Optimal Throughput Fairness Trade-offs for Downlink Non-Orthogonal Multiple Access over Fading Channels

Hong Xing, *Member, IEEE,* Yuanawei Liu, *Member, IEEE,* Arumugam. Nallanathan, *Fellow, IEEE,* Zhiguo Ding, *Member, IEEE,* and H. Vincent Poor *Fellow, IEEE*

***Abstract*—Recently, *non-orthogonal multiple access (NOMA)* has attracted considerable interest as one of the 5G-enabling techniques. However, users with better channel conditions in downlink communicaitons intrinsically benefits from NOMA thanks to successive decoding, judicious designs are required to guarantee user fairness. In this paper, a two-user downlink NOMA system over fading channels is considered. For delay-tolerant transmission, the average sum-rate is maximized subject to both average and peak power constraints as well as a minimum average user rate constraint. The optimal resource allocation is obtained using Lagrangian dual decomposition under full channel state information at the transmitter (CSIT), while an effective power allocation policy under partial CSIT is also developed based on analytical results. In parallel, for delay-limited transmission, the sum of delay-limited throughput (DLT) is maximized subject to a maximum allowable user outage constraint under full CSIT, and the analysis for the sum of DLT is also performed under partial CSIT. Furthermore, an optimal orthogonal multiple access (OMA) scheme is also studied as a benchmark to prove the superiority of NOMA over OMA under full CSIT. Finally, the theoretical analysis is verified by simulations via different trade-offs for the average sum-rate (sum-DLT) versus the minimum (maximum) average user rate (outage) requirement.**

***Index Terms*—Non-orthogonal multiple access, orthogonal multiple access, fairness, fading channel, ergodic rate, outage probability, Lagrangian dual decomposition, strong duality,**

I. Introduction

As the incoming fifth generation (5G) wireless communications features massive connectivity among heterogeneous types of users in the Internet of Things (IoT), non-orthogonal multiple access (NOMA) has been envisioned as a promising candidate for 5G networks [2–4], due to its advantage in enabling high spectral efficiency via non-orthogonal resource allocations over other orthogonal multiple access (OMA) techniques, such as time-division multiple access (TDMA) and frequency-division multiple access (FDMA) (see [5] and the references therein). Hence, it has recently sparked widespread interest in both industry [6, 7] and academia [8–17]. Variation forms of NOMA, namely, multi-user superposition transmission (MUST) and layer division multiplexing (LDM), have been included in the 3rd Generation Partnership Project Long Term Evolution Advanced (3GPP-LTE-A) [6] and the next general digital TV standard ATSC 3.0 [7], respectively.

Among a variety of studies addressing the challenges posed by NOMA, a general NOMA downlink framework was proposed in [8] in which a base station (BS) is capable of simultaneously communicating with several randomly deployed users. To increase the throughput of cell-edge users in multi-cell NOMA networks, coordinated multi-point (CoMP) transmission techniques were adopted in [9] and [10] with the BS equipped with a single antenna and multiple antennas, respectively.

On another front, far users that suffer from severe path-loss attenuation are usually disadvantaged in competing for resources enhancing the sum throughput of the system, and therefore their performance could be substantially compromised without proper design. Multi-user proportional fairness was adopted in [2] as a scheduling metric to achieve a good trade-off between system throughput and user fairness. There are mainly three types of countermeasures against such unfairness in NOMA networks. The *first strategy* is to invoke cooperative NOMA [11, 12], in which a nearby user is regarded as a relay to assist a distant user. It is demonstrated in [11] that by utilizing the proposed cooperative protocol, all users experience the same diversity order. In [12] the nearby NOMA users are equipped with wireless energy harvesting capability to assist far users. The *second strategy* is to enhance the worst user performance [13–15]. The max-min power allocation problem that maximizes the minimum achievable user rate was studied for single-input-single-output (SISO) NOMA systems in [13], and for clustered multiple-input-multiple-output (MIMO) NOMA systems in [14]. In [15], the authors provided a mathematical proof for NOMA's superiority over conventional OMA transmission in terms of the optimum sum rate subject to a minimum rate constraint. The *third strategy* is to introduce additional factors to guarantee fairness. Weighted sum-rate is an effective metric to reflect the priority of users in resource allocation [16, 17]. [16] considered a mutli-carrier downlink network, in which each sub-channel can be shared by multiple users by adopting NOMA. Joint sub-channel

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) of UK under Grant EP/N005651/2. Part of this paper was accepted by IEEE Wireless Communications and Networking Conference (WCNC), Barcelona, Spain, April, 2018 [1].

H. Xing, Y. Liu, and A. Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, E1 4NS, U.K. (e-mails: h.xing@qmul.ac.uk, yuanwei.liu@qmul.ac.uk, nallanathan@ieee.org).

Z. Ding and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA (e-mail: poor@princeton.edu).

Z. Ding is also with the School of Computing and Communications, Lancaster University, Lancaster, LA1 4YW, U.K. (e-mail: z.ding@lancaster.ac.uk).

and power allocation was formulated as a weighted sum-rate maximization problem, and iteratively solved by leveraging a matching problem with externalities. multi-carrier NOMA systems employing a full-dupex (FD) BS was considered in [17], and an optimal joint sub-carrier and power policy for maximizing the weighted sum-rate was developed by applying monotonic optimization.

A. Related Work

Fairness issues were studied for NOMA over fading channels in the above work. However, they were either considered in a long term with fixed power allocations, e.g., in [8, 12], or investigated exploiting adaptive allocation of power and/or bandwidth in a short term, e.g., in [16, 17]. By contrast, we consider adaptive resource allocations to channel dynamics for a two-user downlink NOMA over the whole fading process, the system design of which requires satisfying long-term constraints for quality-of-service (QoS) thus posing new challenges compared with short-term objectives. The information theoretic study of fading broadcast channels (BCs) can be traced back to [18] and [19]. Assuming perfect channel state information (CSI) at both the transmitter (Tx) and the receivers (Rxs), dynamic power and rate allocations for various transmission schemes including code division (CD) with and without successive decoding, time division, and frequency division over different fading states were studied for the ergodic capacity region (ECR) and the (zero-) outage capacity region (OCR) in [18] and [19], respectively. The boundaries of the ECRs have been characterized in [18] by solving equivalent weighted sum-rate problems each corresponding to one set of weights. The (zero-) OCRs were inexplicitly characterized by deriving the outage probability regions given a rate vector in [19]. The boundaries of these regions were also obtained by solving equivalent sum-reward maximization problems [19].

While [18] studied the boundary of the ergodic capacity region by solving an equivalent average weighted-sum rate problem subject to an average total power constraint, the optimal throughput fairness trade-off region that we characterize in this paper is obtained by maximizing the average sum rate subject to a minimum average rate constraint in addition to average and/or peak power constraints. Therefore, the single-variable Lagrangian multiplier employed to decide the "water-filing" power level therein is not readily applicable to our proposed problem. With more Lagrangian multipliers involved, our formulated Lagrangian can be decoupled into many (equal to the total number of fading states) subproblems, which can thus be solved in a parallel fashion with high efficiency. On the other hand, in [19], assuming that the transmission to each user is independent, for each joint fading state, an outage was declared when a given rate vector cannot be maintained for all the users using CD either with or without successive decoding. By contrast, we considered a more general scenario in which, for example, under full CSI at the Tx (CSIT), even if the user with better channel condition fails to decode the weak user's message using successive decoding, it is still possible to directly retrieve its own treating the interference from the weak user as noise. Furthermore, unlike [19] that defined the usage probability via the power set of the users, we equivalently reformulate this continuous variable by arithmetic operation over multiple discrete variables via an indicator function [20].

B. Motivation and Contributions

Since the performance of users with disadvantage channel conditions over multi-user fading BC tends to be compromised for the objective of mere sum-throughput maximization, we aim for maximizing the sum throughput of these systems while satisfying the QoS of the worst user. The classical results derived in the above work are nevertheless not readily extendible to problems with minimum ergodic rate constraints in delay-tolerant scenarios or those with maximum outage constraints in delay-limited scenarios. Although [21] investigated the minimum-rate capacity region taking fairness into account, it imposes the minimum rate constraint in every fading state, which may require quite complex encoding/coding design (see Section IV. B of [21]). Furthermore, other than the modified "water-filling" based optimal power allocation procedure [18, 22] that requires iteratively selecting the "best" Rx for each fading state, in this paper we are interested in optimal solution that can be obtained more efficiently, e.g., by solving a series of subproblems in parallel.

Motivated by these new challenges, we study the average sum-rate and/or the sum of delay-limited throughput (DLT) maximization subject to user fairness for a two-user downlink NOMA system over fading channels. The main contributions of this paper are summarized as follows. We 1) solve the ergodic sum-rate (ESR) maximization problem ensuring a minimum average user rate by optimally adapting the power and rate allocations to fading states with full CSIT for both NOMA and an optimal OMA scheme; 2) obtain the optimal power control to the sum of DLT maximization problem, which is subject to a maximum permissive user outage constraint, with full CSIT for both NOMA and the optimal OMA scheme; 3) under full CSIT, prove the superiority of NOMA over OMA in terms of the considered metrics; 4) under partial CSIT, analyse the ESR and the DLT, respectively, in closed-form with the static power allocation and/or proportion of orthogonal resources designed; and 5) characterize the optimal average sum-rate (sum-DLT) versus min-rate (max-outage) trade-offs for different transmission schemes via simulations.

The remainder of the paper is organized as follows: Section II introduces the system model and the corresponding performance metrics. In Section III, the average sum-rate is maximized subject to transmit power constraints as well as a minimum average user rate constraint under full and partial CSIT, respectively, while in Section IV, the sum of DLT is maximized subject to transmit power constraints as well as a maximum user outage constraint. Numerical results are provided in V. Finally, Section VI concludes the paper.

Notation—We use upper-case boldface letters for matrices and lower-case boldface letters for vectors. $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ denotes the gradient of $f(\boldsymbol{x})$ with respect to (w.r.t.) $\boldsymbol{x}$. $\mathbb{E}_x[\cdot]$ stands for the statistical expectation w.r.t. the random variable (RV) x. $\sim$ represents "distributed as" and $\triangleq$ means "denoted by". The circularly symmetric complex Gaussian (CSCG) distribution with mean u and variance σ^2 is denoted by $\mathcal{CN}(u, \sigma^2)$. $\mathrm{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t} \mathrm{d}t$ $(x < 0)$ is the exponential integral

function of argument x. In addition, $(x)^+ = \max(0, x)$ and $[x]_a^b = \max(\min(x, b), a)$.

II. System Model

We consider a simplified single-carrier downlink cellular system that consists of one BS and two users[1], denoted by $\mathcal{U}_k$, $k \in \{1, 2\}$, as shown in Fig. 1. Both the BS and the users are assumed to be equipped with single antenna. We assume that the complex channel coefficient from the BS to $\mathcal{U}_k$, $h_k(\nu)$ experiences block fading with a continuous joint probability density function (pdf), where ν represents a fading state. The channel remains constant during each transmission block, but may vary from block to block as ν changes[2]. The channel gain $|h_k(\nu)|^2$ is assumed to consist of multiplicative small scale and large scale fading given by $|h_k(\nu)|^2 = \frac{|\bar{h}_k(\nu)|^2}{\lambda_k}$, in which $\bar{h}_k(\nu)$ is a complex Gaussian RV denoted by $\bar{h}_k(\nu) \sim \mathcal{CN}(0, 1)$, and λ_k is a distant-dependent constant. Hence, $|h_k(\nu)|^2$ is an exponentially distributed RV with its mean value specified by $1/\lambda_k$.

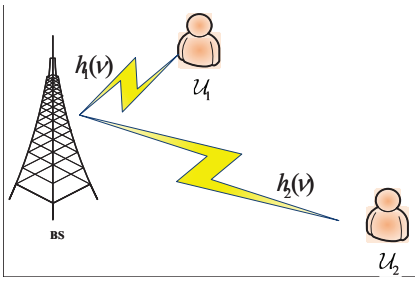


Fig. 1. System model for a two-user downlink NOMA.

A. Full CSIT

In this paper, we investigate two types of CSIT, i.e., full CSIT and partial CSIT, while CSI at the Rxs is assumed to be perfectly known. When full CSIT is available, the BS can adapt its power and rate of the transmit signal intended for each user to the channel $h_k(\nu)$'s in each fading state. On the other hand, when only partial CSIT including the order of the two channel gains and their channel distribution information (CDI) is available, due to some reasons like limited feedback from the users to the BS or reducing signalling for the purpose of reducing overhead, the BS can only determine its power allocation policy at each fading state based on this order. We also consider two different multiple access transmission schemes, viz., NOMA and optimal OMA. In the NOMA transmission scheme, the two users non-orthogonally access the channel by enabling superposition coding (SC) at the BS and SIC at the users. For optimal OMA transmission, we consider power and (continuous) time/frequency allocation both in an adaptive manner, which is referred as *OMA-TYPE-II* [15]. (Another benchmark scheme, *OMA-Type-I*, will be introduced in Section V.)

1) NOMA: For NOMA transmission, the received signal at the downlink user $\mathcal{U}_k$ is given by

$$y_k(\nu) = \sqrt{p_k(\nu)} h_k(\nu) s_k + \sqrt{p_{\bar{k}}(\nu)} h_k(\nu) s_{\bar{k}} + n_k, \quad (1)$$

where $\bar{k}$ denotes the element in the complementary set of $\{1, 2\}$ w.r.t. k; s_k's is the transmit signal intended for $\mathcal{U}_k$'s, denoted by $s_k \sim \mathcal{CN}(0, 1)$[3]; $p_k(\nu)$'s denotes $\mathcal{U}_k$'s transmit power; and n_k's is the AWGN at $\mathcal{U}_k$'s Rx, denoted by $n_k \sim \mathcal{CN}(0, \sigma_k^2)$.

We also define $\mathcal{U}_k$'s achievable rate for decoding $\mathcal{U}_{\bar{k}}$'s message at fading state ν in bits/sec/Hz, $k \in \{1, 2\}$, treating interference as noise (TIAN), as follows.

$$R_{k\to\bar{k}}^{\mathrm{NOMA}}(\nu) = \log_2\left(1 + \frac{p_{\bar{k}}(\nu) g_k(\nu)}{p_k(\nu) g_k(\nu) + 1}\right), \quad (2)$$

where $g_k(\nu)$ is the normalized channel gain given by $\frac{|h_k(\nu)|^2}{\sigma_k^2} \triangleq g_k(\nu)$. Similarly, the achievable rate for $\mathcal{U}_k$ to decode its own message by TIAN is given by

$$R_{k\to k}^{\mathrm{NOMA}}(\nu) = \log_2\left(1 + \frac{p_k(\nu) g_k(\nu)}{p_{\bar{k}}(\nu) g_k(\nu) + 1}\right). \quad (3)$$

If $g_k(\nu) > g_{\bar{k}}(\nu)$, it implies that $R_{k\to\bar{k}}^{\mathrm{NOMA}}(\nu) > R_{\bar{k}\to\bar{k}}^{\mathrm{NOMA}}(\nu)$. In other words, under this condition, the achievable rate for $\mathcal{U}_k$ (the stronger user) to decode the message of $\mathcal{U}_{\bar{k}}$ (the weaker user) is larger than that intended for $\mathcal{U}_{\bar{k}}$'s transmission, and therefore $\mathcal{U}_k$ is able to successfully perform SIC. Otherwise, $\mathcal{U}_k$ is only able to decode its own message by TIAN. To sum up, the instantaneous achievable rate for $\mathcal{U}_k$'s is thus given by [18]

$$R_k^{\mathrm{NOMA}}(\nu) = \begin{cases} \log_2\left(1 + p_k(\nu) g_k(\nu)\right), & \text{if } g_k(\nu) > g_{\bar{k}}(\nu), \\ R_{k\to k}^{\mathrm{NOMA}}(\nu), & \text{otherwise.} \end{cases} \quad (4)$$

Moreover, similar to [21], we simultaneously consider two types of transmit power constraints on p_k's, namely, average power constraint (APC) and peak power constraint (PPC), in which the former constrains the total transmit power in the long term, i.e., $\mathbb{E}_\nu[p_k(\nu) + p_{\bar{k}}(\nu)] \le \bar{P}$, and the latter limits the instantaneous total transmit power below $\hat{P}$, i.e., $p_k(\nu) + p_{\bar{k}}(\nu) \le \hat{P}$, $\forall \nu$. It is assumed that $\bar{P} \le \hat{P}$ without loss of generality (w.l.o.g.).

[1]We consider a single-carrier multi-user downlink NOMA with only two users for the following two factors. First, the two-user case is practically favourable to industry [23], since the delay incurred in successive interference cancellation (SIC) is significantly reduced. Second, insights for system design can be drawn easily from the two-user solution, while general solution with more than two users can also be obtained under full CSIT without much difficulty. In addition, more complex design for multi-carrier NOMA can be applied to each transmission block considered herein, but is beyond the scope of this paper. The interested reader can refer to [16, 17] for multi-carrier based transmission schemes in NOMA.

[2]Note that the "block fading" herein refers to slow fading scenarios in which the channel remains constant within each block length such that short-length coding schemes are applicable.

[3]Note that in real communications system with transmitted signals drawn from finite-alphabet (i.e., discrete) constellations and uniform distribution, the associated encoding/decoding schemes must be judiciously designed such that SIC detector is performed to satisfied level [24, 25]. However, the associated design is beyond the scope of this paper, and is left as an interesting future direction.

2) OMA-Type-II: For OMA-Type-II transmission, each user receives its information over $\alpha_k(\nu)$ of the time/frequency dedicated to it in fading state ν, such that $\alpha_k(\nu)+\alpha_{\bar{k}}(\nu)=1$, where $\alpha_k(\nu) \in [0,1]$. The same sets of transmit power constraints as in its NOMA counterpart, i.e., APC and PPC, are taken into account as well. Accordingly, the instantaneous achievable rate for $\mathcal{U}_k$'s in fading state ν is given by[4]

$$R_k^{\text{OMA-II}}(\nu) = \alpha_k(\nu)\log_2\left(1+\frac{p_k(\nu)g_k(\nu)}{\alpha_k(\nu)}\right). \quad (5)$$

Note that (5) applies to both TDMA and FDMA transmission in the sense that the total energy consumed for the two users in fading state ν over time remains the same as that over frequency, which is given by $\alpha_k(\nu)\frac{p_k(\nu)}{\alpha_k(\nu)}+\alpha_{\bar{k}}(\nu)\frac{p_{\bar{k}}(\nu)}{\alpha_{\bar{k}}(\nu)} = p_k(\nu)+p_{\bar{k}}(\nu)$, $\forall\nu$.

B. Partial CSIT

1) NOMA: Under partial CSIT, for NOMA transmission, the BS does not know the exact CSI of the two users due to insufficient channel estimation but their relation, i.e., whether $g_k(\nu) \le g_{\bar{k}}(\nu)$ or $g_k(\nu) < g_{\bar{k}}(\nu)$, and statistical characteristics, and therefore the Tx cannot dynamically adjust the allocation of power, rate and/or time/frequency resources to each fading state as in full CSIT. Hence, we adopt a binary power allocation strategy depending on which user has better CSI[5]. Specifically, in each fading state ν, an amount of power p_s is always assigned to the stronger user while p_w is assigned to the other weaker user. We also assume that p_s and p_w are static over all fading states, and therefore only APC applies, i.e., $p_s+p_w \le \bar{P}$. In this case, the instantaneous rate $R_k'^{\text{NOMA}}(\nu)$'s for $\mathcal{U}_k$'s is expressed as

$$R_k'^{\text{NOMA}}(\nu) = \begin{cases} \log_2\left(1+p_s g_k(\nu)\right), & \text{if } g_k(\nu) > g_{\bar{k}}(\nu), \\ \log_2\left(1+\frac{p_w g_k(\nu)}{p_s g_k(\nu)+1}\right), & \text{otherwise.} \end{cases} \quad (6)$$

2) OMA-Type-II: Similarly, for OMA-Type-II transmission, the binary allocation policy with a fixed sharing of time/frequency between the two users is adopted. Specifically, the signal intended for $\mathcal{U}_k$ is transmitted with power p_s if its channel gain from the BS is stronger than $\mathcal{U}_{\bar{k}}$'s, and with power p_w otherwise. The fixed proportion of time/frequency assigned to $\mathcal{U}_k$ and $\mathcal{U}_{\bar{k}}$ is α_k and $\alpha_{\bar{k}}$, respectively. Consequently, the instantaneous achievable rate for $\mathcal{U}_k$'s is expressed as

$$R_k'^{\text{OMA-II}}(\nu) = \begin{cases} \alpha_k\log_2\left(1+\frac{p_s g_k(\nu)}{\alpha_k}\right), & \text{if } g_k(\nu) > g_{\bar{k}}(\nu), \\ \alpha_k\log_2\left(1+\frac{p_w g_k(\nu)}{\alpha_k}\right), & \text{otherwise.} \end{cases} \quad (7)$$

C. System Throughput

1) Delay-Tolerant Transmission: First, for "delay-tolerant" transmission, we refer it to the scenario in which no delay constraints are imposed for decoding, and thus the codeword can be designed arbitrarily long (approaching infinity in theory) spanning over all the fading states, and decoded until it is received in its full length. The associated performance metric for each user is *ergodic rate* [18], at which the Tx delivers the intended data for each user over the entire fading process. Consequently, the ergodic sum-rate (ESR) of the two users are given by $\mathbb{E}_\nu[R_k^{\text{NOMA}}(\nu)+R_{\bar{k}}^{\text{NOMA}}(\nu)]$ ($\mathbb{E}_\nu[R_k'^{\text{NOMA}}(\nu)+R_{\bar{k}}'^{\text{NOMA}}(\nu)]$), and $\mathbb{E}_\nu[R_k^{\text{OMA-II}}(\nu)+R_{\bar{k}}^{\text{OMA-II}}(\nu)]$ ($\mathbb{E}_\nu[R_k'^{\text{OMA-II}}(\nu)+R_{\bar{k}}'^{\text{OMA-II}}(\nu)]$), for NOMA and OMA-Type-II transmission, respectively[6].

2) Delay-Limited Transmission: Next, consider the delay-limited types of transmission for downlink NOMA and/or OMA-Type-II system. We relax the classical information theoretic "zero-outage" definition in [26]. Other than maintain a constant rate vector at all fading states via power control, we refer this notion to the scenario in which delay-sensitive data such as video streaming requires to be correctly decoded at a constant rate at the end of every fading state[7]. The associated performance metric for each user is *outage probability*, which measures the percentage of fading states at which a predefined constant rate cannot be supported.

Specifically, under full CSIT, the outage probability for user $\mathcal{U}_k$ with the target rate $\bar{R}_k$, $k\in\{1,2\}$, is introduced as below.
Case 1: $g_k(\nu) > g_{\bar{k}}(\nu)$

$$\zeta_k^{\text{NOMA}} = \Pr\left\{R_{k\to\bar{k}}^{\text{NOMA}}(\nu) < \bar{R}_{\bar{k}}, R_{k\to k}^{\text{NOMA}}(\nu) < \bar{R}_k\right\} + \Pr\left\{R_{k\to\bar{k}}^{\text{NOMA}}(\nu) \ge \bar{R}_{\bar{k}}, R_k^{\text{NOMA}}(\nu) < \bar{R}_k\right\}. \quad (8)$$

Case 2: $g_k(\nu) \le g_{\bar{k}}(\nu)$

$$\zeta_k^{\text{NOMA}} = \Pr\left\{R_k^{\text{NOMA}}(\nu) < \bar{R}_k\right\}. \quad (9)$$

As seen from (8), when $\mathcal{U}_k$ has better channel condition, whether its signal-to-noise ratio (SNR) or signal-to-interference-plus-noise ratio (SINR) leads to its outage depends on whether or not it manages to recover $\mathcal{U}_{\bar{k}}$'s message. If it fails to retrieve $\mathcal{U}_{\bar{k}}$'s message at the predefined transmission rate for $\mathcal{U}_{\bar{k}}$, i.e., $\bar{R}_{\bar{k}}$, it has to decode its own by TIAN. Otherwise, if it succeeds in decoding $\mathcal{U}_{\bar{k}}$'s message, SIC is performed before it decodes its own interference-free. On the other hand, when $\mathcal{U}_k$ has worse channel condition, it always decodes its own by TIAN (c.f. (9)).

In addition, at each fading state ν, an outage indicator function is defined as follows [20].
Case 1: $g_k(\nu) > g_{\bar{k}}(\nu)$

$$X_k^{\text{NOMA}}(\nu) = \begin{cases} 1, & \text{if } R_{k\to\bar{k}}^{\text{NOMA}}(\nu) < \bar{R}_{\bar{k}},\ R_{k\to k}^{\text{NOMA}}(\nu) < \bar{R}_k, \\ 1, & \text{if } R_{k\to\bar{k}}^{\text{NOMA}}(\nu) \ge \bar{R}_{\bar{k}},\ R_k^{\text{NOMA}}(\nu) < \bar{R}_k, \\ 0, & \text{otherwise.} \end{cases} \quad (10)$$

[4]If $\alpha_k(\nu)=0$, we define $R_k^{\text{OMA-II}}(\nu)=0$, since $\lim_{\alpha_k\to 0^+} R_k^{\text{OMA-II}}=0$.

[5]Such power policy under partial CSIT is not necessarily optimal but provided as performance lower-bound in comparison with its counterpart under full CSIT.

[6]Since the analysis developed for "delay-tolerant" transmission may also apply to scenarios, in which the short-length codewords are detected at each user on a block basis and the average sum-rate is used to measure the achievable sum-rate in the long term, we do not explicitly differentiate the two terms, "ESR" and "average sum-rate", throughout the paper.

[7]We assume in the "delay-limited" transmission that SIC can be perfectly performed during one block, which is hardly true in practice and thus provides theoretical upper-bound for the achievable sum of DLT. This assumption may be lifted by explicitly considering imperfect SIC as in [27] in our future work.

Case 2: $g_k(\nu) \leq g_{\bar{k}}(\nu)$

$$X_k^{\mathrm{NOMA}}(\nu) = \begin{cases} 1, & \text{if } R_{k\to k}^{\mathrm{NOMA}}(\nu) < \bar{R}_k, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Combining (10) (c.f. (8)) and (11) (c.f. (9)), it is easily verified that $\mathbb{E}_\nu[X_k^{\mathrm{NOMA}}(\nu)] = \zeta_k^{\mathrm{NOMA}}$, $k \in \{1,2\}$.

For OMA-Type-II transmission, the outage probability of $\mathcal{U}_k$'s is defined independent of the other as follows:

$$\zeta_k^{\text{OMA-II}} = \Pr\left\{R_k^{\text{OMA-II}}(\nu) < \bar{R}_k\right\}. \tag{12}$$

By analogy, we introduce the following indicator function for $\mathcal{U}_k$ w.r.t. the target rate $\bar{R}_k$:

$$X_k^{\text{OMA-II}}(\nu) = \begin{cases} 1, & \text{if } R_k^{\text{OMA-II}}(\nu) < \bar{R}_k, \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

It also follows that $\mathbb{E}_\nu[X_k^{\text{OMA-II}}(\nu)] = \zeta_k^{\text{OMA-II}}$, $k \in \{1,2\}$.

Accordingly, one relevant metric to assess the overall performance in delay-limited case is the *sum of DLT* expressed as $\bar{R}_k(1-\zeta_k^{\mathrm{NOMA}})+\bar{R}_{\bar{k}}(1-\zeta_{\bar{k}}^{\mathrm{NOMA}})$, and $\bar{R}_k(1-\zeta_k^{\text{OMA-II}})+\bar{R}_{\bar{k}}(1-\zeta_{\bar{k}}^{\text{OMA-II}})$, for NOMA and OMA-Type-II transmission under full CSIT, respectively. The sum of DLT for NOMA and OMA-Type-II transmission under partial CSIT is also similarly given by $\bar{R}_k(1-\zeta_k'^{\mathrm{NOMA}})+\bar{R}_{\bar{k}}(1-\zeta_{\bar{k}}'^{\mathrm{NOMA}})$, and $\bar{R}_k(1-\zeta_k'^{\text{OMA-II}})+\bar{R}_{\bar{k}}(1-\zeta_{\bar{k}}'^{\text{OMA-II}})$, respectively.

III. Optimum Delay-Tolerant Transmission

In delay-tolerant scenarios, to maximize the ESR of the system while guaranteeing certain level of fairness, a minimum achievable ergodic rate requirement for each user, namely, $\mathbb{E}_\nu[R_k^{\mathrm{XX}}(\nu)] \geq \bar{R}$ ($\mathbb{E}_\nu[R_k'^{\mathrm{XX}}(\nu)] \geq \bar{R}'$), $k \in \{1,2\}$, is imposed, where $(\cdot)^{\mathrm{XX}}$ denotes the multiple access scheme that is specified in the context throughout the paper. In this section, the optimal trade-off between the system ESR and user fairness is pursued in the case of full and partial CSIT, respectively. Particularly, under full CSIT, the ESR maximization problems are solved using Lagrangian dual decomposition levering "time-sharing" conditions, while under partial CSIT, individual user's ergodic rate needs to be first analysed in closed form by means of CDFs of the related SNR and/or SNRs.

A. Full CSIT

In the case of full CSIT, the design objective is to maximize the system ESR by jointly optimizing the power and/or orthogonal resource allocations, and the two users' instantaneous rate at each fading state, subject to both APC and PPC at the BS, as well as a minimum ergodic rate constraint for the two users. As a result, the optimization problem is formulated as follows[8].

$$\begin{aligned} (\text{P1-XX}): \quad & \underset{\{p_k(\nu),p_{\bar{k}}(\nu),\alpha_k(\nu)\}}{\texttt{Maximize}} \ \mathbb{E}_\nu[R_k^{\mathrm{XX}}(\nu)+R_{\bar{k}}^{\mathrm{XX}}(\nu)] \\ & \texttt{Subject to} \\ & \mathbb{E}_\nu[p_k(\nu)+p_{\bar{k}}(\nu)] \leq \bar{P}, && \text{(14a)} \\ & p_k(\nu)+p_{\bar{k}}(\nu) \leq \hat{P}, \ \forall\nu, && \text{(14b)} \\ & p_k(\nu) \geq 0, \ p_{\bar{k}}(\nu) \geq 0, \ \forall\nu, && \text{(14c)} \\ & \mathbb{E}_\nu[R_k^{\mathrm{XX}}(\nu)] \geq \bar{R}, \ \forall k, && \text{(14d)} \end{aligned}$$

[8]Note that $\mathbb{E}[\cdot]$ in (P1-XX) is evaluated by the sum of the associated instantaneous function of ν divided by the total number of fading states N, assuming that N is large enough such that $N \to \infty$.

where the exclusive parameters for OMA-Type-II, $\{\alpha_k(\nu)\}$'s, are only valid when XX refers to OMA-Type-II. In the following, we develop optimal solution to (P1-NOMA) and (P1-OMA-II), respectively.

1) Optimal Solution to (P1-NOMA)*:* Problem (P1-NOMA) is non-convex due to the non-convex objective function (c.f. (4)), and therefore no immediate solution can be given. However, for channel fading following continuous distributions, (P1-NOMA) proves to satisfy the "time-sharing" condition[9]. Note that if (P1-NOMA) satisfies the "time-sharing condition", then it has a zero duality gap between the primal and the dual problem using Lagrangian duality [28, Theorem 1], i.e., *strong duality* [29] holds, despite of the convexity of the problem itself. Hence, we can still optimally solve it via its dual problem.

Next, we apply the Lagrangian dual method to solve (P1-NOMA), the Lagrangian of which is given by

$$\begin{aligned} &\mathcal{L}_1^{\mathrm{NOMA}}(\{p_k(\nu)\},\{p_{\bar{k}}(\nu)\},\lambda,\delta,\mu) = \\ &\mathbb{E}_\nu[(1+\delta)R_k^{\mathrm{NOMA}}(\nu)+(1+\mu)R_{\bar{k}}^{\mathrm{NOMA}}(\nu)-\lambda(p_k(\nu)+p_{\bar{k}}(\nu))] \\ &\qquad + \lambda\bar{P} - \delta\bar{R} - \mu\bar{R}, \end{aligned} \tag{15}$$

where λ is the Lagrangian multiplier associated with the APC given in (14a); δ and μ are those associated with the ergodic rate constraints given in (14d) for $\mathcal{U}_k$ and $\mathcal{U}_{\bar{k}}$, respectively. The dual function of (P1-NOMA) corresponding to (15) is accordingly given by

$$\begin{aligned} &g(\lambda,\delta,\mu) = \max \mathcal{L}_1^{\mathrm{NOMA}}(\{p_k(\nu)\},\{p_{\bar{k}}(\nu)\},\lambda,\delta,\mu), \\ &\texttt{s.t.} \ \ p_k(\nu) \geq 0, p_{\bar{k}}(\nu) \geq 0, p_k(\nu)+p_{\bar{k}}(\nu) \leq \hat{P}, \ \forall\nu. \end{aligned} \tag{16}$$

The dual problem of (P1-NOMA) is thus formulated as

$$(\text{P1-NOMA-dual}): \ \underset{\lambda\geq 0,\delta\geq 0,\mu\geq 0}{\texttt{Minimize}} \quad g(\lambda,\delta,\mu).$$

It is observed that $g(\lambda,\delta,\mu)$ is obtained by maximizing the Lagrangian given in (15), which can be decoupled into as many subproblems as the number of fading states all sharing the same structure. The index ν is now safely dropped for the ease of exposition. Taking one particular fading state as an example, the associated subproblem given a triple (λ,δ,μ) can be expressed as

$$\begin{aligned} (\text{P1-NOMA-sub}): \ & \underset{p_k\geq 0,p_{\bar{k}}\geq 0}{\texttt{Maximize}} \ \ \bar{\mathcal{L}}_1^{\mathrm{NOMA}}(p_k,p_{\bar{k}}) \\ & \texttt{Subject to} \ \ p_k+p_{\bar{k}} \leq \hat{P}, \end{aligned}$$

where $\bar{\mathcal{L}}_1^{\ \mathrm{NOMA}}(p_k,p_{\bar{k}}) = (1+\delta)R_k^{\mathrm{NOMA}}+(1+\mu)R_{\bar{k}}^{\mathrm{NOMA}}-\lambda(p_k+p_{\bar{k}})$. Since these problems are independent of each other, they can be solved in parallel each for one fading state. Therefore, w.l.o.g., we focus on solving (P1-NOMA-sub) in the sequel.

[9]The original definition of "time-sharing" condition is given by [28, Definition 1], which essentially implies that the maximum value of the optimization problem (P1-NOMA) is a joint concave function of $\bar{P}$ and $\bar{R}$. The proof is rather standard and thus omitted herein for brevity.

Proposition 3.1: The optimal power allocation to Problem (P1-NOMA-sub) assuming $g_1 > g_2$ is given by[10]

$$(p_1^*, p_2^*) = \arg\max\{\bar{\mathcal{L}}_1^{\text{NOMA}}(0,0), \bar{\mathcal{L}}_1^{\text{NOMA}}(0,\hat{P}), \bar{\mathcal{L}}_1^{\text{NOMA}}(\hat{P},0), \bar{\mathcal{L}}_1^{\text{NOMA}}(p_{i,1},p_{i,2})\},\ i \in \{1,2,3,4\}, \quad (17)$$

where $(p_{i,1}, p_{i,2})$, $i \in \{1,2,3,4\}$, are given at the top of the next page with each corresponding to one solution pair given by (19).

Proof: Since $\bar{\mathcal{L}}_1^{\text{NOMA}}(p_1,p_2)$ is a continuous function over $\Psi = \{(p_1,p_2)|p_1 \geq 0, p_2 \geq 0, p_1+p_2 \leq \hat{P}\}$, its maximum proves to be either at the stationary point, denoted by $(p_{4,1}, p_{4,2})$, or on the boundary of Ψ depending on whether $(p_{4,1}, p_{4,2}) \in \Psi$ or not. We calculate $(p_{4,1}, p_{4,2})$ as follows:

$$(p_{4,1}, p_{4,2}) = \arg\left\{\nabla_{(p_1,p_2)} \bar{\mathcal{L}}_1^{\text{NOMA}}(p_1,p_2) = \mathbf{0}\right\}. \quad (18)$$

If $(p_{4,1}, p_{4,2}) \in \Psi$, the maximum is $\bar{\mathcal{L}}_1^{\text{NOMA}}(p_{4,1}, p_{4,2})$, otherwise the maximum can be attained by restricting (p_1,p_2) to the lines $p_1 = 0$, $p_2 = 0$ or $p_1 + p_2 = \hat{P}$. The stationary points on these lines are denoted by $(p_{i,1}, p_{i,2})$'s, $i = 1$, 2 and 3, respectively. ∎

Note that Proposition 3.1 assumes $g_1 > g_2$ for the ease of exposition though, its results also apply to the fading states where $g_1 < g_2$ by simply exchanging δ, p_1, and g_1 with μ, p_2, and g_2, respectively, in (19). Some optimal system design insights are gained from Proposition 3.1. Considering an extreme case in favour of $\mathcal{U}_2$, in which $\delta \ll \mu$, it is observed from (19) that $p_{4,1}$ monotonically decreases with μ while $p_{4,2}$ monotonically increases with μ, which suggests that when μ associated with $\mathcal{U}_2$'s QoS requirement is sufficiently large, the optimal power allocation policy tends to suppress $\mathcal{U}_1$'s transmission while supporting $\mathcal{U}_2$'s despite of $\mathcal{U}_1$'s channel condition better than $\mathcal{U}_2$.

Thanks to Proposition 3.1, given a triple (λ,δ,μ), $g(\lambda,\delta,\mu)$ is obtained efficiently by solving (P1-NOMA-sub) in parallel over all fading states. (P1-NOMA-dual) can thus be iteratively solved using sub-gradient based methods, e.g., deep-cut ellipsoid method (with constraints) [30, Localization methods]. The required sub-gradient for updating (λ,δ,μ) turns out to be $(\bar{P} - \mathbb{E}_\nu[p_k^*(\nu) + p_{\bar{k}}^*(\nu)], \mathbb{E}_\nu[R_k^{*\text{NOMA}}(\nu)] - \bar{R}, \mathbb{E}_\nu[R_{\bar{k}}^{*\text{NOMA}}(\nu)] - \bar{R})^T$, where $(p_k^*(\nu), p_{\bar{k}}^*(\nu))$ is the optimal solution to (P1-NOMA-sub) at fading state ν, and $R_k^{*\text{NOMA}}(\nu)$'s is obtained by substituting $(p_k^*(\nu), p_{\bar{k}}^*(\nu))$ into (4).

Note that a feasible $\bar{R}$ in (14d) ensures the successful implementation of the ellipsoid method, and thus it is important to consider a reasonable $\bar{R}$ that does not exceed $\bar{R}_{\max}$. We can obtain $\bar{R}_{\max}$ by replacing the objective function of (P1-NOMA) with a variable $\bar{R}$ and then solving the feasibility problem by bi-section over $\bar{R}$. Since the involved procedure is quite similar to that for solving (P1-NOMA), we omit it herein for brevity.

[10]In fact, under full CSIT, following the same method utilized to develop Proposition 3.1, the two-user results can be generalized to more general cases with $K > 2$, since the difficulty of solving (18) does not increase with K.

2) Optimal Solution to (P1-OMA-II)*:* First, (P1-OMA-II) is a convex problem, since (5) as the perspective of the jointly concave function $\log_2(1+p_k(\nu)g_k(\nu))$ proves to be jointly concave w.r.t. $\alpha_k(\nu)$ and $p_k(\nu)$, $k \in \{1,2\}$, $\forall \nu$. As such, we can solicit the Lagrangian dual method to solve (P1-OMA-II) due to strong duality.

The Lagrangian of (P1-OMA-II) is expressed as

$$\mathcal{L}_1^{\text{OMA-II}}(\{p_k(\nu)\},\{p_{\bar{k}}(\nu)\},\{\alpha_k(\nu)\},\lambda,\delta,\mu) = \mathbb{E}_\nu[(1+\delta)R_k^{\text{OMA-II}}(\nu)+(1+\mu)R_{\bar{k}}^{\text{OMA-II}}(\nu)-\lambda(p_k(\nu)+p_{\bar{k}}(\nu))] + \lambda\bar{P} - \delta\bar{R} - \mu\bar{R}, \quad (20)$$

where λ, δ and μ are Lagrangian multipliers associated with the same constraints as those for (P1-NOMA). Similar to the previous section, $\mathcal{L}_1^{\text{OMA-II}}(\{p_k(\nu)\},\{p_{\bar{k}}(\nu)\},\{\alpha_k(\nu)\},\lambda,\delta,\mu)$ can also be decoupled into parallel sub-Lagrangian all having the same structure. We define $\bar{\mathcal{L}}_1^{\text{OMA-II}}(p_k,p_{\bar{k}},\alpha_k) = (1+\delta)R_k^{\text{OMA-II}}+(1+\mu)R_{\bar{k}}^{\text{OMA-II}}-\lambda(p_k+p_{\bar{k}})$. Then the associated subproblem one particular fading state is formulated as

$$\text{(P1-OMA-II-sub)}: \quad \underset{p_k\geq 0, p_{\bar{k}}\geq 0, \alpha_k}{\texttt{Maximize}} \quad \bar{\mathcal{L}}_1^{\text{OMA-II}}(p_k,p_{\bar{k}},\alpha_k)$$
$$\texttt{Subject to} \quad p_k + p_{\bar{k}} \leq \hat{P},$$
$$0 \leq \alpha_k \leq 1,\ \forall k,$$

where the index ν has been dropped for the ease of exposition. To solve (P1-OMA-II-sub), the following two lemmas are required.

Lemma 3.1: If the maximum of $\bar{\mathcal{L}}_1^{\text{OMA-II}}(p_1,p_2,\alpha_1)$ is achieved by its jointly stationary point, it is necessary to have the following conditions satisfied:

$$h(\lambda,\delta,\mu) = 0, \quad (21a)$$
$$c_1 \geq 0, \quad (21b)$$
$$c_2 \geq 0, \quad (21c)$$
$$\begin{cases} \frac{\hat{P}-c_2}{c_1-c_2} \geq 0, & \text{if } c_1 > c_2 \\ \frac{\hat{P}-c_2}{c_1-c_2} \leq 1, & \text{otherwise} \end{cases}, \quad (21d)$$

where $c_1 = \frac{1+\delta}{\lambda \ln 2} - \frac{1}{g_1}$, $c_2 = \frac{1+\mu}{\lambda \ln 2} - \frac{1}{g_2}$, and $h(\lambda,\delta,\mu)$ is given by

$$h(\lambda,\delta,\mu) = (1+\delta)\log_2\left(\frac{1+\delta}{\lambda \ln 2} g_1\right) - (1+\mu)\log_2\left(\frac{1+\mu}{\lambda \ln 2} g_2\right) - \lambda c_1 + \lambda c_2. \quad (22)$$

The corresponding stationary point is given by

$$p_1^* = c_1\alpha_1^*,\ p_2^* = c_2(1-\alpha_1^*), \quad (23)$$

and

$$\alpha_1^* = \begin{cases} \forall \in [0, \min\{\frac{\hat{P}-c_2}{c_1-c_2}, 1\}], & \text{if } c_1 > c_2, \\ \forall \in [(\frac{\hat{P}-c_2}{c_1-c_2})^+, 1], & \text{otherwise.} \end{cases} \quad (24)$$

Proof: First, solve $\nabla_{(p_1,p_2,\alpha_1)} \bar{\mathcal{L}}_1^{\text{OMA-II}}(p_1,p_2,\alpha_1) = \mathbf{0}$ to obtain the jointly stationary point. Next, by plugging $p_1 = c_1\alpha_1$ and $p_2 = c_2(1-\alpha_1)$ into the partial derivative of $\bar{\mathcal{L}}_1^{\text{OMA-II}}(p_1,p_2,\alpha_1)$ w.r.t. α_1, (21a) is obtained. Finally, constrain $P_1 \geq 0$, $p_2 \geq 0$, and $p_1+p_2 \leq \hat{P}$, we arrive at (21b),

$$\begin{cases} p_{1,1}=0, & p_{1,2}=\left[\frac{1+\mu}{\lambda\ln 2}-\frac{1}{g_2}\right]_0^{\hat{P}} \\ p_{2,1}=\left[\frac{1+\delta}{\lambda\ln 2}-\frac{1}{g_1}\right]_0^{\hat{P}}, & p_{2,2}=0 \\ p_{3,1}=\left[\frac{(1+\mu)/g_1-(1+\delta)/g_2}{\delta-\mu}\right]_0^{\hat{P}}, & p_{3,2}=\left[\hat{P}-\frac{(1+\mu)/g_1-(1+\delta)/g_2}{\delta-\mu}\right]_0^{\hat{P}} \\ \begin{cases} p_{4,1}=\frac{(1+\mu)/g_1-(1+\delta)/g_2}{\delta-\mu} \\ p_{4,2}=\frac{1+\mu}{\lambda\ln 2}-\frac{1}{g_2}-\frac{(1+\mu)/g_1-(1+\delta)/g_2}{\delta-\mu} \end{cases}, & \text{if } p_{4,1}\geq 0, p_{4,2}\geq 0, p_{4,1}+p_{4,2}\leq\hat{P} \\ & \text{N/A. otherwise} \end{cases} \tag{19}$$

(21c), and the feasible range for α_1 given in (24), respectively. ∎

Lemma 3.2: If the maximum of $\bar{\mathcal{L}}_1^{\text{OMA-II}}(p_1,p_2,\alpha_1)$ is achieved by points on the boundary $p_1+p_2=\hat{P}$, the optimum (p_1,p_2,α_1) turns out to be

$$\begin{cases} p_1^*=0,\ p_2^*=\hat{P},\ \alpha_1^*=0, & \text{if } \frac{1+\mu}{1+\delta}>\frac{\log_2(1+\hat{P}g_1)}{\log_2(1+\hat{P}g_2)} \\ p_1^*=\hat{P},\ p_2^*=0,\ \alpha_1^*=1. & \text{otherwise} \end{cases} \tag{25}$$

Proof: Please refer to a longer version of this paper [31, Appendix A]. ∎

Based on Lemma 3.1 and Lemma 3.2, the following proposition is derived.

Proposition 3.2: The optimal power as well as time/frequency allocation to (P1-OMA-II-sub) is given by

$$\begin{aligned}(p_1^*,p_2^*,\alpha_1^*)=\arg\max\{&\bar{\mathcal{L}}_1^{\text{OMA-II}}(0,0,0),\\ &\bar{\mathcal{L}}_1^{\text{OMA-II}}(0,\hat{P},0),\bar{\mathcal{L}}_1^{\text{OMA-II}}(\hat{P},0,1),\\ &\bar{\mathcal{L}}_1^{\text{OMA-II}}(0,c_2,0)\mathbb{1}_{c_2},\bar{\mathcal{L}}_1^{\text{OMA-II}}(c_1,0,1)\mathbb{1}_{c_1}\},\end{aligned} \tag{26}$$

where $\mathbb{1}_{(\cdot)}$ is an indicator function defined as

$$\mathbb{1}_x=\begin{cases} 1, & \text{if } 0\leq x\leq\hat{P} \\ 0. & \text{otherwise} \end{cases} \tag{27}$$

Proof: Please refer to a longer version of this paper [31, Appendix B]. ∎

Remark 3.1: When there are only two users, the optimal solution given by (26) shares some philosophy in common with that achieves the boudary of the time division (TD) capacity region discussed in [18, *Thoerem 3*]. We focus on solving (P1-OMA-II-sub) in any fading state given a triple (λ,δ,μ), while [18] maximized the total weighted sum-rate in any fading state by determining how to distribute $P(\boldsymbol{n})$ among $M=2$ users such that the instantaneous total power constraint $\sum_{j=1}^{2}\tau_j P_j(\boldsymbol{n})=P(\boldsymbol{n})$ (c.f. [18, Eqn. (11)]) is satisfied. The optimal solutions both suggest that with probability 1, at most one single user transmits in any fading state. This is because the probability measure of any subset of $\{((g_k(\nu),g_{\bar{k}}(\nu))):h(\lambda,\delta,\mu)=0\}$ (c.f. (22)) assuming continuously joint distribution of $(g_k(\nu),g_{\bar{k}}(\nu))$ is zero, and so is that of $L\triangleq\{\boldsymbol{n}:h(\lambda,\boldsymbol{n})=0\}$ in [18, *Thoerem 3*]. This also explains why the maximum of $\bar{\mathcal{L}}_1^{\text{OMA-II}}(p_1,p_2,\alpha_1)$ cannot be achieved by its jointly stationary point in probability.

With Proposition 3.2, given a triple (λ,δ,μ), (P1-OMA-II-sub) is first solved state by state; then by updating (λ,δ,μ) in accordance with the associated sub-gradient $(\bar{P}-\mathbb{E}_\nu[p_k(\nu)+p_{\bar{k}}(\nu)],\mathbb{E}_\nu[R_k^{\text{OMA-II}}(\nu)]-\bar{R},\mathbb{E}_\nu[R_{\bar{k}}^{\text{OMA-II}}(\nu)]-\bar{R})^T$, (P1-OMA-II) is iteratively solved.

Next, we rigorously prove that the ESR achieved by OMA-Type-II cannot perform better than that achieved by NOMA. To prove so, we denote the optimal power and time/frequency allocation to (P1-OMA-II) by $\{p_1^*(\nu),p_2^*(\nu),\alpha_1^*(\nu)\}$. Then let $p_1=p_1^*$ and $p_2=p_2^*$ in each fading state for NOMA transmission. If an alternative user is selected to transmit, then the optimum α_k associated with it is seen to be 1 (c.f. (26)). Hence, assuming $p_1^*>0$ ($\alpha_1^*=1$) and $p_2^*=0$ ($\alpha_2^*=0$), R_1^{NOMA} turns out be $\log_2(1+p_1^*g_1)$ and $R_2^{\text{NOMA}}=0$, which is exactly equal to $R_1^{*\text{OMA-II}}$ and $R_2^{*\text{OMA-II}}$, respectively; vice versa when $p_1^*=0$ ($\alpha_1^*=0$) and $p_2^*>0$ ($\alpha_2^*=1$). The other trivial case is that $R_k^{\text{NOMA}}=R_k^{*\text{OMA-II}}=0$, $\forall k$, when $p_1^*=p_2^*=0$. Hence, with $p_1=p_1^*$ and $p_2=p_2^*$ in each fading state, $\mathbb{E}_\nu[R_k^{\text{NOMA}}(\nu)]=\mathbb{E}_\nu[R_k^{*\text{OMA-II}}(\nu)]\geq\bar{R}$, $\forall k$, is met. It is also easily examined that (14d) is satisfied. To sum up, the optimal solution to (P1-OMA-II) proves to be feasible to (P1-NOMA), the former of which thus yields an optimum value no more than the latter.

B. Partial CSIT

By analogy, the partial CSIT counterpart of Problem (P1-XX) is formulated as below:

$$\begin{aligned}(\text{P1}'\text{-XX}):\ \underset{p_s,p_w,\alpha_k}{\texttt{Maximize}}\ & \mathbb{E}_\nu[R_k'^{\text{XX}}(\nu)+R_{\bar{k}}'^{\text{XX}}(\nu)] \\ \texttt{Subject to}\ & p_s+p_w\leq\bar{P}, && \text{(28a)}\\ & p_s\geq 0,\ p_w\geq 0, && \text{(28b)}\\ & \mathbb{E}_\nu[R_k'^{\text{XX}}(\nu)]\geq\bar{R}',\ \forall k, && \text{(28c)}\end{aligned}$$

where α_k's is only valid in the transmission adopting OMA-Type-II. Similar to Problem (P1-XX), (28c) constrain the minimum average user rate achieved by the two users. In the following, we provide optimal solution to (P1′-NOMA) and (P1′-OMA-II), respectively.

1) Optimal Solution to (P1′-NOMA)*:* Since only the relation between the two users' channel gains at each fading state and their CDI are known to the BS, we first derive the expectation of $R_k'^{\text{NOMA}}(\nu)$'s as function of p_s and p_w, and then solve (P1′-NOMA) in accordance with these expectation results.

First, denote the RV $|h_k(\nu)|^2$ ($|h_{\bar{k}}(\nu)|^2$) by X (Y)[11]. Also, denote the SNR $p_sg_k(\nu)$ and the SINR $p_wg_k(\nu)/(p_sg_k(\nu)+1)$

[11]Note that we assume $\sigma_k^2=\sigma^2$, $\forall k$, throughout the paper such that the relation between the effective channels of the two users , i.e., $g_k(\nu)$ and $g_{\bar{k}}(\nu)$, is equivalent to that between $|h_k(\nu)|^2$ and $|h_{\bar{k}}(\nu)|^2$.

(c.f. (6)) by Γ_k and $\tilde{\Gamma}_k$, respectively. It thus follows that the conditional cumulative density functions (CDFs) of Γ_k and $\tilde{\Gamma}_k$ are given by[12]

$$\begin{aligned} F_{\Gamma_{k|X\geq Y}}(z) =& \frac{\Pr\{p_s X/\sigma_k^2 \leq z, X \geq Y\}}{\Pr\{X \geq Y\}} \\ =& 1 - \frac{\lambda_k + \lambda_{\bar{k}}}{\lambda_{\bar{k}}} e^{-\lambda_k \varepsilon_k} + \frac{\lambda_k}{\lambda_{\bar{k}}} e^{-(\lambda_k+\lambda_{\bar{k}})\varepsilon_k}, \quad (29) \\ F_{\tilde{\Gamma}_k|X<Y}(z) =& \frac{\Pr\{p_w X/\left(p_s X + \sigma_k^2\right) \leq z, X < Y\}}{\Pr\{X < Y\}} \\ =& \begin{cases} 1, & \text{if } p_w - p_s z \leq 0 \\ 1 - e^{-(\lambda_k+\lambda_{\bar{k}})\tilde{\varepsilon}_k}, & \text{otherwise} \end{cases}, \quad (30) \end{aligned}$$

respectively, where $\varepsilon_k \triangleq \frac{\sigma_k^2 z}{p_s}$ and $\tilde{\varepsilon}_k \triangleq \frac{\sigma_k^2 z}{p_w - p_s z}$. In accordance with (29) and (30), $\mathbb{E}_\nu[R_k'^{\mathrm{NOMA}}(\nu)]$'s can be obtained by the following proposition.

Proposition 3.3: The ergodic rate for NOMA user $\mathcal{U}_k$, $k \in \{1,2\}$, under partial CSIT is given by

$$\begin{aligned} \mathbb{E}_\nu[R_k'^{\mathrm{NOMA}}(\nu)] =& \frac{2}{\ln 2} \frac{\lambda_k}{\lambda_k + \lambda_{\bar{k}}} f\left(\frac{(\lambda_k + \lambda_{\bar{k}})\sigma_k^2}{p_s}\right) \\ -& \frac{1}{\ln 2} \frac{\lambda_k}{\lambda_k + \lambda_{\bar{k}}} f\left(\frac{(\lambda_k + \lambda_{\bar{k}})\sigma_k^2}{p_s + p_w}\right) - \frac{1}{\ln 2} f\left(\frac{\lambda_k \sigma_k^2}{p_s}\right), \quad (31) \end{aligned}$$

where $f(\cdot)$ denotes the function $f(x) = e^x \mathrm{Ei}(-x)$ $(x > 0)$.

Proof: Please refer to a longer version of this paper [31, Appendix C]. ∎

Since the optimization variable p_w only contributes to $f(\frac{(\lambda_k+\lambda_{\bar{k}})\sigma_k^2}{p_s+p_w})$, we examine the property of $\mathbb{E}_\nu[R_k'^{\mathrm{NOMA}}(\nu)]$ in terms of p_w by studying $f(\frac{(\lambda_k+\lambda_{\bar{k}})\sigma_k^2}{p_s+p_w})$ as follows:

$$\begin{aligned} \frac{\partial f\left(\frac{(\lambda_k+\lambda_{\bar{k}})\sigma_k^2}{p_s+p_w}\right)}{\partial p_w} =& \frac{\partial\left(e^{\frac{(\lambda_k+\lambda_{\bar{k}})\sigma_k^2}{p_s+p_w}} \mathrm{Ei}\left(-\frac{(\lambda_k+\lambda_{\bar{k}})\sigma_k^2}{p_s+p_w}\right)\right)}{\partial p_w} \\ =& \frac{1}{p_s + p_w}\left(u \mathrm{E1}(u) e^u - 1\right) \\ \overset{(a)}{<}& \frac{1}{p_s + p_w}\left(u \ln\left(1 + \frac{1}{u}\right) - 1\right) < 0, \end{aligned} \quad (32)$$

where $\frac{(\lambda_k+\lambda_{\bar{k}})\sigma_k^2}{p_s+p_w} \triangleq u$, $\mathrm{E1}(x) = -\mathrm{Ei}(-x)$ $(x > 0)$, and (a) is due to the inequality $\mathrm{E1}(x)e^x < \ln(1 + \frac{1}{x})$ $(X > 0)$ [33, Eq. (5.1.20)]. Hence, $\mathbb{E}_\nu[R_k'^{\mathrm{NOMA}}(\nu)]$'s proves to monotonically increase with p_w.

Next, we solicit this monotonicity for solving (P1′-NOMA). As it is easily seen that given any p_s, (P1′-NOMA) attains its optimum value when $\mathbb{E}_\nu[R_k'^{\mathrm{NOMA}}(\nu)]$'s takes on its maximum w.r.t. p_w, i.e., when $p_w = \bar{P} - p_s$ (c.f. (28a)), (P1′-NOMA) is thus related with only one optimization variable p_s. Hence, one-dimension search over $p_s \in [0, \bar{P}]$ can be implemented to find the optimum solution (up to numerical accuracy) to (P1′-NOMA).

[12]These results are not explicitly applicable to $K > 2$ case. A better approach to deal with the more general cases with $K > 2$ is to solicit order statistics [8, 32], which is beyond the scope of this treatise.

2) Optimal Solution to (P1′-OMA-II)*:* Denoting the SNR of $\mathcal{U}_k$ in the case of $X \geq Y$ by Γ_k, and that in the case of $X < Y$ by $\tilde{\Gamma}_k$ (c.f. (7)), the conditional CDFs of Γ_k and $\tilde{\Gamma}_k$ are given by

$$\begin{aligned} F_{\Gamma_{k|X\geq Y}}(z) =& 1 - \frac{\lambda_k + \lambda_{\bar{k}}}{\lambda_{\bar{k}}} e^{-\lambda_k \varphi_k} + \frac{\lambda_k}{\lambda_{\bar{k}}} e^{-(\lambda_k+\lambda_{\bar{k}})\varphi_k}, \quad (33) \\ F_{\tilde{\Gamma}_{k|X<Y}}(z) =& 1 - e^{-(\lambda_k+\lambda_{\bar{k}})\tilde{\varphi}_k}, \quad (34) \end{aligned}$$

where $\varphi_k \triangleq \frac{\alpha_k \sigma_k^2 z}{p_s}$ and $\tilde{\varphi}_k \triangleq \frac{\alpha_k \sigma_k^2 z}{p_w}$. With (33) and (34), we have the following proposition.

Proposition 3.4: The ergodic rate for user $\mathcal{U}_k$ operating with OMA-type-II, $k \in \{1,2\}$, under partial CSIT is given by

$$\begin{aligned} &\mathbb{E}_\nu[R_k'^{\text{OMA-II}}(\nu)] = \\ &\frac{\alpha_k}{\ln 2}\left(-f\left(\frac{\lambda_k \alpha_k \sigma_k^2}{p_s}\right) + \frac{\lambda_k}{\lambda_k + \lambda_{\bar{k}}} f\left(\frac{(\lambda_k + \lambda_{\bar{k}})\alpha_k \sigma_k^2}{p_s}\right)\right. \\ &\left. - \frac{\lambda_k}{\lambda_k + \lambda_{\bar{k}}} f\left(\frac{(\lambda_k + \lambda_{\bar{k}})\alpha_k \sigma_k^2}{p_w}\right)\right). \quad (35) \end{aligned}$$

Proof: Please refer to a longer version of this paper [31, Appendix D]. ∎

Similar as is done in (32), $\mathbb{E}_\nu[R_k'^{\text{OMA-II}}(\nu)]$'s can be shown to monotonically increase with p_w as well. Therefore it implies that the optimal solution to (P1′-OMA-II) satisfies $p_s + p_w = \bar{P}$. As a result, there are two optimization variables (p_s and α_k) remaining for (P1′-OMA-II), which can be solved (up to numerical accuracy) by two-dimension search over $\{(p_s, \alpha_k) | p_s \in [0, \bar{P}], \alpha_k \in [0,1]\}$ such that $\mathbb{E}_\nu[R_k'^{\text{OMA-II}}(\nu)] \geq \bar{R}'$, $k \in \{1,2\}$.

IV. Optimum Delay-Limited Transmission

In delay-limited scenarios, each user attempts to maintain their respective prescribed rate in as much fading states as possible so as to reduce their outage probability (c.f. (8), (9), and (12)). When the users compete for power and/or time/frequency resources to get their intended data transmitted at the target rate in each fading state, the combined effects of outage probability and individual target rate accounts for the DLT of each user, which causes the solution to the sum of DLT maximization non-trivial. In this section, the optimal trade-offs between the system sum-DLT and the maximum outage probability requirement for the users is investigated for different multiple access schemes under full and partial CSIT, respectively. Particularly, under full CSIT, the DLT maximization problems are solved using Lagrangian dual decomposition levering "time-sharing" conditions, while under partial CSIT, the individual user's outage probability needs to be first analysed in closed form by means of CDI.

A. Full CSIT

In the case of full CSIT, we aim for maximizing the system sum of DLT by jointly optimizing the individual transmit power as well as time/frequency allocation over different fading states, subject to a given pair of APC and PPC at the

BS, and a maximum user outage probability constraint. The optimization problem is thus formulated as below.

$$\text{(P2-XX)}: \underset{\{p_k(\nu),p_{\bar{k}}(\nu),\alpha_k(\nu)\}}{\texttt{Maximize}} \bar{R}_k(1-\zeta_k^{\mathrm{XX}})+\bar{R}_{\bar{k}}(1-\zeta_{\bar{k}}^{\mathrm{XX}})$$
$$\texttt{Subject to}$$
$$\mathbb{E}_\nu[p_k(\nu)+p_{\bar{k}}(\nu)]\le \bar{P}, \tag{36a}$$
$$p_k(\nu)+p_{\bar{k}}(\nu)\le \hat{P},\ \forall\nu, \tag{36b}$$
$$p_k(\nu)\ge 0,\ p_{\bar{k}}(\nu)\ge 0,\ \forall\nu, \tag{36c}$$
$$\mathbb{E}_\nu[X_k^{\mathrm{XX}}(\nu)]\le \bar{\zeta},\ \forall k, \tag{36d}$$

where $\{\alpha_k(\nu)\}$'s are only valid when the two users access the channel by OMA-Type-II. It is worthy of noting that given the same target rate intended for each user, i.e., $\bar{R}_k=\bar{R}_{\bar{k}}=\bar{R}$, even if $\mathcal{U}_k$ and $\mathcal{U}_{\bar{k}}$ suffer from "near-far" physical condition, the far user can still successfully decode its data at this constant rate for more than $1-\bar{\zeta}$ proportion of the fading states, thanks to the constraints (36d). As seen from (10) and (11) ((13)), the discrete value of $X_k^{\mathrm{NOMA}}(\nu)$ ($X_k^{\text{OMA-II}}(\nu)$)'s renders non-convexity w.r.t. the optimization variables $p_k(\nu)$, $k\in\{1,2\}$, and thus Problem (P2-XX) is also non-convex. Therefore we exploit the similar "time-sharing" condition aforementioned to find their optimal solutions in subsection IV-A1 and IV-A2, respectively. In the following, we aim for solving (P2-NOMA) and (P2-OMA-II), respectively.

1) Optimal Solution to (P2-NOMA)*:* Adopting Lagrangian dual decomposition method, the Lagrangian of Problem (P2-NOMA) is given by

$$\mathcal{L}_2^{\mathrm{NOMA}}(\{p_k(\nu)\},\{p_{\bar{k}}(\nu)\},\lambda,\delta,\mu)=$$
$$\mathbb{E}_\nu[-\bar{R}_kX_k^{\mathrm{NOMA}}(\nu)-\bar{R}_{\bar{k}}X_{\bar{k}}^{\mathrm{NOMA}}(\nu)-\lambda(p_k(\nu)+p_{\bar{k}}(\nu))-$$
$$\delta X_k^{\mathrm{NOMA}}(\nu)-\mu X_{\bar{k}}^{\mathrm{NOMA}}(\nu)]+\lambda\bar{P}+\delta\bar{\zeta}+\mu\bar{\zeta}, \tag{37}$$

where λ is the Lagrangian multiplier associated with the APC; δ and μ are those associated with the maximum user outage probability constraints given in (36d) for $\mathcal{U}_k$ and $\mathcal{U}_{\bar{k}}$, respectively. In line with the principle of dual decomposition, (37) can be maximized by decoupling it into independent subproblems each for one fading state and solving those subproblems in parallel. Define $\bar{\mathcal{L}}_2^{\mathrm{NOMA}}(p_k,p_{\bar{k}})=\bar{R}_kX_k^{\mathrm{NOMA}}+\bar{R}_{\bar{k}}X_{\bar{k}}^{\mathrm{NOMA}}+\lambda(p_k+p_{\bar{k}})+\delta X_k^{\mathrm{NOMA}}+\mu X_{\bar{k}}^{\mathrm{NOMA}}$. With the fading index ν safely dropped, given the dual variables' triple (λ,δ,μ), the following problem is typical of the subproblems sharing the same structure:

$$\text{(P2-NOMA-sub)}: \underset{p_k\ge0,p_{\bar{k}}\ge0}{\texttt{Minimize}} \quad \bar{\mathcal{L}}_2^{\mathrm{NOMA}}(p_k,p_{\bar{k}})$$
$$\texttt{Subject to} \quad p_k+p_{\bar{k}}\le\hat{P}.$$

Then, we investigate the possible combinations of outage occurrences for $\mathcal{U}_k$ and $\mathcal{U}_{\bar{k}}$. Assuming $g_k>g_{\bar{k}}$, $k\in\{1,2\}$, the possible combinations of indicator function X_k^{NOMA} and the corresponding decoding strategies adopted by $\mathcal{U}_k$ are summarized in Table I, where $\mathcal{U}_{\bar{k}}\to\mathcal{U}_{\bar{k}}$ represents that $\mathcal{U}_{\bar{k}}$ directly decodes its own information TIAN; $\mathcal{U}_k\to\mathcal{U}_{\bar{k}}\to\mathcal{U}_k$ denotes $\mathcal{U}_k$'s attempt to perform SIC[13]; $\not\to$ indicates failure of decoding. Specifically, if the first step succeeds, $\mathcal{U}_k$ is able to cancel the interference from $\mathcal{U}_{\bar{k}}$, otherwise $\mathcal{U}_k$ continues to decode its own treating $\mathcal{U}_{\bar{k}}$'s as interference. Based on Table I, we derive the optimal solution to (P2-NOMA-sub) in the following proposition.

TABLE I

$\mathcal{U}_k$	$\mathcal{U}_k\not\to\mathcal{U}_{\bar{k}}\not\to\mathcal{U}_k,\ X_k^{\mathrm{NOMA}}=1$ (I.A)	$\mathcal{U}_k\to\mathcal{U}_{\bar{k}}\not\to\mathcal{U}_k,\ X_k^{\mathrm{NOMA}}=1$ (II.A)
	$\mathcal{U}_k\not\to\mathcal{U}_{\bar{k}}\to\mathcal{U}_k,\ X_k^{\mathrm{NOMA}}=0$ (I.B)	$\mathcal{U}_k\to\mathcal{U}_{\bar{k}}\to\mathcal{U}_k,\ X_k^{\mathrm{NOMA}}=0$ (II.B)
	$\mathcal{U}_k\to\mathcal{U}_{\bar{k}}\not\to\mathcal{U}_k,\ X_k^{\mathrm{NOMA}}=1$ (I.C)	
	$\mathcal{U}_k\to\mathcal{U}_{\bar{k}}\to\mathcal{U}_k,\ X_k^{\mathrm{NOMA}}=0$ (I.D)	
$\mathcal{U}_{\bar{k}}$	$\mathcal{U}_{\bar{k}}\not\to\mathcal{U}_{\bar{k}},\ X_k^{\mathrm{NOMA}}=1$	$\mathcal{U}_{\bar{k}}\to\mathcal{U}_{\bar{k}},\ X_k^{\mathrm{NOMA}}=0$

Proposition 4.1: The optimal power allocation to Problem (P2-NOMA-sub) assuming $g_k>g_{\bar{k}}$ is given by

$$(p_k^*,p_{\bar{k}}^*)=\arg\min_{i\in\{1,2,3,4\}}\{\bar{\mathcal{L}}_2^{\mathrm{NOMA}}(p_{i,k},p_{i,\bar{k}})\mathbb{1}_{p_{i,k}+p_{i,\bar{k}}}\},$$

where $p_{i,k}$'s and $p_{i,\bar{k}}$'s are given by[14]

$$\begin{cases} p_{1,k}=0,\ p_{1,\bar{k}}=0\\ p_{2,k}=\frac{2^{\bar{R}_k}-1}{g_k},\ p_{2,\bar{k}}=0\\ p_{3,k}=0,\ p_{3,\bar{k}}=\frac{2^{\bar{R}_{\bar{k}}}-1}{g_{\bar{k}}}\\ p_{4,k}=\frac{2^{\bar{R}_k}-1}{g_k},\ p_{4,\bar{k}}=\left(2^{\bar{R}_{\bar{k}}}-1\right)\left(\frac{2^{\bar{R}_k}-1}{g_k}+\frac{1}{g_{\bar{k}}}\right)\end{cases}, \tag{38}$$

and the indicator function $\mathbb{1}_{(\cdot)}$ is defined the same as (27).

Proof: To minimize $\bar{\mathcal{L}}_2^{\mathrm{NOMA}}(p_k,p_{\bar{k}})$, we need to examine every case of combination regarding $\mathcal{U}_k$'s and $\mathcal{U}_{\bar{k}}$'s outage occurrences so as to find the one that minimizes $\bar{\mathcal{L}}_2^{\mathrm{NOMA}}(p_k,p_{\bar{k}})$. First, we show that the cases I.C and I.D can be safely removed since they are always outperformed by other cases. Take I.C as an example, if $\mathcal{U}_k$ succeeds in decoding $\mathcal{U}_{\bar{k}}$'s message at rate $\bar{R}_{\bar{k}}$, it inexplicitly suggests that $\mathcal{U}_{\bar{k}}$'s message is transmitted at $p_{\bar{k}}>0$. Therefore, the corresponding $\bar{\mathcal{L}}_2^{\mathrm{NOMA}}(p_k,p_{\bar{k}})=\bar{R}_k+\bar{R}_{\bar{k}}+\lambda p_{\bar{k}}+\delta+\mu$ is strictly larger than $\bar{\mathcal{L}}_2^{\mathrm{NOMA}}(0,p_{\bar{k}})=\bar{R}_k+\bar{R}_{\bar{k}}+\delta+\mu$ in Case I.A. Similarly, Case I.D can be shown to be strictly outperformed by Case I.B. With the remaining four cases, $p_{i,k}$ ($p_{i,\bar{k}}$), $i\in\{1,2,3,4\}$, is the minimum power required for $\mathcal{U}_k$ ($\mathcal{U}_{\bar{k}}$) to succeed in transmission associated with the case I.A, I.B, II.A, and II.B, respectively. Next, select the minimizer out from these four cases, which depends on how the required transmit power weighs $\bar{R}_k$'s as well as the given multipliers (λ,δ,μ). ■

Note from Proposition 4.1 that the optimal power policy allocates either the minimum required power to support $\mathcal{U}_k$ and/or $\mathcal{U}_{\bar{k}}$'s transmission at their respective target rate or completely shuts down the transmission. For example, when $\mathcal{U}_k$ suspends its transmission in Cases I.A and II.A, Case II.A outperforms Case I.A iff $p_{\bar{k}}\le\frac{\bar{R}_{\bar{k}}+\mu}{\lambda}$. From the perspective of fairness, when μ is large enough appealing for smaller outage, this condition is easier to be satisfied and thus $X_{\bar{k}}^{\mathrm{NOMA}}$ is more likely to be 0, and vice versa.

With Proposition 4.1, given any multiplier-triple (λ,δ,μ), the maximum of the Lagrangian in (37) is obtained by

[13]Given $g_k>g_{\bar{k}}$, we assume the decoding order $\mathcal{U}_k\to\mathcal{U}_{\bar{k}}\to\mathcal{U}_k$, since in delay-limited NOMA, it is optimum to have only the user with better CSI perform SIC for the sake of saving the total transmit power.

[14]In fact, under full CSIT, following the same method utilized to develop Proposition 4.1, the two-user results can be generalized to cases with $K>2$ by mathematical induction.

solving (P2-NOMA-sub) state by state in parallel. Finally, (P2-NOMA) is solved by updating (λ, δ, μ) in accordance with the ellipsoid method.

2) Optimal Solution to (P2-OMA-II)*:* Despite of its non-convexity due to the same reason as that for (P2-NOMA), we can still find the optimal solution to (P2-OMA-II) thanks to the "time-sharing" condition that (P2-OMA-II) meets.

Similar to Section IV-A1, ((P2-OMA-II) can also be decoupled into as many subproblems as the number of fading states each for one fading state, which is expressed as

$$\begin{aligned}(\text{P2-OMA-II-sub}): \quad & \underset{p_k\geq 0, p_{\bar{k}}\geq 0, \alpha_k}{\texttt{Minimize}} \ \bar{\mathcal{L}}_2^{\text{OMA-II}}(p_k, p_{\bar{k}}, \alpha_k) \\ & \texttt{Subject to} \\ & p_k + p_{\bar{k}} \leq \hat{P}, \\ & 0 \leq \alpha_k \leq 1, \ \forall k,\end{aligned}$$

where the objective function is defined as $\bar{\mathcal{L}}_2^{\text{OMA-II}}(p_k, p_{\bar{k}}, \alpha_k) = \bar{R}_k X_k^{\text{OMA-II}} + \bar{R}_{\bar{k}} X_{\bar{k}}^{\text{OMA-II}} + \lambda(p_k + p_{\bar{k}}) + \delta X_k^{\text{OMA-II}} + \mu X_{\bar{k}}^{\text{OMA-II}}$ with the fading index ν dropped for brevity.

Since each $\mathcal{U}_k$ only needs to decode its own information without seeing interference in the orthogonal transmission, the possible combinations of outage occurrences for $\mathcal{U}_k$ and $\mathcal{U}_{\bar{k}}$ are easily shown in Table II, where $\mathcal{U}_k \to \mathcal{U}_k$ denotes $\mathcal{U}_k$'s direct decoding of its own message, $k \in \{1, 2\}$. Based on Table II,

TABLE II

$\mathcal{U}_k$	$\mathcal{U}_k \not\to \mathcal{U}_k,\ X_k^{\text{OMA-II}} = 1$	$\mathcal{U}_k \not\to \mathcal{U}_k,\ X_k^{\text{OMA-II}} = 1$
	$\mathcal{U}_k \to \mathcal{U}_k,\ X_k^{\text{OMA-II}} = 0$	$\mathcal{U}_k \to \mathcal{U}_k,\ X_k^{\text{OMA-II}} = 0$
$\mathcal{U}_{\bar{k}}$	$\mathcal{U}_{\bar{k}} \not\to \mathcal{U}_{\bar{k}},\ X_{\bar{k}}^{\text{OMA-II}} = 1$	$\mathcal{U}_{\bar{k}} \to \mathcal{U}_{\bar{k}},\ X_{\bar{k}}^{\text{OMA-II}} = 0$

we obtain the optimal solution to (P2-OMA-II-sub) in the following proposition.

Proposition 4.2: The optimal power allocation to Problem (P2-OMA-II-sub) is given by

$$\begin{aligned}&(p_k^*, p_{\bar{k}}^*, \alpha_k^*) = \\ &\quad \arg \min_{i\in\{1,2,3,4\}} \{\bar{\mathcal{L}}_2^{\text{OMA-II}}(p_{i,k}, p_{i,\bar{k}}, \alpha_{i,k}) \mathbb{1}_{p_{i,k}+p_{i,\bar{k}}}\}, \quad (39)\end{aligned}$$

where $p_{i,k}$'s and $p_{i,\bar{k}}$'s are given by

$$\begin{cases} p_{1,k} = 0, \ p_{1,\bar{k}} = 0, \ \alpha_{1,k} = 0; \\ p_{2,k} = \frac{2^{\bar{R}_k}-1}{g_k}, \ p_{2,\bar{k}} = 0, \ \alpha_{2,k} = 1; \\ p_{3,k} = 0, \ p_{3,\bar{k}} = \frac{2^{\bar{R}_{\bar{k}}}-1}{g_{\bar{k}}}, \ \alpha_{3,k} = 0; \\ p_{4,k} = \frac{\alpha_k^*(2^{\frac{\bar{R}_k}{\alpha_k^*}}-1)}{g_k}, \ p_{4,\bar{k}} = \frac{\alpha_{\bar{k}}^*(2^{\frac{\bar{R}_{\bar{k}}}{\alpha_{\bar{k}}^*}}-1)}{g_{\bar{k}}}, \ \alpha_{4,k} = \alpha_k^*. \end{cases} \quad (40)$$

In (40), α_k^*'s denotes the optimum proportion of time/frequency resource allocated to $\mathcal{U}_k$'s to minimize the instantaneous total transmit power, which is obtained by solving the following (convex) problem.

$$\begin{aligned}&(\text{P2-OMA-II-MP}): \\ &\underset{\alpha_k}{\texttt{Minimize}} \ \frac{\alpha_k(2^{\frac{\bar{R}_k}{\alpha_k}}-1)}{g_k} + \frac{\alpha_{\bar{k}}(2^{\frac{\bar{R}_{\bar{k}}}{\alpha_{\bar{k}}}}-1)}{g_{\bar{k}}} \\ &\texttt{Subject to} \ \ 0 \leq \alpha_k \leq 1.\end{aligned}$$

In addition, the indicator function is given by (27).

Proof: Please refer to a longer version of this paper [31, Appendix E]. ∎

As a result, given any triple (λ, δ, μ), (P2-OMA-II-sub) is solved. (P2-OMA-II) is then solved by updating (λ, δ, μ) using the ellipsoid method similarly as solving (P2-NOMA).

Next, given the same set of $\bar{R}_k$'s, we provide mathematical proof for the superiority of NOMA over OMA-Type-II in terms of the optimum sum of DLT. To prove so, we introduce the following lemma.

Lemma 4.1: Given $g_1 \geq g_2 \geq \ldots \geq g_K > 0$, denoting $\min_{\sum_{i=1}^K \alpha_i = 1} \sum_{i=1}^K \frac{(2^{\frac{\bar{R}_i}{\alpha_i}}-1)\alpha_i}{g_i}$ by P_{O2}^*, and $\sum_{i=0}^{K-1} \frac{(2^{\bar{R}_{K-i}}-1)2^{\sum_{j=0}^{i-1}\bar{R}_{K-j}}}{g_{K-i}}$ by P_{N}^*, where $\alpha_i \geq 0$ and $\bar{R}_i \geq 0$, $\forall i$, then it follows that $P_{\text{O2}}^* \geq P_{\text{N}}^*$.

Proof: Please refer to a longer version of this paper [31, Appendix F]. ∎

Then, assuming $\{p_1^*(\nu), p_2^*(\nu), \alpha_1^*(\nu)\}$ as the optimal solution to (P2-OMA-II), we construct a solution to (P2-NOMA) based on $\{p_1^*(\nu), p_2^*(\nu)\}$ as follows. We need to modify solution (c.f. (40)) to (P2-OMA-II-sub) corresponding to each of the four cases. It is straightforward to check that for the first three cases in (40), setting $p_1 = p_1^*$ and $p_2 = p_2^*$ corresponds to the first three cases in (38), and thus $X_k^{\text{NOMA}} = X_k^{*\text{OMA-II}}$, $\forall k$. If (p_1^*, p_2^*) falls in the last case of (40), we reallocate them among the two users such as $p_1 = \frac{2^{\bar{R}_1}-1}{g_1}$ and $p_2 = (2^{\bar{R}_2}-1)(\frac{2^{\bar{R}_1}-1}{g_1} + \frac{1}{g_2})$ assuming $g_1 > g_2$ w.l.o.g., and therefore it follows that $X_k^{\text{NOMA}} = X_k^{*\text{OMA-II}} = 1$, $\forall k$. The modification is feasible, since $p_1^* + p_2^* \geq p_1 + p_2$. This is because when $K = 2$, it follows from Lemma 4.1 that

$$\min_{\sum_{i=1}^2 \alpha_k = 1} \sum_{i=1}^2 \frac{(2^{\frac{\bar{R}_i}{\alpha_i}}-1)\alpha_i}{g_i} \geq \frac{2^{\bar{R}_2}-1}{g_2} + \frac{2^{\bar{R}_2}(2^{\bar{R}_1}-1)}{g_1},$$

in which the left-hand side and right right-hand side corresponds to $p_1^* + p_2^*$ and $p_1 + p_2$, respectively.

To sum up, with the constructed solution (p_1, p_2) in each fading state, it follows that $\mathbb{E}_\nu[X_k^{\text{NOMA}}(\nu)] = \mathbb{E}_\nu[X_k^{*\text{OMA-II}}(\nu)] \leq \bar{\zeta}$, $\forall k$. Moreover, benefiting from the saved power by NOMA, (14a) now becomes inactive, which suggests that $\mathbb{E}_\nu[X_k^{\text{NOMA}}(\nu)]$'s are potentially to be further reduced should the power be fully allocated. Hence, (P2-NOMA) is shown to be able to achieve larger sum of DLT than (P2-OMA-II).

B. Partial CSIT

Problem (P2-XX) under partial CSIT is recast as follows:

$$\begin{aligned}(\text{P2}'\text{-XX}): \quad & \underset{p_s, p_w, \alpha_k}{\texttt{Maximize}} \ \ \bar{R}_k(1-\zeta_k'^{\text{XX}}) + \bar{R}_{\bar{k}}(1-\zeta_{\bar{k}}'^{\text{XX}}) \\ & \texttt{Subject to} \ \ p_s + p_w \leq \bar{P}, \quad (41a) \\ & \qquad\qquad p_s \geq 0, \ p_w \geq 0, \quad (41b) \\ & \qquad\qquad \mathbb{E}_\nu[X_k'^{\text{XX}}(\nu)] \leq \bar{\zeta}', \ \forall k, \quad (41c)\end{aligned}$$

where α_k's is only valid when "XX" is replaced by OMA-Type-II, and (41c) constrain the maximum user outage probability of the two below $\bar{\zeta}'$. We present in the sequel how

to optimally solve (P2′-NOMA) and (P2′-OMA-II), respectively.

1) Optimal Solution to (P2′-NOMA)*:* In line with the same notation for channel gains as defined in (29) and (30), replace $p_k(\nu)$ with p_s, $p_{\bar{k}}(\nu)$ with p_w, $\forall \nu$, when $X > Y$, and otherwise do this reversely. As a result, $\mathbb{E}_\nu[X_k'^{\mathrm{NOMA}}(\nu)] = \zeta_k'^{\mathrm{NOMA}}$ can be recast as follows:

$$\begin{aligned}\mathbb{E}_\nu[X_k'^{\mathrm{NOMA}}(\nu)] = \Pr\left\{\log_2\left(1+\frac{p_s X}{p_w X+\sigma_k^2}\right) < \bar{R}_k,\right.\\ \left.\log_2\left(1+\frac{p_w X}{p_s X+\sigma_k^2}\right) < \bar{R}_{\bar{k}}, X>Y\right\} + \Pr\left\{\right.\\ \left.\log_2\left(1+\frac{p_w X}{p_s X+\sigma_k^2}\right) \geq \bar{R}_{\bar{k}}, X>Y\right\} +\\ \Pr\left\{\log_2\left(1+\frac{p_w X}{p_s X+\sigma_k^2}\right) < \bar{R}_k, X\leq Y\right\}. \quad (42)\end{aligned}$$

With CDI regarding X and Y given in Section II, $\mathbb{E}_\nu[X_k'^{\mathrm{NOMA}}(\nu)]$'s can be derived based upon (42) shown in the following proposition.

Proposition 4.3: The outage probability for NOMA user $\mathcal{U}_k$ given the prescribed transmit rate $\bar{R}_k$, $k \in \{1,2\}$, under partial CSIT is given by

$$\begin{aligned}&\mathbb{E}_\nu[X_k'^{\mathrm{NOMA}}(\nu)] =\\ &\begin{cases} 1-e^{-\lambda_k\varepsilon_{k,j}} + \frac{\lambda_k}{\lambda_k+\lambda_{\bar{k}}}e^{-(\lambda_k+\lambda_{\bar{k}})\varepsilon_{k,j}} - \\ \quad \frac{\lambda_k}{\lambda_k+\lambda_{\bar{k}}}e^{-(\lambda_k+\lambda_{\bar{k}})\varepsilon_{k,3}}, \ (\mathbf{a}) \\ 1-e^{-\lambda_k\varepsilon_{k,j}} + \frac{\lambda_k}{\lambda_k+\lambda_{\bar{k}}}e^{-(\lambda_k+\lambda_{\bar{k}})\varepsilon_{k,j}}, \ (\mathbf{b}) \\ 1-\frac{\lambda_k}{\lambda_k+\lambda_{\bar{k}}}e^{-(\lambda_k+\lambda_{\bar{k}})\varepsilon_{k,3}}, \ (\mathbf{c}) \\ \frac{\lambda_{\bar{k}}}{\lambda_k+\lambda_{\bar{k}}} - e^{-\lambda_k\varepsilon_{k,1}} + \frac{\lambda_k}{\lambda_k+\lambda_{\bar{k}}}e^{-(\lambda_k+\lambda_{\bar{k}})\varepsilon_{k,1}}, \ (\mathbf{d}) \\ 1, \ (\mathbf{e}) \end{cases}\end{aligned} \quad (43)$$

where $\varepsilon_{k,1} \triangleq \frac{\sigma_k^2\tau_k}{p_s}$, $\varepsilon_{k,2} \triangleq \frac{\sigma_k^2\tau_k}{p_s - p_w\tau_k}$, $\varepsilon_{k,3} \triangleq \frac{\sigma_k^2\tau_k}{p_w - p_s\tau_k}$, and $\varepsilon_{k,4} \triangleq \frac{\sigma_k^2\tau_{\bar{k}}}{p_w - p_s\tau_{\bar{k}}}$, with $\tau_k \triangleq 2^{\bar{R}_k}-1$. The conditions in (a)-(e) of (43) corresponds to

$$\begin{cases} \alpha_{k,c_j^1} \leq \frac{p_s}{p_w} \leq \beta_{k,c_j^1}, \ j\in\{1,2,3\}, \\ \alpha_{k,c_j^2} \leq \frac{p_s}{p_w} \leq \beta_{k,c_j^2}, \ j\in\{1,2,3\}, \\ \alpha_{k,c^3} \leq \frac{p_s}{p_w} \leq \beta_{k,c^3}, \\ \alpha_{k,c^4} \leq \frac{p_s}{p_w} \leq \beta_{k,c^4}, \\ \alpha_{k,c^5} \leq \frac{p_s}{p_w} \leq \beta_{k,c^5}, \end{cases} \quad (44)$$

respectively. In (44), α_{k,c_j^1} and β_{k,c_j^1}, $j\in\{1,2,3\}$, are given by[15]

$$\begin{cases} \alpha_{k,c_1^1} = 0 \\ \alpha_{k,c_2^1} = \max\left\{\frac{\tau_k(\tau_{\bar{k}}+1)}{\tau_{\bar{k}}(\tau_k+1)}, \tau_k\right\} \text{ or } \max\left\{\tau_k, \frac{1}{\tau_{\bar{k}}}\right\} \\ \alpha_{k,c_3^1} = \max\left\{\frac{\tau_k}{\tau_{\bar{k}}(\tau_k+1)}, \tau_k\right\} \text{ or } \frac{\tau_k}{\tau_{\bar{k}}(\tau_k+1)} \end{cases}$$

and

$$\begin{cases} \beta_{k,c_1^1} = \min\left\{\frac{\tau_k}{\tau_{\bar{k}}(\tau_k+1)}, \frac{1}{\tau_k}\right\} \\ \beta_{k,c_2^1} = \min\left\{\frac{1}{\tau_k}, \frac{1}{\tau_{\bar{k}}}\right\} \text{ or } \frac{1}{\tau_k} \\ \beta_{k,c_3^1} = \min\left\{\frac{\tau_k(\tau_{\bar{k}}+1)}{\tau_{\bar{k}}(\tau_k+1)}, \frac{1}{\tau_k}, \frac{1}{\tau_{\bar{k}}}\right\} \text{ or } \min\left\{\tau_k, \frac{1}{\tau_k}, \frac{1}{\tau_{\bar{k}}}\right\} \end{cases}.$$

[15]The parameter values preceding and coming after "or" form a pair, respectively.

Moreover, α_{k,c_j^2} and β_{k,c_j^2}, $j\in\{1,2,3\}$, are given by

$$\begin{cases} \alpha_{k,c_1^2} = \max\left\{\tau_k, \frac{1}{\tau_k}\right\} \\ \alpha_{k,c_2^2} = \max\left\{\frac{\tau_k(\tau_{\bar{k}}+1)}{\tau_{\bar{k}}(\tau_k+1)}, \tau_k, \frac{1}{\tau_k}\right\} \text{ or } \max\left\{\tau_k, \frac{1}{\tau_k}, \frac{1}{\tau_{\bar{k}}}\right\} \\ \alpha_{k,c_3^2} = \max\left\{\frac{\tau_k}{\tau_{\bar{k}}(\tau_k+1)}, \tau_k, \frac{1}{\tau_k}\right\} \text{ or } \max\left\{\frac{\tau_k}{\tau_{\bar{k}}(\tau_k+1)}, \frac{1}{\tau_k}\right\} \end{cases}$$

and

$$\begin{cases} \beta_{k,c_1^2} = \frac{\tau_k}{\tau_{\bar{k}}(\tau_k+1)} \\ \beta_{k,c_2^2} = \frac{1}{\tau_{\bar{k}}} \text{ or } +\infty \\ \beta_{k,c_3^2} = \min\left\{\frac{\tau_k(\tau_{\bar{k}}+1)}{\tau_{\bar{k}}(\tau_k+1)}, \frac{1}{\tau_{\bar{k}}}\right\} \text{ or } \min\left\{\tau_k, \frac{1}{\tau_{\bar{k}}}\right\} \end{cases}.$$

Finally, α_{k,c^l} and β_{k,c^l}, $l\in\{3,4,5\}$, are given by

$$\begin{cases} \alpha_{k,c^3} = \frac{1}{\tau_{\bar{k}}} \\ \alpha_{k,c^4} = \frac{1}{\tau_k} \\ \alpha_{k,c^5} = \max\left\{\frac{1}{\tau_k}, \frac{1}{\tau_{\bar{k}}}\right\} \end{cases}, \quad \begin{cases} \beta_{k,c^3} = \min\left\{\tau_k, \frac{1}{\tau_k}\right\} \\ \beta_{k,c^4} = \min\left\{\frac{\tau_k}{\tau_{\bar{k}}(\tau_k+1)}, \tau_k\right\} \\ \beta_{k,c^5} = \tau_k \end{cases}.$$

Proof: Please refer to a longer version of this paper [31, Appendix G]. ■

In accordance with Proposition 4.3, we are able to derive the sum of the DLT $\bar{R}_k(1-\zeta_k'^{\mathrm{NOMA}}) + \bar{R}_{\bar{k}}(1-\zeta_{\bar{k}}'^{\mathrm{NOMA}})$. Problem (P2′-NOMA) can also be characterized by only one optimization variable p_s by replacing p_w with $\bar{P}-p_s$, and then optimally solved (up to numerical accuracy) via one-dimension search over $p_s \in [0, \bar{P}]$.

2) Optimal Solution to (P2′-OMA-II)*:* In line with the principle of power and time/frequency allocations for OMA-Type-II transmission described above (7), replace $p_k(\nu)$ with p_s, $p_{\bar{k}}(\nu)$ with p_w, $\forall \nu$, when $X > Y$, and the reverse when $X \leq Y$ in (12). $\zeta_k'^{\mathrm{OMA\text{-}II}} = \mathbb{E}_\nu[X_k'^{\mathrm{OMA\text{-}II}}(\nu)]$, is derived as follows:

$$\begin{aligned}\mathbb{E}_\nu[X_k'^{\mathrm{OMA\text{-}II}}(\nu)] =\\ \Pr\left\{\alpha_k\log_2\left(1+\frac{p_s X}{\alpha_k\sigma_k^2}\right) < \bar{R}_k, X>Y\right\} +\\ \Pr\left\{\alpha_k\log_2\left(1+\frac{p_w X}{\alpha_k\sigma_k^2}\right) < \bar{R}_{\bar{k}}, X\leq Y\right\}. \quad (45)\end{aligned}$$

We are thus able to derive $\mathbb{E}_\nu[X_k'^{\mathrm{OMA\text{-}II}}(\nu)]$ in the following proposition.

Proposition 4.4: The outage probability for OMA-Type-II user $\mathcal{U}_k$ given the prescribed transmit rate $\bar{R}_k$, $k\in\{1,2\}$, under partial CSIT is given by

$$\begin{aligned}\mathbb{E}_\nu[X_k'^{\mathrm{OMA\text{-}II}}(\nu)] = 1 + \frac{\lambda_k}{\lambda_k+\lambda_{\bar{k}}}\big(e^{(\lambda_k+\lambda_{\bar{k}})\varphi_{k,1}} -\\ e^{(\lambda_k+\lambda_{\bar{k}})\varphi_{k,2}}\big) - e^{-\lambda_k\varphi_{k,1}}, \quad (46)\end{aligned}$$

where $\varphi_{k,1} \triangleq \frac{\alpha_k\sigma_k^2\xi_k}{p_s}$ and $\varphi_{k,2} \triangleq \frac{\alpha_k\sigma_k^2\xi_k}{p_w}$, with $\xi_k \triangleq 2^{\frac{\bar{R}_k}{\alpha_k}}-1$.

Proof: With CDI of X and Y known, the derivation of $\mathbb{E}_\nu[X_k'^{\mathrm{OMA\text{-}II}}(\nu)]$ from (45) is straightforward and thus omitted here for brevity. ■

Based on Proposition 4.4, (P2′-OMA-II) can be solved similarly as (P2′-NOMA), the detail of which is omitted herein for brevity.

V. Numerical Results

In this section, we verify the theoretical analysis for the considered two-user downlink NOMA system via numerical results. As a performance bench mark, we also provide one classical type of OMA transmission scheme, referred as *OMA-Type-I*, which assigns equal amount of time (in TDMA) or frequency (in FDMA) resources among users over all fading states, i.e., $\alpha_k(\nu) = \frac{1}{2}$, $\forall k$, $\forall \nu$ in (5), and $\alpha_k = \frac{1}{2}$, $\forall k$, in (7). The corresponding optimal power policies to OMA-Type-I are easily seen to be special cases of OMA-Type-II, which has already been solved. $\mathcal{U}_1$ and $\mathcal{U}_2$ are assumed to be located with a distance of d_1 and d_2 away from the BS, respectively. The large-scale path loss model of the channel is given by $128.1 + 37.6 \log 10(D)$ in dB, where D in kilometer (km) denotes the distance from the BS to the user. The small-scale fading is assumed to be independent and identically distributed ($i.i.d.$) Rayleigh fading. The AWGNs at the users' Rxs are both assumed to be -169dBm/Hz over 10MHz bandwidth. The infinite number of fading states is approximated by 10^7. Other simulation parameters are set as follows: $d_1 = 0.1$km, $d_2 = 0.5$km, $\hat{P} = 5$Watt and $\bar{P} = 1$Watt unless otherwise specified.

A. Delay-Tolerant Transmission

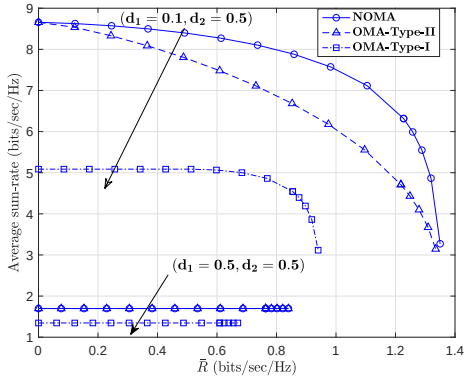


Fig. 2. The average sum-rate of the system versus the minimum rate constraints under full CSIT.

Fig. 2 depicts the optimal trade-offs between the average sum-rate of the system and the minimum rate constraints under full CSIT, i.e., $\bar{R}$, achieved by NOMA and the OMA schemes with different distance settings. It is seen that with the near-far distance setting, NOMA outperforms OMA-Type-II transmission in most cases, while the gap shrinks when $\bar{R}$ is very little and/or approaches $\bar{R}_{\max}$, respectively. Moreover, both NOMA and OMA-Type-II achieve substantially larger optimal trade-off than OMA-Type-I, although OMA-Type-I is seen more robust against increase in $\bar{R}$. This is because OMA-Type-I is intrinsically of fairness in view of equal time/frequency assigned to each user irrespective of their CSI. It is also worth noting that when there is no difference between the two users in terms of large-scale fading, the average sum-rate versus min-rate trade-offs almost vanish, since the average sum-rate w/o the minimum rate constraint has already achieved certain fairness, i.e., $\mathbb{E}_\nu[R_k^{\mathrm{NOMA}}(\nu)] \approx \mathbb{E}_\nu[R_{\bar{k}}^{\mathrm{NOMA}}(\nu)]$ ($\mathbb{E}_\nu[R_k^{\mathrm{OMA\text{-}II}}(\nu)] \approx \mathbb{E}_\nu[R_{\bar{k}}^{\mathrm{OMA\text{-}II}}(\nu)]$) due to their statistically similar channel distribution.

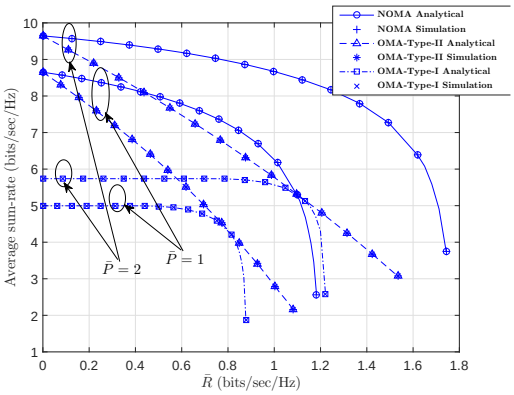


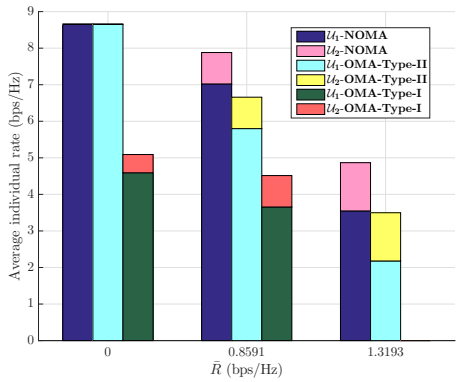
Fig. 3. The average sum-rate of the system versus the minimum rate constraints under partial CSIT.

Fig. 3 shows the optimal trade-offs between the average sum-rate of the system versus the minimum rate constraints under partial CSIT, i.e., $\bar{R}'$, achieved by various multiple access schemes with different APC. The optimal trade-off regions between the average-sum rate of the system and the fairness are expectedly seen to enlarge with increasing limit on the transmit power $\bar{P}$. While the superiority of the proposed power allocation policies for NOMA against OMA-Type-II is obviously seen, the contrast is more sharply observed for NOMA against OMA-Type-I in Fig. 2.

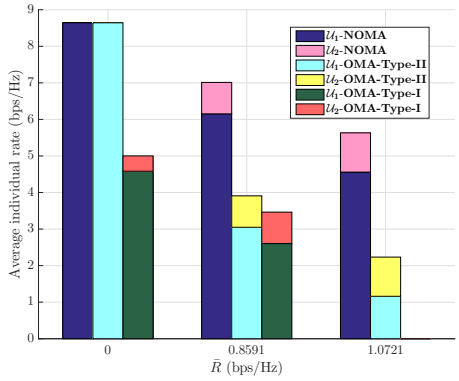
The comparison between the individual ergodic rate subject to varied minimum average rate constraints is demonstrated in Fig. 4(a) (Fig. 4(b)) for NOMA, OMA-Type-I, and OMA-Type-II, respectively, under full (partial) CSIT. First, we see that OMA-Type-II achieves almost the same ergodic rate for $\mathcal{U}_1$ as NOMA with $\mathcal{U}_2$'s ergodic rate both as little as zero, when there is no minimum rate requirement. This can be intuitively explained as follows. Since $\mathcal{U}_1$ is the near user who enjoys better CSI in most of the fading states, the optimal power policy that maximizes the average sum-rate for both NOMA and OMA-Type-II is to allocate power only to $\mathcal{U}_1$ in such states. Moreover, the advantage of NOMA begins promising when the system requires a larger $\bar{R}$ ($\bar{R}'$), in that NOMA guarantees the minimum average rate achieved by $\mathcal{U}_2$ while keeping $\mathcal{U}_1$'s average rate the maximum.

B. Delay-Limited Transmission

Fig. 5(a) shows the optimal trade-offs between the sum of DLT and the maximum permissive outage probability, i.e., $\bar{\zeta}$, under full CSIT given the same prescribed rate $\bar{R}_k = 2$bits/sec/Hz for each user. It is seen that when the two users suffer from near-far unfairness, the optimum sum-DLT versus max-outage trade-off achieved by NOMA outperforms that achieved by OMA-Type-II and OMA-Type-I. However, this superiority almost disappears when $\mathcal{U}_1$ and $\mathcal{U}_2$ are both 0.5km away from the BS. This is because in this case $g_1(\nu) \approx g_2(\nu)$ in most fading states, thanks to which the total amount of

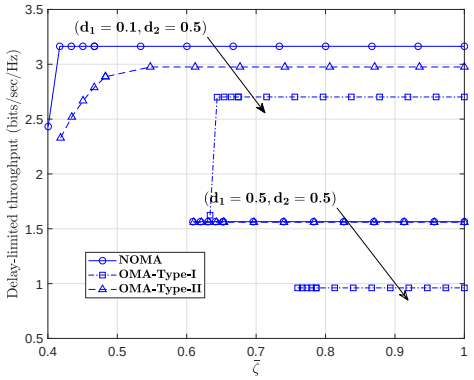


(a) Full CSIT.

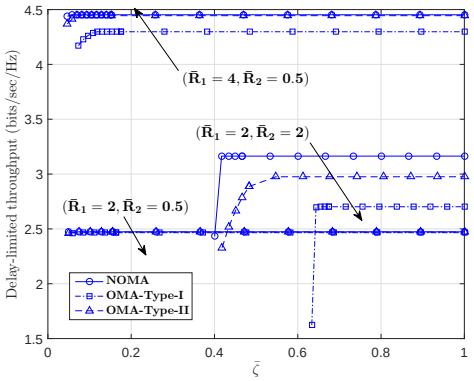


(b) Partial CSIT.

Fig. 4. The average achievable rate allocation between the two users with different minimum rate requirements, under full and partial CSIT, respectively.



(a)



(b)

Fig. 5. The DLT of the system versus the maximum permissive outage probability under full CSIT.

transmit power saved by NOMA tends to be less. Furthermore, no much trade-off is seen for the sum of DLT versus user fairness, as the two users hold similar chances to be the stronger user, and therefore when ζ_k^{XX} is minimized, $\zeta_{\bar{k}}^{\mathrm{XX}}$ is nearly minimized as well, where $(\cdot)^{\mathrm{XX}}$ stands for NOMA or OMA-Type-II.

On the other hand, in near-far channel conditions, the impact of different $\bar{R}_k$s on the optimum sum-DLT versus max-outage trade-off is demonstrated in Fig. 5(b). With the same intended rate $\bar{R}_1 = \bar{R}_2 = 2$bits/sec/Hz, the optimum trade-off achieved by NOMA outperforms that achieved by the OMA schemes. By contrast, when $\bar{R}_2$ reduces to 0.5bits/sec/Hz, the trade-off becomes trivial, since in this case the stronger user's advantage in saving power is compromised by its higher target rate.

Fig. 6 shows the optimum sum-DLT versus max-outage trade-offs achieved by various schemes with different settings of $\bar{R}_1$ and $\bar{R}_2$. Unlike in Fig. 5(b), the superiority of NOMA over the other OMA schemes is significantly seen in Fig. 6. Moreover, the minimum max-outage achieved by NOMA is significantly lower than that attained by other schemes. For example, with $\bar{R}_1 = \bar{R}_2 = 1$bist/sec/Hz, $\min\limits_{p_s, p_w} \{\max\limits_k \zeta_k'^{\mathrm{NOMA}}\}$ falls below 0.47 while that achieved by OMA-Type-I and OMA-Type-II is as large as about 0.49 and 0.60, respectively.

The DLT allocation between the two users subject to different maximum permissive outage is reflected by their

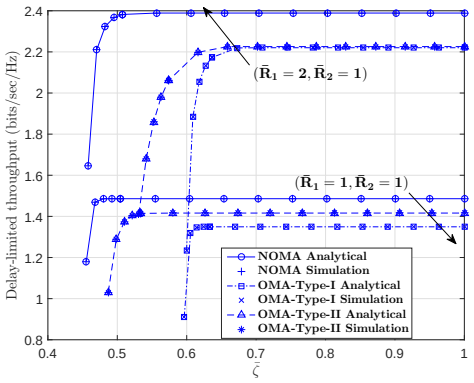
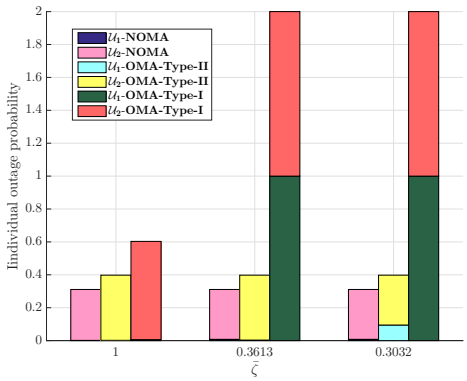
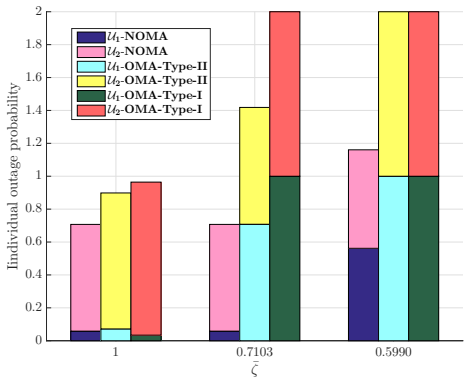


Fig. 6. The DLT of the system versus the maximum permissive user outage probability under partial CSIT.



(a) Full CSIT.



(b) Partial CSIT.

Fig. 7. The DLT allocation between the two users with different maximum outage requirements, under full and partial CSIT.

outage probability allocation in Fig. 7 under full and partial CSIT, respectively. Under full CSIT, it is seen from Fig. 7(a) that with $\bar{R}_1 = \bar{R}_2 = 2$bits/sec/Hz and $\bar{P} = 2$Watt, $\mathcal{U}_1$ achieves almost negligible outage while $\mathcal{U}_2$ can achieve an outage probability as low as 0.3032 by NOMA. OMA-Type-II follows the same trend unless $\mathcal{U}_1$ has to claim more outage states to reserve power for $\mathcal{U}_2$'s transmission. $\mathcal{U}_1$'s outage probability compromised by satisfying a lower maximum outage constraint is larger in the case of partial CSIT than in the case of full CSIT.

VI. Conclusion

In this paper, we have investigated the average sum-rate and/or the sum of DLT maximization for a two-user downlink NOMA over fading channels imposing QoS constraints on the worst user performance. Under full CSIT, the non-convex resource allocation problems have been solved using the technique of dual decomposition leveraging "time-sharing" conditions. Under partial CSIT, the individual ergodic rate and/or outage probability have been characterized in closed-form, based on which the optimal power policies have been numerically obtained. Simulation results have unveiled that the optimal NOMA-based power allocation schemes in general outperform the optimal OMA-based ones in terms of various throughput versus fairness trade-offs, especially when the two users' channels experience contrasting fading gains.

References

[1] H. Xing, Y. Liu, A. Nallanathan, and Z. Ding, "Sum-rate maximization guaranteeing user fairness for NOMA in fading channels," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Barcelona, Spain, Apr. 2018.

[2] 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. IEEE Vehicular Technology Conference (VTC Spring)*, Dresden, Germany, June 2013.

[3] H. Xie, B. Wang, F. Gao, and S. Jin, "A full-space spectrum-sharing strategy for massive MIMO cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 10, pp. 2537–2549, Oct. 2016.

[4] L. Dai, B. Wang, Y. Yuan, S. Han, C. l. I, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sept. 2015.

[5] 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 Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[6] 3rd Generation Partnership Project (3GPP), "Study on downlink multiuser superposition transmission for LTE," Mar. 2015.

[7] L. Zhang, W. Li, Y. Wu, X. Wang, S. I. Park, H. M. Kim, J. Y. Lee, P. Angueira, and J. Montalban, "Layered-division-multiplexing: theory and practice," *IEEE Trans. on Broadcast.*, vol. 62, no. 1, pp. 216–232, Mar. 2016.

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

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

[10] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Coordinated beamforming for multi-cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 84–87, Jan. 2017.

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

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

[13] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, Oct. 2015.

[14] Y. Liu, M. Elkashlan, Z. Ding, and G. K. Karagiannidis, "Fairness of user clustering in MIMO non-orthogonal multiple access systems," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1465–1468, Jul. 2016.

[15] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5191–5202, Oct 2017.

[16] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[17] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[18] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels .I. ergodic capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1083–1102, Mar. 2001.

[19] ——, "Capacity and optimal resource allocation for fading broadcast channels .II. outage capacity," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 1103–1127, Mar. 2001.

[20] H. Xing, L. Liu, and R. Zhang, "Secrecy wireless information and power transfer in fading wiretap channel," *IEEE Trans. Veh. Technol.*, vol. 65, no. 1, pp. 180–190, Jan. 2016.

[21] N. Jindal and A. Goldsmith, "Capacity and optimal power allocation for fading broadcast channels with minimum rates," *IEEE Trans. Inf. Theory*, vol. 49, no. 11, pp. 2895–2909, Nov. 2003.

[22] D. Hughes-Hartogs, "The capacity of a degraded spectral gaussian broadcast channel," Ph.D. dissertation, Inform. Syst. Lab., Ctr. Syst. Res., Stanford Univ., Stanford, CA, Jul. 1975.

[23] *Study on Downlink Multiuser Superposition Transmission (MUST) for LTE (Release 13)*, 3GPP document TR 36.859, Dec. 2015.

[24] Y. Wu, C. K. Wen, C. Xiao, X. Gao, and R. Schober, "Linear precoding for the MIMO multiple access channel with finite alphabet inputs and statistical CSI," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 983–997, Feb. 2015.

[25] Z. Dong, H. Chen, J. K. Zhang, and L. Huang, "On non-orthogonal multiple access with finite-alphabet inputs in Z-channels," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2829–2845, Dec. 2017.

[26] S. V. Hanly and D. N. C. Tse, "Multiaccess fading channels. ii. delay-limited capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2816–2831, Nov 1998.

[27] H. Tabassum, E. Hossain, and J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access in large-scale cellular networks using Poisson cluster processes," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3555–3570, Aug. 2017.

[28] W. Yu and R. Lui, "Dual methods for nonconvex spectrum optimization of multicarrier systems," *IEEE Trans. Commun.*, vol. 54, no. 7, pp. 1310–1322, July 2006.

[29] R. T. Rockafellar, *Convex Analysis*. Princeton Univ. Press, 1997.

[30] S. Boyd, "Lecture notes for EE364b: Convex Optimization II." [Online]. Available: https://stanford.edu/class/ee364b/lectures.html

[31] H. Xing, Y. Liu, A. Nallanathan, Z. Ding, and H. V. Poor, "Optimal throughput fairness trade-offs for downlink non-orthogonal multiple access over fading channels," 2017. [Online]. Available: https://arxiv.org/abs/1708.01403

[32] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Non-orthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10 152–10 157, Dec. 2016.

[33] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, 9th ed. Mineola, NY, USA: Dover Publication, Inc., 1972.