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

Hong Xing, Yuanwei Liu, A. Nallanathan, Zhiguo Ding and H. Vincent Poor

Abstract

Recently, non-orthogonal multiple access (NOMA) has attracted considerable interest as one of the 5G-defining techniques. However, as NOMA is intrinsically in favour of the transmission of strong users who are capable of carrying out 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 performed under partial CSIT. Furthermore, a sophisticated orthogonal multiple access (OMA) scheme is also studied as a benchmark to prove the superiority of NOMA over OMA with full CSIT. Finally, the theoretical analysis is verified via simulations by means of various trade-offs for the average sum-rate (sum-DLT) versus the minimum (maximum) user rate (outage) requirement.

Index Terms

Non-orthogonal multiple access, orthogonal multiple access, fading channel, fairness, strong duality, dual decomposition

Part of this paper has been submitted to the IEEE Wireless Communications and Networking Conference (WCNC), 2018.

H. Xing, Y. Liu, and A. Nallanathan are with the Department of Informatics, King’s College London, London, WC2R 2LS, U.K. (e-mails: hong.xing@kcl.ac.uk; yuanwei.liu@kcl.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).
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 [1–3], 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 [4] and the references therein). Hence, it has recently sparked widespread interest in both industry [5, 6] and academia [7–16]. Variation forms of NOMA, namely, multiuser superposition transmission (MUST) and layer division multiplexing (LDM), have been included in the 3rd Generation Partnership Project Long Term Evolution Advanced (3GPP-LTE-A) [5] and the next general digital TV standard ATSC 3.0 [6], respectively.

Among a variety of studies addressing the challenges posed by NOMA, a general NOMA downlink framework was proposed in [7] 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 multipoint (CoMP) transmission techniques were adopted in [8] and [9] 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. There are mainly three types of countermessures against such unfairness in NOMA networks. The first strategy is to invoke cooperative NOMA [10, 11], in which a nearby user is regarded as a relay to assist a distant user. It is demonstrated in [10] that by utilizing the proposed cooperative protocol, all users experience the same diversity order. In [11] 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 [12–14]. The max-min power allocation problem that maximizes the minimum achievable user rate was studied for single-input-single-output (SISO) NOMA systems in [12], and for clustered multiple-input-multiple-output (MIMO) NOMA systems in [13]. In [14], 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 [15, 16]. [15] considered a multicarrier downlink network, in which each sub-channel can be shared by multiple users by adopting NOMA. Joint sub-channel and power allocation was formulated as a weighted sum-rate maximization problem, and iteratively solved by leveraging a matching problem with externalities. multicarrier NOMA systems employing a full-duplex (FD) BS was considered in [16], and an optimal joint sub-carrier and power policy for maximizing the weighted sum-rate was developed by applying monotonic optimization.

A. Related Work

The above fairness issues were considered for NOMA operating over added white Gaussian Noise (AWGN) channels. In this paper, by contrast, we consider a two-user downlink NOMA network operating over a fading channel, in which optimal power and rate allocation strategies are applied adaptively to channel dynamics for delay-tolerant transmission subject to a minimum average rate constraint, and optimal power controls are derived to maintain a target rate vector at as many fading states as possible for delay-limited transmission subject to a maximum outage constraint.

The information theoretic study of fading broadcast channels (BCs) can be traced back to [17] and [18]. Assuming perfect CSI at both the transmitter (Tx) and the receivers (Rxs), dynamic power and rate allocations for various transmission schemes including code division 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 [17] and [18], respectively. The boundaries of the ECRs have been characterized in [17] 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 [18]. The boundaries of these regions were also obtained by solving equivalent sum-reward maximization problems [18].

However, these schemes do not favour far users that are susceptible to deep fading, since these users are either allocated little power in characterization of the ECRs or prohibited from transmission due to large power expenditure at the boundary of the outage probability regions. [19] investigated the optimal power allocation over fading BCs taking fairness into account. Specifically, the authors showed that the minimum-rate capacity region is equal to the sum of
the zero-outage minimum-rate vector plus the ECR of a related BC with effective noise. The boundary of this region was then derived based on known results for the ECR in [17].

Our work differs from the above mainly in three aspects. First, instead of an instantaneous minimum rate constraint that cannot be violated in every fading state as considered in [19], we consider an average minimum rate constraint in delay-tolerant applications, and a maximum outage constraint in delay-limited applications, respectively. Hence, we cannot directly relate our proposed trade-off regions to the well investigated ergodic capacity regions in [17]. Second, the key step of weighted sum-utility reformulation in [17] and [18] cannot be readily used for our problems due to the extra quality-of-service (QoS) constraints guaranteeing worst user performance in ours. Further, the optimal solutions are obtained using different techniques. We tackle the non-convexity of the problems using Lagrangian dual decomposition together with the ellipsoid method leveraging “time-sharing” conditions [20], while [17] adopted a multi-level “water-filling” algorithm and [18] derived a multi-user generalization of the threshold-decision based power policy.

The main contributions of this paper are summarized as follows. We 1) solve the ergodic sum-rate (ESR) maximization problem ensuring a minimum average rate by optimally adapting the power and rate allocations to fading states with full CSI at the Tx (CSIT) for both NOMA and an optimal OMA scheme; 2) obtain the optimal power control to the sum of delay-limited throughput (DLT) maximization problem, which is subject to a maximum 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 optimal fixed 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.

**Notation**—We use upper-case boldface letters for matrices and lower-case boldface letters for vectors. \( \nabla_x f(x) \) denotes the gradient of \( f(x) \) with respect to (w.r.t) \( x \). \( \mathbb{E}_x[\cdot] \) 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 \( \sigma^2 \) is denoted by \( \mathcal{CN}(u, \sigma^2) \). \( \text{Ei}(x) = \int_{-\infty}^{x} \frac{e^t}{t} \, dt \) (\( x < 0 \)) 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 downlink cellular system that consists of one BS and two users, which are denoted by $U_k$, $k \in \{1, 2\}$, with each of their transceivers equipped with a single antenna. We assume that the complex channel coefficient from the BS to $U_k$, $h_k(\nu)$ experiences block fading with a continuous joint probability density function (pdf), where $\nu$ represents a fading state. The channel remains constant during each transmission block, but may vary from block to block as $\nu$ changes. The channel gain $|h_k(\nu)|^2$ is assumed to consist of multiplicative small scale and large scale fading given by $|h_k(\nu)|^2 = \frac{|\tilde{h}_k(\nu)|^2}{\lambda_k}$, in which $\tilde{h}_k(\nu)$ is a complex Gaussian RV denoted by $\tilde{h}_k(\nu) \sim CN(0, 1)$, and $\lambda_k$ is a distant-dependent constant. Hence, $|h_k(\nu)|^2$ is an exponentially distributed RV with its mean value specified by $1/\lambda_k$.

In this paper, we investigate two types of CSIT, i.e., full CSIT and partial CSIT, while CSI at the Rx's 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 successive interference cancellation (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 [14]. (Another benchmark scheme, OMA-Type-I, will be introduced in Section V.)

A. Full CSIT

1) NOMA: For NOMA transmission, the received signal at the downlink user $U_k$ is given by

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

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 $U_k$'s, denoted by $s_k \sim CN(0, 1)$; $p_k(\nu)$'s denotes $U_k$'s transmit power; $n_k$'s is the AWGN at $U_k$'s Rx, denoted by $n_k \sim CN(0, \sigma_k^2)$. Moreover, similar to [19], 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., \(\mathbb{E}_\nu [p_k(\nu) + p_{\hat{k}}(\nu)] \leq \hat{P}\), and the latter limits the instantaneous total transmit power below \(\hat{P}\), i.e., \(P_k(\nu) + p_{\hat{k}}(\nu) \leq \hat{P}\), \(\forall \nu\). It is assumed that \(\hat{P} \leq \hat{P}\) without loss of generality (w.l.o.g.).

If \(g_k(\nu) > g_{\hat{k}}(\nu)\), where \(\frac{|h_k(\nu)|^2}{\sigma_k^2} \triangleq g_k(\nu)\), it implies that \(\log_2(1 + \frac{p_k(\nu)g_k(\nu)}{p_{\hat{k}}(\nu)g_{\hat{k}}(\nu)+1}) > \log_2(1 + \frac{p_k(\nu)g_k(\nu)}{p_k(\nu)g_k(\nu)+1})\). In other words, in this CSI condition, the achievable rate for \(U_k\) (the stronger user) to decode \(U_{\hat{k}}\) (the weaker user)’s message is larger than the rate intended for \(U_{\hat{k}}\’\)s transmission, and therefore \(U_{\hat{k}}\) is able to successfully perform SIC. Otherwise, \(U_k\) directly decodes its own message treating \(U_{\hat{k}}\’\)s as interference. The instantaneous achievable rate for \(U_k\) at fading state \(\nu\) in bits/sec/Hz is thus given by [17]

\[
P_{k}^{\text{NOMA}}(\nu) = \begin{cases} 
\log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{\rho_k(\nu)g_k(\nu)+1} \right), & \text{if } g_k(\nu) > g_{\hat{k}}(\nu) \\
\log_2 \left( 1 + \frac{p_{\hat{k}}(\nu)g_{\hat{k}}(\nu)}{\rho_k(\nu)g_k(\nu)+1} \right), & \text{otherwise}
\end{cases}, \quad k \in \{1, 2\}.
\]  

(2)

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 \(\nu\), where \(\alpha_k(\nu) \in [0, 1]\) denotes the proportion of time/frequency shared by \(U_k\). The same sets of transmit power constraints as its NOMA counterpart, i.e., APC and PPC, are taken into account as well. Accordingly, the instantaneous achievable rate for \(U_k\) in fading state \(\nu\) is given by

\[
R_k^{\text{OMA-II}}(\nu) = \alpha_k(\nu) \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{\alpha_k(\nu)} \right), \quad k \in \{1, 2\}.
\]  

(3)

Note that (3) applies to both TDMA and FDMA transmission in the sense that the total energy consumed for the two users in fading state \(\nu\) over time remains the same as over frequency, which is given by \(\alpha_k(\nu)p_k(\nu) + \alpha_{\hat{k}}(\nu)p_{\hat{k}}(\nu) = p_k(\nu) + p_{\hat{k}}(\nu)\). Hence, we do not differentiate TDMA from FDMA throughout the paper unless otherwise specified.

B. Partial CSIT

1) NOMA: Under partial CSIT, for NOMA transmission, the BS does not know the exact CSI of the two users but their relation and statistical characteristics, and thus we assume a binary power allocation strategy as follows. In each fading state \(\nu\), 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 fixed over all fading states, and therefore only APC applies, i.e., \(p_s + p_w \leq \hat{P}\).
On the other hand, thanks to the perfect CSI at the receiver, the stronger user can still perform SIC, which yields the instantaneous rate $R_{k}^{\text{NOMA}}(\nu)$ for $U_k$ expressed as

$$R_{k}^{\text{NOMA}}(\nu) = \begin{cases} \log_2 \left( 1 + p_s g_k(\nu) \right), & \text{if } g_k(\nu) > g_k(\nu) \\ \log_2 \left( 1 + \frac{p_w g_k(\nu)}{p_s g_k(\nu) + 1} \right), & \text{otherwise} \end{cases}, \quad k \in \{1, 2\}. \quad (4)$$

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 $U_k$ is transmitted with power $p_s$ if its channel gain from the BS is stronger than $U_k$’s, and with power $p_w$ otherwise. The fixed proportion of time/frequency assigned to $U_k$ and $U_k$ is $\alpha_k$ and $\alpha_k$, respectively. Consequently, the instantaneous achievable rate for $U_k$ is expressed as

$$R_{k}^{\text{OMA-II}}(\nu) = \begin{cases} \alpha_k \log_2 \left( 1 + \frac{p_w g_k(\nu)}{\alpha_k} \right), & \text{if } g_k(\nu) > g_k(\nu) \\ \alpha_k \log_2 \left( 1 + \frac{p_w g_k(\nu)}{\alpha_k} \right), & \text{otherwise} \end{cases}, \quad k \in \{1, 2\}. \quad (5)$$

C. **System Throughput**

1) **Delay-Tolerant Transmission:** First, we define the proper metric accounted for the overall performance of downlink NOMA and/or OMA-Type-II system in delay-tolerant types of transmission, where the users will not decode their individual data until the end of all the transmission blocks. A relevant metric in this case is *ergodic sum-rate (ESR)* [17] that is the achievable sum-rate of the two users averaged over all fading states given by $\mathbb{E}_\nu[R_{k}^{\text{NOMA}}(\nu) + R_{k}^{\text{OMA}}(\nu)]$ ($\mathbb{E}_\nu[R_{k}^{\text{NOMA}}(\nu) + R_{k}^{\text{OMA}}(\nu)]$), and $\mathbb{E}_\nu[R_{k}^{\text{OMA-II}}(\nu) + R_{k}^{\text{OMA-II}}(\nu)]$ ($\mathbb{E}_\nu[R_{k}^{\text{OMA-II}}(\nu) + R_{k}^{\text{OMA-II}}(\nu)]$), for NOMA and OMA-Type-II transmission, respectively.

2) **Delay-Limited Transmission:** Next, consider the delay-limited types of transmission for downlink NOMA and/or OMA-Type-II system, in which the BS concurrently transmits to $U_k$ and $U_k$ at a predefined rate of $\bar{R}_k$ and $\bar{R}_k$, respectively. As the user attempts to recover its message instantaneously at the end of every fading state, to properly measure the performance

\(^1\)The terms “average sum-rate” and the “ESR” are used interchangeably throughout the paper.
of each user in such case, the outage probability with which the NOMA user $\mathcal{U}_k$ fails to decode its intended data w.r.t the target rate $\bar{R}_k$, $k \in \{1, 2\}$, under full CSIT is introduced as follows:

$$
\zeta_k^{\text{NOMA}} =
\begin{cases}
\Pr \left\{ \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{P_k(\nu)g_k(\nu)+1} \right) < \bar{R}_k, \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{P_k(\nu)g_k(\nu)+1} \right) < \bar{R}_k \right\}, & \text{if } g_k(\nu) > g_k(\nu) \\
\Pr \left\{ \log_2 \left( 1 + p_k(\nu)g_k(\nu) \right) < \bar{R}_k, \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{P_k(\nu)g_k(\nu)+1} \right) \geq \bar{R}_k \right\}, & \text{otherwise}
\end{cases}
$$

(6)

As seen from (6), 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 $\mathcal{U}_k$’s message. If it fails to decode $\mathcal{U}_k$’s message at a transmission rate of $\bar{R}_k$, it has to decode its own by treating $\mathcal{U}_k$’s as interference. Otherwise, if it successfully recovers $\mathcal{U}_k$’s, SIC is then 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 treating interference as noise. In each fading state $\nu$, an outage indicator function is defined as follows [21].

$$
X_k^{\text{NOMA}}(\nu) =
\begin{cases}
1, & g_k(\nu) > g_k(\nu), \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{P_k(\nu)g_k(\nu)+1} \right) < \bar{R}_k, \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{P_k(\nu)g_k(\nu)+1} \right) < \bar{R}_k \\
1, & g_k(\nu) > g_k(\nu), \log_2 \left( 1 + p_k(\nu)g_k(\nu) \right) < \bar{R}_k, \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{P_k(\nu)g_k(\nu)+1} \right) \geq \bar{R}_k \\
1, & g_k(\nu) \leq g_k(\nu), \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{P_k(\nu)g_k(\nu)+1} \right) < \bar{R}_k \\
0, & \text{otherwise}
\end{cases}
$$

(7)

Hence, it is easily checked that $\mathbb{E}_\nu[X_k^{\text{NOMA}}(\nu)] = \zeta_k^{\text{NOMA}}$, $k \in \{1, 2\}$.

For OMA-Type-II transmission, the outage probability of one user is defined independent of the other as follows:

$$
\zeta_k^{\text{OMA-II}} = \Pr \left\{ \alpha_k(\nu) \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{\alpha_k(\nu)} \right) < \bar{R}_k \right\}, k \in \{1, 2\}.
$$

(8)

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

$$
X_k^{\text{OMA-II}}(\nu) =
\begin{cases}
1, & \alpha_k(\nu) \log_2 \left( 1 + \frac{p_k(\nu)g_k(\nu)}{\alpha_k(\nu)} \right) < \bar{R}_k \\
0, & \text{otherwise}
\end{cases}
$$

(9)

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 $\bar{R}_k(1 - \zeta_k^{\text{NOMA}}) + \bar{R}_k(1 - \zeta_k^{\text{OMA}})$, and $\bar{R}_k(1 - \zeta_k^{\text{OMA-II}}) + \bar{R}_k(1 - \zeta_k^{\text{OMA-II}})$, for NOMA and OMA-Type-II transmission under full CSIT, respectively. The sum of
DLT for NOMA and OMA-Type-II transmission under partial CSI is also similarly given by 
\( \bar{R}_k(1 - \zeta_k^{NOMA}) + \bar{R}_k(1 - \zeta_k^{OMA-II}) \), and \( \bar{R}_k(1 - \zeta_k^{OMA-II}) + \bar{R}_k(1 - \zeta_k^{OMA-II}) \), respectively.

### III. Optimum Delay-Tolerant Transmission

In delay-tolerant scenarios, ESR is maximized at the cost of compromising fairness among users by allocating the resources to the user of better channel condition over all fading states. To maximize the ESR of the system while guaranteeing certain level of fairness, a minimum achievable ergodic rate requirement for each user, namely, \( \mathbb{E}_\nu[R_{XX}^k(\nu)] \geq \bar{R} \left( \mathbb{E}_\nu[R_{XX}^{\bar{k}}(\nu)] \geq \bar{R}' \right) \), \( k \in \{1, 2\} \), is imposed, where \((\cdot)^{XX}\) 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.

#### 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:

\[
(P1-XX): \quad \text{Maximize} \quad \mathbb{E}_\nu[R_{XX}^k(\nu) + R_{XX}^{\bar{k}}(\nu)]
\]

Subject to
\[
\mathbb{E}_\nu[p_k(\nu) + p_{\bar{k}}(\nu)] \leq \bar{P}, \quad (10a)
\]
\[
p_k(\nu) + p_{\bar{k}}(\nu) \leq \hat{P}, \quad \forall \nu, \quad (10b)
\]
\[
p_k(\nu) \geq 0, \quad p_{\bar{k}}(\nu) \geq 0, \quad \forall \nu, \quad (10c)
\]
\[
\mathbb{E}_\nu[R_{XX}^k(\nu)] \geq \bar{R}, \quad \forall k, \quad (10d)
\]

where the exclusive parameters for OMA-Type-II, \( \{\alpha_k(\nu)\} \)'s, are only valid when XX is replaced by OMA-Type-II.

It is worthy of noting that the constraints (10d) guarantee that even under circumstances of extremely asymmetric CSI, e.g., when \( U_1 \) is located nearer to the BS and thus enjoys much better CSI than \( U_2 \) at most fading states, the BS can still delivery information to \( U_2 \) at an average rate larger than \( \bar{R} \). However, this average rate for \( U_2 \) cannot be achieved without (w/o) the minimum ergodic rate constraint.
1) Optimal Solution to (P1-NOMA): In this section, we aim for solving (P1-NOMA). Problem (P1-NOMA) is non-convex due to the non-convex objective function (c.f. (2)), and therefore no immediate solution can be given. However, for channel fading following continuous distributions, (P1-NOMA) proves to satisfy the “time-sharing” condition in an asymptotic sense [20]. Hence, we can still optimally solve it via its dual problem [21], since strong duality holds under this condition [22]. Next, we apply the Lagrangian dual method to solve (P1-NOMA), the Lagrangian of which is given by

$$\mathcal{L}_{1}^{\text{NOMA}} (\{p_k(\nu)\}, \{p_\bar{k}(\nu)\}, \lambda, \delta, \mu) = \mathbb{E}_\nu [(1 + \delta)R_k^{\text{NOMA}}(\nu) + (1 + \mu)R_\bar{k}^{\text{NOMA}}(\nu) - \lambda(p_k(\nu) + p_\bar{k}(\nu))] + \lambda \bar{P} - \delta \bar{R} - \mu \bar{R}, \quad (11)$$

where \(\lambda\) is the Lagrangian multiplier associated with the APC given in (10a); \(\delta\) and \(\mu\) are those associated with the ergodic rate constraints given in (10d) for \(U_k\) and \(U_\bar{k}\), respectively. The dual function of (P1-NOMA) corresponding to (11) is accordingly given by

$$g(\lambda, \delta, \mu) = \max_{p_k(\nu) \geq 0, p_\bar{k}(\nu) \geq 0, \{p_k(\nu) + p_\bar{k}(\nu) \leq \bar{P}\}} \mathcal{L}_{1}^{\text{NOMA}} (\{p_k(\nu)\}, \{p_\bar{k}(\nu)\}, \lambda, \delta, \mu). \quad (12)$$

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

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

It is observed that \(g(\lambda, \delta, \mu)\) is obtained by maximizing the Lagrangian given in (11), which can be decoupled into as many subproblems as the number of fading states all sharing the same structure. The index \(\nu\) is now safely dropped for the ease of exposition. Taking one particular fading state as an example, the associated subproblem given a triple \((\lambda, \delta, \mu)\) can be expressed as

$$(\text{P1-NOMA-sub}) : \text{Maximize}_{p_k \geq 0, p_\bar{k} \geq 0} \quad \tilde{\mathcal{L}}_{1}^{\text{NOMA}}(p_k, p_\bar{k})$$

Subject to \(p_k + p_\bar{k} \leq \bar{P}\),

where \(\tilde{\mathcal{L}}_{1}^{\text{NOMA}}(p_k, p_\bar{k}) = (1 + \delta)R_k^{\text{NOMA}} + (1 + \mu)R_\bar{k}^{\text{NOMA}} - \lambda(p_k + p_\bar{k})\). 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.
If \((0, U)\) where \(U\) lines are denoted by \(
abla \Psi\) or on the boundary of \(U\) in favour of power allocation policy advocates fairness. To elaborate further, assuming the system design is apply to the fading states where \(g\) is given by (15). Some insights can be drawn from Proposition 3.1 on how the optimal power allocation to Problem (P1-NOMA-sub) assuming \(g_1 > g_2\) is given by

\[
\begin{aligned}
(p_1^*, p_2^*) &= \arg \max \{ \mathcal{L}_1^{\text{NOMA}}(0, 0), \mathcal{L}_1^{\text{NOMA}}(0, \hat{P}), \mathcal{L}_1^{\text{NOMA}}(\hat{P}, 0),
\mathcal{L}_1^{\text{NOMA}}(p_{i,1}, p_{i,2}) \}, \forall i \in \{1, 2, 3, 4\},
\end{aligned}
\]  

where \(p_{i,1}\)'s and \(p_{i,2}\)'s are given by (15) at the top of next page.

**Proof:** Since \(\mathcal{L}_1^{\text{NOMA}}(p_1, p_2)\) 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 \(\Psi\) depending on whether \((p_{4,1}, p_{4,2}) \in \Psi\) or not. We calculate \((p_{4,1}, p_{4,2})\) as follows:

\[
(p_{4,1}, p_{4,2}) = \arg \left\{ \nabla (p_1, p_2) \mathcal{L}_1^{\text{NOMA}}(p_1, p_2) = 0 \right\}.
\]  

If \((p_{4,1}, p_{4,2}) \in \Psi\), the maximum is \(\mathcal{L}_1^{\text{NOMA}}(p_{4,1}, p_{4,2})\), 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, \(i = 1, 2, 3\), 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 \(\delta, p_1\), and \(g_1\) with \(\mu, p_2\), and \(g_2\), respectively, in (15). Some insights can be drawn from Proposition 3.1 on how the optimal power allocation policy advocates fairness. To elaborate further, assuming the system design is in favour of \(U_2\), i.e., \(\delta \ll \mu\), it is observed from (15) that \(p_{i,1}\) monotonically decreases with \(\mu\) while \(p_{i,2}\) monotonically increases with \(\mu\), which suggests that when the multiplier associated with \(U_2\)'s QoS requirement gets larger, the optimal power allocation policy tends to suppress \(U_1\)'s transmission while supporting \(U_2\)'s despite of a possible case of \(U_1\)'s CSI better than \(U_2\).

Thanks to Proposition 3.1, given a triple \((\lambda, \delta, \mu)\), \(g(\lambda, \delta, \mu)\) 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) [23, Localization methods]. The required sub-gradient for updating \((\lambda, \delta, \mu)\) turns out to be \((\tilde{P} - \mathbb{E}_\nu[p_k(\nu) + p_k(\nu)], \mathbb{E}_\nu[R_k^{\text{NOMA}}(\nu)] - \tilde{R}, \mathbb{E}_\nu[R_k^{\text{NOMA}}(\nu)] - \tilde{R})^T\), where \((p_k(\nu), p_k(\nu))\) is the optimal solution to (P1-NOMA-sub) at fading state \(\nu\), and \(R_k^{\text{NOMA}}(\nu)\)'s is obtained by substituting \((p_k(\nu), p_k(\nu))\) into (2).

Note that a feasible \(\tilde{R}\) in (10d) ensures the successful implementation of the ellipsoid method, and thus it is important to consider a reasonable \(\tilde{R}\) that does not exceed \(\tilde{R}_{\text{max}}\). We can obtain \(\tilde{R}_{\text{max}}\) by replacing the objective function of (P1-NOMA) with a variable \(\tilde{R}\) and then solving the feasibility problem by bi-section over \(\tilde{R}\). Since the involved procedure is quite similar to that for solving (P1-NOMA), we omit it herein for brevity.

2) Optimal Solution to (P1-OMA-II): In this section, we investigate Problem (P1-OMA-II).

First, (P1-OMA-II) is a convex problem, since (3) as the perspective of the jointly concave function \(\log_2(1 + g_k(\nu)p_k(\nu))\) 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

\[
\mathcal{L}_{1}^{\text{OMA-II}}(\{p_k(\nu)\}, \{p_k(\nu)\}, \{\alpha_k(\nu)\}, \lambda, \delta, \mu) = \mathbb{E}_\nu[(1 + \delta)R_k^{\text{OMA-II}}(\nu) + (1 + \mu)R_k^{\text{OMA-II}}(\nu) - \lambda(p_k(\nu) + p_k(\nu))] + \lambda\tilde{P} - \delta\tilde{R} - \mu\tilde{R},
\]

(16)

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

\[
\text{(P1-OMA-II-sub): Maximize } \mathcal{L}_{1}^{\text{OMA-II}}(p_k, p_k, \alpha_k) \text{ Subject to } p_k + p_k \leq \hat{P}, 0 \leq \alpha_k \leq 1, \forall k,
\]

where the index \(\nu\) 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 $\hat{L}_1^{\text{OMA-II}}(p_1, p_2, \alpha_1)$ is achieved by its jointly stationary point, it is necessary to have the following conditions satisfied:

\[
\begin{align*}
    h(\lambda, \delta, \mu) &= 0, \quad (17a) \\
    c_1 &\geq 0, \quad (17b) \\
    c_2 &\geq 0, \quad (17c) \\
    \begin{cases}
        \frac{\hat{p} - c_2}{c_1 - c_2} &\geq 0, \text{ if } c_1 > c_2 \\
        \frac{\hat{p} - c_2}{c_1 - c_2} &< 1, \text{ otherwise}
    \end{cases}, \quad (17d)
\end{align*}
\]

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

\[
    h(\lambda, \delta, \mu) = (1 + \delta) \log_2 \left(\frac{1 + \delta}{\lambda\ln 2 g_1}\right) - (1 + \mu) \log_2 \left(\frac{1 + \mu}{\lambda\ln 2 g_2}\right) - \lambda c_1 + \lambda c_2. \quad (18)
\]

The corresponding stationary point is given by

\[
    p_1^* = c_1 \alpha_1^*, \quad p_2^* = c_2 (1 - \alpha_1^*), \quad \alpha_1^* = \begin{cases}
        \forall \in [0, \min\{\frac{\hat{p} - c_2}{c_1 - c_2}, 1\}], &\text{if } c_1 > c_2 \\
        \forall \in [\frac{\hat{p} - c_2}{c_1 - c_2}, 1]. &\text{otherwise}
    \end{cases} \quad (19)
\]

Proof: First, solve $\nabla_{(p_1, p_2, \alpha_1)} \overline{L}_1^{\text{OMA-II}}(p_1, p_2, \alpha_1) = 0$ 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 $\overline{L}_1^{\text{OMA-II}}(p_1, p_2, \alpha_1)$ w.r.t $\alpha_1$, (17a) is obtained. Finally, constrain $P_1 \geq 0$, $p_2 \geq 0$, and $p_1 + p_2 \leq \hat{P}$, we arrive at (17b), (17c), and the feasible range for $\alpha_1$ given in (19), respectively. ■

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

\[
    \begin{cases}
        p_1^* = 0, \quad p_2^* = \hat{P}, \quad \alpha_1^* = 0, &\text{if } \frac{1+\mu}{1+\delta} > \frac{\log_2(1+\hat{P} g_1)}{\log_2(1+\hat{P} g_2)} \\
        p_1^* = \hat{P}, \quad p_2^* = 0, \quad \alpha_1^* = 1, &\text{otherwise}
    \end{cases} \quad (20)
\]

Proof: Please refer to 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

\[
    (p_1^*, p_2^*, \alpha_1^*) = \arg \max \{ \overline{L}_1^{\text{OMA-II}}(0, 0, 0), \overline{L}_1^{\text{OMA-II}}(0, \hat{P}, 0), \overline{L}_1^{\text{OMA-II}}(\hat{P}, 0, 1), \overline{L}_1^{\text{OMA-II}}(0, c_2, 0) \mathbb{1}_{c_2}, \overline{L}_1^{\text{OMA-II}}(c_1, 0, 1) \mathbb{1}_{c_1}\}, \quad (21)
\]

Proof: First, solve $\nabla_{(p_1, p_2, \alpha_1)} \overline{L}_1^{\text{OMA-II}}(p_1, p_2, \alpha_1) = 0$ 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 $\overline{L}_1^{\text{OMA-II}}(p_1, p_2, \alpha_1)$ w.r.t $\alpha_1$, (17a) is obtained. Finally, constrain $P_1 \geq 0$, $p_2 \geq 0$, and $p_1 + p_2 \leq \hat{P}$, we arrive at (17b), (17c), and the feasible range for $\alpha_1$ given in (19), respectively. ■
where \( \mathbb{1}_x \) is an indicator function defined as
\[
\mathbb{1}_x = \begin{cases} 
1, & \text{if } 0 \leq x \leq \hat{P} \\
0, & \text{otherwise}
\end{cases}
\]  
(22)

**Proof:** Please refer to Appendix B.

**Remark 3.1:** When there are only two users, the optimal solution given by (21) shares some philosophy in common with that achieves the boundary of the time division (TD) capacity region discussed in [17, Thoerem 3]. We focus on solving (P1-OMA-II-sub) in any fading state given a triple \((\lambda, \delta, \mu)\), while [17] maximized the total weighted sum-rate in any fading state by determining how to distribute \( P(n) \) among \( M = 2 \) users such that the instantaneous total power constraint \( \sum_{j=1}^{2} \tau_j P_j(n) = P(n) \) (c.f. [17, 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 \( \{(g_k(\nu), g_k(\nu)) : h(\lambda, \delta, \mu) = 0\} \) (c.f. (18)) assuming continuously joint distribution of \((g_k(\nu), g_k(\nu))\) is zero, and so is that of \( L \triangleq \{n : h(\lambda, n) = 0\} \) in [17, Thoerem 3]. This also explains why the maximum of \( L_{1OMA-II}(p_1, p_2, \alpha_1) \) cannot be achieved by its jointly stationary point in probability.

With Proposition 3.2, given a triple \((\lambda, \delta, \mu)\), (P1-OMA-II-sub) is first solved state by state; then by updating \((\lambda, \delta, \mu)\) in accordance with the associated sub-gradient \((\hat{P} - \mathbb{E}_\nu[p_k(\nu) + p_k(\nu)], \mathbb{E}_\nu[R_k^{OMA-II}(\nu)] - \hat{R}, \mathbb{E}_\nu[R_k^{OMA-II}(\nu)] - \hat{R})^T\), (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 \( \{p_1^*(\nu), p_2^*(\nu), \alpha_k^*(\nu)\} \). 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 \( \alpha_k \) associated with it is seen to be 1 (c.f. (21)). Hence, assuming \( p_1^* > 0 \) (\( \alpha_k^* = 1 \)) and \( p_2^* = 0 \) (\( \alpha_k^* = 0 \)), \( R_1^{NOMA} \) turns out be \( \log_2(1 + p_1^* g_1) \) and \( R_2^{NOMA} = 0 \), which is exactly equal to \( R_1^{OMA-II} \) and \( R_2^{OMA-II} \), respectively; vice versa when \( p_1^* = 0 \) (\( \alpha_k^* = 0 \)) and \( p_2^* > 0 \) (\( \alpha_k^* = 1 \)). The other trivial case is that \( R_k^{NOMA} = R_k^{OMA-II} = 0 \), \( \forall k \), when \( p_1^* = p_2^* = 0 \). Hence, with \( p_1 = p_1^* \) and \( p_2 = p_2^* \) in each fading state, \( \mathbb{E}_\nu[R_k^{NOMA}(\nu)] = \mathbb{E}_\nu[R_k^{OMA-II}(\nu)] \geq \hat{R}, \forall k \), is met. It is also easily examined that (10d) 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.
By analogy, the partial CSIT counterpart of Problem (P1-XX) is formulated as below:

\[(P1' - XX): \text{Maximize} \quad \mathbb{E}_\nu[R'^{XX}_k(\nu) + \bar{R}'^{XX}_k(\nu)] \]

\[\text{Subject to} \quad p_s + p_w \leq \bar{P}, \quad (23a)\]

\[p_s \geq 0, \; p_w \geq 0, \quad (23b)\]

\[\mathbb{E}_\nu[R'^{XX}_k(\nu)] \geq \bar{R}', \quad \forall k, \quad (23c)\]

where \(\alpha_k\)'s is only valid in the transmission adopting OMA-Type-II. Similar to Problem (P1-XX), (23c) constrain the minimum average user rate achieved by the two users.

1) Optimal Solution to (P1'-NOMA): In this section, we develop 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 \(R'^{NOMA}_k(\nu)\)'s as function of \(p_s\) and \(p_w\), and then solve (P1'-NOMA) in accordance with these expectation results.

First, denote the RV \(|h_k(\nu)|^2 (|h_k(\nu)|^2)\) by \(X (Y)^2\). Also, denote the SNR \(p_s g_k(\nu)\) and the SINR \(p_w g_k(\nu)/(p_s g_k(\nu) + 1)\) (c.f. (4)) by \(\Gamma_k\) and \(\tilde{\Gamma}_k\), respectively. It thus follows that the conditional cumulative density functions (CDFs) of \(\Gamma_k\) and \(\tilde{\Gamma}_k\) are given by

\[F_{\Gamma_k|X \geq Y}(z) = \frac{\Pr\{p_s X/\sigma^2_k \leq z, X \geq Y\}}{\Pr\{X \geq Y\}} = 1 - \frac{\lambda_k + \lambda_{\bar{k}}}{\lambda_k} e^{-(\lambda_k + \lambda_{\bar{k}})\varepsilon_k}, \quad (24)\]

\[F_{\Gamma_k|X < Y}(z) = \frac{\Pr\{p_w X/(p_s X + \sigma^2_k) \leq z, X < Y\}}{\Pr\{X < Y\}} = \begin{cases} 1, & \text{if } p_w - p_s z \leq 0 \\ 1 - e^{-(\lambda_k + \lambda_{\bar{k}})\varepsilon_k}, & \text{otherwise} \end{cases} \quad (25)\]

respectively, where \(\varepsilon_k \triangleq \frac{\sigma^2_k}{p_s}\) and \(\tilde{\varepsilon}_k \triangleq \frac{\sigma^2_k}{p_w - p_s z}\). In accordance with (24) and (25), \(\mathbb{E}_\nu[R'^{NOMA}_k(\nu)]\)'s can be obtained by the following proposition.

\[\text{Note that we assume } \sigma^2_k = \sigma^2, \; \forall k, \text{ throughout the paper such that the relation between the effective channels of the two users , i.e., } g_k(\nu) \text{ and } g_{\bar{k}}(\nu), \text{ is equivalent to that between } |h_k(\nu)|^2 \text{ and } |h_{\bar{k}}(\nu)|^2.\]
**Proposition 3.3:** The ergodic rate for NOMA user $\mathcal{U}_k$, $k \in \{1, 2\}$, under partial CSIT is given by

$$
\mathbb{E}_\nu[R_k^\text{NOMA}(\nu)] = \frac{2}{\ln 2} \frac{\lambda_k}{\lambda_k + \tilde{\Gamma}_k} f \left( \frac{(\lambda_k + \lambda_{\tilde{k}}) \sigma_k^2}{p_s} \right) - \frac{1}{\ln 2} \frac{\lambda_k}{1 + \lambda_{\tilde{k}}} f \left( \frac{(\lambda_k + \lambda_{\tilde{k}}) \sigma_k^2}{p_s + p_w} \right) - \frac{1}{\ln 2} f \left( \frac{\lambda_{\tilde{k}} \sigma_k^2}{p_{\tilde{s}}} \right),
$$

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

**Proof:** Please refer to Appendix C.

Since the optimization variable $p_w$ only contributes to $\mathbb{E}_\nu[R_k^\text{NOMA}(\nu)]$ via $A_1$, we examine the property of $\mathbb{E}_\nu[R_k^\text{NOMA}(\nu)]$ in terms of $p_w$ by studying $f(\frac{(\lambda_k + \lambda_{\tilde{k}}) \sigma_k^2}{p_s + p_w})$ (c.f. $A_1$) as follows:

$$
\frac{\partial f \left( \frac{(\lambda_k + \lambda_{\tilde{k}}) \sigma_k^2}{p_s + p_w} \right)}{\partial p_w} = \frac{\partial}{\partial p_w} \left( e^{\frac{(\lambda_k + \lambda_{\tilde{k}}) \sigma_k^2}{p_s + p_w}} \text{Ei} \left( -\frac{(\lambda_k + \lambda_{\tilde{k}}) \sigma_k^2}{p_s + p_w} \right) \right)
$$

$$
= \frac{1}{p_s + p_w} \left( u \text{Ei}(u) e^u - 1 \right) \left( u \ln \left( 1 + \frac{1}{u} \right) - 1 \right),
$$

where $u \triangleq \frac{(\lambda_k + \lambda_{\tilde{k}}) \sigma_k^2}{p_s + p_w}$, $\text{Ei}(x) = -\text{Ei}(-x)$ ($x > 0$). (a) is because of the inequality $\text{Ei}(x)e^x < \frac{\ln(1 + \frac{1}{x})}{x}$ ($X > 0$) [25, Eq. (5.1.20)], and (b) is due to $\ln(1 + \frac{1}{u}) \leq \frac{1}{u}$. Therefore $\mathbb{E}_\nu[R_k^\text{NOMA}(\nu)]$’s proves to monotonically increase with $p_w$. Next, we facilitate solving Problem (P1'-NOMA) with this monotonicity. As it is easily seen that given any $p_s$, (P1'-NOMA) attains its optimum value when $\mathbb{E}_\nu[R_k^\text{NOMA}(\nu)]$’s takes on its maximum w.r.t $p_w$, i.e., when $p_w = \bar{P} - p_s$ (c.f. (23a)), (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). In addition, the feasibility of (P1'-NOMA) can be similarly studied as that of (P1-NOMA).

2) **Optimal Solution to (P1'-OMA-II):** In this section, we optimally solve (P1'-OMA-II) by first characterizing $\mathbb{E}_\nu[R_k^\text{OMA-II}(\nu)]$’s in the form of function w.r.t $p_s, p_w$ and $\alpha_k$, and then optimizing these variables.

Denoting the SNR of $\mathcal{U}_k$ in the case of $X \geq Y$ by $\Gamma_k$, and that in the case of $X < Y$ by $\tilde{\Gamma}_k$ (c.f. (5)), the conditional CDFs of $\Gamma_k$ and $\tilde{\Gamma}_k$ are given by

$$
F_{\Gamma_k|X \geq Y}(z) = 1 - \frac{\lambda_k + \lambda_{\tilde{k}}}{\lambda_k} e^{-\lambda_k \varphi_k} + \frac{\lambda_k}{\lambda_{\tilde{k}}} e^{-(\lambda_k + \lambda_{\tilde{k}}) \varphi_k},
$$

(28)
\begin{equation}
F_{\Gamma_k|X<Y}(z) = 1 - e^{-(\lambda_k + \lambda_k)\tilde{\phi}_k},
\end{equation}

where \(\varphi_k \triangleq \frac{\alpha_k\sigma_k^2}{p_s}\) and \(\tilde{\varphi}_k \triangleq \frac{\alpha_k\sigma_k^2}{p_w}\). With (28) and (29), we have the following proposition.

**Proposition 3.4:** The ergodic rate for user \(U_k\) operating with OMA-type-II, \(k \in \{1, 2\}\), under partial CSIT is given by

\begin{equation}
\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_k} f\left( \frac{(\lambda_k + \lambda_k)\alpha_k\sigma_k^2}{p_s} \right) - \frac{\lambda_k}{\lambda_k + \lambda_k} f\left( \frac{(\lambda_k + \lambda_k)\alpha_k\sigma_k^2}{p_w} \right) \right).
\end{equation}

**Proof:** Please refer to Appendix D.

Thanks to the results in (27), \(\mathbb{E}_\nu[R_k^{\text{OMA-II}}(\nu)]\)'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 \((p_s, \alpha_k)\) 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 defined in (6)/(8). 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.

**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.

\[
(P2-XX) : \text{Maximize} \quad \bar{R}_k(1 - \xi^X_k) + \bar{R}_k(1 - \xi^X_k)
\]

Subject to
- \( \mathbb{E}_\nu[p_k(\nu) + p_\bar{k}(\nu)] \leq \bar{P} \) \quad (31a)
- \( p_k(\nu) + p_\bar{k}(\nu) \leq \hat{P}, \quad \forall \nu \) \quad (31b)
- \( p_k(\nu) \geq 0, \quad p_\bar{k}(\nu) \geq 0, \quad \forall \nu \) \quad (31c)
- \( \mathbb{E}_\nu[X_k^{\text{NOMA}}(\nu)] \leq \bar{\zeta}, \quad \forall k \) \quad (31d)

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., \( \bar{R}_k = \bar{R}_k = \bar{R} \), even if \( U_k \) and \( U_\bar{k} \) 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 (31d). As seen from (7) ((9)), the discrete value of \( X_k^{\text{NOMA}}(\nu) (X_k^{\text{OMA-II}}(\nu)) \)'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 asymptotically optimal solutions in subsection IV-A1 and IV-A2, respectively.

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

\[
L_2^{\text{NOMA}}(\{p_k(\nu)\}, \{p_\bar{k}(\nu)\}, \lambda, \delta, \mu) = \mathbb{E}_\nu[-\bar{R}_kX_k^{\text{NOMA}}(\nu) - \bar{R}_kX_\bar{k}^{\text{NOMA}}(\nu) - \lambda(p_k(\nu) + p_\bar{k}(\nu)) - \delta X_k^{\text{NOMA}}(\nu) - \mu X_\bar{k}^{\text{NOMA}}(\nu)] + \lambda\bar{P} + \delta\bar{\zeta} + \mu\bar{\zeta}, \quad (32)
\]

where \( \lambda \) is the Lagrangian multiplier associated with the APC; \( \delta \) and \( \mu \) are those associated with the maximum user outage probability constraints given in (31d) for \( U_k \) and \( U_\bar{k} \), respectively. In line with the principle of dual decomposition, (32) can be maximized by decoupling it into independent subproblems each for one fading state and solving those subproblems in parallel. Define \( L_2^{\text{NOMA}}(p_k, p_\bar{k}) = \bar{R}_kX_k^{\text{NOMA}} + \bar{R}_kX_\bar{k}^{\text{NOMA}} + \lambda(p_k + p_\bar{k}) + \delta X_k^{\text{NOMA}} + \mu X_\bar{k}^{\text{NOMA}} \). With the fading index \( \nu \) safely dropped, given the dual variables’ triple \( (\lambda, \delta, \mu) \), the following problem is typical of the subproblems sharing the same structure:

\[
(P2-\text{NOMA-sub}) : \text{Minimize} \quad L_2^{\text{NOMA}}(p_k, p_\bar{k}) \quad \text{Subject to} \quad p_k + p_\bar{k} \leq \hat{P}.
\]
Then, we investigate the possible combinations of outage occurrences for $\mathcal{U}_k$ and $\mathcal{U}_k$. Assuming $g_k > g_k$, $k \in \{1, 2\}$, the possible combinations of indicator function $X_k^{\text{NOMA}}$ and the corresponding decoding strategies adopted by $\mathcal{U}_k$ are summarized in Table I, where $\mathcal{U}_k \rightarrow \mathcal{U}_k$ represents that $\mathcal{U}_k$ directly decodes its own information treating interference as noise; $\mathcal{U}_k \rightarrow \mathcal{U}_k \rightarrow \mathcal{U}_k$ denotes $\mathcal{U}_k$’s attempt to perform SIC; $\not\rightarrow$ indicates failure of decoding. Specifically, if the first step succeeds, $\mathcal{U}_k$ is able to cancel the interference from $\mathcal{U}_k$, otherwise $\mathcal{U}_k$ continues to decode its own treating $\mathcal{U}_k$’s as interference. Based on Table I, we derive the optimal solution to (P2-NOMA-sub) in the following proposition.

**Table I**

| $\mathcal{U}_k$ | $\mathcal{U}_k \not\rightarrow \mathcal{U}_k \not\rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 1$ (I.A) | $\mathcal{U}_k \rightarrow \mathcal{U}_k \not\rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 1$ (I.IA) | $\mathcal{U}_k \not\rightarrow \mathcal{U}_k \rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 0$ (I.B) | $\mathcal{U}_k \rightarrow \mathcal{U}_k \rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 0$ (I.B) | $\mathcal{U}_k \not\rightarrow \mathcal{U}_k \not\rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 1$ (I.C) | $\mathcal{U}_k \rightarrow \mathcal{U}_k \not\rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 0$ (I.D) | $\mathcal{U}_k \rightarrow \mathcal{U}_k \rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 0$ (I.D) |
|----------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| $\mathcal{U}_k$ | $\mathcal{U}_k \not\rightarrow \mathcal{U}_k \not\rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 1$ (I.A) | $\mathcal{U}_k \rightarrow \mathcal{U}_k \not\rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 1$ (I.IA) | $\mathcal{U}_k \not\rightarrow \mathcal{U}_k \rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 0$ (I.B) | $\mathcal{U}_k \rightarrow \mathcal{U}_k \rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 0$ (I.B) | $\mathcal{U}_k \not\rightarrow \mathcal{U}_k \not\rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 1$ (I.C) | $\mathcal{U}_k \rightarrow \mathcal{U}_k \not\rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 0$ (I.D) | $\mathcal{U}_k \rightarrow \mathcal{U}_k \rightarrow \mathcal{U}_k$, $X_k^{\text{NOMA}} = 0$ (I.D) |

**Proposition 4.1:** The optimal power allocation to Problem (P2-NOMA-sub) assuming $g_k > g_k$ is given by

$$
(p^*_k, p^*_k) = \arg \min_i \{ \mathcal{L}_{2}^{\text{NOMA}}(p_{i,k}, p_{i,k}) \mathbb{1}_{p_{i,k} + p_{i,k}} \}, \ \forall i \in \{1, 2, 3, 4\},
$$

(33)

where $p_{i,k}$’s and $p_{i,k}$’s are given by

$$
\begin{align*}
    p_{1,k} &= 0, & p_{1,k} &= 0, \\
    p_{2,k} &= \frac{2^{R_k-1}}{g_k}, & p_{2,k} &= 0, \\
    p_{3,k} &= 0, & p_{3,k} &= \frac{2^{R_k-1}}{g_k}, \\
    p_{4,k} &= \frac{2^{R_k-1}}{g_k}, & p_{4,k} &= (2^{R_k} - 1) \left( \frac{2^{R_k-1}}{g_k} + \frac{1}{g_k} \right) \\
\end{align*}
$$

(34)

and the indicator function $\mathbb{1}(\cdot)$ is defined the same as (22).

**Proof:** To minimize $\mathcal{L}_{2}^{\text{NOMA}}(p_k, p_k)$, we need to examine every case of combination regarding $\mathcal{U}_k$’s and $\mathcal{U}_k$’s outage occurrences so as to find the one that minimizes $\mathcal{L}_{2}^{\text{NOMA}}(p_k, p_k)$. 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 $\mathcal{U}_k$’s message at rate $R_k$,

$^3$Given $g_k > g_k$, we assume the decoding order $\mathcal{U}_k \rightarrow \mathcal{U}_k \rightarrow \mathcal{U}_k$, 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.
it inexplicitly suggests that $U_k$’s message is transmitted at $p_k > 0$. Therefore, the corresponding $\mathcal{L}_{2}^{\text{NOMA}}(p_k, p_{\bar{k}}) = \bar{R}_k + \bar{R}_{\bar{k}} + \lambda p_k + \delta + \mu$ is strictly larger than $\mathcal{L}_{2}^{\text{NOMA}}(0, p_k) = \bar{R}_k + \bar{R}_{\bar{k}} + \delta + \mu$ in Case I.A. Similarly, Case I.D can be shown to be strictly outperformed by Case I.B. With the remaining four cases, $p_{i,k} (p_{i,\bar{k}}), i = 1, 2, 3, 4$, is the minimum power required for $U_k (U_{\bar{k}})$ 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 $(\lambda, \delta, \mu)$.

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

With Proposition 4.1, given any multiplier-triple $(\lambda, \delta, \mu)$, the maximum of the Lagrangian in (32) is obtained by solving (P2-NOMA-sub) state by state in parallel. Finally, (P2-NOMA) is solved by updating $(\lambda, \delta, \mu)$ in accordance with the ellipsoid method.

It is also worthy of noting that $\zeta$ must not fall below $\zeta_{\text{min}}$ to ensure feasibility of Problem (P2-NOMA) (c.f. (31d)). To obtain the minimum possible $\zeta$, a variable $\zeta$ is set as the new objective of (P2-NOMA) to be minimized. This feasibility problem can be solved by bi-section over $\zeta$.

2) Optimal Solution to (P2-OMA-II): In this section, we aim for solving (P2-OMA-II). Despite of its non-convexity due to the same reason as that for (P2-NOMA), we can still find asymptotically 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

\[
(P2-\text{OMA-II-sub}): \quad \text{Minimize} \quad \mathcal{L}_{2}^{\text{OMA-II}}(p_k, p_{\bar{k}}, \alpha_k)
\]

Subject to

\[
p_k + p_{\bar{k}} \leq \hat{P},
\]

\[
0 \leq \alpha_k \leq 1, \quad \forall k,
\]

where the objective function is defined as $\mathcal{L}_{2}^{\text{OMA-II}}(p_k, p_{\bar{k}}, \alpha_k) = \bar{R}_k X_k^{\text{OMA-II}} + \bar{R}_{\bar{k}} X_{\bar{k}}^{\text{OMA-II}} + \lambda (p_k + p_{\bar{k}}) + \delta X_k^{\text{OMA-II}} + \mu X_{\bar{k}}^{\text{OMA-II}}$ with the fading index $\nu$ dropped for brevity.
Since each $U_k$ only needs to decode its own information without seeing interference in the orthogonal transmission, the possible combinations of outage occurrences for $U_k$ and $U_{\bar{k}}$ are easily shown in Table II, where $U_k \to U_k$ denotes $U_k$’s direct decoding of its own message, $k \in \{1, 2\}$. Based on Table II, 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

\[
(p^*_k, \alpha^*_k) = \arg \min_i \left\{ \sum_{i} L_{\text{OMA-II}}(p_{i,k}, p_{i,\bar{k}}, \alpha_{i,k}) | p_{i,k} + p_{i,\bar{k}} \right\}, \quad \forall i \in \{1, 2, 3, 4\},
\]

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

\[
\begin{align*}
P_{1,k} &= 0, & P_{1,\bar{k}} &= 0, & \alpha_{1,k} &= 0 \\
P_{2,k} &= \frac{2^\beta_{k-1}}{\gamma_k}, & P_{2,\bar{k}} &= 0, & \alpha_{2,k} &= 1 \\
P_{3,k} &= 0, & P_{3,\bar{k}} &= \frac{2^\beta_{\bar{k}-1}}{\gamma_k}, & \alpha_{3,k} &= 0 \\
P_{4,k} &= \frac{2^\beta_k (2^\alpha_k - 1)}{\gamma_k}, & P_{4,\bar{k}} &= \frac{2^\beta_{\bar{k}} (2^\alpha_{\bar{k}} - 1)}{\gamma_k}, & \alpha_{4,k} &= \alpha^*_k
\end{align*}
\]

In (36), $\alpha^*_k$, which denotes the optimum proportion of time/frequency resource allocated to $U_k$ to minimize the total transmit power at one particular fading state, is obtained by solving the following (convex) problem.

\[
(P2-\text{OMA-II-MP}): \quad \text{Minimize} \quad \alpha_k \left( \frac{2^\beta_k (2^\alpha_k - 1)}{\gamma_k} \right) + \alpha_{\bar{k}} \left( \frac{2^\beta_{\bar{k}} (2^\alpha_{\bar{k}} - 1)}{\gamma_{\bar{k}}} \right)
\]

Subject to $0 \leq \alpha_k \leq 1$.

In addition, the indicator function is given by (22).

**Proof:** Please refer to Appendix E.

As a result, given any triple $(\lambda, \delta, \mu)$, (P2-OMA-II-sub) is solved. (P2-OMA-II) is then solved by updating $(\lambda, \delta, \mu)$ 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 first introduce the following lemma.

**Lemma 4.1:** Given $0 < g_1 \geq g_2 \geq \ldots \geq g_K$, denoting \( \min_{\sum_{i=1}^{K} \alpha_i = 1} \sum_{i=1}^{K} \frac{R_i}{g_i} \) by \( P_{02}^* \), and 
\[ \sum_{i=0}^{K-1} \frac{(2^{\alpha_i} - 1)\bar{R}_{K-j}}{g_{K-1}} \] by \( P_{N}^* \), where \( \alpha_i \geq 0 \) and \( \bar{R}_i \geq 0, \forall i \), then we have \( P_{02}^* \geq P_{N}^* \).

**Proof:** Please refer to Appendix F.

Second, 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. Since there are four possible types of solution (c.f. (36)) for \((P2-OMA-II-sub)\), we need to modify them in each case. It is straightforward to check that for the first three cases in (36), setting \( p_1 = p_1^* \) and \( p_2 = p_2^* \) corresponds to the first three cases in (34), and thus \( X_k^{\text{NOMA}} = X_k^{\text{OMA-II}}, \forall k \). In the sequel, we review the case that \( \{p_1^*, p_2^*\} \) falls in the last case of (36). Given \( p_1^* = \frac{\alpha_1(2^{\alpha_1} - 1)}{g_1} \) and \( p_2^* = \frac{\alpha_2(2^{\alpha_2} - 1)}{g_2} \), we reallocate them among the two users as \( p_1 = \frac{2^{\bar{R}_2}}{g_1} - 1 \) and \( p_2 = (2^{\bar{R}_2} - 1)(\frac{2^{\bar{R}_1} - 1}{g_1} + \frac{1}{g_2}) \) assuming \( g_1 \geq g_2 \) w.l.o.g., and therefore \( X_k^{\text{NOMA}} = X_k^{\text{OMA-II}} = 1, \forall k \). This modification is feasible, since \( p_1^* + p_2^* \geq p_1 + p_2 \) in accordance with Lemma 4.1. (Considering a special case of \( K = 2 \), it holds true that \( \min_{\sum_{i=1}^{2} \alpha_i = 1} \sum_{i=1}^{2} \frac{2^{\bar{R}_j} - 1}{g_i} \geq \frac{2^{\bar{R}_2} - 1}{g_2} + \frac{2^{\bar{R}_1} - 1}{g_1} \), the left-hand side (LHS) and the RHS of which corresponds to \( p_1^* + p_2^* \) for OMA-Type-II and \( p_1 + p_2 \) for NOMA, respectively.) To sum up, with the constructed solution \( \{p_1, p_2\} \) in each fading state, it follows that \( E_\nu[X_k^{\text{NOMA}}(\nu)] = E_\nu[X_k^{\text{OMA-II}}(\nu)] \leq \bar{\zeta}, \forall k \). Moreover, benefiting from the saved power by NOMA, (10a) now becomes inactive, which suggests that \( E_\nu[X_k^{\text{NOMA}}(\nu)] \)'s are likely 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 accordingly recast as follows:

\[(P2'-XX) : \text{Maximize} \quad \bar{R}_k(1 - \zeta_k^{XX}) + \bar{R}_k(1 - \zeta_k^{XX}) \]

Subject to

\[ p_s + p_w \leq \bar{P}, \quad (37a) \]
\[ p_s \geq 0, \quad p_w \geq 0, \quad (37b) \]
\[ E_\nu[X_k^{XX}(\nu)] \leq \bar{\zeta}', \forall k, \quad (37c) \]
where $\alpha_k$’s is only valid when “XX” is replaced by OMA-Type-II, and (37c) constrain the maximum user outage probability of the two below $\bar{\zeta}$.

1) **Optimal Solution to (P2'-NOMA):** In this section, we derive $\mathbb{E}_\nu[X_{k}^{\text{NOMA}}]$’s as well as the optimal solution to (P2'-NOMA).

In line with the same notation for channel gains as defined in (24) and (25), replace $p_k(\nu)$ with $p_s$, $p_k(\nu)$ with $p_w$, $\forall \nu$, when $X > Y$, and otherwise do this reversely. As a result, $\mathbb{E}_\nu[X_{k}^{\text{NOMA}}(\nu)] = \epsilon_k^{\text{NOMA}}$ can be recast as follows:

$$
\mathbb{E}_\nu[X_{k}^{\text{NOMA}}(\nu)] = \Pr \left\{ \log_2 \left( 1 + \frac{p_s X}{p_w X + \sigma_k^2} \right) < \bar{R}_k, \log_2 \left( 1 + \frac{p_w X}{p_s X + \sigma_k^2} \right) < \bar{R}_k, X > Y \right\} 
+ \Pr \left\{ \log_2 \left( 1 + \frac{p_s X}{\sigma_k^2} \right) < \bar{R}_k, \log_2 \left( 1 + \frac{p_w X}{p_s X + \sigma_k^2} \right) \geq \bar{R}_k, X > Y \right\} 
+ \Pr \left\{ \log_2 \left( 1 + \frac{p_w X}{p_s X + \sigma_k^2} \right) < \bar{R}_k, X \leq Y \right\}. 
$$

(38)

With CDI regarding $X$ and $Y$ given in Section II, $\mathbb{E}_\nu[X_{k}^{\text{NOMA}}(\nu)]$’s can be derived based upon (38) shown in the following proposition.

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

$$
\mathbb{E}_\nu[X_{k}^{\text{NOMA}}(\nu)] =
\begin{cases}
1 - e^{-\lambda_k \epsilon_{k,1}} + \frac{\lambda_k}{\lambda_k + \lambda_k} e^{-\lambda_k + \lambda_k} \epsilon_{k,2} - \frac{\lambda_k}{\lambda_k + \lambda_k} e^{-\lambda_k + \lambda_k} \epsilon_{k,3}, & \alpha_{k,c_1} \leq \frac{p_s}{p_w} \leq \beta_{k,c_1}, \ j = 1, 2, 4 \\
1 - e^{-\lambda_k \epsilon_{k,1}} + \frac{\lambda_k}{\lambda_k + \lambda_k} e^{-\lambda_k + \lambda_k} \epsilon_{k,2}, & \alpha_{k,c_2} \leq \frac{p_s}{p_w} \leq \beta_{k,c_2}, \ j = 1, 2, 4 \\
1 - \frac{\lambda_k}{\lambda_k + \lambda_k} e^{-\lambda_k + \lambda_k} \epsilon_{k,1}, & \alpha_{k,c_3} \leq \frac{p_s}{p_w} \leq \beta_{k,c_3} \\
1, & \alpha_{k,c_4} \leq \frac{p_s}{p_w} \leq \beta_{k,c_4} \\
\end{cases}
$$

(39)

where $\epsilon_{k,1} \triangleq \frac{\sigma_k^2}{p_s}, \epsilon_{k,2} \triangleq \frac{\sigma_k^2}{p_s - p_w - p_s \tau_k}, \epsilon_{k,3} \triangleq \frac{\sigma_k^2}{p_w - p_s \tau_k},$ and $\epsilon_{k,4} \triangleq \frac{\sigma_k^2}{p_w - p_s \tau_k},$ in which $\tau_k \triangleq \frac{2\bar{R}_k}{\bar{R}_k} - 1.$

In addition, $\alpha_{k,c_1}$ and $\beta_{k,c_1}, j = 1, 2, 4,$ are given by

$$
\begin{align*}
\alpha_{k,c_1} &= 0, \quad \beta_{k,c_1} = \min \left\{ \frac{\tau_k}{\tau_k + 1}, 1 \right\} \\
\alpha_{k,c_2} &= \max \left\{ \frac{\tau_k}{\tau_k + 1}, \tau_k \right\} \text{ or } \max \left\{ \tau_k, 1 \right\}, \quad \beta_{k,c_2} = \min \left\{ \frac{1}{\tau_k}, 1 \right\} \text{ or } \frac{1}{\tau_k} \right\} \\
\alpha_{k,c_3} &= \max \left\{ \frac{\tau_k}{\tau_k + 1}, \tau_k \right\} \text{ or } \frac{\tau_k}{\tau_k + 1}, \quad \beta_{k,c_3} = \min \left\{ \frac{\tau_k}{\tau_k + 1}, \frac{1}{\tau_k} \right\} \text{ or } \min \left\{ \frac{\tau_k}{\tau_k + 1}, \frac{1}{\tau_k} \right\}
\end{align*}
$$
\( \alpha_{k,c^j} \) and \( \beta_{k,c^j} \), \( j = 1, 2, 4 \), are given by

\[
\begin{align*}
\alpha_{k,c^2} &= \max \left\{ \frac{\tau_k}{r_k}, \frac{1}{r_k} \right\}, \\
\alpha_{k,c^4} &= \max \left\{ \frac{\tau_k(r_k+1)}{r_k}, \tau_k, \frac{1}{r_k} \right\} \text{ or } \max \left\{ \frac{\tau_k}{r_k}, \frac{1}{r_k} \right\}, \\
\alpha_{k,c^5} &= \max \left\{ \frac{\tau_k}{r_k}, \frac{1}{r_k} \right\},
\end{align*}
\]

\( \beta_{k,c^2} = \frac{1}{r_k} \) or \( +\infty \);

\( \beta_{k,c^4} = \min \left\{ \frac{\tau_k(r_k+1)}{r_k}, \frac{1}{r_k} \right\} \) or \( \min \left\{ \tau_k, \frac{1}{r_k} \right\} \);

and \( \alpha_{k,c^l} \) and \( \beta_{k,c^l} \), \( l = 3, 4, 5 \), are given by

\[
\begin{align*}
\alpha_{k,c^3} &= \frac{1}{r_k}, \\
\beta_{k,c^3} &= \min \left\{ \frac{\tau_k}{r_k} \right\}, \\
\alpha_{k,c^4} &= \frac{1}{r_k}, \\
\beta_{k,c^4} &= \min \left\{ \frac{\tau_k}{r_k(r_k+1)}, \tau_k \right\}, \\
\alpha_{k,c^5} &= \frac{1}{r_k}, \\
\beta_{k,c^5} &= \tau_k
\end{align*}
\]

**Proof:** Please refer to Appendix G.

In accordance with Proposition 4.3, we are able to derive the sum of the DL \( \bar{R}_k(1-c_k^{\text{NOMA}}) + \bar{R}_k(1-\zeta_k^{\text{NOMA}}) \). Problem (P2'-NOMA) can also be characterized by only one optimization variable \( p_s \) by replacing \( p_w \) with \( \bar{P} - p_s \), which can be 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 this section, we characterize \( \mathbb{E}_\nu[X_k^{\text{OMA-II}}] \), \( k \in \{1, 2\} \), before optimally solve Problem (P2'-OMA-II).

In line with the principle of power and time/frequency allocations for OMA-Type-II transmission described above (5), replace \( p_k(\nu) \) with \( p_s, p_k(\nu) \) with \( p_w, \forall \nu \), when \( X > Y \), and the reverse when \( X \leq Y \) in (8). The partial CSIT counterpart of (8), \( c_k^{\text{OMA-II}} = \mathbb{E}_\nu[X_k^{\text{OMA-II}}(\nu)] \), is derived as follows:

\[
\mathbb{E}_\nu[X_k^{\text{OMA-II}}(\nu)] = \mathbb{P} \left\{ \alpha_k \log_2 \left( 1 + \frac{p_s X Y}{\alpha_k \sigma_n^2} \right) < \bar{R}_k, X > Y \right\} + \\
\mathbb{P} \left\{ \alpha_k \log_2 \left( 1 + \frac{p_w X Y}{\alpha_k \sigma_n^2} \right) < \bar{R}_k, X \leq Y \right\}.
\]

(40)

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

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

\[
\mathbb{E}_\nu[X_k^{\text{OMA-II}}(\nu)] = 1 + \frac{\lambda_k}{\lambda_k + \lambda_k} \left( e^{(\lambda_k + \lambda_k)\varphi_{k,1}} - e^{(\lambda_k + \lambda_k)\varphi_{k,2}} - e^{-\lambda_k \varphi_{k,1}} \right),
\]

(41)

where \( \varphi_{k,1} \triangleq \frac{\alpha_k \sigma_n^2 \xi_k}{\rho_s} \) and \( \varphi_{k,2} \triangleq \frac{\alpha_k \sigma_n^2 \xi_k}{\rho_w} \), in which \( \xi_k \triangleq \frac{\bar{R}_k}{\alpha_k} - 1 \).

\(^4\)The parameter values preceding and coming after “or” form a pair, respectively.
**Proof:** With CDI of $X$ and $Y$ known, the derivation of $E_\nu[X_k^{OMA-II}(\nu)]$ from (40) 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 thus omitted herein.

V. Numerical Results

In this section, we verify the theoretical analysis for the considered two-user downlink NOMA system via numerical results. As a benchmark, we also provide a suboptimal OMA transmission scheme, referred as *OMA-Type-I*, which assigns equal orthogonal resource (time/frequency) to the users over all fading states, i.e., $\alpha_k(\nu) = \frac{1}{2}, \forall k, \forall \nu$ in (3), and $\alpha_k = \frac{1}{2}, \forall k$, in (5). The corresponding optimal power policies to OMA-Type-I are easily seen to be special cases of $(P1-OMA-II)$ ($(P2-OMA-II)$) and $(P1'-OMA-II)$ ($(P2'-OMA-II)$), which can thus be accordingly obtained. $U_1$ and $U_2$ 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 $-169$dBm/Hz over $10$MHz 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. 1 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 worthy of noting that the horizontal axe in Fig. 1 is given by $[0, \bar{R}_{max}]$, and it has been discussed in Section III-A1 regarding how to obtain $\bar{R}_{max}$. 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 $\bar{R}$ 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 $\bar{R}$ until $\bar{R}$ is larger than $0.65$bits/sec/Hz. This is because OMA-Type-I is intrinsically of fairness in terms 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., $E_{\nu}[R_{k}^{\text{NOMA}}(\nu)] \approx E_{\nu}[R_{k}^{\text{NOMA}}(\nu)]$ ($E_{\nu}[R_{k}^{\text{OMA-II}}(\nu)] \approx E_{\nu}[R_{k}^{\text{OMA-II}}(\nu)]$) due to their statistically similar channel distribution. Furthermore, the average sum-rate achieved by NOMA is seen to be achieved by OMA-Type-II with negligible gap, since in this case, the achievable average sum-rate for both schemes remains approximately the same as that w/o minimum average constraint, the latter of which achieved by OMA-Type-II is in theory the same as that achieved by NOMA.

Fig. 1. The average sum-rate of the system versus the minimum rate constraints under full CSIT.

Fig. 2 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. 1.

![Bar chart](image1)

(a) Full CSIT.

![Bar chart](image2)

(b) Partial CSIT.

Fig. 3. The average achievable rate allocation between the two users with different minimum rate requirements, under full and partial CSIT.

The comparison between the individual ergodic rate subject to varied minimum average rate constraints is demonstrated in Fig. 3(a) (Fig. 3(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 $U_1$ as NOMA with $U_2$’s ergodic rate both as little as zero, when there is no minimum rate requirement. This can be intuitively explained as follows. Since $U_1$ 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 $U_1$ in all 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 $U_2$ while keeping $U_1$’s average rate the maximum.

B. Delay-Limited Transmission

Fig. 4(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 $\bar{R}_e = 2$bits/sec/Hz
Fig. 4. The DLT of the system versus the maximum permissive outage probability under full CSIT.

for each user. It is seen that when the two users suffer from near-far unfairness, the optimum sum of DLT versus max-outage trade-off achieved by NOMA outperforms that achieved by OMA-Type-II and OMA-Type-I, respectively. However, this superiority almost disappears when \( \mathcal{U}_1 \) and \( \mathcal{U}_2 \) are both 0.5km away from the BS. This is because with the same large-scale fading, \( g_1(\nu) \approx g_2(\nu) \) in most of fading states, thanks to which the difference between \( P_{O2}^* \) and \( P_N^* \) in Lemma 4.1 given \( K = 2 \) becomes less. This implies that the total amount of transmit power saved by NOMA tends to be trivial. Furthermore, in this case, 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 \( \zeta_{XX}^k \) (c.f. (6) and (8)) is minimized, \( \bar{\zeta}_{XX}^k \) is nearly minimized as well, where \( (\cdot)^{XX} \) stands for NOMA or OMA-Type-II.

On the other hand, the impact of different \( \bar{R}_k \)'s on the optimum sum-DLT versus max-outage trade-offs in near-far channel conditions is demonstrated in Fig. 4(b). With the same intended rate \( \bar{R}_1 = \bar{R}_2 = 2 \)bits/sec/Hz, the optimum trade-off achieved by NOMA outperforms those 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. When \( \bar{R}_1 \) continues increasing to 4bits/sec/Hz, the trade-off is seen come back with superior sum of DLT over that in other settings of \( \bar{R}_1 \) and \( \bar{R}_2 \).

Fig. 5(a) illustrates the optimum sum-DLT versus max-outage trade-offs achieved by different schemes with varied \( \bar{P} \) under partial CSIT given \( \bar{R}_k = 2 \)bits/sec/Hz, \( k \in \{1, 2\} \). It is seen that the optimum trade-offs between the sum of DLT and the maximum permissive outage
probability achieved by NOMA significantly outperform those achieved by the OMA schemes with different $\bar{P}$. Moreover, the minimum max-outage achieved by NOMA is significantly lower than that attained by other schemes. For example, with $\bar{P} = 2$Watt, $\min_{p_s,p_w} \max_{k} \xi_{k}^{\text{NOMA}}$ falls below 0.60 while that achieved by OMA-Type-I and OMA-Type-II is as large as about 0.71 and 0.90, respectively. Fig. 5(b) 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. 4(b), the superiority of NOMA over the other OMA schemes is significantly seen in Fig. 5(b).

![Fig. 5. The DLT of the system versus the maximum permissive user outage probability under partial CSIT.](image)

![Fig. 6. The DLT allocation between the two users with different maximum outage requirements, under full and partial CSIT.](image)

When $\bar{R}_k$, $k \in \{1, 2\}$, are set equal to each other, the individually achievable DLT under full
(partial) CSIT, namely, $\bar{R}_k(1 - \zeta_k^{XX})(\bar{R}_k(1 - \zeta_k^{XX}))$, can be equivalently characterized by their individual outage probability $\zeta_k^{XX}(\zeta_k^{XX})$. The DLT allocation between the two users subject to different maximum permissive outage is reflected by their outage probability allocation in Fig. 6 under full (c.f. Fig. 6(a)) and partial (c.f. Fig. 6(b)) CSIT, respectively. Under full CSIT, it is seen from Fig. 6(a) that with $\bar{R}_1 = \bar{R}_2 = 2$bits/sec/Hz and $\bar{P} = 2$Watt, $U_1$ achieves almost negligible outage while $U_2$ can achieve an outage probability as low as 0.3032 by NOMA. OMA-Type-II follows the same trend unless $U_1$ has to claim more outage states to reserve power for $U_2$’s transmission. OMA-Type-I is the least favourable in the sense that it does not support any outage requirement lower than 0.36 (0.71). $U_1$’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 due to loss of optimality (only $p_s$, $p_w$, and/or $\alpha_k$ to be optimized).

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.

Appendix A

As $p_1 + p_2 = \hat{P}$ is fixed, (P1-OMA-II-sub) reduces to the following problem irrespective of $\lambda$:

Maximize $\sum_{p_1 \geq 0, p_2 \geq 0, \alpha_1} (1 + \delta)\alpha_1 \log_2 \left( 1 + \frac{p_1 g_1}{\alpha_1} \right) + (1 + \mu)\alpha_2 \log_2 \left( 1 + \frac{p_2 g_2}{\alpha_2} \right)$

Subject to $p_1 + p_2 \leq \hat{P}$, 

$0 \leq \alpha_1 \leq 1$, (42a)

where $\alpha_2 = 1 - \alpha_1$. Note that $=$ has been relaxed into $\leq$ in (42a), since it is easy to check that the above problem obtains its optimum value when (42a) is active.
To facilitate solving (42), we introduce a new Lagrangian multiplier $\lambda'$ associated with the constraint (42a). Then given $\lambda'$, we aim for maximizing the Lagrangian regarding Problem (42) as follows:

\[
\begin{align*}
\text{Maximize} & \quad \bar{L}_{OMA-II}(p_1, p_2, \alpha_1, \lambda') \\
\text{Subject to} & \quad 0 \leq \alpha_1 \leq 1,
\end{align*}
\]

in which $\bar{L}_{OMA-II}(p_1, p_2, \alpha_1, \lambda') = (1+\delta)\alpha_1 \log_2(1+\frac{p_1 g_1}{\alpha_1}) + (1+\mu)(1-\alpha_1) \log_2(1+\frac{p_2 g_2}{1-\alpha_1}) - \lambda'(p_1 + p_2)$. It is worthy of noting that Problem (43) shares the same structure as (P1-OMA-II-sub) except that the constraint $p_1 + p_2 \leq \hat{P}$ is now removed. Therefore, the rationale behind solving it has been similarly given in the proof for Proposition 3.2. As a result, given $\lambda'$, the maximizer of $\bar{L}_{OMA-II}(p_1, p_2, \alpha_1, \lambda')$ proves to be either $(0, (\frac{1+\mu}{\lambda'\ln 2} - \frac{1}{g_2})^+, 0)$ or $((\frac{1+\delta}{\lambda'\ln 2} - \frac{1}{g_1})^+, 0, 1)$, i.e., alternative transmission. Therefore, by updating $\lambda'$ via bi-section until (42a) is active, the optimal solution to Problem (42) ends up with $(0, \hat{P}, 0)$ or $(\hat{P}, 0, 1)$ depending on which leads to a larger objective value. Hence, we complete the proof for Lemma 3.2.

**APPENDIX B**

As the maximum of $\bar{L}_{OMA-II}(p_1, p_2, \alpha_1)$ must be either at the stationary point or on the boundary of $\Psi$ as defined in Proposition 3.1, the vertexes of $(0, 0, 0), (0, \hat{P}, 0), \text{and } (\hat{P}, 0, 1)$ are included in (21) for sure.

As for another case that the maximum of $\bar{L}_{OMA-II}(p_1, p_2, \alpha_1)$ is achieved on $p_1 = 0$, the corresponding optimum value of $\bar{L}_{OMA-II}(p_1, p_2, \alpha_1)$ takes on $\bar{L}_{OMA-II}(0, c_2, 0)\mathbb{1}_{c_2}$ because of the following reasons. Plugging $p_1 = 0$ into $\bar{L}_{OMA-II}(p_1, p_2, \alpha_1)$, its derivative w.r.t $p_2$ and $\alpha_1$ are, respectively, expressed as:

\[
\begin{align*}
\frac{\partial \bar{L}_{OMA-II}(p_1, p_2, \alpha_1)}{p_2} &= \left(1 + \mu\right)\frac{\alpha_2 g_2}{(\alpha_2 + p_2 g_2)\ln 2} - \lambda, \\
\frac{\partial \bar{L}_{OMA-II}(p_1, p_2, \alpha_1)}{\alpha_1} &= \frac{\left(1 + \mu\right)}{\ln 2} \left[ -\ln \left(1 + \frac{p_2 g_2}{1 - \alpha_1}\right) + 1 - \frac{1}{1 + p_2 g_2/(1 - \alpha_1)} \right].
\end{align*}
\]

(44)

(45)

It is then easily seen that the optimal $p_2$ admits the form of $p_2^* = [c_2(1 - \alpha_1)]^\frac{\hat{P}}{\alpha_2}$. Further, $\bar{L}_{OMA-II}(0, p_2, \alpha_1)$ turns out to monotonically decrease w.r.t $\alpha_1$ by observing that the right-hand side (RHS) of (45) is always negative in view of the inequality $1 - \frac{1}{x} \leq \ln x$. Hence, $\alpha_1^\ast = 0$, and $(0, c_2, 0)$ is the optimum if and only if (iff) $0 \leq c_2 \leq \hat{P}$, which leads to $\bar{L}_{OMA-II}(0, c_2, 0)\mathbb{1}_{c_2}$. Similarly, $\bar{L}_{OMA-II}(c_1, 0, 1)\mathbb{1}_{c_1}$ can be justified considering another boundary of $p_2 = 0$. 
At last, the reasons why the jointly stationary point (c.f. (19)) cannot be the optimal solution to (P1-OMA-II-sub) has been explained in Remark 3.1.

**APPENDIX C**

\[ \mathbb{E}_\nu[R_{k}^{\text{NOMA}}(\nu)] = \Pr\{X \geq Y\} \mathbb{E}_\nu[\log_2 (1 + \Gamma_k) | X \geq Y] + \Pr\{X < Y\} \mathbb{E}_\nu \left[ \log_2 \left( 1 + \tilde{\Gamma}_k \right) \right] X < Y \]

\[ = \frac{1}{\ln 2} \Pr\{X \geq Y\} \int_0^\infty \frac{1 - F_{\Gamma_k | X \geq Y}(z)}{1 + z} dz + \frac{1}{\ln 2} \Pr\{X < Y\} \int_0^\infty \frac{1 - F_{\Gamma_k | X < Y}(z)}{1 + z} dz \]

\[ \equiv \frac{1}{\ln 2} \frac{\lambda_k}{\lambda_k + \lambda_k} A_1 + \frac{1}{\ln 2} B_1 - \frac{1}{\ln 2} \frac{\lambda_k}{\lambda_k + \lambda_k} C_1, \quad (46) \]

where \( A_1 \equiv \int_0^\infty \frac{\alpha_k \log_2 (1 + \Gamma_k)}{1 + z} e^{-\lambda_k \frac{1 + z}{\lambda_k + \lambda_k}} dz \)

\[ \text{and} \quad B_1 \equiv \int_0^\infty \frac{\alpha_k \log_2 (1 + \Gamma_k)}{1 + z} e^{-\lambda_k \frac{1 + z}{\lambda_k + \lambda_k}} dz, \text{ and } C_1 \equiv \int_0^\infty \frac{\alpha_k \log_2 (1 + \Gamma_k)}{1 + z} e^{-\lambda_k \frac{1 + z}{\lambda_k + \lambda_k}} dz. \]

Since \( \Pr\{X \geq Y\} = \frac{\lambda_k}{\lambda_k + \lambda_k} \), substituting (24) and (25) for \( F_{\Gamma_k | X \geq Y}(z) \) and \( F_{\Gamma_k | X < Y}(z) \) in (46), respectively, (a) is derived. Then after some manipulations, by applying [24, Eq. (3.352.4)], we have \( A_1 = f\left(\frac{(\lambda_k + \lambda_k) \sigma^2}{p_w + p_w} \right) - f\left(\frac{\lambda_k \sigma^2}{p_w + p_w} \right) \). It also immediately follows that \( B_1 = -f\left(\frac{\lambda_k \sigma^2}{p_w} \right) \) and \( C_1 = -f\left(\frac{(\lambda_k + \lambda_k) \sigma^2}{p_w} \right) \). Plugging \( A_1, B_1 \) and \( C_1 \) into (46), Proposition 3.3 is proved.

**APPENDIX D**

\[ \mathbb{E}_\nu[R_{k}^{\text{OMA-II}}(\nu)] = \Pr\{X \geq Y\} \mathbb{E}_\nu[\alpha_k \log_2 (1 + \Gamma_k) | X \geq Y] + \Pr\{X < Y\} \mathbb{E}_\nu \left[ \alpha_k \log_2 \left( 1 + \tilde{\Gamma}_k \right) \right] X < Y \]

\[ = \frac{\alpha_k}{\ln 2} \frac{\lambda_k}{\lambda_k + \lambda_k} \int_0^\infty \frac{1 - F_{\Gamma_k | X \geq Y}(z)}{1 + z} dz + \frac{\alpha_k}{\ln 2} \frac{\lambda_k}{\lambda_k + \lambda_k} \int_0^\infty \frac{1 - F_{\Gamma_k | X < Y}(z)}{1 + z} dz \]

\[ \equiv \frac{\alpha_k}{\ln 2} \frac{\lambda_k}{\lambda_k + \lambda_k} A'_1 + \frac{\alpha_k}{\ln 2} B'_1 + \frac{\alpha_k}{\ln 2} \frac{\lambda_k}{\lambda_k + \lambda_k} C'_1, \quad (47) \]

where \( A'_1 \equiv \int_0^\infty \frac{e^{-\lambda_k \frac{1 + z}{\lambda_k + \lambda_k}}}{1 + z} dz \), \( B'_1 \equiv \int_0^\infty \frac{\alpha_k \log_2 (1 + \Gamma_k)}{1 + z} e^{-\lambda_k \frac{1 + z}{\lambda_k + \lambda_k}} dz \)

\[ \text{and} \quad C'_1 \equiv \int_0^\infty \frac{\alpha_k \log_2 (1 + \Gamma_k)}{1 + z} e^{-\lambda_k \frac{1 + z}{\lambda_k + \lambda_k}} dz, \]

which are obtained by substituting (28) and (29) for \( F_{\Gamma_k | X \geq Y}(z) \) and \( F_{\Gamma_k | X < Y}(z) \), respectively. Then by directly applying [24, Eq. (3.352.4)], \( A'_1 = -f\left(\frac{(\lambda_k + \lambda_k) \alpha_k \sigma^2}{p_w + p_w} \right) \), \( B'_1 = -f\left(\frac{\lambda_k \alpha_k \sigma^2}{p_w + p_w} \right) \) and \( C'_1 = -f\left(\frac{(\lambda_k + \lambda_k) \alpha_k \sigma^2}{p_w + p_w} \right) \) are derived, which completes the proof for Proposition 3.4.


**APPENDIX E**

Following similar analysis as for Proposition 4.1, the minimum of $L_{2}^{OMA-II}(p_k, p_k, \alpha_k)$ is obtained by comparing all possible combinations of outage occurrences for $U_k$ and $U_k$ (c.f. Table II). In (36), $p_{i,k}$ and $p_{i,k}$, $i = 1, 2, 3, 4$, are respectively the minimum power required to have both of the users suspend their transmission, only $U_k$ or $U_k$ supported, and both of the users simultaneously served.

**APPENDIX F**

Note from Lemma 4.1 that to prove $P_{O2}^* \geq P_N^*$, it is sufficient to show that $\sum_{i=1}^{K} \frac{(2^{\alpha_i} - 1)\alpha_i}{g_i} \geq P_N^*$ holds for any $\alpha_i$'s such that $\sum_{i=1}^{K} \alpha_i = 1$. By variable transformation of $i \leftarrow (K - i)$, it follows that $P_N^* = \sum_{i=1}^{K} \frac{(2^{R_i - 1})2^{\sum_{l=1}^{i+1} R_l}}{g_i}$.

First, denoting $(2^{R_i - 1})\alpha_i$ by $a_i$, and $(2^{R_i - 1})2^{\sum_{l=i+1}^{K} R_l}$ by $b_i$, $\forall i$, we prove that $\sum_{i=j}^{K} a_i \geq \sum_{i=j}^{K} b_i$ holds for $\forall j = 1, \ldots, K$. Expand $\sum_{i=j}^{K} b_i$ as follows:

\[
\begin{align*}
2^{\sum_{i=j}^{K} R_i} - 2^{\sum_{i=j+1}^{K} R_i}, & \quad i = j \\
2^{\sum_{i=j+1}^{K} R_i} - 2^{\sum_{i=j+2}^{K} R_i}, & \quad i = j + 1 \\
\vdots & \quad i = j + 2, \ldots, K - 1 \\
2^{R_K - 1}. & \quad i = K
\end{align*}
\]

(48)

By summing-up the LHS of (48), $\sum_{i=j}^{K} b_i$ is simplified as $2^{\sum_{i=j}^{K} R_i} - 1$. Defining a function $f_0(x) = 2^{x} - 1$, let $x_i = 0$ for $i = 1, \ldots, j - 1$, and $x_i = \frac{R_i}{\alpha_i}$ for $i = j, \ldots, K$. Then by applying Jensen’s inequality due to the convexity of $f_0(x)$, it follows that

\[
\sum_{i=j}^{K} a_i = \sum_{i=1}^{K} \alpha_i f_0(x_i) \geq f_0 \left( \sum_{i=1}^{K} \alpha_i x_i \right) = f_0 \left( \sum_{i=j}^{K} \alpha_i x_i \right) = \sum_{i=j}^{K} b_i.
\]

(49)

Next, define $\frac{1}{g_i}$ by $c_i$, $\forall i$, it is easily verified that $0 < c_1 \leq c_2 \leq \ldots \leq c_K$. By applying [14, Lemma 4], we conclude that $\sum_{i=1}^{K} \frac{(2^{R_i - 1})\alpha_i}{g_i} \geq \sum_{i=1}^{K} \frac{(2^{R_i - 1})2^{\sum_{l=i+1}^{K} R_l}}{g_i} = P_N^*$, which completes the proof for Lemma 4.1.
APPENDIX G

The piece-wise presentation of \( \mathbb{E}_\nu [X_k^{\text{NOMA}}(\nu)] \) is caused by the range of the parameters. To illustrate as an example, take the second term of (38) as an example and express it as follows:

\[
\Pr \left\{ \log_2 \left( 1 + \frac{p_s X}{\sigma_k^2} \right) < \bar{R}_k, \log_2 \left( 1 + \frac{p_w X}{p_s X + \sigma_k^2} \right) \geq \bar{R}_k, X > Y \right\} 
\]

\[
= \Pr \left\{ X < \varepsilon_{k,1}, (p_w - p_s \tau_k) X \geq \sigma_k^2 \tau_k, X > Y \right\} 
\]

\[
= \begin{cases} 
\Pr \left\{ X < \varepsilon_{k,1}, X \geq \varepsilon_{k,4}, X > Y \right\}, & \frac{p_w}{p_s} < \frac{1}{\tau_k} \\
0, & \text{otherwise} 
\end{cases} 
\]

(50)

Further, the first case in (50) implies the following two sub-cases:

\[
\Pr \left\{ X < \varepsilon_{k,1}, X \geq \varepsilon_{k,4}, X > Y \right\} 
\]

\[
= \begin{cases} 
\Pr \left\{ \varepsilon_{k,4} \leq X < \varepsilon_{k,1}, X > Y \right\}, & \frac{p_w}{p_s} < \min \left\{ \frac{\tau_k}{\tau_k (\tau_k + 1)}, \frac{1}{\tau_k} \right\} \\
0, & \text{otherwise} 
\end{cases} 
\]

(51)

Hence, in the case of \( \frac{p_w}{p_s} < \min \{ \tau_k / \tau_k (\tau_k + 1), 1 / \tau_k \} \), the second term of (38) turns out to be \( \Pr \{ \varepsilon_{k,4} \leq X < \varepsilon_{k,1}, X > Y \} \), which is equal to \( e^{-\lambda_k \varepsilon_{k,4}} - e^{-\lambda_k \varepsilon_{k,1}} + \lambda_k (e^{-\lambda_k + \lambda_k})^{e_{k,1}} - e^{-(\lambda_k + \lambda_k) e_{k,4}} \) after some manipulations, and otherwise zero. By analogy, the first and the third term of (38) can also be analysed piece-wisely, and finally we arrive at (39) combining all possible cases.

REFERENCES

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, “Non-orthogonal multiple access (noma) for cellular future radio access,” in Proc. IEEE Vehicular Technology Conference (VTC Spring), Dresden, Germany, June 2013, pp. 1–5.

[2] 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.

[3] L. Dai, B. Wang, Y. Yuan, S. Han, C. I. 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.

[4] 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.

[5] 3rd Generation Partnership Project (3GPP), “Study on downlink multiuser superposition transmission for LTE,” Mar. 2015.

[6] 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.

[7] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Z. Chen, Z. Ding, X. Dai, and R. Zhang, “An optimization perspective of the superiority of NOMA compared to conventional OMA,” to appear in *IEEE Trans. Signal Process.*, 2017.

[15] 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.

[16] 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.

[17] 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.

[18] ——, “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.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products*, 6th ed. New York, NY, USA: Academic Press, 2000.

[25] 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.