose a layered-algorithm to efficiently compute the optimal power allocation solution. Numerical results show that the proposed NOMA relay-transmission can increase the throughput up to 30% compared with the conventional time division multiple access (TDMA) scheme, and we find that increasing the relays' power capacity can increase the throughput gain of the NOMA relay against the TDMA relay. Furthermore, to improve the throughput under weak channel power gains, we propose a hybrid NOMA (HB-NOMA) relay that adaptively exploits the benefit of NOMA relay and that of the interference-free TDMA relay. By using the throughput provided by the HB-NOMA relay for each individual MU, we study the multi-MUs scenario and investigate the multi-MUs scheduling problem over a long-term period to maximize the overall utility of all MUs. Numerical results demonstrate the performance advantage of the proposed multi-MUs scheduling that adopts the HB-NOMA relay-transmission.**

I. Introduction

The proliferation of smart and media-hungry mobile services have led to a rapid growth of mobile data traffic, which have increasingly overloaded radio access networks (RANs) and degraded mobile users' (MUs') quality of services (QoS). How to efficiently accommodate such a heavy traffic pressure has imposed a crucial challenge to cellular network operators. The emerging non-orthogonal multiple access (NOMA) [1], which enables a cluster of MUs to share a same frequency channel simultaneously and adopts the successive interference cancellation (SIC) to mitigate the co-channel the MUs' interference, has been considered as a very promising scheme to tackle the aforementioned challenge. Thanks to the potential advantage compared with conventional orthogonal multiple access (OMA), NOMA has been envisioned to play an important role in a variety of scenarios such as 5G Long-term evolution (LTE) [2]–[4], Internet of Things [6], and vehicular networks [7]. Meanwhile, lots of studied have been denoted to evaluating the performance of NOMA [8]–[11], and studying the resource management [12]–[16].

In addition to the advances in the multi-access scheme, current cellular RANs have been evolving towards a heterogeneous multi-tier architecture, in which dedicated relays are deployed to improve transmission quality and resource utilization. Transmission via relays avoids the resource-consuming long-distance transmission between the macro base station (BS) and the MU, and thus can effectively enhance the throughput by exploiting the close-proximity between the BS and the relays and that between the relays and the MU. In general, there exist two types of benefits brought by the relay-assisted transmission, namely, exploiting the *spatial multiplexing* and exploiting the *spatial diversity*. To exploit the spatial multiplexing, different spatially deployed relays are used to *forward* independent traffic flows, which thus improves the overall throughput [30]–[32]. In comparison, to exploit the spatial diversity, different relays are used to transmit the copies of a same traffic flow, which thus improves the quality of received signal (e.g., signal to interference plus noise ratio (SINR)) at the receiver [33], [34].

In this paper, we investigate the NOMA relay-transmission through a group of half-duplex relays, in which the BS first uses NOMA to transmit to a group of relays in the first portion of the time-slot, and all relays then use NOMA to forward their respectively received data to the MU in the remainder of the time-slot. We focus on the downlink transmission and exploit the benefit of spatial multiplexing, i.e., different relays forward independent traffic flows from the BS to the MU, with the objective of maximizing the aggregate throughput via all relays. However, several challenges need to be addressed. First, due to co-channel interference incurred by NOMA, the throughput carried by different BS-relay paths are coupled, and so are the throughput carried by different relay-BS paths. Moreover, from each relay's perspective, relaying traffic requires the achievable output-throughput no smaller than the input-throughput (otherwise, part of the input throughput will be lost). To address these challenges, the main contributions of this paper are summarized as follows.

- Exploiting relay-transmission through a group of dedicated relays, we propose an optimal power allocation for the BS and relays to maximize the overall throughput for an arbitrary MU. The co-channel interference among different BS-relay paths and that among different relay-MU paths lead to the power allocation problem strictly non-convex. To tackle with this issue, we exploit the decomposable structure of the problem and propose a layered-algorithm that efficiently computes the optimal power allocation solution. The layered-algorithm is comprised of solving the top-problem regarding the relays' throughput-capacity assignment and solving a series of parameterized subproblems regarding the BS's achievable SINR-assignment. Specifically, for each parameterized SINR-assignment subproblem, we identify its convexity and analytically characterize the optimal SINR-assignment. For the top-problem, we identify its hidden monotonicity and propose a poly-block approximation algorithm to compute the optimal relays' throughput-capacity assignment. The layered-algorithm computes the maximum throughput through the relays for the targeted MU.
- We provide extensive numerical results to evaluate the performance of the proposed NOMA relay-transmission. The results

Y. Wu, L. Qian, H. Mao, and X. Yang are with College of Information Engineering, Zhejiang University of Technology, Hangzhou, China (emails: iewuy@zjut.edu.cn, lpqian@zjut.edu.cn). L. Qian is the corresponding author.

H. Zhou and X. Shen is with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (email: h53zhou@uwaterloo.ca, xshen@bbcr.uwaterloo.ca).

show that the proposed NOMA relay-transmission can increase the throughput up to 30% compared with the conventional time division multiple access (TDMA) scheme. Meanwhile, we find that increasing the relays' power capacity can effectively increase the throughput gain of the NOMA relay against the TDMA relay, since a larger power capacity can provide a larger variation in different relay-MU paths, which yields a larger gain from executing the SIC in NOMA. Furthermore, to address the influence of weak channel power gains that impair the performance of SIC in NOMA, we propose a hybrid NOMA (HB-NOMA) relay. The HB-NOMA attains the benefit of NOMA relay (when executing the SIC is beneficial) and adaptively reaps the benefit of TDMA relay.

- Based on the optimal throughput provided by the HB-NOMA relay for each individual MU, we further extend to consider a scenario of multiple MUs and investigate the multi-MUs scheduling problem over a long-term period to maximize the overall utility of all MUs. To this end, we adopt the throughput provided by the HB-NOMA relay for each scheduled MU, and adopt the framework of the gradient-projection (GP) scheduling scheme which is able to yield the maximum overall utility for all MUs. Numerical results demonstrate that the GP-based scheduling can effectively increase the total utility compared with the conventional round-robin (RR) scheduling. Moreover, using the HB-NOMA relay for each scheduled MU can jointly reap the throughput gain provided by the NOMA relay and address the impact of weak channel power gain by adaptively exploiting the TDMA relay.

The remainder of this paper is organized as follows. We first review the related studies in Section II. We present the relay-transmission model and the throughput maximization formulation in Section III. We show the decomposition of the problem in Section IV, and propose the layered-algorithm to solve the problem in Section V. Numerical results for the proposed NOMA relay are presented in Section VI, in which we further propose the HB-NOMA relay scheme. We discuss the extension to the multi-MUs scheduling scheme with the HB-NOMA relay in Section VII. We conclude this work in Section VIII.

II. Related Studies

(Studies about power-domain NOMA): We mainly review the power-domain NOMA that utilizes superposition coding at the transmitter and successive interference cancellation at the receiver-side [5]. Many efforts have been devoted to analyzing the performance of NOMA under different scenarios. In [8], Ding *et. al.* analyzed the performance of downlink NOMA with randomly deployed MUs. In [9], Zhang *et. al.* analyzed the performance of both downlink and uplink NOMA for multi-cell systems. In [10], Liu *et. al.* analyzed the performance of the hybrid massive MIMO and NOMA heterogeneous networks. The impact of user-pairing on NOMA has been analyzed in [11]. In addition to performance analysis, mitigating the co-channel interference in NOMA transmission necessitates proper radio resource allocations. In [12] and [13], efficient power allocation algorithms have been proposed to optimize the energy-efficiency of NOMA-transmission. Joint optimization of users' transmit-power and sub-channel allocations have been investigated in [14], [15]. In [16], the joint optimization of uplink and downlink resource optimization, mode selection, and power allocation was studied for NOMA cellular systems with in-band full-duplex BSs.

(Studies focusing on analyzing NOMA relay): Assuming relatively fixed resource allocations, many research efforts have been devoted to analyzing the NOMA relay-transmission. Several studies have focused on the scenario of single-relay. In [17], Liang *et. al.* considered a scenario comprised of one BS, one relay, and two MUs (in which one MU has direct transmission to the BS, and another MU has no direct transmission to the BS and thus requires the relay's help), and analyzed the corresponding NOMA transmission outage performance. In [18], Men *et. al.* analyzed the transmission outage probability of NOMA-assisted relay transmission through one relay to a group of MUs (but without the direct path from the BS to the MUs), in which the BS and all MUs are equipped with multiple antennas. In [19], Kim *et. al.* considered a three-node scenario comprised the source, relay, and destination using NOMA and analyzed the corresponding transmission capacity. Performance of the NOMA based relaying under Rician channels has been investigated in [20]. In [21], a full-duplex (FD) cooperative NOMA system with dual users and a dedicated FD relay assisting the transmission to the weak user has been studied. In [22], Luo *et. al.* considered a buffer-aided NOMA relaying model. For the scenario of multiple relay, in [24], Ding *et. al.* analyzed the impact of the relay-selection on the performance NOMA relay-transmission for two MUs. In [23], Zhang *et. al.* proposed a novel non-regenerative massive-MIMO NOMA relay system model incorporating the MMSE-SIC decoding scheme and multiple sub-band frequency regime and investigated the consequent performance such as system capacity and sum-rate. In [25], a cooperative NOMA networks where multiple users transmit messages to two destinations by utilizing multiple amplify-and-forward relays has been studied.

(Studies about resource allocation for NOMA relay): Proper radio resource allocation plays a crucial role in reaping the benefits of NOMA relay. In [26], Zhang *et. al.* considered a single-relay scenario in which one relay forwards traffic for multiple pairs of source-destination, and proposed a joint allocation of transmit-power and sub-carrier allocation for the source-destination pairs using NOMA transmission. In [27], Liu *et. al.* proposed a collaborative NOMA assisted-relaying systems, in which one user is selected as a relay to relay traffic for other users with NOMA transmission. A power allocation scheme was also proposed to minimize the transmission outage probability. In [28], Xue *et. al.* proposed a joint power allocation and relay beam-forming problem for NOMA amplify-and-forward relay network to maximize the achievable rate of the destination which has the best channel condition. In [29], Zhang *et. al.* studied the resource allocation problem for a single-cell NOMA relay network where an OFDM amplify-and-forward relay allocates the spectrum and power resources to the source-destination pairs.

Our study in this paper belongs to the last group of related studies, In particular, we consider a downlink transmission from the BS to the MU via a group of dedicated half-duplex relays. As a key feature different from the existing studies, we consider that both the BS and all relays use NOMA to exploit spatial multiplexing, namely, the BS first uses NOMA to transmit independent traffic flow to a group of relays in the first portion of the time-slot, and all relays then use NOMA to forward their respectively received traffic flow to the MU in the remainder of the time-slot. Our objective is to maximize the total throughput from the BS to the MU via all relays.

III. System Model and Problem Formulation

A. System Model and Relay Protocol with Half-Duplex

(System model): We consider a system model as follows. There exists a group of I relays that relay traffic for a group of MUs $\mathcal{K} = \{1, 2, ..., K\}$ at the edge of the cell. We assume that there exists no direct-path from the BS to the MU, and all relay operate

in a half-duplex manner to forward traffic from the BS to the targeted MU. Figure 1 plots an example of the system model comprised of three relays (please note that we will explain how we index the relays and the associated channel power gains soon in the next subsection).

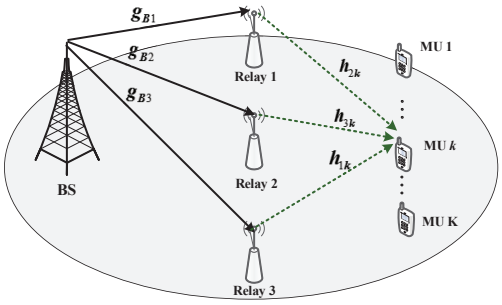


Fig. 1: System model comprised of three relays who relay traffic to MU k. Parameter g_{Bi} denotes the channel power gain from the BS to relay i, and parameter h_{mk} denotes the channel power gain from relay m to MU k. We will explain how we index the subscribes i and m soon in the next subsection.

(Relay-transmission protocol for an arbitrary MU): Corresponding to Figure 1, Figure 2 shows the relay-based transmission protocol for one MU within one time-slot. Without loss of generality, we consider that the time-slot is of one-unit length, and it includes two phases, i.e., Phase-I of duration $\theta \in (0,1)$ and Phase-II of duration $1-\theta$. In Phase-I, the BS uses NOMA to transmit to all relays. Then, in Phase-II, all relays then use NOMA to relay their respectively received traffic to MU k. We consider that the BS transmits different data flows to the relays, who then relay their respectively received to MU k. Hence, MU k's throughput is equal to the sum of the throughput from the BS through all relays.

In Sections III-VI, we focus on studying the maximum throughput via all relay to one targeted MU (i.e., MU k) within one time-slot. Using this maximum throughput for each MU, we will extend to investigate the transmission scheduling for a group of MUs over a large number of time-slots in Section VII.

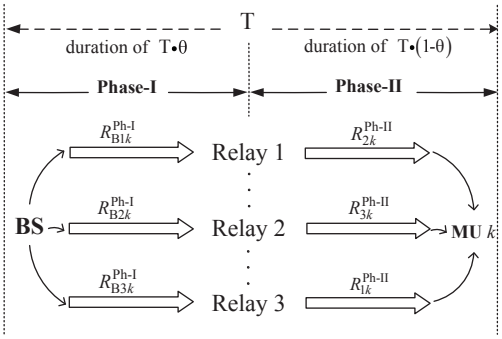


Fig. 2: Illustration of relay-transmission for MU k within a time-slot. Phase-I: transmission from the BS to all relays (i.e., BS-relay paths) with duration $T\theta$. Phase-II: transmission from all relays to MU k (i.e., relay-MU paths) with duration $T(1-\theta)$. We assume T of one-unit in the remainder of this paper.

B. NOMA Relay-Transmission

Due to the operations of successive interference cancellation in NOMA-transmission [2-5], the data reception at the relays requires to order the channel power gains from the BS to all relays. Similarly, the data reception at MU k requires to order the channel power gains from all relays to MU k. We thus use index-set $\mathcal{I}_{\text{B}}$ for ordering the channel gains from BS to all relays, and use index $\mathcal{I}_{\text{MU}k}$ for ordering the channel gains from all relays to MU k.

(BS's transmission to all relays in Phase-I) In Phase-I, for transmission from the BS to all relays, we introduce the index-set $\mathcal{I}_{\text{B}} = \{1, 2, ..., i, j, ..., I\}$ which is given by:

$$g_{B1} > g_{B2} > ... > g_{Bi} > g_{Bj} > ... > g_{BI}, \tag{1}$$

where g_{Bi} denotes the channel power gain from the BS to relay $i \in \mathcal{I}_{\text{B}}$ (i.e., the i-th relay in $\mathcal{I}_{\text{B}}$). In the remainder of the paper, we use index i (or j) to denote relay $i \in \mathcal{I}_{\text{B}}$ (or relay $j \in \mathcal{I}_{\text{B}}$) in the ordering (1). Moreover, we consider that all relays have been indexed according to (1) as shown in Figure 1. We assume that the BS knows the channel state information to all relay, and all relays know their respective channel state information to the MU.

In Phase-I, the BS uses NOMA to transmit to all relays (recall that the relays will relay these traffic to MU k in Phase-II). Let p_{Bik} denote the BS's transmit-power to relay i. Thus, according to [2]–[5], the operations of NOMA lead to the achievable data rate from the BS to relay $i \in \mathcal{I}_{\text{B}}$ (for data-delivery to MU k) as follows:

$$R^{\text{Ph-I}}_{Bik} = W_B \log_2 \left(1 + \frac{g_{Bi} p_{Bik}}{g_{Bi} \sum_{j=1}^{i-1} p_{Bjk} + n_B}\right), \forall i \in \mathcal{I}_{\text{B}}, \tag{2}$$

where W_B denotes the BS's channel bandwidth, and n_B denotes the power of the background noise at all relays. For simplicity, we introduce β_{Bik} to denote the received SINR of the transmission from the BS to relay $i \in \mathcal{I}_{\text{B}}$, i.e.,

$$\beta_{Bik} = \frac{g_{Bi} p_{Bik}}{g_{Bi} \sum_{j=1}^{i-1} p_{Bjk} + n_B}, \forall i \in \mathcal{I}_{\text{B}}. \tag{3}$$

(Relays' transmissions to MU k in Phase-II): In Phase-II, for transmissions from all relays to MU k, we introduce the index-set $\mathcal{I}_{\text{MU}k} = \{1, 2, ..., m, n, ..., I\}$ which is given by:

$$h_{1k} > h_{2k} > ... > h_{mk} > h_{nk} ... > h_{Ik}, \tag{4}$$

where h_{mk} denotes the channel power gain from relay $m \in \mathcal{I}_{\text{MU}k}$ to MU k[1]. In the remainder of the paper, we use index m (or n) to denote relay $m \in \mathcal{I}_{\text{MU}k}$ (or relay $n \in \mathcal{I}_{\text{MU}k}$) in the ordering (4).

In Phase-II, all relays use NOMA and relay their respectively received data to MU k. Let q_{mk} denote relay m's transmit-power for MU k. Due to NOMA, the achievable data rate from relay m to MU k can be given by:

$$R^{\text{Ph-II}}_{mk} = W_S \log_2 \left(1 + \frac{q_{mk} h_{mk}}{\sum_{n=1}^{m-1} q_{nk} h_{nk} + n_0}\right), \forall m \in \mathcal{I}_{\text{MU}k}, \tag{5}$$

where W_S denotes the bandwidth used for the relays, and n_0 denotes the power of the background noise at the MU. Moreover, we introduce γ_{mk} to denote the received SINR of the transmission from relay $m \in \mathcal{I}_{\text{MU}k}$ to MU k, i.e.,

$$\gamma_{mk} = \frac{q_{mk} h_{mk}}{\sum_{n=1}^{m-1} q_{nk} h_{nk} + n_0}, \forall m \in \mathcal{I}_{\text{MU}k}. \tag{6}$$

Remark 1: (Connection between $\mathcal{I}_B$ and $\mathcal{I}_{MUk}$) Notice that both $\mathcal{I}_{\text{B}}$ (given by (1)) and $\mathcal{I}_{\text{MU}k}$ (given by (4)) denote the same group of relays. In other words, each relay has an index-tuple (i, m), i.e., the relay is the i-th relay (or relay i) in $\in \mathcal{I}_{\text{B}}$ from the perspective of the BS's transmission, and it is also the m-th relay (or relay m) in $\mathcal{I}_{\text{MU}k}$ from the perspective of the relays' transmission to MU k. To present such a connection, we introduce a mapping $\Phi_k(i)$ that finds the index

[1]Notice that for two different targeted MU k and MU k', the associated $\mathcal{I}_{\text{MU}k}$ and $\mathcal{I}_{\text{MU}k'}$ might be different, which thus requires us to use MUk in the subscript of $\mathcal{I}_{\text{MU}k}$.

in $\mathcal{I}_{\text{MU}k}$ for relay $i \in \mathcal{I}_B$, i.e., $m = \Phi_k(i)$ (in particular, due to the randomness in the channel power gains, we consider that $\Phi_k(i)$ is an one-to-one mapping between $\mathcal{I}_B$ and $\mathcal{I}_{\text{MU}k}$). The mapping $\Phi_k(i)$ will be used in our following problem formulation (i.e., constraint (8)). In this paper, we consider that the BS knows the instantaneous channel power gains $\{g_{Bi}\}_{i\in\mathcal{I}_\text{B}}$ and $\{h_{mk}\}_{m\in\mathcal{I}_{\text{MU}k}}$ (e.g., based on the relays' reports), and the BS can obtain the mapping $\Phi_k(i), \forall i \in \mathcal{I}_\text{B}$. Figure 1 shows an illustrative example of three relays, with $\Phi_k(1) = 2$, $\Phi_k(2) = 3$, and $\Phi_k(3) = 1$.

In addition to eqs. (2) and (5), we can also express the considered relay-transmission as follows. Let $\boldsymbol{s}_\text{B} = (s_1, s_2, ..., s_i, ...s_I)^\dagger$ (where $\dagger$ means the transpose-operation) denote the symbols transmitted from the BS to the relays in $\mathcal{I}_\text{B}$, with s_i denoting the symbol transmitted from the BS to rely i. Then, the received symbols at the relays can be expressed as $\boldsymbol{s}_\text{R} = \boldsymbol{G}_{\text{B-R}}\boldsymbol{s}_\text{B} + \boldsymbol{n}$, where $\boldsymbol{G}_{\text{B-R}}$ denotes the I-by-I matrix of the channel power gains from the BS to all relays. Meanwhile, let $\tilde{\boldsymbol{s}}_\text{R} = (\tilde{s}_1, \tilde{s}_2, ...\tilde{s}_m, ..., \tilde{s}_I)^\dagger$ denote the symbols transmitted from the respective relays (in $\mathcal{I}_{\text{MU}k}$) to the targeted MU k, with $\tilde{s}_m$ denoting the symbol transmitted from relay $m \in \mathcal{I}_{\text{MU}k}$ to MU k. Then, the received symbols at MU k can be expressed as: $\tilde{\boldsymbol{s}}_{\text{MU}-k} = \boldsymbol{H}_{\text{R-MU}k}\tilde{\boldsymbol{s}}_\text{R} + \boldsymbol{n}$, where $\boldsymbol{H}_{\text{R-MU}k}$ denotes the I-by-I matrix of the channel power gains from all relays to MU k.

C. Problem Formulation for Single-MU Throughput Maximization

Considering the data rate (2) with duration θ, the throughput from the BS to relay i is given by $\theta R^{\text{Ph-I}}_{Bik}$. Similarly, considering the data rate (5) with duration $1-\theta$, the throughput from relay $m = \Phi_k(i)$ to MU k is given by $(1-\theta)R^{\text{Ph-II}}_{\Phi_k(i)k}$. Thus, we formulate the following optimization that aims at maximizing the effective throughput from the BS to MU k via all relays (here, we use (TTM-k) to denote "Total Throughput Maximization for MU k")[2]:

$$\textbf{(TTM-}k\textbf{):} \quad V^*_{k,\text{NOMA}} = \max \sum_{i\in\mathcal{I}_\text{B}} \theta R^{\text{Ph-I}}_{Bik} \tag{7}$$

$$\text{subject to:} \quad \theta R^{\text{Ph-I}}_{Bik} \le (1-\theta) R^{\text{Ph-II}}_{\Phi_k(i)k}, \forall i \in \mathcal{I}_\text{B}, \tag{8}$$

$$\theta \sum_{i\in\mathcal{I}_\text{B}} p_{Bik} \le P^{\text{tot}}_B, \tag{9}$$

$$(1-\theta) q_{mk} \le Q^{\text{tot}}_m, \forall m \in \mathcal{I}_{\text{MU}k}, \tag{10}$$

$$\text{variables:} \quad p_{Bik} \ge 0, \forall i \in \mathcal{I}_\text{B},\ q_{mk} \ge 0, \forall m \in \mathcal{I}_{\text{MU}k}.$$

Constraint (8) means that from the perspective of each relay i's (i.e., equivalently relay $\Phi_k(i)$), the throughput from the BS to relay i should be no greater than the throughput from relay $\Phi_k(i)$ to MU k, i.e., each relay's output throughput should be no smaller than its input-throughput. Constraint (9) means that BS's the total transmit-power cannot exceed its power capacity P^{tot}_B. Notice that for two different MU k and MU k', the correspondingly two optimal solutions $\{p_{Bik}\}_{\forall i\in\mathcal{I}_\text{B}}$ and $\{p_{Bik'}\}_{\forall i\in\mathcal{I}_\text{B}}$ are different in general. Thus, in constraint (9), we include the subscript k in the BS's transmit-powers. Constraint (10) means that the transmit-power of relay $m \in \mathcal{M}$ cannot exceed its power capacity Q^{tot}_m.

IV. Problem Transformation and Decomposition

Directly solving Problem (TTM-k) is very challenging due to the non-convexity of Problem (TTM-k). To efficiently solve Problem (TTM-k), we first characterize two important properties regarding the BS's transmit-power and all relays' transmit-powers in Section IV-A. With these properties, we transform Problem (TTM-k) into an equivalent SINR-assignment in Section IV-B, and propose a decomposition of the SINR-assignment problem in Section IV-C.

A. Important Properties

We firstly analyze the BS's transmission to all relays in Phase-I. Suppose that $\{\beta_{Bik}\}_{i\in\mathcal{I}_\text{B}}$ is given in advance. With (3), we can use the following iterative calculation to compute the BS's minimum required transmit-power to achieve β_{Bik} at relay $i \in \mathcal{I}_\text{B}$:

$$p_{Bik} = \beta_{Bik} \sum_{j=1}^{i-1} p_{Bjk} + \beta_{Bik}\frac{n_B}{g_{Bi}}, \forall i \in \mathcal{I}_\text{B}. \tag{11}$$

By using (11), we can have the following important results:

Proposition 1: (Proposition 1 in [35]) Given the profile of SINR $\{\beta_{Bik}\}_{i\in\mathcal{I}_\text{B}}$ required to achieve at the relays $\mathcal{I}_\text{B}$, the BS's total minimum transmit-power can be given by

$$p^{\text{tot,min}}_{Bk}\left(\{\beta_{Bik}\}_{i\in\mathcal{I}}\right) = \sum_{i=1}^{I}\left(\frac{n_B}{g_{Bi}} - \frac{n_B}{g_{Bi-1}}\right)\prod_{j=i}^{I}(1+\beta_{Bjk}) - \frac{n_B}{g_{BI}}, \tag{12}$$

where we set g_{B0k} sufficiently large such that $\frac{n_B}{g_{B0k}} = 0$.

For the relays' transmissions in Phase-II, with (6), we can use the following iterative calculation to compute relay m's ($m \in \mathcal{I}_{\text{MU}k}$) minimum required transmit-power q_{mk} to achieve γ_{mk} at MU k:

$$q_{mk} = \frac{\gamma_{mk}}{h_{mk}} \sum_{n=1}^{m-1} q_{nk}h_{nk} + \gamma_{mk}\frac{n_0}{h_{mk}}, \forall m \in \mathcal{I}_{\text{MU}k}. \tag{13}$$

As shown in (13), only relay $n \le m, n \in \mathcal{I}_{\text{MU}k}$ will influence the transmit-power of relay $m \in \mathcal{I}_{\text{MU}k}$. Specifically, with (13), we can have the following important result.

Proposition 2: Given the profile of SINR $\{\gamma_{nk}\}_{n\le m}$ to achieve at MU k, each relay m's required minimum transmit-power to MU k can be given by

$$q^{\min}_{mk}\left(\{\gamma_{nk}\}_{n\le m}\right) = \frac{n_0}{h_{mk}}\gamma_{mk}\prod_{n=1}^{m-1}\left(1+\gamma_{nk}\right), \forall m \in \mathcal{I}_{\text{MU}k}. \tag{14}$$

Proof: The proof is essentially based on deduction. Specifically, it can be verified that (14) holds for relay 1 in $\mathcal{I}_{\text{MU}k}$. Then, supposing that (14) holds for relay $m \in \mathcal{I}_{\text{MU}k}$, our objective is to prove the following equation holds:

$$q^{\min}_{m+1,k} = \frac{n_0}{h_{m+1,k}}\gamma_{m+1,k}\prod_{n=1}^{m}(1+\gamma_{n,k}). \tag{15}$$

To prove (15), we substitute (14) into (13) and obtain

$$q^{\min}_{m+1,k} = \frac{n_0}{h_{m+1,k}}\gamma_{m+1,k}\left(1 + \sum_{n=1}^{m}\gamma_{n,k}\prod_{l=1}^{n-1}(1+\gamma_{l,k})\right), \tag{16}$$

which means that we just need to prove that the following equation always holds

$$\prod_{n=1}^{m}(1+\gamma_{n,k}) = 1 + \sum_{n=1}^{m}\gamma_{n,k}\prod_{l=1}^{n-1}(1+\gamma_{l,k}). \tag{17}$$

Again, we prove (17) based on deduction. Specifically, it is easy to show that (17) holds for $m = 1$. Next, supposing that (17), we need to prove the following equation holds:

$$\prod_{n=1}^{m+1}(1+\gamma_{n,k}) = 1 + \sum_{n=1}^{m+1}\gamma_{n,k}\prod_{l=1}^{n-1}(1+\gamma_{l,k}). \tag{18}$$

[2]We emphasize that although we consider that the value of θ is fixed, the proposed NOMA relay-transmission can be further extended to adjust θ by treating the maximum throughput $V^*_{k,\text{NOMA}}$ as an implicit function of θ.

To prove (18), we show the following derivation:

$$\begin{aligned}\prod_{n=1}^{m+1}(1+\gamma_{n,k}) &= (1+\gamma_{m+1,k})\prod_{n=1}^{m}(1+\gamma_{n,k})\\ &= \gamma_{m+1,k}\prod_{n=1}^{m}(1+\gamma_{n,k}) + \prod_{n=1}^{m}(1+\gamma_{n,k})\\ &= \gamma_{m+1,k}\prod_{n=1}^{m}(1+\gamma_{n,k}) + \sum_{n=1}^{m}\gamma_{n,k}\prod_{l=1}^{n-1}(1+\gamma_{l,k}) + 1\\ &= 1+\sum_{n=1}^{m+1}\gamma_{n,k}\prod_{l=1}^{n-1}(1+\gamma_{l,k}),\end{aligned}$$

which finishes the proof of (18). We thus finish the proof of (17), which consequently ensures the consistence between (15) and (16). We therefore finish proving (14). ■

B. Equivalent Form of Problem (TTM-k) as an SINR-assignment Problem

By using Propositions 1 and 2, we can use the introduced $\{\beta_{Bik}\}_{i\in\mathcal{I}_\text{B}}$ and $\{\gamma_{mk}\}_{m\in\mathcal{I}_{\text{MU}k}}$ as the decision variables and transform Problem (TTM) into the following SINR-assignment problem:

$$\textbf{(TTM-}k\textbf{-E):}\ \max\sum_{i=1}^{I}\theta W_B\log_2(1+\beta_{Bik})$$

subject to:

$$\sum_{i=1}^{I}\Big(\frac{n_B}{g_{Bi}}-\frac{n_B}{g_{Bi-1}}\Big)\prod_{j=i}^{I}(1+\beta_{Bjk})-\frac{n_0}{g_{BI}}\le\frac{P_B^\text{tot}}{\theta}, \tag{19}$$

$$\frac{n_0}{h_{mk}}\gamma_{mk}\prod_{n=1}^{m-1}(1+\gamma_{nk})\le\frac{Q_m^\text{tot}}{1-\theta},\forall m\in\mathcal{I}_{\text{MU}k}, \tag{20}$$

$$\theta W_B\log_2(1+\beta_{Bik})\le(1-\theta)W_S\log_2(1+\gamma_{\Phi_k(i)k}),\quad\forall i\in\mathcal{I}_\text{B}, \tag{21}$$

variables: $\beta_{Bik}\ge 0,\forall i\in\mathcal{I}_\text{B}$, and $\gamma_{mk}\ge 0,\forall m\in\mathcal{I}_{\text{MU}k}$.

Constraint (19) is from constraint (9), and constraint (20) is from constraint (10). Constraint (21) is from constraint (8).

(Connection between the optimal solution of Problem (TTM-k-E) and Problem (TTM-k)): Let $\{\beta_{Bik}^*\}_{i\in\mathcal{I}_\text{B}}$ and $\{\gamma_{mk}^*\}_{m\in\mathcal{I}_{\text{MU}k}}$ denote the optimal solution of Problem (TTM-k-E). Then, with (11) and (13), the optimal power allocations $\{p_{Bik}^*\}_{i\in\mathcal{I}_\text{B}}$ and $\{q_{mk}^*\}_{m\in\mathcal{I}_{\text{MU}k}}$ for Problem (TTM-k) can be given by the following iterative calculations:

$$p_{Bik}^*=\beta_{Bik}^*\sum_{j=1}^{i-1}p_{Bjk}^*+\beta_{Bik}^*\frac{n_B}{g_{Bi}},\forall i\in\mathcal{I}_\text{B}. \tag{22}$$

$$q_{mk}^*=\frac{\gamma_{mk}^*}{h_{mk}}\sum_{n=1}^{m-1}q_{nk}^*h_{nk}+\gamma_{mk}^*\frac{n_0}{h_{mk}},\forall m\in\mathcal{I}_{\text{MU}k}. \tag{23}$$

C. A Vertical Decomposition of Problem (TTM-k-E)

However, Problem (TTM-k-E) is still a non-convex optimization problem which is difficult to solve. To solve it efficiently, we need to explore its special property. To this end, we introduce a set of auxiliary variables $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$, whose purpose is to separate constraint (21) as follows:

$$\theta W_B\log_2(1+\beta_{Bik})\le z_{ik}\le(1-\theta)W_S\log_2(1+\gamma_{\Phi_k(i)k}),\quad\forall i\in\mathcal{I}_\text{B}. \tag{24}$$

In particular, the introduced $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ can be regarded as the relays' throughput-capacities.

With $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ and constraint (24), we propose a vertical decomposition of Problem (TTM-k-E) into a subproblem to optimize $\{\beta_{Bik}\}_{i\in\mathcal{I}_\text{B}}$ under given $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$, and a top-problem to further optimize $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$. The detailed decomposition is as follows.

(Subproblem to optimize $\{\beta_{Bik}\}_{i\in\mathcal{I}_\text{B}}$): Suppose that $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ is given in advance. Then, Problem (TTM-k-E) under the given $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ turns into the following subproblem to optimize $\{\beta_{Bik}\}_{i\in\mathcal{I}_\text{B}}$, i.e., the BS's SINR-assignment:

$$\textbf{(Sub-P):}F(\{z_{ik}\}_{i\in\mathcal{I}})=\max\sum_{i=1}^{I}\theta W_B\log_2(1+\beta_{Bik})$$

subject to:

$$\sum_{i=1}^{I}\Big(\frac{n_B}{g_{Bi}}-\frac{n_B}{g_{Bi-1}}\Big)\prod_{j=i}^{I}(1+\beta_{Bjk})-\frac{n_0}{g_{BI}}\le\frac{P_B^\text{tot}}{\theta}, \tag{25}$$

$$\theta W_B\log_2(1+\beta_{Bik})\le z_{ik},\forall i\in\mathcal{I}_\text{B}, \tag{26}$$

variables: $\beta_{Bik}\ge 0,\forall i\in\mathcal{I}_\text{B}$.

Problem (Sub-P) can be considered as the BS's optimal SINR-assignment problem to maximize the total throughput from the BS to all relays, under the given relays' throughput-capacities $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$. Notice that we use $F(\{z_{ik}\}_{i\in\mathcal{I}_\text{B}})$ to denote the optimal value of the objective function, which depends on the given $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$, and will be used in the following top-problem.

(Top-problem to optimize $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$): By treating $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ as a controllable vector and using (20) and (24), we can derive the feasible region (which is denoted by Ω_k) for $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ in eq. (28) on the top of the next page. Notice that in Ω_k, for each $m\in\mathcal{I}_{\text{MU}k}$, we have set its corresponding γ_{mk} as:

$$\gamma_{mk}=2^{\frac{z_{ik}}{(1-\theta)W_S}}-1,\ \text{when}\ m=\Phi_k(i),\forall m\in\mathcal{I}_{\text{MU}k}, \tag{27}$$

which is obtained by setting the second inequality in (24) strictly binding. Notice that with the one-to-one mapping $\Phi_k(i)$ that connects $\mathcal{I}_\text{B}$ and $\mathcal{I}_{\text{MU}k}$, there exists an one-to-one connection between $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ and $\{\gamma_{mk}\}_{\forall m\in\mathcal{I}_{\text{MU}k}}$, as shown in (27). This favorable property enables us to only treat $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ as a controllable variable in Ω_k, which eases our following presentation.

Further with $F(\{z_{ik}\}_{i\in\mathcal{I}})$ output by Problem (Sub-P), we consider a top-problem to optimize $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ as follows:

$$\begin{aligned}&\textbf{(Top-P):}\quad&&\max F(\{z_{ik}\}_{i\in\mathcal{I}_\text{B}})\\ &\text{subject to:}&&\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}\in\Omega_k,\\ &\text{variables:}&&z_{ik},\forall i\in\mathcal{I}_\text{B}.\end{aligned}$$

Problem (Top-P) corresponds to properly assigning all relays' throughput-capacities $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$ within the feasible region Ω_k (given in (28)) to maximize the overall relay-assisted throughput from the BS to MU k over all relays.

Figure 3 shows how we equivalently transform the original Problem (TTM-k) and decompose it into Problem (Top-P) and Problem (Sub-P). In particular, Problem (Sub-P-E) is another equivalent form of Problem (Sub-P-E) which we will explain in Section V-A.

(Rationale for the proposed vertical decomposition): The key reason for proposing the above decomposition is to efficiently solve Problem (TTM-k-E). As we will show soon in the next section, for Problem (Sub-P), we can identify its convexity and characterize the structural property of the corresponding optimal solution, i.e., the BS's optimal SINR-assignment under the given $\{z_{ik}\}_{i\in\mathcal{I}}$. Exploiting this important structural property, we can propose an efficient algorithm to find the BS's optimal SINR-assignment for

$$\Omega_k = \left\{ \{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}} | \gamma_{mk} = 2^{\frac{z_{ik}}{(1-\theta)W_S}} - 1, \text{ when } m = \Phi_k(i), \forall m \in \mathcal{I}_{\mathrm{MU}k}, \text{ and } \frac{n_0}{h_{mk}}\gamma_{mk}\prod_{n=1}^{m-1}(1+\gamma_{nk}) \le \frac{Q_m^{\mathrm{tot}}}{1-\theta}, \forall m \in \mathcal{I}_{\mathrm{MU}k} \right\}. \quad (28)$$

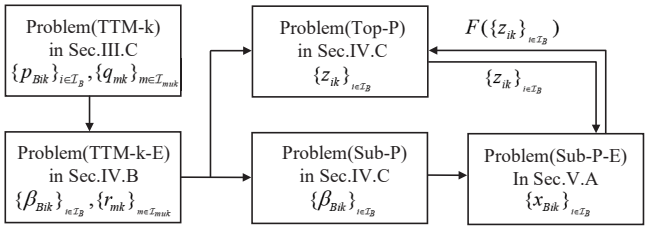


Fig. 3: Decomposition of Problem (TTM-k) into Problem (Top-P) and Problem (Sub-P).

the given $\{z_{ik}\}_{i\in\mathcal{I}}$ (the details are presented in Section V-A). With the output of Problem (Sub-P), i.e., $F(\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}})$, we then continue to optimize the relays' throughput-capacities $\{z_{ik}\}_{i\in\mathcal{I}}$ in Problem (Top-P). As we will discuss in Section V-B, although we cannot analytically derive $F(\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}})$, we identify the hidden monotonicity of Problem (Top-P) which enables us to use a very efficient global-search scheme to find the optimal relays' throughput-capacities to maximize the overall relay-assisted throughput via all relays.

Notice that solving the top problem (Top-P) will complete solving the original Problem (TTM-k). Specifically, let $\{z_{ik}^*\}$ denote the optimal solution of Problem (Top-P). We can obtain the optimal solution of Problem (TTM-k) as follows.

Proposition 3: Given $\{z_{ik}^*\}_{i\in\mathcal{I}_\mathrm{B}}$ as the optimal solution of Problem (Top-P), the optimal solution of Problem (TTM-k-E) can be given as follows. First, $\{\gamma_{mk}^*\}_{m\in\mathcal{I}_{\mathrm{MU}k}}$ can be given by:

$$\gamma_{mk}^* = 2^{\frac{z_{ik}^*}{(1-\theta)W_s}} - 1, \text{ with } m = \Phi_k(i), \forall m \in \mathcal{I}_{\mathrm{MU}k}. \quad (29)$$

Meanwhile, $\{\beta_{Bik}^*\}_{i\in\mathcal{I}_\mathrm{B}}$ is given by solving:

$$\begin{aligned} \{\beta_{Bik}^*\}_{i\in\mathcal{I}_\mathrm{B}} = \quad & \arg\max \sum_{i=1}^{I} \theta W_B \log_2\left(1+\beta_{Bik}\right) \\ \text{subject to:} \quad & \text{constraint } (25), \\ \text{variables:} \quad & 0 \le \beta_{Bik} \le 2^{\frac{z_{ik}^*}{\theta W_B}} - 1, \forall i \in \mathcal{I}_\mathrm{B}. \end{aligned} \quad (30)$$

Proof: The proof is based on the rationale of vertical decomposition. Notice that the objective function of Problem (Sub-P) is same as that of Problem (TTM-k-E). Thus, given the assigned relays' capacity-throughput $\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}}$ within Ω_k (in (28)), $F(\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}})$ corresponds to the *maximum* achievable throughput from the BS to all relays. Hence, by further searching $\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}}$ within Ω_k, we can find the globally maximum of the objective function $F(\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}})$, i.e., the maximum relay-assisted throughput from the BS to the targeted MU k, which completes solving Problem (TTM-k-E). ■

V. Proposed Algorithms to Solve Top-problem and Sub-problem

Based on the proposed decomposition in Section IV-C, we propose algorithms to solve the sub-problem and the top-problem. Specifically, in Section V-A, we firstly analyze Problem (Sub-P) and propose an algorithm (i.e., SubSol-Algorithm) to find the BS's optimal SINR-assignment under the given $\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}}$. Then, in Section V-B, we further analyze Problem (Top-P). By using SubSol-Algorithm as a subroutine, we propose an algorithm (i.e., TopSol-Algorithm) to solve Problem (Top-P), which thus completes solving the very original Problem (TTM-k). Figure 4 shows the connections between our proposed two algorithms.

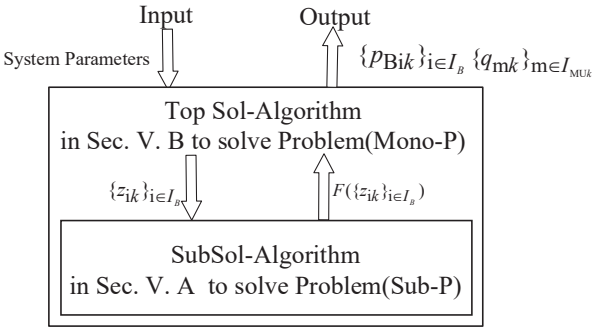


Fig. 4: Connection between TopSol-Algorithm and SubSol-Algorithm

A. Proposed Algorithm to Solve Sub-problem

We first propose an algorithm to solve Problem (Sub-P) (under given $\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}}$), and evaluate the corresponding value $F(\{z_{ik}\}_{i\in\mathcal{I}_\mathrm{B}})$. We introduce the following change of variables:

$$x_{Bik} = \log_2(1+\beta_{Bik}), \forall i \in \mathcal{I}_\mathrm{B}. \quad (31)$$

With (31), we replace β_{Bik} with x_{Bik} in Problem (Sub-P), which turns into the following equivalent form:

$$\textbf{(Sub-P-E):} F(\{z_{ik}\}_{i\in\mathcal{I}}) = \max \theta W_B \sum_{i=1}^{I} x_{Bik}$$

$$\text{subject to: } \sum_{i=1}^{I}\left(\frac{n_B}{g_{Bi}} - \frac{n_B}{g_{Bi-1}}\right)2^{\sum_{j=i}^{I} x_{Bjk}} - \frac{n_0}{g_{BI}} \le \frac{P_B^{\mathrm{tot}}}{\theta}, \quad (32)$$

$$\text{variables: } 0 \le x_{Bik} \le \frac{z_{ik}}{\theta W_B}, \forall i \in \mathcal{I}_\mathrm{B}. \quad (33)$$

Recall that we have shown Problem (Sub-P-E) in Figure 3 before. We identify the following important property for Problem (Sub-P-E):

Proposition 4: *(Convexity of Subproblem)*: Problem (Sub-P-E) is a convex optimization with respect to $\{x_{Bik}\}_{i\in\mathcal{I}}$.

Proof: The objective function is linear with respect to $\{x_{Bik}\}_{i\in\mathcal{I}}$, and constraint (32) is convex with respect to $\{x_{Bik}\}_{i\in\mathcal{I}}$. Thus, Problem (Sub-P-E) is a convex optimization problem [41]. ■

Proposition 4 enables us to efficiently solve Problem (Sub-P-E). For the sake of clear presentation, we introduce the following function

$$G_k(\{x_{Bik}\}_{i\in\mathcal{I}_\mathrm{B}}) = \sum_{i=1}^{I}\left(\frac{n_B}{g_{Bi}} - \frac{n_B}{g_{Bi-1}}\right)2^{\sum_{j=i}^{I} x_{Bjk}}. \quad (34)$$

We identify the following structural property of the optimal solution of Problem (Sub-P-E).

Proposition 5: *(Structural property of the optimal solution of Problem (Sub-P-E))*: The optimal solution $\{x_{Bik}^*\}_{i\in\mathcal{I}_\mathrm{B}}$ of Problem (Sub-P-E) can be given by one of the following two cases.

- (Case-I): if the following condition

$$G_k(\{\frac{z_{ik}}{\theta W_B}\}_{i\in\mathcal{I}_\mathrm{B}}) - \frac{n_0}{g_{BI}} \le \frac{P_B^{\mathrm{tot}}}{\theta} \quad (35)$$

is met, then we have $x^*_{Bik} = \frac{z_{ik}}{\theta W_B}, \forall i \in \mathcal{I}_\text{B}$.

- (Case-II): if the following condition

$$G_k(\{\frac{z_{ik}}{\theta W_B}\}_{i\in\mathcal{I}_\text{B}}) - \frac{n_0}{g_{BI}} > \frac{P_B^\text{tot}}{\theta} \tag{36}$$

is met, then there exists a unique index $r \in \mathcal{I}_\text{B}$. With r, we have

$$x^*_{Bik} = \frac{z_{ik}}{\theta W_B}, \text{ for each } i < r \text{ and } i \in \mathcal{I}_\text{B}, \tag{37}$$

$$x^*_{Bik} = 0, \text{ for each } i > r \text{ and } i \in \mathcal{I}_\text{B}. \tag{38}$$

Meanwhile, the value of x^*_{Brk} is uniquely determined by the following equation:

$$G_k(\{x^*_{Bik}\}_{i<r,i\in\mathcal{I}_\text{B}} \cup \{x^*_{Brk}\} \cup \{x^*_{Bik}\}_{i>r,i\in\mathcal{I}_\text{B}}) - \frac{n_0}{g_{BI}} = \frac{P_B^\text{tot}}{\theta}. \tag{39}$$

Proof: Since the objective of Problem (Sub-P-E) is increasing, the optimal solution in Case-I directly follows. We next prove the optimal solution in Case-II by exploiting the convexity of Problem (Sub-P-E). The details are as follows.

If Condition (36) is met, then constraint (32) is actively binding at the optimal solution of Problem (Sub-P-E). Further exploiting the convexity of Problem (Sub-P-E), we derive the optimal solution by using the Lagrangian multiplier method. Specifically, we use λ to denote the Lagrangian multiplier for (32) as follows:

$$L\big(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}}, \lambda\big) = \theta W_B \sum_{i=1}^{T} x_{Bik} + \lambda\big(\frac{n_0}{g_{BI}} + \frac{P_B^\text{tot}}{\theta} - G_k(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}})\big). \tag{40}$$

By taking the derivative of $L\big(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}}, \lambda\big)$ with respect to $\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}}$, we have:

$$\frac{\partial L\big(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}}, \lambda\big)}{\partial x_{Bik}} = \theta W_B - \lambda(\ln 2) \sum_{j=1}^{i} A_i, \forall i \in \mathcal{I}_\text{B}, \tag{41}$$

where A_i is given by:

$$A_i = (\frac{n_B}{g_{Bi}} - \frac{n_B}{g_{Bi-1}}) 2^{\sum_{j=i}^{I} x_{Bjk}}, \forall i \in \mathcal{I}_\text{B}. \tag{42}$$

Since $A_i \geq 0, \forall i \in \mathcal{I}_\text{B}$ always holds, there *at most* one relay (let us say relay r) with $\frac{\partial L\left(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}}, \lambda\right)}{\partial x_{Brk}} = 0$ (according to (41)) when reaching the optimum of Problem (Sub-P-E). While for each relay $i < r$ and $i \in \mathcal{I}_\text{B}$, we have

$$\frac{\partial L\big(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}}, \lambda\big)}{\partial x_{Bik}} > 0 \Rightarrow x^*_{Bik} = \frac{z_{ik}}{\theta W_B}. \tag{43}$$

For each relay $i > r$ and $i \in \mathcal{I}_\text{B}$, we have

$$\frac{\partial L\big(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}}, \lambda\big)}{\partial x_{Bik}} < 0 \Rightarrow x^*_{Bik} = 0. \tag{44}$$

Finally, regarding relay r, we can uniquely determine the value of x^*_{Brk} according to constraint (32), which is actively binding and leads to equation (39). In particular, exploiting $x_{Brk} \in (0, \frac{z_{rk}}{\theta W_\text{B}}]$ and the property that $G_k(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}})$ is increasing in x_{Brk}, we can use the bisection search method to efficiently find the value of x^*_{Brk}, with the total number of iterations required by the bisection search no greater than $\log_2\left(\frac{z_{rk}}{\epsilon\theta W_\text{B}}\right)$, where ϵ is the tolerable computation-error in the bisection search. Notice that the value of r in Case-II must be unique, due to the structural property of the optimal solution $\{x^*_{Bik}\}_{i\in\mathcal{I}_\text{B}}$ and the property that function $G(\{x_{Bik}\}_{i\in\mathcal{I}_\text{B}})$ is monotonically increasing. We thus finish the whole proof. ■

Based on Proposition 5, we can gradually evaluate $r \in \mathcal{I}_\text{B}$ one-by-one until finding the optimal solution of Problem (Sub-P-E). The details are illustrated in the following SubSol-Algorithm. In SubSol-Algorithm, Steps 2-3 address Case-I, and Steps 5-23 address Case-II. As explained before, the key idea of Steps 5-23 is that we increase r one by one. If the currently evaluated value of r can make (39) strictly binding, then we compute the corresponding optimal solution of Problem (Sub-P-E) (i.e., Steps 9-11 for the case of $r = 1$ and Steps 16-19 for the case of $r > 1$) according to (37), (38), and (39).

We emphasize that SubSol-Algorithm is very efficient, since it invokes the operation of the bisection search only one time, i.e., in either Step 11 or Step 19, which requires no more than $O\left(\frac{z_{rk}}{\theta W_B \epsilon}\right)$ rounds of iterations to converge (given the tolerable computation-error ϵ).

SubSol-Algorithm: to solve Problem (Sub-P-E) and compute $F(\{z_{ik}\}_{i\in\mathcal{I}_\text{B}})$

1: **Input:** $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$.
2: **if** Condition (35) is met. **then**
3: Set $x^*_{Bik} = \frac{z_{ik}}{\theta W_B}, \forall i \in \mathcal{I}_\text{B}$.
4: **else**
5: Set $r = 1$.
6: **while** $r \leq I$ **do**
7: **if** $r == 1$ **then**
8: Set $x^\text{test}_{Brk} = \frac{z_{rk}}{\theta W_B}$ and $x^\text{test}_{Bik} = 0$ for each $i > r, i \in \mathcal{I}_B$.
9: **if** $G(\{x^\text{test}_{Bik}\}_{i\in\mathcal{I}_\text{B}}) - \frac{n_0}{g_{BI}} > \frac{P_B^\text{tot}}{\theta}$ **then**
10: Set $x^*_{Bik} = 0$ for each $i > r, i \in \mathcal{I}_B$.
11: Use the bisection search to find $x^*_{Brk} \in [0, \frac{z_{rk}}{\theta W_B}]$ such that (39) holds. Jump to the step of Output.
12: **end if**
13: **else**
14: Set $x^\text{test}_{Bik} = \frac{z_{ik}}{\theta W_B}$ for each $i \leq r, i \in \mathcal{I}_B$.
15: Set $x^\text{test}_{Bik} = 0$ for each $i > r, i \in \mathcal{I}_B$.
16: **if** $G(\{x^\text{test}_{Bik}\}_{i\in\mathcal{I}_\text{B}}) - \frac{n_0}{g_{BI}} > \frac{P_B^\text{tot}}{\theta}$ **then**
17: Set $x^*_{Bik} = \frac{z_{rk}}{\theta W_B}$ for each $i < r, i \in \mathcal{I}_B$.
18: Set $x^*_{Bik} = 0$ for each $i > r, i \in \mathcal{I}_B$.
19: Use the bisection search to find $x^*_{Brk} \in [0, \frac{z_{rk}}{\theta W_B}]$ such that (39) holds. Jump to the step of Output.
20: **end if**
21: Update $r = r + 1$.
22: **end if**
23: **end while**
24: **end if**
25: **Output**: $F(\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}) = \theta W_B \sum_{i=1}^{I} x^*_{Bik}$ and $\{x^*_{Bik}\}_{i\in\mathcal{I}_\text{B}}$ for Problem (Sub-P-E).

B. Proposed Algorithm to Solve Top-problem

With the above proposed SubSol-Algorithm to solve Problem (Sub-P) and evaluate $F(\{z_{ik}\}_{i\in\mathcal{I}_\text{B}})$ for each given $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}}$, we then can continue to solve Problem (Top-P) in this subsection. However, solving Problem (Top-P) is still difficult, since we cannot analytically derive $F(\{z_{ik}\}_{i\in\mathcal{I}_\text{B}})$. In other words, Problem (Top-P) is an optimization problem without an analytical objective function. To tackle with this difficulty, we exploit the monotonic property of Problem (Top-P). To illustrate this property, we first illustrate the following property regarding Ω_k.

Proposition 6: *(Normal Set in Monotonicity Theory [36], [37])*: Set Ω_k is a normal set.

Proof: The proof is based on the definition of normal set [36]. Specifically, suppose that there exists a $\{z_{ik}\}_{i\in\mathcal{I}_\text{B}} \in \Omega_k$. Then, let us consider another $\{z'_{ik}\}_{i\in\mathcal{I}_\text{B}}$, in which $z'_{ik} \leq z_{ik}, \forall i \in \mathcal{I}$ and with at least one j of $z'_{jk} \leq z_{jk}$. In this case, $\{z'_{ik}\}_{i\in\mathcal{I}} \in \Omega_k$ always holds, since eq. (14) shows that relay m's minimum required

transmit-power $q_{mk}^{\min}$ is monotonically increasing with respect to $\{\gamma_{nk}\}_{n\leq m, n\in\mathcal{I}_{\text{MU}k}}$. ■

Followed by Proposition 6, we have the result below:

Proposition 7: *(Monotonicity of Problem (Top-P))* Problem (Top-P) is a monotonic optimization problem.

Proof: According to [36], [37], the monotonic optimization problem refers to a special optimization problem that aims at maximizing an increasing function subject to the feasible region constructed by the intersection of normal set and reversed normal set. It can be identified that the optimal value of the objective function of Problem (Sub-P), i.e., $F(\{z_{ik}\}_{i\in\mathcal{I}_{\text{B}}})$, is non-decreasing in the relays' throughput-capacities $\{z_{ik}\}_{i\in\mathcal{I}_{\text{B}}}$. Further due to the relationship between Problems (Top-P) and (Sub-P), the objective function of Problem (Top-P) is non-decreasing in $\{z_{ik}\}_{i\in\mathcal{I}_{\text{B}}}$. Meanwhile, the feasible region of Problem (Top-P) is a normal set according to Proposition 6. Hence, our Problem (Top-P) is a monotonic optimization problem. ■

In the framework of monotonic optimization, the monotonicity of the objective function and the nature of normal-set of the feasible region Ω_k indicate that the optimal solution can only occur on the boundary of the feasible region. This important property enables us to exploit the so-called *poly-block outer-approximation algorithm* [36] to solve Problem (Top-P) and find the globally optimal solution $\{z_{ik}^*\}_{i\in\mathcal{I}_{\text{B}}}$. The details are shown in our proposed TopSol-Algorithm in the next subsection. Specifically, by using the monotonicity of the constraints, we iteratively construct a group of poly-blocks to approximate Ω_k with an increasing precision. Thanks to the monotonicity of the objective function, the optimal solution is guaranteed to locate at one of the vertices of the constructed poly-blocks, as long as this vertex is close to the feasible region with the specified precision.

TopSol-Algorithm: to solve Problem (Top-P) and compute $\boldsymbol{z}^*$ and $V_{k,\text{NOMA}}^*$

1: The BS initializes the current best solution $CBS=\emptyset$, the current best value $CBV=-\infty$, the iteration-index $l=1$, and δ as a small positive number (e.g., $\delta=0.001$). The BS also sets the flag for stopping as $f_{\text{stop}}=0$, and initializes the vertex-set $\mathcal{T}_1$ to have a single vertex as $\mathcal{T}_1=\{\{\hat{z}_{ik}\}_{i\in\mathcal{I}_{\text{B}}}\}$.
2: **while** $f_{\text{stop}}=0$ **do**
3: The BS selects the *current best vertex* $\boldsymbol{z}^l\in\arg\max\{F(\boldsymbol{z})|\boldsymbol{z}\in\mathcal{T}_l\}$. Specifically, for each vertex $\boldsymbol{z}\in\mathcal{T}_l$, the BS uses SubSol-Algorithm to compute the corresponding value of $F(\boldsymbol{z})$.
4: The BS constructs a line between origin and $\boldsymbol{z}^l$, and finds the intersection point $\boldsymbol{y}^l$ between the above constructed line and the upper boundary given in $\mathcal{G}$ (via using bisection search). The BS uses SubSol-Algorithm to compute the value of $F(\boldsymbol{y}^l)$.
5: **if** $C(\boldsymbol{y}^l)>CBV$ **then**
6: The BS updates $CBV=F(\boldsymbol{y}^l)$ and sets $CBS=\boldsymbol{y}^l$.
7: **end if**
8: **if** $\|\boldsymbol{y}^l-\boldsymbol{z}^l\|<\delta$ **then**
9: The BS sets $f_{\text{stop}}=1$.
10: **end if**
11: The BS updates the vertex-set as $\mathcal{T}_{l+1}=(\mathcal{T}_l\backslash\{\boldsymbol{z}^l\})\cup\{\boldsymbol{z}^l+(y_j^l-z_j^l)\boldsymbol{e}_j\}$. The BS then remove all vertexes $\boldsymbol{z}\in\mathcal{T}_{l+1}\backslash\mathcal{H}$ (which removes those infeasible vertices).
12: **if** $\mathcal{T}_{l+1}$ is empty **then**
13: The BS sets $f_{\text{stop}}=1$.
14: **end if**
15: The BS sets $l=l+1$.
16: **end while**
17: **Output**: The BS sets $\boldsymbol{z}^*$ according to CBS and $V_{k,\text{NOMA}}^*=F(\{z_{ik}^*\}_{i\in\mathcal{I}_{\text{B}}})$

C. Detailed Explanations about TopSol-Algorithm

We explain our proposed TopSol-Algorithm as follows.

(Initial setting of vertex-set): In our TopSol-Algorithm, for each relay i, we can firstly set the upper-bound for z_{ik} as follows:

$$\hat{z}_{ik}=(1-\theta)W_S\log_2\left(1+\frac{Q_{\Phi_k(i)}^{\text{tot}}}{1-\theta}\frac{h_{\Phi_k(i)k}}{n_0}\right),\forall i\in\mathcal{I}_{\text{B}}. \quad (45)$$

Without taking into account the co-channel interference among the BS-relay paths, $\{\hat{z}_{ik}\}_{i\in\mathcal{I}_{\text{B}}}$ represents a reasonable set of upper-bounds for $\{z_{ik}\}_{i\in\mathcal{I}_{\text{B}}}$. Correspondingly, we initialize the vertex-set $\mathcal{T}_1=\{\{\hat{z}_{ik}\}_{i\in\mathcal{I}_{\text{B}}}\}$, which will be used in the following iterative process.

(Iterative process): The key component of TopSol-Algorithm is the While-Loop (Lines 2-16), whose purpose is to iteratively construct the poly-blocks that approximate the upper-boundary of Ω_k in (28) as much as possible. In the l-th iteration, set $\mathcal{T}_l$ denotes the current set of vertexes. In $\mathcal{T}_l$, the BS finds a best vertex $\boldsymbol{z}^l$ that yields the largest objective value (for Problem (Top-P)) in Step 3. With $\boldsymbol{z}^l$, the BS executes two tasks as follows:

- *Task i): to update the current best solution (CBS) and the current best value (CBV).* The BS first constructs a line from origin to $\boldsymbol{z}^l$. It then finds the intersection point (denoted by $\boldsymbol{y}^l$) between the constructed line and the upper-boundary of the feasible region (Line 4). The BS uses $F(\boldsymbol{y}^l)$ and $\boldsymbol{y}^l$ to update the CBV and the CBS in the l-th iteration (Lines 5-7).
- *Task ii): to construct poly-blocks $\mathcal{T}_{l+1}$ for the next round iteration*. The BS uses vertex $\boldsymbol{z}^l$ and the intersection $\boldsymbol{y}^l$ to construct the new poly-blocks that approximate Ω_k with increasing precisions (Line 11)[3]. The purpose of Line 11 is to remove the region in which the optimal solution does not exist for sure.

As a key feature of TopSol-Algorithm, for each evaluated $\{z_{ik}\}_{\in\mathcal{I}_{\text{B}}}$, we need to invoke SubSol-Algorithm to obtain the value of $F(\{z_{ik}\}_{i\in\mathcal{I}_{\text{B}}})$. Figure 4 shows the whole algorithm to solve Problem (TTM-k-E), in which TopSol-Algorithm uses SubSol-Algorithm as a subroutine to compute $F(\{z_{ik}\}_{i\in\mathcal{I}_{\text{B}}})$ under a given $\{z_{ik}\}_{i\in\mathcal{I}_{\text{B}}}$. Based on the convexity of Problem (Sub-P-E) (Proposition 4) and its structural property (Proposition 5), SubSol-Algorithm is guaranteed to find the optimal solution of Problem (Sub-P) within I rounds of iterations. Further exploiting the monotonicity of Problem (Top-P) (Proposition 7), TopSol-Algorithm, which is based on monotonic optimization and the poly-block outer-approximation algorithm, is guaranteed to find the optimal solution of Problem (Top-P).

Finally, after TopSol-Algorithm outputs $\{z_{ik}^*\}_{i\in\mathcal{I}_{\text{B}}}$, we then can compute the optimal solution of the original problem (TTM-k) as follows.

- First, we use Proposition 3 to compute the optimal solution of Problem (TTM-k-E), i.e., $\{\beta_{Bik}^*\}_{i\in\mathcal{I}_{\text{B}}}$ according to (30) and $\{\gamma_{mk}^*\}_{m\in\mathcal{I}_{\text{MU}k}}$ according to (29).
- Then, with $\{\beta_{Bik}^*\}_{i\in\mathcal{I}_{\text{B}}}$ and $\{\gamma_{mk}^*\}_{m\in\mathcal{I}_{\text{MU}k}}$, we can further use (22) and (23) to iteratively compute the optimal BS's power allocation p_{Bik}^* and each relay m's power allocation q_{mk}^*.

We thus finish solving Problem (TTM-k), and the maximum throughput of the NOMA relay is given by $V_{k,\text{NOMA}}^*=F(\{z_{ik}^*\}_{i\in\mathcal{I}_{\text{B}}})$.

We discuss the complexity of our proposed algorithm as follows. The overall computational complexity of our proposed TopSol-Algorithm grows quadratically with respect to the number of relays for the following two reasons. *First*, SubSol-Algorithm (i.e., the subroutine) requires no more than a total number of I iterations. *Second*, in each round of TopSol-Algorithm, at most I new vertices are generated, which correspondingly invoke SubSol-Algorithm I

[3]In Line 11, scalar y_j^l (or z_j^l) denotes the j-th element of vector $\boldsymbol{y}^l$ (or vector $\boldsymbol{z}^l$). Vector $\boldsymbol{e}_j$ denotes a vector with the j-th element equal to 1, and all the other elements are equal to 0. In PBBA-Algorithm, all vectors are of dimension $1\times I$, with I denoting the number of the MUs.

times. Thus, the overall complexity of our proposed algorithm grows quadratically with respect to the number of relays.

D. TDMA Relay-Transmission: An Interference-free Scheme

For the purpose of comparison, we also consider another TDMA relay-transmission scheme. In the TDMA scheme, all relays provide traffic-relay from the BS to MU k in a TDMA manner, namely, the BS transmits to relay i for duration $\frac{\theta}{I}$, and relay i relays its received data to MU k for duration $\frac{1-\theta}{I}$. Compared with the NOMA scheme, the TDMA scheme is interference-free by avoiding the co-channel interference among all BS-relay paths and that among all relay-MU paths.

The total throughput of such a TDMA scheme $V^*_{k,\text{TDMA}}$ can be calculated by solving the following optimization problem:

(TDMA-k):

$$V^*_{k,\text{TDMA}} = \max \sum_{i\in\mathcal{I}_\text{B}} \frac{1}{I}\theta W_B \log_2(1+\frac{p_{Bik}g_{Bi}}{n_B})$$

$$\text{subject to:} \quad \theta W_B \log_2(1+\frac{p_{Bik}g_{Bi}}{n_B}) \leq (1-\theta)W_S \log_2(1+\frac{q_{ik}h_{ik}}{n_k}), \forall i \in \mathcal{I}_\text{B}, \quad (46)$$

$$\sum_{i\in\mathcal{I}_\text{B}} \frac{\theta}{I} p_{Bik} \leq P_B^\text{tot}, \quad (47)$$

$$\frac{1-\theta}{I} q_{ik} \leq Q_i^\text{tot}, \forall i \in \mathcal{I}_\text{B}, \quad (48)$$

$$\text{variables: } p_{Bik} \geq 0, \text{ and } q_{ik} \geq 0, \forall i \in \mathcal{I}_\text{B}.$$

Since the TDMA scheme does not require to order the MUs according to their channel power gains, we just use $\mathcal{I}_\text{B}$ to index the group of relays. Here, p_{Bik} denotes the BS's transmit-power to relay i, and q_{ik} denotes relay i's transmit-power to MU k.

We next derive $V^*_{k,\text{TDMA}}$ for Problem (TDMA-k). To this end, for relay i, based on constraints (46) and (48), we define the upper-bound of the BS's transmit-power to relay i as:

$$M_{ik} = \frac{n_B}{g_{Bi}} \left((1+Q_i^\text{tot}\frac{I}{1-\theta}\frac{h_{ik}}{n_k})^{\frac{1-\theta}{\theta}\frac{W_S}{W_B}} - 1 \right), \forall i \in \mathcal{I}_\text{B}. \quad (49)$$

With $\{M_{ik}\}_{i\in\mathcal{I}_\text{B}}$, we can characterize the optimal solution of Problem (TDMA-k) as follows.

Proposition 8: The optimal power allocation of Problem (TDMA-k) can be given by the following two cases.

- (Case-I): if $\sum_{i\in\mathcal{I}_\text{B}} M_{ik} \leq \frac{I}{\theta}P_B^\text{tot}$, then we derive the optimal solution as follows:

$$p^*_{Bik} = M_{ik}, \forall i \in \mathcal{I}_\text{B}. \quad (50)$$

$$q^*_{ik} = \left((1+\frac{M_{ik}g_{Bi}}{n_B})^{\frac{\theta}{1-\theta}\frac{W_B}{W_S}} - 1 \right) \frac{n_k}{h_{ik}}, \forall i \in \mathcal{I}_\text{B}. \quad (51)$$

- (Case-II): if $\sum_{i\in\mathcal{I}_\text{B}} M_{ik} > \frac{I}{\theta}P_B^\text{tot}$, then we derive the optimal solution as follows:

$$p^*_{Bik} = \min\left(\max\left(\frac{1}{\lambda}\frac{\theta W_B}{I\ln 2} - \frac{n_B}{g_{Bi}}, 0 \right), M_{ik} \right), \forall i \in \mathcal{I}_\text{B}. \quad (52)$$

$$q^*_{ik} = \left((1+\frac{p^*_{Bik}g_{Bi}}{n_B})^{\frac{\theta}{1-\theta}\frac{W_B}{W_S}} - 1 \right) \frac{n_k}{h_{ik}}, \forall i \in \mathcal{I}_\text{B}, \quad (53)$$

where parameter λ is set[4] such that $\sum_{i\in\mathcal{I}_\text{B}} p^*_{Bik} = \frac{I}{\theta}P_B^\text{tot}$.

[4]Notice that such a value of λ is unique due to that p^*_{Bik} is monotonically decreasing with respect to λ as shown in (52). Moreover, exploiting this decreasing property, we can execute a bisection search of λ within the interval $\left[\min_{i\in\mathcal{I}_\text{B}} \frac{W_B}{I\ln 2}\frac{g_{Bi}}{n_B+M_{ik}g_{Bi}}, \max_{i\in\mathcal{I}_\text{B}} \frac{W_B}{I\ln 2}\frac{g_{Bi}}{n_B}\right]$ to find the value of λ such that $\sum_{i\in\mathcal{I}_\text{B}} p^*_{Bik} = \frac{I}{\theta}P_B^\text{tot}$ is satisfied. Recall that M_{ik} is given in (49).

Proof: We notice that (46) and (47) are independent for different MUs, which helps us transform Problem (TDMA-k) into an equivalent convex optimization problem. We thus can exploit the Karush-Kuhn-Tucker conditions [41] to derive the optimal solution. We skip the details due to the limited space. ■

Using the optimal solution $\{p^*_{Bik}\}_{i\in\mathcal{I}_\text{B}}$ of Problem (TDMA-k) given by Proposition 8, we can derive the maximum throughput of the TDMA relay as follows:

$$V^*_{k,\text{TDMA}} = \max \sum_{i\in\mathcal{I}_\text{B}} \frac{1}{I}\theta W_B \log_2\left(1+\frac{p^*_{Bik}g_{Bi}}{n_B}\right). \quad (54)$$

In the next section, we will compare $V^*_{k,\text{NOMA}}$ with $V^*_{k,\text{TDMA}}$ and show the throughput gain of the NOMA relay.

VI. Numerical Results for Optimal NOMA Relay-Transmission

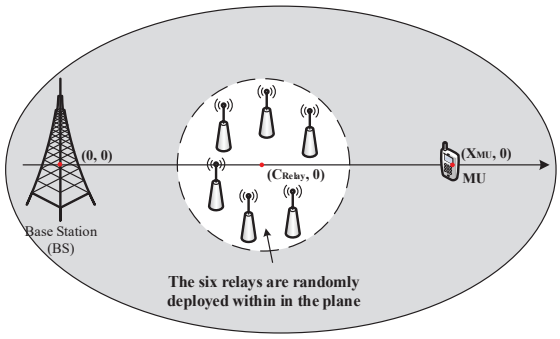
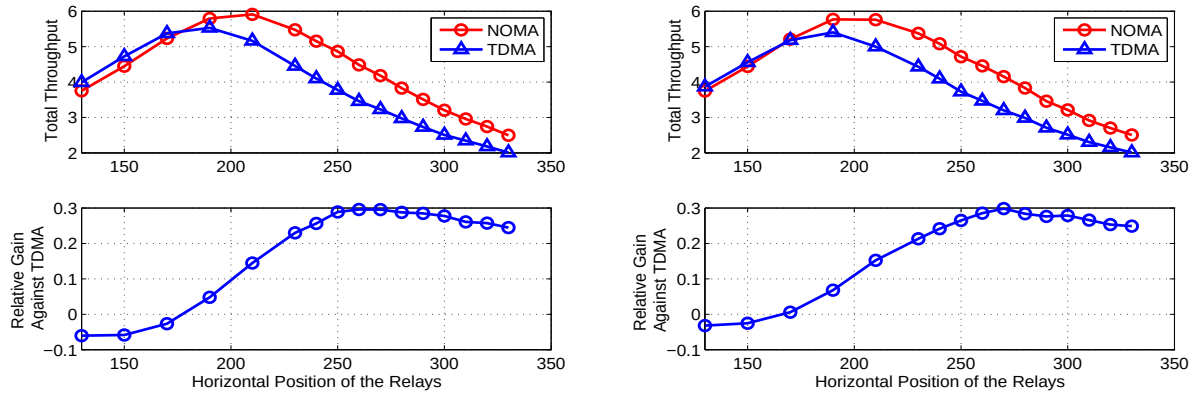


Fig. 5: A scenario of six relays which are randomly deployed within a plane according to a uniform distribution.

We show the performance of the optimal NOMA relay-transmission in this section. All the numerical tests are obtained by using MATLAB on a PC of Intel(R) Core(TM) i5-4590 CPU@3.3GHz.

(Topology and parameter-setting): We use a scenario of six relays as shown in Figure 5. Specifically, the BS is located at the origin $(0,0)$m, and the six relays are randomly deployed within a plane whose center position is $(\text{C}_\text{Relay}, 0)$m and the radius is 20m. The MU is fixed at $(\text{X}_\text{MU}, 0)$m. We will specify the values of C_Relay and X_MU soon. We use the similar method in [40] to model the channel power gain, i.e., $g_{Bi} = \frac{\varrho_{Bi}}{l^\kappa_{Bi}}$, where l_{Bi} denotes the distance between the BS and relay $i \in \mathcal{I}_\text{B}$, and κ denotes the power-scaling factor for the path-loss (we set $\kappa = 3$). To capture the fading and shadowing effects, we assume that ϱ_{iB} follows an exponential distribution with a unit mean[5]. We set the BS's channel bandwidth $W_B = 8$MHz and the relays' channel bandwidth $W_S = 4$MHz. Regarding the half-duplex relay-transmission, without loss of any generality, we set $\theta = 0.5$, i.e., Phase-I and Phase-II shown in Figure 2 have an equal duration. Notice that since we consider that the locations of the relays are randomly generated, every point in the following Figures 6-8 denotes the average result of 400 random realizations of the relays' locations.

[5]After knowing the detailed values of the set of the channel power gains from the BS to all relays, we can order order these channel power gains from the BS to all relays in the descending order (as in eq. (1)), e.g., by using the MATLAB "sort" function. Consequently, we obtain the ordered relay-indexes $\mathcal{I}_\text{B}$ (from the BS's the perspective) as well as the set of the ordered channel power gains $\{g_{Bi}\}_{i\in\mathcal{I}_\text{B}}$. Similarly, after knowing the detailed values of the set of the channel power gains from all relays to MU k, we can order these channel power gains from all relays to MU k (as in eq. (4)). Consequently, we obtain the order relay-indices $\mathcal{I}_{\text{MU}k}$ (from MU k's perspective) as well as the set of ordered channel power gains $\{h_{mk}\}_{m\in\mathcal{I}_{\text{MU}k}}$.



(a) Six relays of homogenous power capacities.

(b) Six relays of heterogeneous power capacities.

Fig. 6: Performance under different central locations of the relays' random distribution. Each point represents the average of 400 random realizations of the six relays' locations. The relative gain in the two bottom-subplots is defined as $\frac{V^*_{k,\text{NOMA}}-V^*_{k,\text{TDMA}}}{V^*_{k,\text{TDMA}}}$.

(Performance under different locations of the relays): Figure 6 shows the performance of the proposed NOMA relay under different horizontal locations of the relays. Specifically, we fix the horizontal position of the MU at $\text{X}_{\text{MU}} = 350$m, and vary C_{Relay} (i.e., the center of the plane in which the relays are randomly deployed) from 130m to 330m. This change corresponds to that the relays move further away from the BS to the MU. In Subplot 6(a), we set all relays of a homogeneous power capacity as $Q_m^{\text{tot}} = 0.2\text{W}, \forall m$. While, in Subplot 6(b), we set all relays of heterogenous power capacities as $\{Q_m^{\text{tot}}\} = (0.1, 0.2, 0.3, 0.1, 0.2, 0.3)\text{W}$.

Figure 6 shows that the optimal (i.e., the maximum) throughput of the NOMA relay-transmission firstly increases when the relays move further away from the BS, and then gradually decreases when the relays move closer to the MU. This result is reasonable. When the relays are too far away from the MU (i.e., too close to the BS), the weak channel gains between the relays and MU k degrade the benefit of using SIC to mitigate the co-channel interference among the relay-MU paths, which consequently lowers the total throughput of the NOMA relay. For the similar result, when the relays too far way from the BS, the total throughput of the NOMA relay also decreases. That is why the throughput of NOMA relay-transmission shows a peak-shape as shown in Figure 6. Notice that the throughput of TDMA relay shows a similar peak-shape due to the influence of the weak channel gains

Meanwhile, compared with the TDMA scheme, Figure 6 shows that the NOMA relay can achieve larger throughput than the TDMA scheme for a wide range of the relays' locations (e.g., when C_{relay} varies from 190m to 330m in Figure 6). In particular, the results show that the NOMA relay can achieve the relative throughput gain against the TDMA scheme (which is defined as $\frac{V^*_{k,\text{NOMA}}-V^*_{k,\text{TDMA}}}{V^*_{k,\text{TDMA}}}$) up to 30%. However, if the relays are deployed too far away from the MU (e.g., when $\text{C}_{\text{relay}} < 170$m in Figure 6), the total throughput of the NOMA relay-transmission decreases and becomes even less than that of the TDMA scheme. Such a result is due to the nature of the SIC in NOMA. Specifically, the very weak channel power gains (due to the too long distance between the relays and the MU) provide a very limited benefit in exploiting the SIC to mitigate the co-channel interference and degrades the advantage of using NOMA. Different from the NOMA scheme, the TDMA relay is interference-free. Thus, the throughput of TDMA scheme is less sensitive to the weak channel power gains compared with the NOMA scheme.

(Performance under different locations of the MU) Figure 7 shows the performance of the proposed NOMA relay-transmission under different values of X_{MU}, i.e., the MU's horizontal position. Specifically, we fix the center of the plane in which the relays are randomly deployed at $\text{C}_{\text{Relay}} = 250$, and vary the MU's horizontal position X_{MU} from 280m to 520m, i.e., the MU moves further away from the relays. We set $P_{\text{B}}^{\text{tot}} = 1\text{W}$ for the BS, and set the relays of homogenous power capacities as $Q_m^{\text{tot}} = 0.2\text{W}, \forall m$.

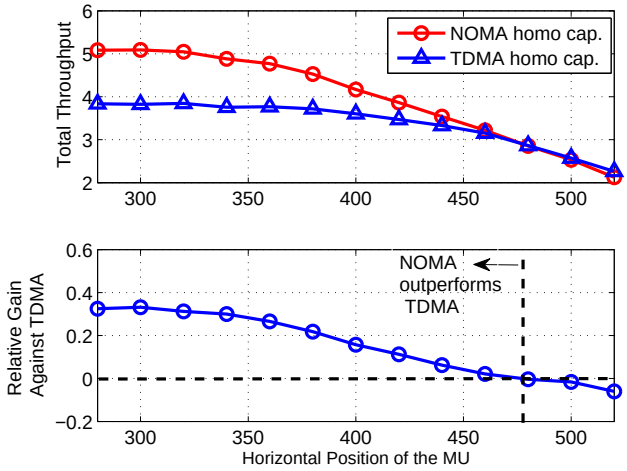


Fig. 7: Performance under different horizontal positions of the MU. Each point denotes the average of 400 random realizations of the six relays' locations.

As shown in Figure 7, when the MU is within the moderate distance of the relays (i.e., when X_{MU} varies from 280m to 460m), the NOMA relay-transmission can outperform the TDMA scheme and achieve the relative throughput gain close to 40%. Such an advantage comes from that the NOMA scheme exploits the SIC to mitigate the co-channel interference, which thus effectively improves the total throughput. However, the throughput gain of NOMA scheme gradually diminishes when the MU becomes further away from the relays. When the MU is too far away from the relays (i.e., when $\text{X}_{\text{MU}} > 480$m), the NOMA scheme becomes worse than the TDMA scheme. This result is consistent with the intuition, i.e., the very weak channel power gains (due to the too large distances between the relays and the MU) limit the benefit in using the SIC to mitigate the co-channel interference and degrade the throughput gain of the NOMA.

(Performance under different power capacities of relays) Figure 8 shows the performance of the NOMA relay-transmission under different power capacities of the relays. Specifically, we fix $P_{\mathrm{B}}^{\mathrm{tot}} = 1$W for the BS, and vary the relays' homogeneous power capacities Q_m^{tot}. We fix the center of plane (in which the relays are randomly deployed) at $\mathrm{C_{Relay}} = 250$m, and we set the horizontal position of the MU at $\mathrm{X_{MU}} = 400$m in Subplot 8(a) and $\mathrm{X_{MU}} = 520$m in Subplot 8(b).

As shown in Figure 8, the throughput of the NOMA relay-transmission increases when the relays' power capacities increase, which is consistent with the intuition. In addition, compared with the TDMA scheme, the relative improvement achieved by the NOMA scheme also gradually increases when we increase the relays' power capacities. This result is reasonable, since increasing the relays' power-capacity provides a larger freedom to reap the benefit of SIC, which thus yields a larger throughput-gain against the TDMA scheme. As shown in Subplot 8(a), when the MU is close to the relays, such a throughput-gain is always positive, meaning that the NOMA scheme always outperforms the TDMA scheme. In comparison, in Subplot 8(b), when the MU is very far away from the relays, the NOMA scheme might fail to outperform the TDMA due to the weak channel gain and the relays' low power capacities. Nevertheless, increasing the relays' power capacities can quickly improve the performance of NOMA relay-transmission and increase the throughput-gain against the TDMA scheme.

(A hybrid NOMA relay to address the influence of weak channel power gains): As we have explained for the results in Figures 6, 7, and 8, the throughput gain of the NOMA relay depends on the channel power gains of the BS-relay paths and the relay-MU paths. Despite the significant throughput gain under a wide range of the locations of the relays and the MU, the throughput of NOMA relay might become less than that of the TDMA scheme when the the relays are too far away from the MU, since the too weak channel gain degrades the benefit of executing SIC in NOMA. To tackle with this issue, we propose a hybrid NOMA (HB-NOMA) relay as follows:

$$V^*_{k,\text{HB-NOMA}} = \max\left\{V^*_{k,\text{NOMA}}, V^*_{k,\text{TDMA}}\right\}. \tag{55}$$

The HB-NOMA relay-transmission enables us to jointly reap the benefit of the SIC in NOMA when the relays experience relatively strong channel power gains to the BS and the MU, and meanwhile to reap the benefit of the interference-free TDMA when the relays experience very weak channel gains with respect to the BS and the MU. Notice that the throughput of the HB-NOMA relay in (55) can be directly computed, since we have proposed TopSol-Algorithm to compute $V^*_{k,\text{NOMA}}$ and provided the analytical solution in (54) for computing $V^*_{k,\text{TDMA}}$.

It is straightforward that the proposed HB-NOMA in (55) must outperform the individual NOMA relay and the individual TDMA relay. We will show the performance of HB-NOMA relay in the next section in which the per-MU maximum throughput provided by the proposed HB-NOMA relay is used for designing an optimal multi-MUs scheduling scheme for maximizing all MUs' overall utility.

VII. Extension to Multi-MUs Scenario and a Multi-MUs Scheduling Algorithm

The previous sections focus on investigating the NOMA relay-transmission for an arbitrary MU k within one time-slot. The proposed NOMA relay (and the HB-NOMA relay) scheme be further utilized for the scenario of multiple MUs, e.g., the group of MUs $\mathcal{K} = \{1, 2, ..., K\}$. Specifically, in this study, for this multi-MUs scenario, we investigate how the BS properly selects different MUs to provide the HB-NOMA relay-transmission via all relays in different time-slots, with the objective of maximizing all MUs' total utility in long-term. Figure 9 shows an example of the considered multi-MUs scheduling for a case of four MUs.

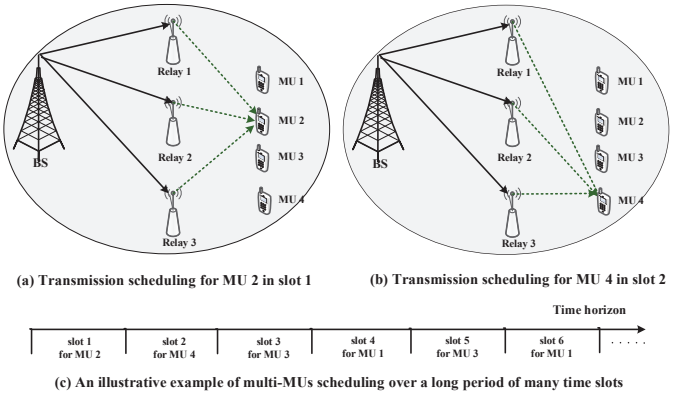


Fig. 9: Illustrative example of three relays and four MUs for the considered multi-MUs scheduling.

A. Illustration of Multi-MUs Scheduling Problem

To study this problem, we consider a long time-horizon comprised of a large number of time-slots as shown in Figure 9. Within each time-slot, the BS selects one of the MUs in $\mathcal{K}$ to send traffic through all relays in $\mathcal{I}_{\mathrm{B}}$, and the throughput for this selected MU is given by $V^*_{k,\text{HB-NOMA}}$, i.e., we use the HB-NOMA relay-transmission in (55) for each scheduled MU[6].

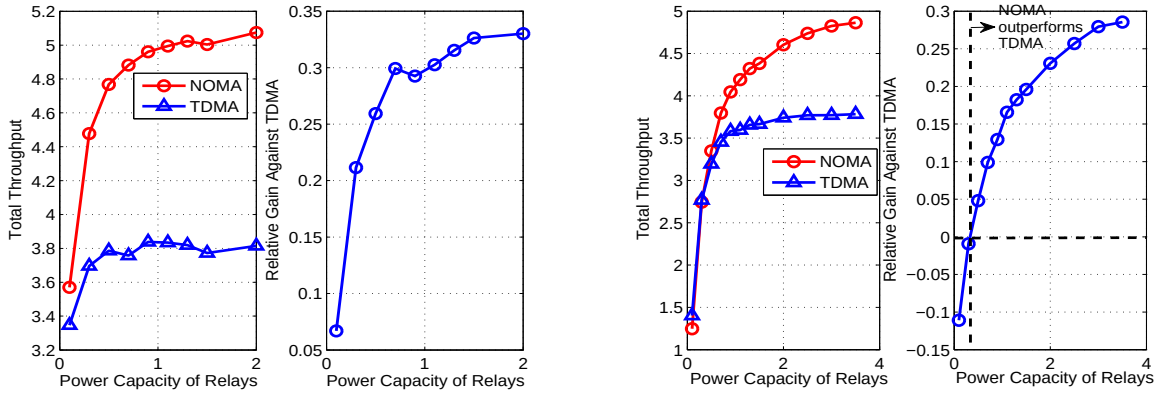
In particular, we consider that the channel power gains, including both the channel gains from the BS to the relays and those from the relays to the MUs, vary over different time-slots. To take into account the time-varying channel gains, we use $V^*_{k,\text{HB-NOMA}}$ to denote MU k's throughput in (55) at the τ-th slot. Therefore, given time-slot t, MU k's average throughput up to t is equal to $\overline{r}^t_k = \frac{1}{t}\sum_{\tau=1}^{t} a^\tau_k V^\tau_{k,\text{HB-NOMA}}$, in which variable $a^\tau_k = 1$ means that the BS selects MU k to serve at time-slot τ (while $a^\tau_k = 0$ means the opposite). We use function $\mathcal{U}_k(.)$ to denote MU k's utility based on its experienced average throughput. Specifically, the following utility function is widely used [38], [39]:

$$\mathcal{U}_k(\overline{r}^t_k) = \begin{cases} \frac{c_k}{\alpha}(\overline{r}_k)^\alpha, & \text{when } 0 < \alpha \leq 1 \\ c_k \log(\overline{r}_k), & \text{when } \alpha = 0 \end{cases} \tag{56}$$

where c_k and α are fixed parameters. The above utility function is concave with respect to MU k's experienced average throughput $\overline{r}^t_k$, which means that the utility increases in MU k's experienced average throughput, and the marginal increase gradually decreases when the MU's average throughput increases. Therefore, the objective of the considered multi-MUs scheduling problem in this section is to maximize a system-wise utility of all MUs, i.e., $\max \frac{1}{K}\sum_{k\in\mathcal{K}} \mathcal{U}_k(\overline{r}^t_k)$ by properly adjusting the scheduling-vector $\boldsymbol{a}^t = (a^t_1, a^t_2, ..., a^t_K)$, subject to the constraint that at most one of the MUs is selected to serve in each time-slot.

Following the gradient-projection (GP) based scheduling framework [38], [39], the scheduling-vector $\boldsymbol{a}^t = (a^t_1, a^t_2, ..., a^t_K)$ that yields the maximum projection on the gradient of the system-wise utility is selected. In other words, after we use (55) to compute

[6]Notice that in the proposed multi-MUs scheduling scheme, we adopt the HB-NOMA relay in (55) for each scheduled MU. This scheme can be modified for the case when we use the NOMA relay (or TDMA relay) for each scheduled MU, i.e., by using $V^*_{k,\text{NOMA}}$ (or $V^*_{k,\text{TDMA}}$).



(a) MU's horizontal position $X_{MU} = 400$m

(b) MU's horizontal position $X_{MU} = 520$m

Fig. 8: Performance under different power capacities of the relays. Each point denotes the average of 400 random realizations of the six relays' locations.

$\{V^t_{k,\text{HB-NOMA}}\}_{k\in\mathcal{K}}$ in slot t, the optimal GP-based scheduling-vector $\boldsymbol{a}^t$ at each slot t corresponds to the optimal solution of the following slot-by-slot optimization:

$$\max \sum_{k\in\mathcal{K}} \mathcal{U}_k^{'}\left(V^t_{k,\text{HB-NOMA}} + \sum_{\tau=1}^{t-1} a_k^\tau V^\tau_{k,\text{HB-NOMA}}\right)(a_k^t V^t_{k,\text{HB-NOMA}})$$

$$\text{subject to: } \sum_{k\in\mathcal{K}} a_k^t \le 1,$$

$$\text{variables: } a_k^t = \{0,1\}, \forall k \in \mathcal{K},$$

where $\mathcal{U}_k^{'}(.)$ is the first order derivative of $\mathcal{U}_k(.)$. Exploiting the simple structure of multi-MUs scheduling problem, we can enumerate the MUs in $\mathcal{K}$ to find the best MU to schedule for transmission at each time-slot t.

B. Performance of the Proposed Multi-MUs Scheduling

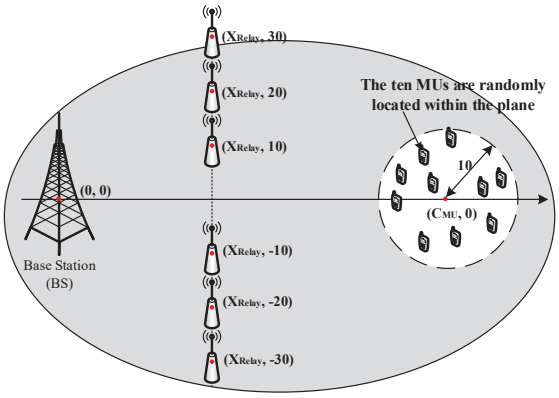


Fig. 10: A scenario of six relays aligned in a vertical line and ten MUs randomly located within a plane according to a uniform distribution.

We next show the performance of the proposed multi-MUs scheduling. Specifically, we setup the scenario of 10 MUs and 6 relays as follows. We place the six relays in a vertical line, i.e., $(X_{\text{Relay}}, -30)$m, $(X_{\text{Relay}}, -20)$m, $(X_{\text{Relay}}, -10)$m, $(X_{\text{Relay}}, 10)$m, $(X_{\text{Relay}}, 20)$m, and $(X_{\text{Relay}}, 30)$m (here we first set $X_{\text{Relay}} = 250$m). The 10 MUs are randomly located within a plane whose center is $(C_{\text{MU}}, 0)$m (here we set $C_{\text{MU}} = 450$m) and radius is 40m. The channel power gains from the BS to the relays and those from the relays to the MUs are all generated in a same manner as that explained in Section VI. Due to studying the scheduling over multiple time-slots, we assume that the channel gains vary independently across different time-slots. We set $P_{\text{B}}^{\text{tot}} = 1$W for the BS, and set the relays of homogenous power capacities as $Q_m^{\text{tot}} = 0.2$W,$\forall m$. Finally, we use $\mathcal{U}_k(\overline{r}_k) = \ln(\overline{r}_k)$ for MU k's utility function, which means that we use the utility function that takes into account different MUs' fairness in their respective achieved average throughput.

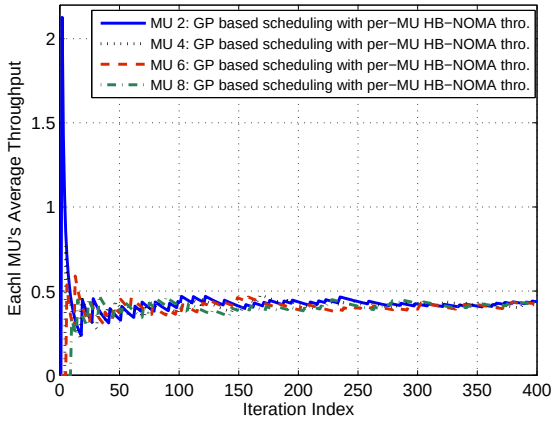
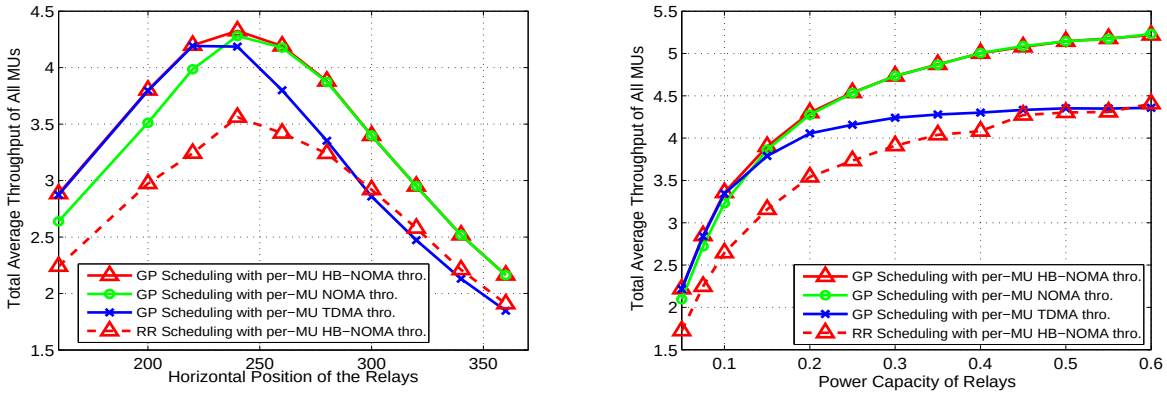


Fig. 11: Convergence of the MUs' average throughput when using the GP-based scheduling with HB-NOMA relay for each scheduled MU.

Figure 11 shows the convergence of the GP-based scheduling scheme with the HB-NOMA relay for each scheduled MU (i.e., using the per-slot throughput $V^{\tau,*}_{k,\text{HB-NOMA}}$ for the scheduled MU k). For simplicity, we plot the average throughput of 4 different MUs (namely, MUs 2,4,6, and 8) in Figure 11. The results show that by using the GP-based scheduling with the HB-NOMA relay, each MU's average throughput quickly converges to a relative stable value, and different MUs' average throughput are close to each other, i.e., keeping a good fairness level.

Figure 12 shows the advantage of the GP-based scheduling with the HB-NOMA throughput $V^{\tau,*}_{k,\text{HB-NOMA}}$ for each scheduled MU. Subplot 12(a) shows all MUs' total throughput under different horizontal positions of the relays (i.e., we vary X_{Relay} from 180m to 360m). The results show that the GP scheduling with the HB-NOMA relay yields a larger total throughput of all MUs than the conventional round-robin (RR) scheduling scheme. Furthermore, to show the advantage of using the HB-NOMA relay, we also plot the GP scheduling with NOMA relay (i.e., the line marked with circle) and the GP scheduling with TDMA relay (i.e., the line marked with "x"). The results show that the HB-NOMA relay can jointly reap the



(a) Different horizontal locations of the relays

(b) Different power capacities of the relays

Fig. 12: Performance advantage of the GP-based scheduling with the HB-NOMA relay. Each point denotes the result after 400 time-slots.

benefit of NOMA relay and that of the TDMA relay. Specifically, when the relays are located within the moderate range of the MUs (i.e., $X_{\text{Relay}} \geq 250$m), executing the SIC is beneficial to increase the total throughput. Hence, the GP scheduling with the HB-NOMA relay achieves the result of the NOMA relay. On the other hand, when the relays are far away from the MUs (i.e., $X_{\text{Relay}} <250$m), the very weak channel gains from the relays to the MUs degrade the effectiveness of SIC. Thus, the GP scheduling with the HB-NOMA relay turns to achieve the result of TDMA relay. Similar to Figure 8 before, Subplot 12(b) shows that increasing the relays' power capacity can not only improve the GP scheduling with the HB-NOMA relay, but also can effectively improve the throughput gain of the NOMA relay against the TDMA relay. Due to the limited space, we skip showing the results about the total utility of all MUs, which actually show the similar trends as the total throughput of all MUs shown in Figure 12.

VIII. Conclusion

In this paper, we have investigated the optimal power allocation and multi-MUs scheduling for NOMA relay-transmission, in which the BS first uses NOMA to transmit data to the relays, and the relays then use NOMA to transmit their respectively received data to the MUs. Targeted for one MU at first, we proposed the optimal power allocation to maximize the overall throughput from the BS to an arbitrary MU via all relays. Despite the non-convexity of the problem, we proposed the layered-algorithm to efficiently compute the optimal power allocation solution and the maximum throughput for the targeted MU. Numerical results have shown that the proposed NOMA relay-transmission can increase the throughput up to 30% compared with the conventional TDMA scheme, and increasing the relays' power capacity can effectively increase the throughput gain of the NOMA relay against the TDMA relay. To address the impact of too weak channel power gains, we further proposed the HB-NOMA relay that adaptively exploits the benefit of NOMA relay and that of the TDMA relay. Based on the maximum throughput of the HB-NOMA relay for each individual MU, we investigated the multi-MUs scheduling problem over a long-term period. We proposed a GP-based multi-MUs scheduling scheme that utilizes the throughput provided by the HB-NOMA relay for each scheduled MU. Numerical results have demonstrated that our proposed multi-MUs scheduling scheme can effectively increase all MUs' total utility compared with the conventional RR scheduling scheme, and the HB-NOMA relay can jointly reap the benefit of NOMA relay and that of the TDMA relay.

For our future work, we will investigate the scenario that different relays select different MUs to provide relay-transmission within a same time-slot. In this situation, we will take into account the interference among the relay-transmissions for different MUs, and investigate the joint optimization of the relay-selection and power allocation for the multi-MUs NOMA-relay transmission. Moreover, we will also investigate the distributed implementation of the proposed NOMA relay-transmission.

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. of IEEE VTC-Spring*, 2013.

[2] L. Dai *et. al.*, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future Research Trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74-81, Sep. 2015.

[3] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C.-L. I, and H.V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Communications Magazine*, vol.55, no.2, pp. 185-191, Feb. 2017.

[4] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo "Non-orthogonal Multiple Access for 5G and Beyond", *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347-2381, Dec. 2017.

[5] S.M. Islam, *et. al.*, "Power-domain non-orthogonal multiple access (NOMA) in 5G Systems: potentials and challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721-742, 2nd quarter 2017.

[6] M. Shirvanimoghaddam, M. Dohler, and S.J. Johnson, "Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations," availabile online at https://arxiv.org/pdf/1612.00552.pdf.

[7] B. Di, L. Song, Y. Li, and Z. Han, "V2X meets NOMA: Non-orthogonal multiple access for 5G enabled vehicular networks," available online at https://arxiv.org/pdf/1705.08709.pdf.

[8] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of nonorthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Processing Letter*, vol. 21, no. 12, pp. 1501-1505, Dec. 2014.

[9] Z. Zhang, H. Sun, and R.Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," to appear in *IEEE Journal on Selected Areas in Communications*, 2016.

[10] Y. Liu, Z. Qin, M. Elkashlan, Y. Gao, and A. Nallanathan "Non-orthogonal multiple access in massive MIMO aided heterogeneous networks," in *Proc. of IEEE GLOBECOM'2016*.

[11] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010-6023, Aug. 2016.

[12] Y. Zhang, H. Wang, T. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2852-2857, Mar. 2017.

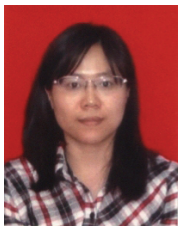
[13] Q. Sun, *et. al.*, "Energy efficiency optimization for fading MIMO non-orthogonal multiple access systems," in *Proc. of IEEE ICC'2015*.

[14] L. Lei, D. Yuan, C.K. Ho, and S. Sun, "Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems" in *Proc. of IEEE GLOBECOM'2015*.

[15] B. Di, S. Bayat, L. Song, and Y. Li "Radio resource allocation for downlink non-orthogonal multiple access (NOMA) networks using matching theory" in *Proc. of IEEE GLOBECOM'2015*.

[16] M.S. Elbamby, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Resource optimization and power allocation in in-band full duplex (IBFD)-enabled non-orthogonal multiple access networks," to appear in *IEEE Journal on Selected Areas in Communications*, 2016.
[17] 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," to appear in *IEEE Communications Letters*, Mar. 2017.
[18] J. Men, and J. Ge, "Non-orthogonal multiple access for multiple-antenna relaying networks," *IEEE Communications Letters*, vol. 19, no. 10, pp.1686-1689, Oct. 2015.
[19] J. Kim, and I. Lee, "Capacity analysis of cooperative relaying systems using non-orthogonal multiple access," *IEEE Communications Letters*, vol. 19, no. 11, pp. 1949-1952, Aug. 2015.
[20] R. Jiao, L. Dai, J. Zhang, R. MacKenzie, and M. Hao, "On the performance of NOMA-based cooperative relaying systems over Rician fading channels," to appear in *IEEE Transactions on Vehicular Technology*, DOI:10.1109/TVT.2017.2728608.
[21] C. Zhong, and Z. Zhang, "Non-orthogonal multiple access with cooperative full-duplex relaying," *IEEE Communications Letters*, vol. 20, no. 12, pp. 2478-2481, Dec. 2016.
[22] S. Luo, and K.C. Teh, "Adaptive transmission for cooperative NOMA system with buffer-aided relaying," *IEEE Communications Letters*, vol. 21, no. 4, pp. 937-940, April 2017.
[23] D. Zhang, Y. Liu, Z. Ding, Z. Zhou, A. Nallanathan, and T. Sato, "Performance analysis of non-regenerative massive-MIMO NOMA relay systems for 5G," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 4777-4790, Nov. 2017.
[24] Z. Ding, *et. al.*, "Relay selection for cooperative NOMA," *IEEE Wireless Communications Letters*, vol. 5, no. 4, pp. 416-419, June 2016.
[25] D. Deng, L. Fan, X. Lei, W. Tan, and D. Xie, "Joint user and relay selection for cooperative NOMA networks," *IEEE Access*, vol. 5, pp. 20220-20227, 2017.
[26] S. Zhang, B. Di, L. Song, and Y. Li, "Radio resource allocation for non-orthogonal multiple access (NOMA) relay network using matching game," in *Proc. of IEEE ICC'2016*.
[27] X. Liu, X. Wang, and Y. Liu, "Power allocation and performance analysis of the collaborative NOMA assisted relaying systems in 5G," *China Communications*, vol. 14, no. 1, pp. 50-60, Jan. 2017.
[28] C. Xue, Q. Zhang, Q. Li, and J. Qin, "Joint power allocation and relay beamforming in non-orthogonal multiple access amplify-and-forward relay networks," to appear in *IEEE Trans. on Vehicular Technology*, Jan. 2017.
[29] S. Zhang, B. Di, L. Song, and Y. Li, "Sub-channel and power allocation for non-orthogonal multiple access relay networks with amplify-and-forward protocol," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2249-2261, April 2017.
[30] Z. Zheng, L.X. Cai, R. Zhang, and X. Shen, "RNP-SA: Joint relay placement and sub-carrier allocation in wireless communication networks with sustainable energy", *IEEE Trans. on Wireless Communications*, vol. 11, no. 10, pp. 3818-3828, 2012.
[31] U. Sethakaset, T.Q.S. Quek, and S. Sun, "Joint source-channel optimization over wireless relay networks," *IEEE Trans. on Communications*, vol. 59, no. 4, pp. 1114-1122, April 2011.
[32] L. Qian, Y. Wu, J. Wang, and W. Zhang, "Energy-efficient distributed user scheduling in relay-assisted cellular networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 4060-4073, June. 2016.
[33] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity part I: system description," *IEEE Trans. Commun.*, vol. 51, no. 11, pp. 1927-1938, Nov. 2003.
[34] M. Tao, and Y. Liu, "A network flow approach to throughput maximization in cooperative OFDMA networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 3, pp. 1138-1148, Mar. 2013.
[35] Y. Wu, and L. Qian, "Energy-efficient NOMA-enabled traffic offloading via dual-connectivity in small-cell networks," *IEEE Communications Letters*, vol. 21, no.7, pp. 1605-1608, July 2017.
[36] H. Tuy, "Monotonic optimization: Problems and solution approaches," SIAM Journal of Optimization, vol. 11, no. 2, 2000.
[37] Y. Zhang, L. Qian, J. Huang, "Monotonic optimization in communication and networking systems," Foundation and Trends in Networking, Now Publisher, October 2013.
[38] J. Huang, V.G. Subramanian, R. Agrawal, and R.A. Berry, "Downlink scheduling and resource allocation for OFDM systems" *IEEE Transactions on Wireless Communications*, vol. 8, no. 1, pp. 288-296, Jan. 2009.
[39] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling policies," in *Proc. of Allerton Conference'2002*.
[40] R. Zhang, "Optimal dynamic resource allocation for multi-antenna broadcasting with heterogeneous delay-constrained traffic," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 2, pp. 243-255, April 2008.
[41] S. Boyd, and L. Vandenberghe, "Convex optimization," Cambridge University Press, 2004.

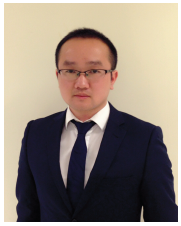
Yuan Wu (S'08-M'10-SM'16) received the Ph.D degree in Electronic and Computer Engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2010. He is an Associate Professor in the College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. During 2016-2017, he was with the Broadband Communications Research (BBCR) group, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests focus on resource management for wireless communications and networks, and smart grid.



Li Ping Qian (S'08-M'10-SM'16) received the Ph.D. degree in information engineering from The Chinese University of Hong Kong, Hong Kong, in 2010. She was with Broadband Communications Research Laboratory, University of Waterloo, from 2016 to 2017. She is currently an Associate Professor with the College of Information Engineering, Zhejiang University of Technology, China. Her research interests lie in the areas of wireless communication and networking, cognitive networks, and smart grids. Dr. Qian was a co-recipient of the IEEE Marconi Prize Paper Award in wireless communications in 2011.

Haowei Mao is currently pursuing his M.S. degree in College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. His research interest focuses on resource management for wireless communications and networks, mobile data offloading, and non-orthogonal multiple access.

Xiaowei Yang is currently pursuing her M.S. degree in College of Information Engineering, Zhejiang University of Technology, Hangzhou, China. Her research interest focuses on resource management for wireless communications and networks, and green communications.



Haibo Zhou received the Ph.D. degree in information and communication engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014. Since 2014, he has been a Post-Doctoral Fellow with the Broadband Communications Research Group, ECE Department, University of Waterloo. He is currently an associate professor with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China. His research interests include resource management and protocol design in cognitive radio networks and vehicular networks.

Xuemin (Sherman) Shen (M'97-SM'02-F'09) is a Professor and University Research Chair, Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research focuses on resource management, wireless network security, social networks, smart grid, and vehicular ad hoc networks. He is an elected member of IEEE ComSoc Board of Governor, and the Chair of Distinguished Lecturers Selection Committee. Dr. Shen served as the Technical Program Committee Chair/Co-Chair for IEEE Globecom16, Infocom14, IEEE VTC10 Fall, and Globecom07, the Symposia Chair for IEEE ICC10, the Tutorial Chair for IEEE VTC'11 Spring and IEEE ICC08, the General Co-Chair for ACM Mobihoc15, the Chair for IEEE Communications Society Technical Committee on Wireless Communications. He also served/serves as the Editor-in-Chief for IEEE Network, Peer-to-Peer Networking and Application, and IET Communications; a Founding Area Editor for IEEE Transactions on Wireless Communications. Dr. Shen received the Excellent Graduate Supervision Award in 2006, and the Outstanding Performance Award from the University of Waterloo, and the Premier's Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. Dr. Shen is a registered Professional Engineer of Ontario, Canada, a Fellow of IEEE, Engineering Institute of Canada, Canadian Academy of Engineering, and Royal Society of Canada, and a Distinguished Lecturer of IEEE Vehicular Technology Society and Communications Society.