Link-Layer Capacity of NOMA Under Statistical Delay QoS Guarantees

Wenjuan Yu, Leila Musavian and Qiang Ni

***Abstract*—In this paper, we study the achievable link-layer rate, namely, effective capacity (EC), under the per-user statistical delay quality-of-service (QoS) requirements, for a downlink non-orthogonal multiple access (NOMA) network with M users. Specifically, the M users are assumed to be divided into multiple NOMA pairs. Conventional orthogonal multiple access (OMA) then is applied for inter-NOMA-pairs multiple access. Focusing on the total link-layer rate for a downlink M-user network, we prove that OMA outperforms NOMA when the transmit signal-to-noise ratio (SNR) is small. On the contrary, simulation results show that NOMA prevails over OMA at high values of SNR. Aware of the importance of a two-user NOMA network, we also theoretically investigate the impact of the transmit SNR and the delay QoS requirement on the individual EC performance and the total link-layer rate for a two-user network. Specifically, for delay-constrained and delay-unconstrained users, we prove that for the user with the stronger channel condition in a two-user network, NOMA prevails over OMA when the transmit SNR is large. On the other hand, for the user with the weaker channel condition in a two-user network, it is proved that NOMA outperforms OMA when the transmit SNR is small. Furthermore, for the user with the weaker channel condition, the individual EC in NOMA is limited to a maximum value, even if the transmit SNR goes to infinity. To confirm these insightful conclusions, the closed-form expressions for the individual EC in a two-user network, by applying NOMA or OMA, are derived for both users and then confirmed using Monte Carlo simulations.**

***Index Terms*—NOMA, Quality-of-service, delay-outage probability constraint, effective capacity, closed-form expressions.**

I. Introduction

Due to the explosive growth of mobile data and the Internet of Things (IoT) applications which exponentially accelerate the demand for high data rates, 5G has been anticipated to offer much higher data rate, less end-to-end latency and a significant reduction in network energy usage [1]. When it comes to the proposed multiple access (MA) techniques for 5G, non-orthogonal multiple access (NOMA) has been attracting a lot of attention as a promising scheme, due to the fact that it can offer improved spectral efficiency [2], higher cell-edge throughput [3] and low transmission latency [4], over conventional orthogonal multiple access (OMA) techniques. Current available NOMA techniques can be broadly divided into two categories, i.e., power-domain and code-domain NOMA [5]. The power-domain NOMA[1] allows multiple users to simultaneously transmit using the same radio resources, either in time, frequency, or in code [5]. At the transmitter side, power-domain user-multiplexing can be enabled using superposition coding [2]. At the receiver side, multiuser separation techniques, such as successive interference cancellation (SIC), can be utilized to decode the signal [6], [7].

Current research work in NOMA-related areas mainly focuses on the topics such as cooperative design [8]–[10], subcarrier assignment and power control policy [11]–[14], physical layer security [15], fairness analysis [16], etc. For example, a cooperative NOMA scheme was analyzed in [8], in which the users with the stronger channel conditions were used as relays to improve the reception reliability for users with poorer connections. In [9], the application of simultaneous wireless information and power transfer (SWIPT) to NOMA networks with randomly located users was investigated. Closed-form expressions for the outage probability and system throughput were derived to characterize the performance of the proposed user selection schemes. Further, considering a downlink NOMA transmission, an energy efficiency (EE) maximization problem was studied in [11], in which both the subcarrier assignment and the power allocation algorithms were provided for multiplexed users. Considering a downlink single-cell space division multiple access (SDMA) network with a multi-antenna base station and randomly deployed users, the performance of NOMA was investigated and optimized in [14], under a general channel state information (CSI) limited feedback framework. By leveraging limited feedback, a dynamic user scheduling and grouping strategy were proposed. The physical layer secrecy issue of NOMA was discussed in [15], in which the secrecy sum rate of a single-input single-output (SISO) NOMA system consisting of a transmitter, multiple legitimate users and an eavesdropper, was maximized subject to per-user minimum data rate requirement. Furthermore, the optimal power allocation technique to maximize the user fairness in a downlink NOMA network was investigated in [16], under two different assumptions: 1) when all users' data rates are adapted to the instantaneous CSI, and 2) when all users have fixed data rates under the average CSI.

However, all the aforementioned studies were based on Shannon limit theory, without taking into consideration the users' delay requirements. For systems with delay-sensitive applications, the physical-layer based performance analysis

Manuscript received October 8, 2017; revised January 18, 2018 and April 5, 2018; accepted April 22, 2018. This work was supported in part by the UK EPSRC under grant number EP/K011693/1, and grant number EP/N032268/1, and the EU FP7 under grant number PIRSES-GA-2013-610524. Part of this work was submitted to IEEE Global Communications Conference (GLOBECOM) 2018. The associate editor coordinating the review of this paper and approving it for publication was Z. Zhang. *(Corresponding author: Qiang Ni)*

W. Yu and Q. Ni are with the School of Computing and Communications, InfoLab21, Lancaster University, LA1 4WA, UK (Emails: {w.yu1, q.ni}@lancaster.ac.uk). L. Musavian is with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ, UK (Email: leila.musavian@essex.ac.uk).

[1]The power-domain NOMA will be simplified as NOMA, in the following sections.

and power adaptive techniques may not be efficient. Due to the random variations experienced in wireless channel conditions, user mobility and changing environment, the best way to consider each user's delay provisioning is to guarantee the per-user delay quality-of-service (QoS) requirement in a statistical way, i.e., to confine the delay bound violation probability to a required range [17]–[22]. Therefore, the effective capacity (EC) theory was introduced in [17], which specifies the maximum arrival rate that can be supported by a wireless channel, given that a target delay-outage probability is guaranteed.

In this paper, we focus on a downlink NOMA network with M users, and theoretically prove the advantage of NOMA over OMA, in terms of either the individual EC or the total link-layer achievable rate. Specifically, we analyze the performance of a two-user downlink NOMA network first. Considering heterogeneous statistical delay QoS constraints for users, we first derive the closed-form expressions for the individual EC in a two-user NOMA network. The impact of the transmit SNR and the delay QoS requirement on the individual link-layer rate and the total EC for the two-user network are then investigated and analyzed. Based on the theoretical analysis for a two-user NOMA network, the advantage of NOMA over OMA, for the whole network with M users is then proposed. To the best of our knowledge, the closed-form expressions for the link-layer rates in NOMA networks, the analytical conclusions regarding to the EC performance in NOMA, and also the comparison of the total EC between NOMA and OMA networks, are not yet studied in the existing literature.

In more detail, this paper has the following contributions:

- Focusing on a downlink M-user network, the individual EC and the total achievable link-layer rate are formulated and investigated. Assuming that M users are divided into multiple NOMA pairs, we prove that OMA achieves higher total EC than NOMA, at small SNRs. Further, simulation results show that NOMA outperforms OMA, at high SNRs.
- Focusing on a downlink M-user network, the total EC difference between NOMA and OMA becomes stable, when the transmit SNR is extremely high.
- Focusing on a two-user network, the closed-form expressions for the link-layer rates for both users, in NOMA and OMA, are derived in Section IV-A1. The accuracy is then confirmed by comparing with the Monte Carlo simulation results in Section V.
- Focusing on a two-user network, the impact of the transmit SNR[2] and the delay QoS exponent on the individual and the total EC is analyzed in two cases. Case 1: consider delay-constrained users; Case 2: consider delay-unconstrained users.
- In Case 1 and Case 2, we characterize the region of the transmit SNR, in which NOMA outperforms OMA, in terms of the individual and the total EC for the two-user system.

The remainder of this paper is organized as follows. The system model is given in Section II. In Section III, the theory of effective capacity is introduced. Then, we start to analyze and investigate the individual EC and the total link-layer rate for a downlink NOMA network in Section IV, which includes the closed-form expressions for the link-layer rates in a two-user network, in NOMA and OMA scenarios, and the theoretical analysis for a two-user network and a downlink NOMA network with multiple NOMA pairs. Simulation results are given in Section V, followed by conclusions in Section VI.

II. System Model

We consider a cellular downlink transmission with one base station (BS) and M single-antenna users. At the BS, the upper layer packets are organized into frames, which are then stored at the transmit buffer[3], in the link layer. After split into bit streams, these frames will be transmitted through the allocated channel. According to the NOMA principle, the BS will send $\sum_{k=1}^{M} \sqrt{\alpha_k P} s_m$ to the destinations, where s_k is the message for the k^{th} user, P is the total transmission power, and α_k denotes the power allocation coefficient for the k^{th} user.

As for each wireless channel from the BS to an individual user, we assume that it is block fading with a bandwidth of B, i.e., the channel gain is invariant during each fading-block, but independently varies from one fading-block to another. The length of each fading-block, denoted by T_{f}, is assumed to be an integer multiple of the symbol duration T_{s}. Meanwhile, the duration of one frame size is assumed to be equal to the length of the fading-block, i.e., T_{f}. The channel gain between the BS and the k^{th} user is denoted by h_k[4], which is modeled according to Rayleigh fading distribution. Without loss of generality, we assume that the users' channels have been sorted so that $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_M|^2$, which indicates that the k^{th} user always holds the k^{th} weakest channel. Henceforth, based on the NOMA protocol, we note that the power coefficients can be ordered as $\alpha_1 \geq \cdots \geq \alpha_M$, and $\sum_{k=1}^{M} \alpha_k = 1$ [8].

The received signal at the k^{th} user is given by $y_k = h_k \sum_{l=1}^{M} \sqrt{\alpha_l P} s_l + n_k$, where n_k denotes the additive white Gaussian noise. By applying the SIC technique, the k^{th} user will detect the i^{th} user's message, when $i < k$, and then remove the i^{th} user's message from its received signal, in a successive manner [8]. The message for the j^{th} user, for $j > k$, however, will be treated as noise at the k^{th} user. Note that the condition under which the k^{th} user can successfully decode the i^{th} user's message is to satisfy $R_{i\to k} \geq \tilde{R}_i$ [23]. Here, $\tilde{R}_i$ is the i^{th} user's target data rate, and $R_{i\to k}$ denotes the k^{th} user's data rate to detect the i^{th} user's message, i.e., $R_{i\to k} = \log_2\left(1 + \frac{\rho|h_k|^2\alpha_i}{\rho|h_k|^2\sum_{l=i+1}^{M}\alpha_l + 1}\right)$, where ρ denotes the transmit SNR, i.e., $\rho = \frac{P}{N_0 B}$, with $N_0 B$ indicating the noise power. Assume that $\tilde{R}_i$ is determined opportunistically by the i^{th} user's channel condition [23], i.e.,

[2]The transmit SNR is defined as the ratio of the transmission power to the noise power, in which the noise is assumed to be the additive white Gaussian noise. Further details will be provided in the next section.

[3]Here, we assume that the BS offers one virtual buffer for every served user.

[4]The time index t is omitted because the channel gains are assumed to be stationary and ergodic random processes.

$\tilde{R}_i = R_i = \log_2\left(1 + \frac{\rho|h_i|^2\alpha_i}{\rho|h_i|^2\sum_{l=i+1}^{M}\alpha_l + 1}\right)$, which means that its target rate equals to the data rate achieved when it decodes its own message. Hence, it is easy to verify that the condition $R_{i\to k} \geq \tilde{R}_i$ always holds since $|h_k|^2 \geq |h_i|^2$, for $k > i$.

Consequently, the achievable data rate[5], in b/s/Hz, for the k^{th} user in a downlink NOMA network, can be formulated as

$$R_k = \log_2\left(1 + \frac{\rho|h_k|^2\alpha_k}{\rho|h_k|^2\sum_{l=k+1}^{M}\alpha_l + 1}\right). \quad (1)$$

III. The Theory of Effective Capacity

In this section, the theory of EC is introduced to incorporate system throughput with the link-layer delay QoS metrics, such as the queue overflow probability, and the delay-outage probability. We take the k^{th} user as an example. At the BS, considering the dynamic queueing system for the k^{th} user, we assume that the buffer size is infinite and the link can serve $R_k(t)$ packets per unit of time, which means that the capacity of the link at time t is $R_k(t)$. Let $a_k(t)$ and $q_k(t)$ be the number of arrivals at time t and the number of packets in the queue at time t, respectively. Further, we assume that $a_k(t)$ and $R_k(t)$ are stationary and ergodic, and $E[a_k(t)] < E[R_k(t)]$, so that $q_k(t)$ converges to a steady rate $q_k(\infty)$ [24], [25].

Let us consider the queue overflow probability first, i.e., the probability of the steady-state queue length exceeding a certain threshold x. From large deviation theory, we note that the buffer overflow probability yields to [25]

$$-\lim_{x\to\infty}\frac{\ln\left(\Pr\{q_k(\infty) > x\}\right)}{x} = \theta_k, \quad (2)$$

where $\Pr\{a > b\}$ shows the probability that $a > b$ holds, and θ_k is the called delay QoS exponent. To satisfy a target buffer overflow probability in (2), it is required that

$$\Lambda_{a_k}(\theta_k) + \Lambda_{R_k}(-\theta_k) = 0, \quad (3)$$

where $\Lambda_{a_k}(\theta_k)$ and $\Lambda_{R_k}(\theta_k)$ are the Gärtner-Ellis limits of the arrival process and the service process, respectively, i.e., $\Lambda_{a_k}(\theta_k) = \lim_{T\to\infty}\frac{1}{T}\ln\left(E\left[e^{\theta_k\sum_{t=1}^{T}a_k(t)}\right]\right)$, and $\Lambda_{R_k}(\theta_k)$ equals to $\lim_{T\to\infty}\frac{1}{T}\ln\left(E\left[e^{\theta_k\sum_{t=1}^{T}R_k(t)}\right]\right)$ [25], [26]. When we assume that the arrival rate is a constant, i.e., $a_k(t) = a_k$, and insert it into (3), we get that $-\frac{\Lambda_{R_k}(-\theta_k)}{\theta_k} = a_k$, where θ_k is the unique delay QoS exponent which satisfies the required queue overflow probability in (2). Hence, $-\frac{\Lambda_{R_k}(-\theta_k)}{\theta_k}$ is called as the effective capacity, denoted by E_{c}^k, which represents the maximum arrival rate that a link can support[6], on the condition that a required delay QoS is satisfied [17].

When the focus is on the delay experienced by a source packet arriving at time t, defined by $D_k(t)$, the expression analogous to (2) can be estimated as [17]

$$P_{\text{delay}}^{\text{out}} = \Pr\{D_k(t) > D_{\max}^k\} \approx \Pr\{q_k(t) > 0\}\text{e}^{-\theta_k\mu_k D_{\max}^k}, \quad (4)$$

where $P_{\text{delay}}^{\text{out}}$ presents the delay-outage probability for the k^{th} user, $D_{\max}^k$ is in the unit of a symbol period, and $\Pr\{q_k(t) > 0\}$ denotes the probability of a non-empty buffer at time t. According to [17], we note that $\mu_k = E_{\text{c}}^k$. Therefore, in order to satisfy a required value of $P_{\text{delay}}^{\text{out}}$, a source needs to limit its maximum arrival rate to the value of μ_k, where μ_k equals to the EC satisfying the statistical delay QoS metrics. Furthermore, from (4), we notice that the parameter θ_k ($\theta_k > 0$) denotes the exponential decay rate of the delay-outage probability, for the k^{th} user. A smaller θ_k represents a slower decay rate, which indicates that the user can tolerate a loose delay QoS guarantee, while a larger θ_k means that a more stringent delay QoS guarantee is required [17], [18]. Specifically, when $\theta_k \to 0$, it indicates that the k^{th} user has no delay requirement. When $\theta_k \to \infty$, it means that the k^{th} user has an extremely stringent delay requirement [21].

By recalling that the wireless channel from the BS to the k^{th} user follows a block fading distribution, hence, the EC of the k^{th} user can be formulated as [17], [18]

$$E_{\text{c}}^k = -\frac{1}{\theta_k T_{\text{f}} B}\ln\left(\mathbb{E}\left[e^{-\theta_k T_{\text{f}} B R_k}\right]\right), \quad \text{(b/s/Hz)}, \quad (5)$$

by assuming that the Gärtner-Ellis limit exists. Here, $\mathbb{E}[\cdot]$ indicates the expectation over the probability density function (PDF) of the allocated channel. Then, by inserting (1) into (5), we can get the achievable link-layer rate for the k^{th} user in a downlink NOMA network, yielding

$$E_{\text{c}}^k = -\frac{1}{\theta_k T_{\text{f}} B}\ln\left(\mathbb{E}\left[\left(1 + \frac{\rho|h_k|^2\alpha_k}{\rho|h_k|^2\sum_{l=k+1}^{M}\alpha_l + 1}\right)^{-\frac{\theta_k T_{\text{f}} B}{\ln 2}}\right]\right). \quad (6)$$

IV. Effective Capacity in a Downlink NOMA Network

Aware of the difficulty of deriving the closed-form expression for the individual EC in (6) when all M users transmit on the same channel, we start to investigate the situation when there are multiple NOMA pairs in a M-user network. Specifically, we consider that the M users are divided into $M/2$ groups[7], so that within each group, NOMA will be implemented for only two users, and the conventional OMA can be used for inter-NOMA-pairs multiple access [8]. Furthermore, we note that a two-user downlink version of NOMA, called as the multiuser superposition transmission (MUST), has been proposed for the Third Generation Partnership Project Long Term Evolution Advanced (3GPP-LTE-A) networks [27]. Inspired by this, we first focus on the link-layer rate performance of a two-user downlink NOMA network, which itself is of great importance, and also paves the way for the performance analysis of multiple NOMA pairs. Closed-form expressions and insightful theoretical conclusions will be provided. Finally, based on the proposed derivations and theoretical insights, we derive and investigate the total EC for

[5]We assume that the distance-based path-loss is uniform for each user.

[6]Although the above analysis is based on the constant arrival rate, the theory of EC can also apply to any stationary arrival processes [25].

[7]To achieve this, we assume that M is an even positive number.

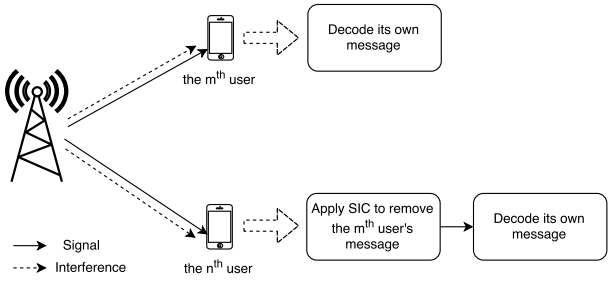

Fig. 1: Two-user downlink NOMA network.

the multiple NOMA pairs, in comparison with the total EC for M OMA users.

A. Effective Capacity of a two-user NOMA network

Without loss of generality, the m^{th} user and the n^{th} user, $m < n$, are assumed to be paired together as a two-user NOMA network, as depicted in Fig. 1. By applying the SIC strategy, the n^{th} user, which has the relatively stronger channel condition, will first decode the message of the user with the weaker channel condition, i.e., the m^{th} user, and then decode its own message by removing the m^{th} user's message. On the other hand, the m^{th} user with the weaker channel condition, will decode its own message by treating the n^{th} user's information as noise. In order to make sure that SIC can be correctly carried out at the n^{th} user, it is required that $R_{m\to n} \geq R_m$, i.e., $\log_2\left(1+\frac{\rho\alpha_m|h_n|^2}{\rho\alpha_n|h_n|^2+1}\right) \geq R_m$. According to the analysis in Section II, we note that this always holds since $|h_n|^2 \geq |h_m|^2$, for $n > m$.

By applying the fixed power allocation, the power allocation coefficients for the m^{th} user and the n^{th} user are denoted by α_m and α_n, respectively, where $\alpha_m \geq \alpha_n$, and $\alpha_m + \alpha_n = 1$, according to the NOMA principle. By assuming that both users experience the same strength of additive white Gaussian noise, then the achievable data rates[8], in b/s/Hz, for the m^{th} user and the n^{th} user in a two-user NOMA network, are respectively formulated as

$$R_m = \log_2\left(1+\frac{\rho\alpha_m|h_m|^2}{\rho\alpha_n|h_m|^2+1}\right), \tag{7a}$$

$$R_n = \log_2\left(1+\rho\alpha_n|h_n|^2\right). \tag{7b}$$

On the other hand, if the m^{th} user and the n^{th} user each have their message transmitted using OMA scheduling, e.g., time division multiple access (TDMA), with total transmit SNR ρ, the achievable data rate of each user can then be given by

$$\bar{R}_i = \frac{1}{2}\log_2\left(1+\rho|h_i|^2\right),\ i \in \{m,n\} \tag{8}$$

where $\frac{1}{2}$ denotes that each user has only half of the available radio resources in OMA networks. Considering the duration of one frame as one time slot, (8) implies that in TDMA networks, each user can only occupy half of the time slot to transmit, while in the other half time slot, it will stay silent[9].

[8]We assume that the distance-based path-loss is uniform for each user.

[9]We note that the way of equally allocating resource is a typical and special case. However, the influence of different resource allocation strategies is beyond the scope of this paper.

Assuming that the Gärtner-Ellis theorem [26] is satisfied, the expressions of EC for the m^{th} user and the n^{th} user in a block fading channel can be respectively given as [17]

$$E_{\text{c}}^m = -\frac{1}{\theta_m T_{\text{f}} B}\ln\left(\mathbb{E}\left[e^{-\theta_m T_{\text{f}} B R_m}\right]\right) \quad \text{(b/s/Hz)}, \tag{9a}$$

$$E_{\text{c}}^n = -\frac{1}{\theta_n T_{\text{f}} B}\ln\left(\mathbb{E}\left[e^{-\theta_n T_{\text{f}} B R_n}\right]\right) \quad \text{(b/s/Hz)}. \tag{9b}$$

By inserting (7a) into (9a) and inserting (7b) into (9b), we then get that

$$E_{\text{c}}^m = -\frac{1}{\theta_m T_{\text{f}} B}\ln\left(\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]\right), \tag{10a}$$

$$E_{\text{c}}^n = -\frac{1}{\theta_n T_{\text{f}} B}\ln\left(\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]\right), \tag{10b}$$

where $\beta_m = -\frac{\theta_m T_{\text{f}} B}{2\ln 2}$, and $\beta_n = -\frac{\theta_n T_{\text{f}} B}{2\ln 2}$.

For an OMA scheme, such as TDMA, the EC expressions for both users can be calculated by inserting (8) into (9a) and (9b), which yield to

$$\bar{E}_{\text{c}}^m = -\frac{1}{\theta_m T_{\text{f}} B}\ln\left(\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m}\right]\right), \tag{11a}$$

$$\bar{E}_{\text{c}}^n = -\frac{1}{\theta_n T_{\text{f}} B}\ln\left(\mathbb{E}\left[\left(1+\rho|h_n|^2\right)^{\beta_n}\right]\right). \tag{11b}$$

In the following subsection, we first derive the closed-form expressions for the link-layer rates for both users, in NOMA and OMA, i.e., E_{c}^m, $\bar{E}_{\text{c}}^m$, E_{c}^n, and $\bar{E}_{\text{c}}^n$. Further, the impact of the transmit SNR ρ and the per-user delay QoS exponent, on the individual EC performance and the total link-layer rates, in both NOMA and OMA scenarios, will be investigated and analyzed for the two-user network.

1) The closed-form expressions for the individual EC in a two-user system

We suppose that $h_1, \ldots, h_M$ are M unordered independent channel gains, modeled according to the unit-variance Rayleigh fading distribution. Set $\gamma_m = \rho|h_m|^2$ and $\gamma_n = \rho|h_n|^2$. When γ_m and γ_n are unordered, the PDF of γ_m and γ_n is denoted by $f(\gamma_m)$ and $f(\gamma_n)$, respectively. Correspondingly, the cumulative distribution function (CDF) of the unordered γ_m and γ_n can be denoted by $F(\gamma_m)$, and $F(\gamma_n)$. Since the unordered channel gains are assumed to be statistically independent and identically distributed, hence, we can notice that $f(\gamma_m) = f(\gamma_n)$, and $F(\gamma_m) = F(\gamma_n), \forall\, m,\ n \in \{1,\ldots,M\}$. However, when we assume that the users' channels are sorted so that $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_M|^2$, the order statistics of different channel power gains will not be the same. In NOMA networks, the users are ordered first according to their channel conditions, therefore the statistical features of the ordered channel power gains fall into the scope of the order statistics [28]. The PDF of the ordered γ_m and γ_n, where $\gamma_m \leq \gamma_n$, are denoted by $f_{(m)}(\gamma_m)$, and $f_{(n)}(\gamma_n)$, respectively. From order statistics [28], $f_{(m)}(\gamma_m)$ and $f_{(n)}(\gamma_n)$ are given by

$$f_{(m)}(\gamma_m) = \psi_m f(\gamma_m) F(\gamma_m)^{m-1}\left(1-F(\gamma_m)\right)^{M-m}, \tag{12a}$$

$$f_{(n)}(\gamma_n) = \psi_n f(\gamma_n) F(\gamma_n)^{n-1}\left(1-F(\gamma_n)\right)^{M-n}, \tag{12b}$$

$$E_{\rm c}^m = -\frac{1}{\theta_m T_{\rm f} B}\ln\Bigg(\frac{\alpha_n^{-2\beta_m}\psi_m}{\rho}\Bigg(\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k\frac{\rho}{M-m+1+k}+\frac{\theta_m(\alpha_n-1)}{\alpha_n\ln 2}\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k e^{\frac{M-m+1+k}{\rho\alpha_n}}$$
$$\times E_i\left(-\frac{M-m+1+k}{\rho\alpha_n}\right)+\sum_{j=2}^{\infty}\binom{2\beta_m}{j}\left(\frac{\alpha_n-1}{\alpha_n}\right)^j\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k\Bigg(\frac{\sum_{i=1}^{j-1}\frac{(i-1)!}{\alpha_n^{-i}}\left(-\frac{M-m+1+k}{\rho}\right)^{j-i-1}}{(j-1)!}$$
$$-\frac{\left(-\frac{M-m+1+k}{\rho}\right)^{j-1}}{(j-1)!}\,e^{\frac{M-m+1+k}{\rho\alpha_n}}E_i\left(-\frac{M-m+1+k}{\rho\alpha_n}\right)\Bigg)\Bigg)\Bigg), \tag{13a}$$

$$\bar{E}_{\rm c}^m = -\frac{1}{\theta_m T_{\rm f} B}\ln\left(\frac{\psi_m}{\rho}\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k U\left(1,2+\beta_m,\frac{M-m+1+k}{\rho}\right)\right), \tag{13b}$$

$$E_{\rm c}^n = -\frac{1}{\theta_n T_{\rm f} B}\ln\left(\frac{\psi_n}{\rho\alpha_n}\sum_{k=0}^{n-1}\binom{n-1}{k}(-1)^k U\left(1,2+2\beta_n,\frac{M-n+1+k}{\rho\alpha_n}\right)\right), \tag{13c}$$

$$\bar{E}_{\rm c}^n = -\frac{1}{\theta_n T_{\rm f} B}\ln\left(\frac{\psi_n}{\rho}\sum_{k=0}^{n-1}\binom{n-1}{k}(-1)^k U\left(1,2+\beta_n,\frac{M-n+1+k}{\rho}\right)\right). \tag{13d}$$

where $\psi_m = \dfrac{1}{B\left(m, M-m+1\right)}$, $\psi_n = \dfrac{1}{B\left(n, M-n+1\right)}$, in which $B(a,b)$ denotes the beta function, according to $B(a,b) = \dfrac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(a) = a!$, as a is a positive integer.

Theorem 1: For the $m^{\rm th}$ user, the closed-form expression for the EC in NOMA, $E_{\rm c}^m$, is given in (13a), where $E_i\left(\cdot\right)$ is the exponential integral. Meanwhile, the EC in OMA, $\bar{E}_{\rm c}^m$, can be expressed in closed-form, given in (13b), where $U\left(a,b,z\right)$ is the confluent hypergeometric function of the second kind [29].

Proof: The proof is provided in Appendix A. ■

Theorem 2: For the $n^{\rm th}$ user, the closed-form expression for the EC in NOMA, $E_{\rm c}^n$, is given in (13c). Meanwhile, the EC in OMA, $\bar{E}_{\rm c}^n$, can be expressed in closed-form, given in (13d).

Proof: The proof is omitted here, but can be found by following similar steps as in Appendix A. ■

The accuracy of the above closed-form expressions will be confirmed by comparing with Monte Carlo simulations in Section V. Then, we start to investigate the impact of the transmit SNR ρ and the per-user delay QoS exponents θ_m, θ_n, on the individual EC performance and the total link-layer rate for a two-user network, in both NOMA and OMA scenarios. Two cases are deliberately analyzed in the following subsections, i.e., Case 1: consider delay-constrained users[10]; Case 2: consider delay-unconstrained users. We note that Case 2 is an extreme case of no delay, in which the individual EC is proved to be equivalent to ergodic capacity[11]. Interestingly, the theoretical and simulation results obtained for this case are indeed novel and not found in the current literature. Further, by including Case 1 and Case 2, the performance of a two-user downlink NOMA network, either delay-constrained or delay-unconstrained, can be comprehensively analyzed and investigated.

2) Case 1: consider delay-constrained users

Lemma 1: Considering the individual EC in NOMA and OMA, for both users, we prove that

(a) When $\rho \to 0$, $E_{\rm c}^m \to 0$, $\bar{E}_{\rm c}^m \to 0$, $E_{\rm c}^m - \bar{E}_{\rm c}^m \to 0$, $E_{\rm c}^n \to 0$, $\bar{E}_{\rm c}^n \to 0$, and $E_{\rm c}^n - \bar{E}_{\rm c}^n \to 0$.

(b) When $\rho \to \infty$[12], $\lim\limits_{\rho\to\infty} E_{\rm c}^m = \log_2\left(\dfrac{1}{\alpha_n}\right)$, $\lim\limits_{\rho\to\infty} \bar{E}_{\rm c}^m \to \infty$, and $\lim\limits_{\rho\to\infty}\left(E_{\rm c}^m - \bar{E}_{\rm c}^m\right) \to -\infty$.

(c) When $\rho \to \infty$, $\lim\limits_{\rho\to\infty} E_{\rm c}^n \to \infty$, $\lim\limits_{\rho\to\infty} \bar{E}_{\rm c}^n \to \infty$, and $\lim\limits_{\rho\to\infty}\left(E_{\rm c}^n - \bar{E}_{\rm c}^n\right) \to \infty$.

Proof: The proof is provided in Appendix B. ■

From Lemma 1.(a), we note that, for both users, either in NOMA or OMA, their individual rates start at the same initial value of 0, at small values of ρ. Lemma 1.(b), on the other hand, indicates that for the weaker user[13], when $\rho \to \infty$, its EC achieved by applying NOMA is limited by $\log_2\left(\dfrac{1}{\alpha_n}\right)$. This means that in a two-user NOMA network, the weaker user can only achieve a limited EC, no matter how large the transmit SNR can be. On the contrary, for the stronger user[14], Lemma 1.(c) indicates that when $\rho \to \infty$, its achievable EC in NOMA approaches infinity. Furthermore, Lemma 1.(b) and Lemma 1.(c) reveal that when $\rho \to \infty$, the EC values achieved by applying OMA approach infinity, for both of the two users.

[10]In this case, finite values of θ_m, θ_n are considered.

[11]The proof and further explanations can be found in Lemma 5.

[12]We note that $\rho \to \infty$ is not practical, but this is only to provide a guideline. From the simulation results in Section V, it shows that the conclusions we proved for the case of $\rho \to \infty$, are valid for values of ρ as big as $\rho = 30$dB.

[13]Hereafter, the user with the weaker channel condition is referred to as the weaker user.

[14]Hereafter, the user with the stronger channel condition is referred to as the stronger user.

Apparently, Lemma 1 only considers two extreme cases of ρ for both users. Henceforth, from Lemma 1, one cannot know how the individual EC will change with respect to ρ on general terms. Will NOMA be always better than OMA for the n^{th} user, at any positive values of ρ? Will OMA be always better than NOMA for the m^{th} user, for any settings of ρ? To answer these questions and to further analyze the impact of ρ on the individual EC, in a two-user NOMA network and in a two-user OMA network, we provide the following lemmas.

Lemma 2: Considering the m^{th} user's EC, in NOMA and OMA, we prove that

(a) At any values of ρ, $\frac{\partial E_{\text{c}}^m}{\partial \rho} \geq 0$, and $\frac{\partial \bar{E}_{\text{c}}^m}{\partial \rho} \geq 0$.

(b) When $\rho \to 0$, $\lim\limits_{\rho \to 0} \frac{\partial \left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right)}{\partial \rho} = \frac{\frac{1}{2} - \alpha_n}{\ln 2} \mathbb{E}\left[|h_m|^2\right] \geq 0$.

(c) When ρ is very large, $\frac{\partial \left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right)}{\partial \rho} \leq 0$, and it approaches 0 when $\rho \to \infty$.

Proof: The proof is provided in Appendix C. ∎

Lemma 2.(b) indicates that, for the weaker user, when the transmit SNR is very small, the EC in NOMA has a faster increasing speed than that in OMA. On the contrary, Lemma 2.(c) shows that for the weaker user, when the transmit SNR is very large, the EC in OMA increases faster than that in NOMA. To further explain these conclusions, we focus on analyzing the EC difference between NOMA and OMA, for the weaker user in a two-user system. From Lemma 1 and Lemma 2, one can conclude that, $E_{\text{c}}^m - \bar{E}_{\text{c}}^m$ starts at the initial value of 0, first increases, and at the end decreases to $-\infty$ with a gradually diminishing speed. This means that, for the weaker user, NOMA can achieve higher EC than OMA, at small values of ρ. When the transmit SNR becomes extremely large, OMA is more beneficial than NOMA, for the weaker user. Finally, when $\rho \to \infty$, the performance gain of OMA over NOMA becomes stable.

Lemma 3: Considering the n^{th} user's EC, in NOMA and OMA, we prove that

(a) At any values of ρ, $\frac{\partial E_{\text{c}}^n}{\partial \rho} \geq 0$, and $\frac{\partial \bar{E}_{\text{c}}^n}{\partial \rho} \geq 0$.

(b) When $\rho \to 0$, $\lim\limits_{\rho \to 0} \frac{\partial \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right)}{\partial \rho} = \frac{\alpha_n - \frac{1}{2}}{\ln 2} \mathbb{E}\left[|h_n|^2\right] \leq 0$.

(c) When ρ is very large, $\frac{\partial \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right)}{\partial \rho} \geq 0$, and it approaches 0 when $\rho \to \infty$.

Proof: The proof is provided in Appendix D. ∎

Lemma 3.(b) indicates that, for the stronger user, when the transmit SNR is very small, the EC in OMA increases faster than that in NOMA. On the contrary, Lemma 3.(c) shows that when the transmit SNR becomes very large, the EC in NOMA increases faster than the one in OMA, for the stronger user. Then we start to analyze the range of ρ, in which NOMA is more beneficial than OMA, for the stronger user in a two-user system. From Lemma 1 and Lemma 3, one can conclude that, $E_{\text{c}}^n - \bar{E}_{\text{c}}^n$ starts at the initial value of 0, first decreases, and finally increases to ∞ with a gradually reducing speed. This means that, for the stronger user, OMA achieves higher EC than NOMA, when the transmit SNR is small. At high values of ρ, NOMA becomes more beneficial than OMA, for the stronger user. Finally, when $\rho \to \infty$, the performance gain of NOMA over OMA becomes stable, for the stronger user.

In order to investigate the impact of the transmit SNR ρ on the performance of the total link-layer achievable rate, we define $\mathrm{T}_N = E_{\text{c}}^m + E_{\text{c}}^n$, which indicates the total EC for the two-user NOMA network. Meanwhile, we define $\mathrm{T}_O = \bar{E}_{\text{c}}^m + \bar{E}_{\text{c}}^n$, which denotes the total achievable link-layer rate for the two-user OMA system.

Lemma 4: Considering the total EC in NOMA, T_N, for the two-user system, we prove that

(a) At any values of ρ, $\frac{\partial \mathrm{T}_N}{\partial \rho} \geq 0$.

(b) When $\rho \to 0$, $\mathrm{T}_N \to 0$, $\lim\limits_{\rho \to 0} \frac{\partial \mathrm{T}_N}{\partial \rho} = \frac{1 - \alpha_n}{\ln 2} \mathbb{E}\left[|h_m|^2\right] + \frac{\alpha_n}{\ln 2} \mathbb{E}\left[|h_n|^2\right] \geq 0$.

(c) When $\rho \to \infty$, $\mathrm{T}_N \to \infty$, $\lim\limits_{\rho \to \infty} \frac{\partial \mathrm{T}_N}{\partial \rho} = 0$.

Considering the total EC in OMA, T_O, for the two-user system, we prove that

(d) At any values of ρ, $\frac{\partial \mathrm{T}_O}{\partial \rho} \geq 0$.

(e) When $\rho \to 0$, $\mathrm{T}_O \to 0$, $\lim\limits_{\rho \to 0} \frac{\partial \mathrm{T}_O}{\partial \rho} = \frac{1}{2 \ln 2} \mathbb{E}\left[|h_m|^2\right] + \frac{1}{2 \ln 2} \mathbb{E}\left[|h_n|^2\right] \geq 0$.

(f) When $\rho \to \infty$, $\mathrm{T}_O \to \infty$, $\lim\limits_{\rho \to \infty} \frac{\partial \mathrm{T}_O}{\partial \rho} = 0$.

Proof: The proof is provided in Appendix E. ∎

Lemma 4.(b) indicates that when the NOMA scheme is applied, the total EC has a constant slope at small values of ρ, in which the constant depends on the average of the channel power gains and the allocated power coefficients. On the contrary, from Lemma 4.(e), we find that the total EC obtained in OMA scheme also shows a constant increasing speed at small values of ρ, in which the constant only depends on the average of the channel power gains. Finally, when $\rho \to \infty$, Lemma 4.(c) and Lemma 4.(f) show that the increasing speed of the total EC, either in NOMA or OMA, gradually diminishes.

3) Case 2: consider delay-unconstrained users

In this subsection, we investigate the delay-unconstrained EC, in a two-user NOMA network and a two-user OMA network, when $\theta_m \to 0$, $\theta_n \to 0$, i.e., $\lim\limits_{\theta_m \to 0} E_{\text{c}}^m$, $\lim\limits_{\theta_m \to 0} \bar{E}_{\text{c}}^m$, $\lim\limits_{\theta_n \to 0} E_{\text{c}}^n$, $\lim\limits_{\theta_n \to 0} \bar{E}_{\text{c}}^n$, and also the EC difference between NOMA and OMA, for both users, i.e., $\lim\limits_{\theta_m \to 0} \left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right)$ and $\lim\limits_{\theta_n \to 0} \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right)$. Further, the impact of ρ in this extreme case is also analyzed and investigated.

Lemma 5: Considering the EC for the m^{th} user with $\theta_m \to 0$, in NOMA and OMA, we prove that

(a) When $\theta_m \to 0$, $\lim\limits_{\theta_m \to 0} E_{\text{c}}^m = \mathbb{E}[R_m]$, $\lim\limits_{\theta_m \to 0} \bar{E}_{\text{c}}^m = \mathbb{E}\left[\bar{R}_m\right]$, $\lim\limits_{\theta_m \to 0} \left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right) = \mathbb{E}\left[R_m\right] - \mathbb{E}\left[\bar{R}_m\right]$.

(b) When $\theta_m \to 0$, $\rho \to \infty$, $\lim\limits_{\substack{\theta_m \to 0 \\ \rho \to \infty}} E_{\text{c}}^m = \log_2\left(\frac{1}{\alpha_n}\right)$, $\lim\limits_{\substack{\theta_m \to 0 \\ \rho \to \infty}} \bar{E}_{\text{c}}^m \to \infty$, $\lim\limits_{\substack{\theta_m \to 0 \\ \rho \to \infty}} \left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right) \to -\infty$.

Considering the EC for the n^{th} user with $\theta_n \to 0$, in NOMA and OMA, we prove that

(c) When $\theta_n \to 0$, $\lim\limits_{\theta_n \to 0} E_{\text{c}}^n = \mathbb{E}\left[R_n\right]$, $\lim\limits_{\theta_n \to 0} \bar{E}_{\text{c}}^n = \mathbb{E}\left[\bar{R}_n\right]$, $\lim\limits_{\theta_n \to 0} \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right) = \mathbb{E}\left[R_n\right] - \mathbb{E}\left[\bar{R}_n\right]$.

(d) When $\theta_n \to 0$, $\rho \to \infty$, $\lim\limits_{\substack{\theta_n \to 0 \\ \rho \to \infty}} E_{\text{c}}^n \to \infty$, $\lim\limits_{\substack{\theta_n \to 0 \\ \rho \to \infty}} \bar{E}_{\text{c}}^n \to \infty$, $\lim\limits_{\substack{\theta_n \to 0 \\ \rho \to \infty}} \left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right) \to \infty$.

Proof: The proof is provided in Appendix F. ∎

From Lemma 5.(a) and Lemma 5.(c), we note that for both users, no matter in NOMA or OMA, when there is no delay requirement, i.e., $\theta_m \to 0$, and $\theta_n \to 0$, the individual achievable link-layer rate is equivalent to the ergodic capacity. Furthermore, from Lemma 1 and Lemma 5, we can find that Case 2, as an extreme case of no delay, follows similar conclusions with Case 1, regarding to the individual EC performance at high SNRs. For example, from Lemma 1.(b) and Lemma 5.(b), we note that, the weaker user in a two-user NOMA system can only achieve a limited EC, no matter how large the transmit SNR can be, or how strict or loose the delay exponent is. Further, one can also conclude that, for the weaker user, either with or without delay constraint, OMA offers higher EC than NOMA, when $\rho \to \infty$. On the contrary, for the stronger user, either with or without delay constraint, NOMA achieves higher EC than OMA at high SNRs. Note that by following similar steps as in Appendix E, one can show that Case 2 follows similar conclusions as in Case 1, regarding to the total EC performance in NOMA and OMA. However, these are omitted in this paper to avoid redundancy.

B. Effective Capacity of multiple NOMA pairs

After analyzing the two-user NOMA network and deriving the closed-form expressions, we investigate the total achievable link-layer rate for multiple NOMA pairs. By considering that the M users are divided into $\frac{M}{2}$ groups, we define $\mathbb{I} = \left\{1, 2, \ldots, \frac{M}{2}\right\}$, which contains the group index. Then, all NOMA pairs can be included in Φ, $\Phi = \{\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_{M/2}\}$, satisfying $\boldsymbol{\phi}_i \cap \boldsymbol{\phi}_j = \varnothing, i \neq j, \forall i, j \in \mathbb{I}$, where $\boldsymbol{\phi}_i = \left\{(m_i, n_i) \mid m_i \neq n_i, |h_{m_i}|^2 \leq |h_{n_i}|^2, \forall i \in \mathbb{I}\right\}$ denotes the i^{th} NOMA pair with two users, i.e., m_i and n_i.

Assume that for the i^{th} NOMA pair, $\forall i \in \mathbb{I}$, NOMA will be implemented for the two users, i.e., m_i and n_i. Meanwhile, for the inter-group multiple access, we assume that TDMA will be applied. Hence, for the two users in the i^{th} NOMA pair, the achievable data rates, in b/s/Hz, can be respectively formulated as

$$R_{m_i} = \frac{2}{M} \log_2 \left(1 + \frac{\rho \alpha_{m_i} |h_{m_i}|^2}{\rho \alpha_{n_i} |h_{m_i}|^2 + 1}\right), \tag{14a}$$

$$R_{n_i} = \frac{2}{M} \log_2 \left(1 + \rho \alpha_{n_i} |h_{n_i}|^2\right). \tag{14b}$$

On the other hand, if the users m_i and n_i each have their message transmitted using TDMA, the achievable data rate for each user can be given by

$$\bar{R}_j = \frac{1}{M} \log_2 \left(1 + \rho |h_j|^2\right), \ j \in \{m_i, n_i\}, \tag{15}$$

where $\frac{1}{M}$ denotes that each user has only $\frac{1}{M}$ of the time slot to transmit, while in the other fractions of the time slot, it will stay silent.

Assuming that the Gärtner-Ellis theorem is satisfied, we can get the EC formulations for the users m_i and n_i in the i^{th} NOMA pair, yielding

$$E_{\text{c}}^{m_i} = -\frac{1}{\theta_{m_i} T_{\text{f}} B} \ln \left(\mathbb{E}\left[\left(\frac{\rho |h_{m_i}|^2 + 1}{\rho \alpha_{n_i} |h_{m_i}|^2 + 1}\right)^{\frac{4}{M} \beta_{m_i}}\right]\right), \tag{16a}$$

$$E_{\text{c}}^{n_i} = -\frac{1}{\theta_{n_i} T_{\text{f}} B} \ln \left(\mathbb{E}\left[\left(1 + \rho \alpha_{n_i} |h_{n_i}|^2\right)^{\frac{4}{M} \beta_{n_i}}\right]\right), \tag{16b}$$

where $\beta_{m_i} = -\frac{\theta_{m_i} T_{\text{f}} B}{2 \ln 2}$, and $\beta_{n_i} = -\frac{\theta_{n_i} T_{\text{f}} B}{2 \ln 2}$. On the contrary, for the TDMA scheme, the EC expressions for both users can also be obtained, which respectively yield to

$$\bar{E}_{\text{c}}^{m_i} = -\frac{1}{\theta_{m_i} T_{\text{f}} B} \ln \left(\mathbb{E}\left[\left(1 + \rho |h_{m_i}|^2\right)^{\frac{2}{M} \beta_{m_i}}\right]\right), \tag{17a}$$

$$\bar{E}_{\text{c}}^{n_i} = -\frac{1}{\theta_{n_i} T_{\text{f}} B} \ln \left(\mathbb{E}\left[\left(1 + \rho |h_{n_i}|^2\right)^{\frac{2}{M} \beta_{n_i}}\right]\right). \tag{17b}$$

Comparing (16a)-(17b) with (10a)-(11b), we can notice that the EC formulations for the two users in the i^{th} NOMA pair, have similar expressions with those proposed for a two-user NOMA network in Section IV-A. Hence, by following similar steps in Appendix A, the closed-form expressions for $E_{\text{c}}^{m_i}$, $E_{\text{c}}^{n_i}$, $\bar{E}_{\text{c}}^{m_i}$, and $\bar{E}_{\text{c}}^{n_i}$ can be easily obtained, which are omitted here for simplicity. Our focus lies on analyzing the total EC of multiple NOMA pairs, denoted by M_N, in comparison with the total EC for the M OMA users, i.e., M_O. Note that M_N can be defined as $\sum\limits_{i=1}^{M/2} \left(E_{\text{c}}^{m_i} + E_{\text{c}}^{n_i}\right)$, and correspondingly, M_O equals to $\sum\limits_{i=1}^{M/2} \left(\bar{E}_{\text{c}}^{m_i} + \bar{E}_{\text{c}}^{n_i}\right)$. To investigate the region of ρ, in which NOMA can offer a higher value of the total link-layer rate for multiple NOMA pairs, in comparison with the OMA scheme, we provide the following lemma.

Lemma 6: Considering the difference of the total EC, between multiple NOMA pairs and M OMA users, we prove that

(a) When $\rho \to 0$, $\text{M}_N - \text{M}_O \to 0$, $\lim\limits_{\rho \to 0} \dfrac{\partial \left(\text{M}_N - \text{M}_O\right)}{\partial \rho} = \sum\limits_{i=1}^{M/2} \dfrac{1 - 2\alpha_{n_i}}{M \ln 2} \left(\mathbb{E}\left[|h_{m_i}|^2\right] - \mathbb{E}\left[|h_{n_i}|^2\right]\right) \leq 0$.

(b) When $\rho \to \infty$, $\text{M}_N - \text{M}_O$ approaches a constant, given in (18), and $\lim\limits_{\rho \to \infty} \dfrac{\partial \left(\text{M}_N - \text{M}_O\right)}{\partial \rho} = 0$.

$$\lim_{\rho \to \infty} \left(\text{M}_N - \text{M}_O\right) = \sum_{i=1}^{M/2} -\frac{1}{\theta_{m_i} T_{\text{f}} B} \ln \left(\frac{\alpha_{n_i}^{-\frac{4}{M} \beta_{m_i}}}{\mathbb{E}\left[\left(|h_{m_i}|^2\right)^{\frac{2}{M} \beta_{m_i}}\right]}\right) - \frac{1}{\theta_{n_i} T_{\text{f}} B} \ln \left(\frac{\alpha_{n_i}^{\frac{4}{M} \beta_{n_i}} \mathbb{E}\left[\left(|h_{n_i}|^2\right)^{\frac{4}{M} \beta_{n_i}}\right]}{\mathbb{E}\left[\left(|h_{n_i}|^2\right)^{\frac{2}{M} \beta_{n_i}}\right]}\right). \tag{18}$$

Proof: The proof is provided in Appendix G. ∎

From Lemma 6, one can conclude that $\text{M}_N - \text{M}_O$ starts at the initial value of 0, first decreases at small values of ρ, and finally approaches a constant, given in (18), when $\rho \to \infty$.

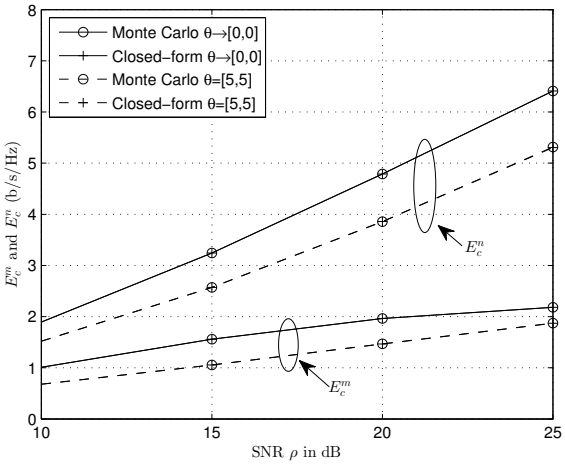

Fig. 2: $E_{\rm c}^m$ and $E_{\rm c}^n$, in NOMA, versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

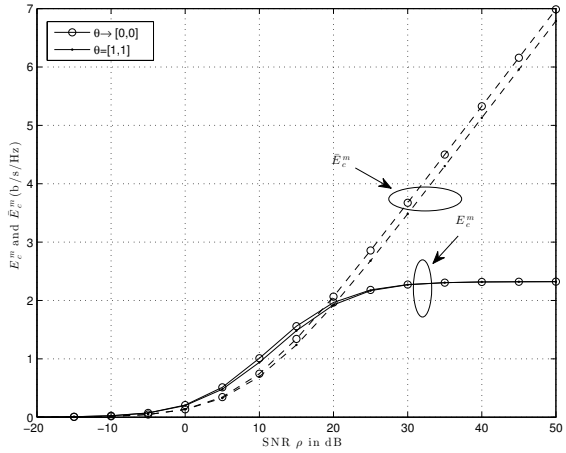

Fig. 3: $E_{\rm c}^m$, in NOMA, and $\bar{E}_{\rm c}^m$, in OMA, versus the transmit SNR ρ for various values of $\boldsymbol{\theta}$.

This indicates that OMA outperforms NOMA on the total link-layer rate performance for a M-user network, at small SNRs. Simulation results in the next section further show that NOMA achieves higher total EC than OMA at high values of SNR. Finally, Lemma 6.(b) indicates that the performance gain of NOMA over OMA becomes stable when the transmit SNR becomes extremely high.

Note that in Section IV-A2 and Section IV-A3, considering delay-constrained and delay-unconstrained users, we have comprehensively investigated the individual link-layer rate and the total EC for a two-user NOMA system, in comparison with the conventional OMA scheme. Then, in Section IV-B, considering that M users are divided into multiple NOMA pairs, we have characterized the regions of ρ, in which NOMA offers higher total EC than the conventional OMA scheme. These insightful conclusions, mathematically derived and theoretically proved, can provide valuable guidelines for the further research, such as the resource allocation design, user pairing/clustering technique and delay analysis in NOMA. Further, the above theoretical conclusions will be confirmed using simulation results in Section V.

V. Numerical Results

In this section, we will numerically confirm all the theorems and the lemmas proposed in Section IV. Further, the impact of the per-user delay QoS exponent, and the transmit SNR ρ on the individual EC performance and the total link-layer rate, in NOMA and OMA scenarios, is numerically analyzed and investigated in this section. Specifically, we start from showing the simulation results for the two-user system, in NOMA and OMA. To consider a two-user NOMA system, the total number of users $M = 10$, and the users with the 2$^{\rm rd}$ and the 8$^{\rm th}$ weakest channels are assumed to be paired together, i.e., $m = 2$, $n = 8$. The corresponding power coefficients for the two users are set as, $\alpha_m = 0.8$, $\alpha_n = 0.2$, unless otherwise indicated. The fading-block duration $T_{\rm f} = 0.01$ ms, and the bandwidth $B = 100$kHz.

To confirm the accuracy of the proposed closed-form expressions for EC in NOMA scheme for both users, we include

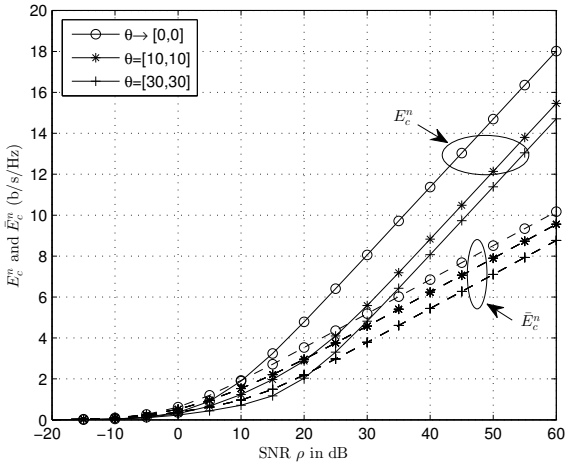

Fig. 4: $E_{\rm c}^n$, in NOMA, and $\bar{E}_{\rm c}^n$, in OMA, versus the transmit SNR ρ for various values of $\boldsymbol{\theta}$.

Fig. 2 which plots the curves of $E_{\rm c}^m$ and $E_{\rm c}^n$ versus the transmit SNR ρ, for various values of the delay QoS exponent vector $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = [\theta_m, \theta_n]$. This figure shows the results calculated in two ways, i.e., by using Monte Carlo simulation method and the proposed closed-form expressions. From Fig. 2, the accuracy of the closed-form expressions for EC in NOMA scheme for both users can be confirmed. For both users, $E_{\rm c}^m$ and $E_{\rm c}^n$ gradually increase with the transmit SNR ρ, which confirms the proposed Lemma 2.(a) and Lemma 3.(a). Further, when the delay QoS exponent vector becomes more stringent, i.e., changing from $\boldsymbol{\theta} \to [0, 0]$ to $\boldsymbol{\theta} = [5, 5]$, the individual link-layer rates in NOMA, for both users, decrease. This phenomenon will be further investigated in Fig. 11.

Fig. 3 includes the plots for $E_{\rm c}^m$ and $\bar{E}_{\rm c}^m$ versus the transmit SNR ρ, for various values of the delay QoS exponent vector $\boldsymbol{\theta}$. This figure first shows that when ρ increases, the link-layer rate for the $m^{\rm th}$ user, either in NOMA or OMA, shows a non-decreasing trend. This confirms the proved Lemma 2.(a). For the $E_{\rm c}^m$ in NOMA scheme, it first increases when ρ is relatively small, then reaches a limit when ρ becomes very large. This observation confirms Lemma 1.(b), since we proved

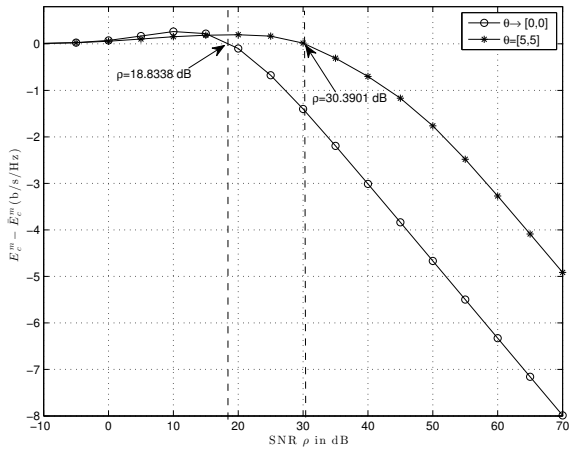

Fig. 5: $E_{\rm c}^m - \bar{E}_{\rm c}^m$ versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

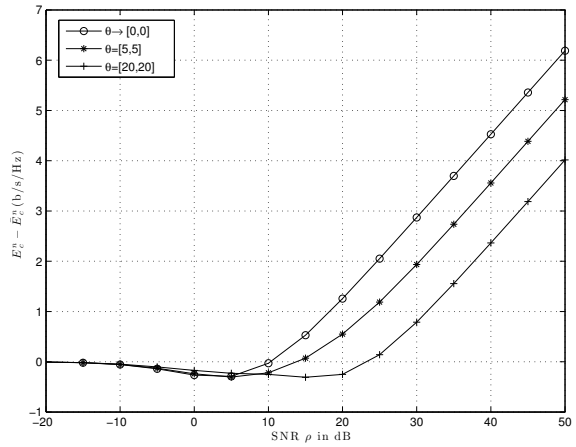

Fig. 6: $E_{\rm c}^n - \bar{E}_{\rm c}^n$ versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

that when $\rho \to \infty$, $E_{\rm c}^m$ approaches a maximum limit which is independent from the transmit SNR and the user's delay QoS requirement. Further, from Fig. 3, we note that $E_{\rm c}^m$ saturates as soon as $\rho \geq 30$dB, although in Lemma 1.(b), the maximum limit of $E_{\rm c}^m$ achieves when $\rho \to \infty$. Finally, Fig. 3 shows that $E_{\rm c}^m$ in NOMA prevails over $\bar{E}_{\rm c}^m$ in OMA, when ρ is small, but with the increase of ρ, OMA outperforms NOMA on the link-layer rate performance, for the $m^{\rm th}$ user, which confirms the analysis and explanations in Lemma 2 and Lemma 5.

Considering the $n^{\rm th}$ user, Fig. 4 plots the curves of $E_{\rm c}^n$ and $\bar{E}_{\rm c}^n$ versus the transmit SNR ρ, for various values of the delay QoS exponent vector $\boldsymbol{\theta}$. From this figure, we note that $E_{\rm c}^n$ and $\bar{E}_{\rm c}^n$ start at the same value of 0, then monotonically increase with respect to the transmit SNR ρ. This confirms Lemma 1.(a) and Lemma 3.(a). Furthermore, for a fixed value of $\boldsymbol{\theta}$, when ρ is small, $\bar{E}_{\rm c}^n$ in OMA is larger than $E_{\rm c}^n$ in NOMA, but with the increase of the transmit SNR, NOMA becomes more beneficial, in terms of the link-layer rate, which is analytically explained in Lemma 3 and Lemma 5. In addition, when the delay QoS exponent vector becomes more stringent, i.e., changing from $\boldsymbol{\theta} \to [0, 0]$ to $\boldsymbol{\theta} = [30, 30]$, the link-layer rate for the $n^{\rm th}$ user, either in NOMA or OMA, decreases, considering a fixed value of ρ.

In order to investigate the advantage of NOMA over OMA, for the $m^{\rm th}$ user and the $n^{\rm th}$ user, we provide Fig. 5 and Fig. 6, which include the plots for $E_{\rm c}^m - \bar{E}_{\rm c}^m$ and $E_{\rm c}^n - \bar{E}_{\rm c}^n$ versus the transmit SNR ρ, respectively, for various values of the delay QoS exponent vector $\boldsymbol{\theta}$. Fig. 5 indicates that for the $m^{\rm th}$ user, $E_{\rm c}^m - \bar{E}_{\rm c}^m$ starts at the initial value of 0, increases slightly at small values of ρ, and then decreases when the transmit SNR ρ further increases. This confirms Lemma 2.(b) and Lemma 2.(c). When the transmit SNR is high and fixed, Fig. 5 further shows that a more stringent delay requirement with $\boldsymbol{\theta} = [5, 5]$, results in a larger value of $E_{\rm c}^m - \bar{E}_{\rm c}^m$ than the delay-unconstrained situation with $\boldsymbol{\theta} \to [0, 0]$. Specifically, in comparison with the delay-unconstrained system, the delay-constrained system with $\boldsymbol{\theta} = [5, 5]$ allows a longer range of ρ, in which NOMA prevails over OMA. For delay-constrained system with $\boldsymbol{\theta} = [5, 5]$, $E_{\rm c}^m - \bar{E}_{\rm c}^m$ becomes negative when $\rho \geq 30.3901$dB. Meanwhile, for delay-unconstrained system, $E_{\rm c}^m - \bar{E}_{\rm c}^m$ becomes negative when $\rho \geq 18.8338$dB. This means that, after this point, OMA performs better than NOMA, for the $m^{\rm th}$ user. On the other hand, for the $n^{\rm th}$ user, Fig. 6 shows that $E_{\rm c}^n - \bar{E}_{\rm c}^n$ first starts at the initial value of 0, slightly decreases when ρ is small, and with the further increase of ρ, it increases. This confirms Lemma 3.(b) and Lemma 3.(c). Furthermore, when the transmit SNR is high and fixed, a more stringent delay requirement with $\boldsymbol{\theta} = [20, 20]$ leads to a smaller value of $E_{\rm c}^n - \bar{E}_{\rm c}^n$, than the delay-unconstrained situation with $\boldsymbol{\theta} \to [0, 0]$.

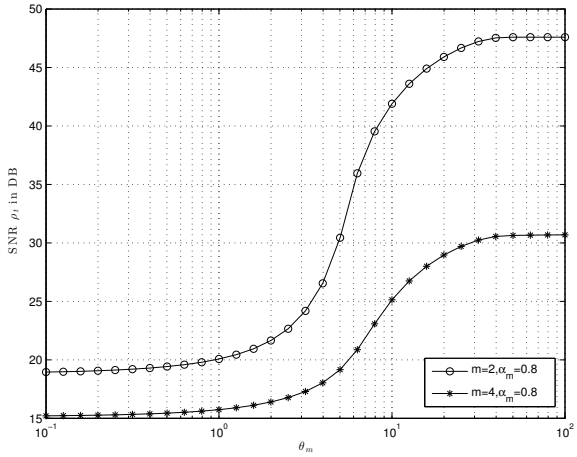

Fig. 7: SNR ρ_t versus θ_m for various values of m and power coefficient α_m.

Note that Fig. 5 shows two SNR transition points, i.e., 30.3901dB when $\boldsymbol{\theta} = [5, 5]$ and 18.8338dB when $\boldsymbol{\theta} \to [0, 0]$, after which OMA becomes better than NOMA for the $m^{\rm th}$ user. By setting the transition point as ρ_t, we include Fig. 7 to show the curves of ρ_t versus the delay QoS exponent θ_m, for various values of m. For a fixed m, ρ_t first stays stable at small values of θ_m, then gradually increases, and finally becomes stable at high θ_m values, i.e., $\theta_m \to 10^2$. This means that when the $m^{\rm th}$ user's delay requirement becomes more stringent, NOMA performs better than OMA for a longer range of SNR.

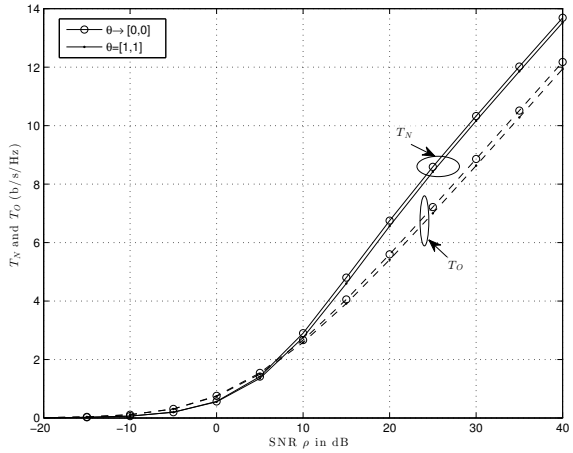

Fig. 8: T_N and T_O versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

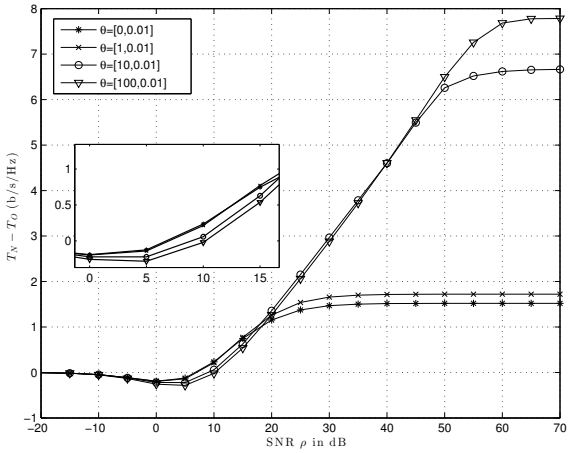

Fig. 9: $\mathrm{T}_N - \mathrm{T}_O$ versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

Meanwhile, at extreme values of θ_m, i.e., when $\theta_m \to 10^{-1}$ or $\theta_m \to 10^2$, the values of ρ_t are stable, which indicates that for very loose or very stringent delay requirements, the range of SNR in which NOMA outperforms OMA is fixed. Furthermore, from Fig. 7, one can also notice that for a fixed θ_m, the value of ρ_t obtained with $m = 4$ is smaller than the one obtained with $m = 2$. This means that when a user's channel conditions become weaker, NOMA outperforms OMA for a wider range of SNR.

To investigate the impact of ρ on the performance of the total link-layer rate for the two-user system, we provide Fig. 8 which includes the plots for T_N in NOMA and T_O in OMA, versus the transmit SNR ρ, for various values of $\boldsymbol{\theta}$. Fig. 8 first indicates that the total EC for the two-user network, either in NOMA or OMA, starts at the initial value of 0, and then gradually increases with the transmit SNR ρ. This confirms Lemma 4.(a) and Lemma 4.(d). Specifically, Fig. 8 shows that when ρ is very small, the total rate for the two-user network in OMA, T_O, has a faster increasing speed than T_N in NOMA. Then, with the increase of ρ, T_N in NOMA gradually becomes higher than T_O in OMA, for the delay-constrained situation with $\boldsymbol{\theta} = [1, 1]$ and the delay-unconstrained situation with $\boldsymbol{\theta} \to [0, 0]$. Furthermore, at high values of ρ, the gap of the total EC between NOMA and OMA, for this two-user network, becomes steady.

To further investigate and analyze the impact of the transmit SNR ρ and the delay QoS exponent vector $\boldsymbol{\theta}$ on the total EC difference, between a two-user NOMA network and a two-user OMA network, we provide Fig. 9 and Fig. 10 which include the plots for $\mathrm{T}_N - \mathrm{T}_O$ versus the transmit SNR ρ, for various settings of the delay QoS exponent vector $\boldsymbol{\theta}$. Specifically, to plot Fig. 9, the delay QoS exponent of the n^{th} user is fixed at $\theta_n = 0.01$. Meanwhile, in Fig. 10, all curves are plotted by fixing the value of θ_m at 0.01. From Fig. 9, we note that for a fixed value of $\boldsymbol{\theta}$, $\mathrm{T}_N - \mathrm{T}_O$ starts at the initial value of 0, first decreases, then increases with the transmit SNR ρ, finally reaches a maximum limit and stabilizes. Further, Fig. 9 indicates that when θ_n is fixed at 0.01, a larger θ_m leads to a higher value of $\mathrm{T}_N - \mathrm{T}_O$ at high SNRs. Correspondingly, Fig.

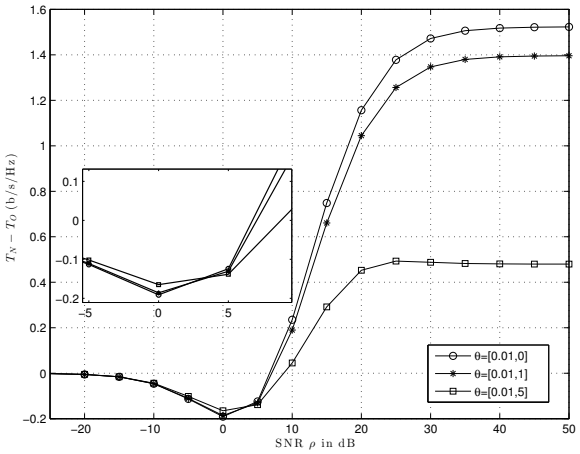

Fig. 10: $\mathrm{T}_N - \mathrm{T}_O$ versus ρ for various values of the delay QoS exponent vector $\boldsymbol{\theta}$.

10 shows that when θ_m is fixed at 0.01, a smaller θ_n results in a higher level of $\mathrm{T}_N - \mathrm{T}_O$ at high SNRs.

To investigate the impact of the delay QoS exponent θ_m on the link-layer rate performance for the m^{th} user, we plot the results of E_{c}^m in NOMA (in solid lines) and $\mathbb{E}[R_m]$ (in dash lines) versus the delay QoS exponent θ_m, for various values of ρ in Fig. 11. This figure first indicates that, when the m^{th} user has a loose delay requirement, i.e., $\theta_m \leq 10^{-1}$, the link-layer rate in NOMA, E_{c}^m, is equivalent to the physical-layer rate $\mathbb{E}[R_m]$, which confirms Lemma 5.(a). When the delay requirement becomes more stringent, E_{c}^m gradually decreases to the minimum value of 0, for various values of ρ. On the contrary, the curves of $\mathbb{E}[R_m]$ versus θ_m always stay high and stable, but this is due to the reason that there is no delay requirement guaranteed when the physical-layer rate is considered. Furthermore, considering a fixed θ_m, when ρ increases from 10 dB to 30 dB, E_{c}^m becomes larger, which indicates that a higher value of ρ will result in a larger value of EC in NOMA, for the m^{th} user.

Then, we focus on the comparison of NOMA and OMA, in terms of the difference of the total link-layer rate, between

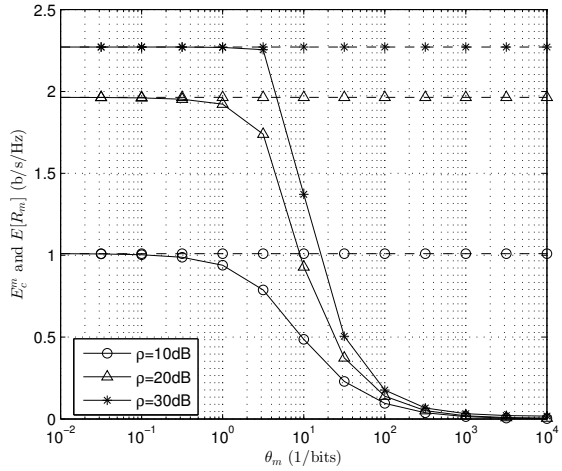

Fig. 11: $E_{\rm c}^m$, in NOMA, versus θ_m for various values of the transmit SNR ρ.

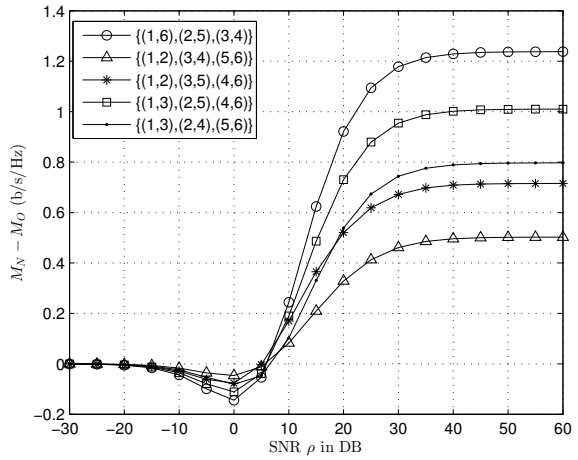

Fig. 12: $\mathrm{M}_N - \mathrm{M}_O$, versus the transmit SNR ρ for various settings of user pairing set Φ.

multiple NOMA pairs and M OMA users, i.e., $\mathrm{M}_N - \mathrm{M}_O$. To investigate the impact of the transmit SNR ρ and the user pairing set Φ on the $\mathrm{M}_N - \mathrm{M}_O$, we provide Fig. 12 which includes the plots for $\mathrm{M}_N - \mathrm{M}_O$ versus the transmit SNR ρ, for various settings of the user pairing set Φ. Specifically, the total number of users $M = 6$, the power coefficients allocated to both users in a NOMA pair are given as $\alpha_{m_i} = 0.8$, $\alpha_{n_i} = 0.2$, $\forall i \in \mathbb{I}$, $(m_i, n_i) \in \Phi$[15], and the delay QoS exponents of all users are assumed to be approaching 0. From Fig. 12, we note that for a fixed setting of Φ, $\mathrm{M}_N - \mathrm{M}_O$ starts at the initial value of 0, first decreases, then increases until it reaches a maximum value. This confirms the proposed Lemma 6 in Section IV-B, which reveals that OMA achieves higher total EC than NOMA at small values of ρ. Fig. 12 also indicates that NOMA is more beneficial than OMA, on the total EC performance for a M-user network, when the transmit SNR becomes extremely high. Furthermore, from Fig. 12, we note that at high SNRs, the user pairing setting of $\Phi = \{(1,6),(2,5),(3,4)\}$ provides the largest level of $\mathrm{M}_N - \mathrm{M}_O$, which means that among all the simulated settings, this case is the best user pairing solution.

[15]We note that different settings of power coefficients can influence the simulation results, but this is beyond the scope of this paper, and can be kept as a future research topic.

VI. Conclusions

The advantage of NOMA over OMA, on the total link-layer rate performance for a downlink NOMA network with M users, was investigated and analyzed in this paper. Specifically, by assuming that the M users are divided into multiple NOMA pairs, simulation results show that NOMA offers higher total EC than OMA at high SNR values. Furthermore, we found that the advantage of NOMA over OMA becomes stable when the transmit SNR is extremely high. This indicates that once above a high level, the increase of transmit SNR cannot guarantee any more performance gain. Focusing on a simple two-user network, we also proved that for the stronger user, either delay-constrained or delay-unconstrained, NOMA prevails over OMA, when the transmit SNR is large. On the contrary, for the weaker user in a two-user network, we proved that NOMA offers higher EC than OMA at small SNR values. To confirm these theoretical conclusions, the closed-form expressions for the individual EC in a two-user network were derived and confirmed by using Monte Carlo simulation results. Further, simulation results also reveal that the user pairing settings and the allocated power coefficients can influence the throughput performance, which can be reserved as potential research topics.

Appendix A: Proof of Theorem 1

By applying the order statistics, the EC in NOMA for the $m^{\rm th}$ user, $E_{\rm c}^m$, can be expanded as

$$E_{\rm c}^m = -\frac{1}{\theta_m T_{\rm f} B} \ln \left(\int_0^\infty \left(\frac{\gamma_m + 1}{\alpha_n \gamma_m + 1} \right)^{2\beta_m} \psi_m f(\gamma_m) \times F(\gamma_m)^{m-1} \left(1 - F(\gamma_m)\right)^{M-m} d\gamma_m \right). \quad (19)$$

By inserting $f(\gamma_m) = \frac{1}{\rho} e^{-\frac{\gamma_m}{\rho}}$, and $F(\gamma_m) = 1 - e^{-\frac{\gamma_m}{\rho}}$ into (19), we have

$$E_{\rm c}^m = -\frac{1}{\theta_m T_{\rm f} B} \ln \left(\frac{\psi_m}{\rho} \int_0^\infty \left(\frac{\gamma_m + 1}{\alpha_n \gamma_m + 1} \right)^{2\beta_m} \times e^{-\frac{(M-m+1)\gamma_m}{\rho}} \left(1 - e^{-\frac{\gamma_m}{\rho}} \right)^{m-1} d\gamma_m \right). \quad (20)$$

To obtain the closed-form expressions, we need to transform and simplify $\left(\frac{\gamma_m + 1}{\gamma_m \alpha_n + 1} \right)^{2\beta_m}$ and $\left(1 - e^{-\frac{\gamma_m}{\rho}} \right)^{m-1}$ first. According to the generalized binomial expansion, we first get that

$$\left(\frac{\gamma_m + 1}{\gamma_m \alpha_n + 1} \right)^{2\beta_m} = \left(\frac{1}{\alpha_n} \right)^{2\beta_m} \left(1 + \frac{\alpha_n - 1}{\gamma_m \alpha_n + 1} \right)^{2\beta_m}, \quad (21)$$

where $\left(1 + \frac{\alpha_n - 1}{\gamma_m \alpha_n + 1} \right)^{2\beta_m}$ can then be expanded as

$\sum_{j=0}^{\infty}\binom{2\beta_m}{j}\left(\frac{\alpha_n-1}{\gamma_m\alpha_n+1}\right)^j$, due to the fact that $(1+x)^s = \sum_{j=0}^{\infty}\binom{s}{j}x^j$, for $|x|<1$, where $\binom{s}{j}$ is defined as follows [29]:

$$\binom{s}{j} = \frac{s(s-1)\dots(s-j+1)}{j!} = \frac{(s)_j}{j!}, \quad \text{if } j \geq 1, \quad (22)$$

where $(\cdot)_j$ is the Pochhammer symbol, and $\binom{s}{0}=1$ [29].

Furthermore, we note that $\left(1-e^{-\frac{\gamma_m}{\rho}}\right)^{m-1}$ in (20) can be replaced with the summation $\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k e^{-\frac{\gamma_m}{\rho}k}$, by applying the binomial expansion [29]. Hence, by replacing $\left(1+\frac{\alpha_n-1}{\gamma_m\alpha_n+1}\right)^{2\beta_m}$ and $\left(1-e^{-\frac{\gamma_m}{\rho}}\right)^{m-1}$, (20) can be transformed into

$$E_{\mathrm{c}}^m = -\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\frac{(\alpha_n)^{-2\beta_m}\psi_m}{\rho}\int_0^{\infty}\left(\underbrace{1}_{\text{when j=0}} + \underbrace{(2\beta_m)\frac{\alpha_n-1}{\gamma_m\alpha_n+1}}_{\text{when j =1}} + \underbrace{\sum_{j=2}^{\infty}\binom{2\beta_m}{j}\left(\frac{\alpha_n-1}{\gamma_m\alpha_n+1}\right)^j}_{\text{when j} \geq 2}\right) \times \left(\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k e^{-\frac{(M-m+1+k)\gamma_m}{\rho}}\right)d\gamma_m\right).$$

Then we can use the following equations from [30], namely, (3.353.2) and (3.352.4).

$$\int_0^{\infty}\frac{e^{-\mu x}}{(x+\beta)^n}dx = \frac{1}{(n-1)!}\sum_{k=1}^{n-1}(k-1)!(-\mu)^{n-k-1}\beta^{-k} - \frac{(-\mu)^{n-1}}{(n-1)!}e^{\beta\mu}E_i(-\beta\mu), [n\geq 2,\ |\arg\beta|<\pi,\ \mathrm{Re}\ \mu>0], \quad (23a)$$

$$\int_0^{\infty}\frac{e^{-\mu x}}{x+\beta}dx = -e^{\beta\mu}E_i(-\beta\mu), [|\arg\beta|<\pi,\ \mathrm{Re}\ \mu>0], \quad (23b)$$

where $E_i(\cdot)$ is the exponential integral. Finally, by applying (23a) and (23b), we can obtain the closed-form expression for E_{c}^m, given in (13a).

Now, let us consider the closed-form expression for the EC in OMA scheme for the m^{th} user. By applying the order statistics, $\bar{E}_{\mathrm{c}}^m$ can be expanded as

$$\bar{E}_{\mathrm{c}}^m = -\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\frac{\psi_m}{\rho}\int_0^{\infty}(1+\gamma_m)^{\beta_m}e^{-\frac{(M-m+1)\gamma_m}{\rho}} \times \left(1-e^{-\frac{\gamma_m}{\rho}}\right)^{m-1}d\gamma_m\right). \quad (24)$$

After applying the binomial expansion for $\left(1-e^{-\frac{\gamma_m}{\rho}}\right)^{m-1}$, we have

$$\bar{E}_{\mathrm{c}}^m = -\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\frac{\psi_m}{\rho}\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k \times\int_0^{\infty}(1+\gamma_m)^{\beta_m}e^{-\frac{(M-m+1+k)\gamma_m}{\rho}}d\gamma_m\right). \quad (25)$$

From (13.2.5) in [29], we note that

$$U(a,b,z) = \frac{1}{\Gamma(a)}\int_0^{\infty}e^{-zt}t^{a-1}(1+t)^{b-a-1}dt, \quad \text{for Re } a,\ \text{Re } z>0, \quad (26)$$

where $U(\cdot)$ is the confluent hypergeometric function of the second kind [29]. By applying (26) to (25), $\bar{E}_{\mathrm{c}}^m$ can be finally expressed as

$$\bar{E}_{\mathrm{c}}^m = -\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\frac{\psi_m}{\rho}\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k \times U\left(1,2+\beta_m,\frac{M-m+1+k}{\rho}\right)\right). \quad (27)$$

Appendix B : Proof of Lemma 1

By inserting $\rho\to 0$ into (10a), (11a), (10b), and (11b), we can prove that $E_{\mathrm{c}}^m - \bar{E}_{\mathrm{c}}^m \to 0$, and $E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n \to 0$. When $\rho\to\infty$, E_{c}^m can be expressed as

$$\lim_{\rho\to\infty} -\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\mathbb{E}\left[\left(\frac{|h_m|^2+\frac{1}{\rho}}{\alpha_n|h_m|^2+\frac{1}{\rho}}\right)^{2\beta_m}\right]\right) = \log_2\left(\frac{1}{\alpha_n}\right).$$

For finite value of θ_m, it can be proved that $\lim_{\rho\to\infty}\bar{E}_{\mathrm{c}}^m \to \infty$, and $\lim_{\rho\to\infty}\left(E_{\mathrm{c}}^m - \bar{E}_{\mathrm{c}}^m\right) \to -\infty$.

As for the n^{th} user, $\lim_{\rho\to\infty}E_{\mathrm{c}}^n \to \infty$, and $\lim_{\rho\to\infty}\bar{E}_{\mathrm{c}}^n \to \infty$ can be easily proved, which are omitted here. To analyze the EC difference of the NOMA and OMA scheme for the n^{th} user when $\rho\to\infty$, we have that

$$\lim_{\rho\to\infty}\left(E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n\right) = \lim_{\rho\to\infty} -\frac{1}{\theta_n T_{\mathrm{f}} B}\ln\left(\frac{\rho^{\beta_n}\mathbb{E}\left[\left(\alpha_n|h_n|^2\right)^{2\beta_n}\right]}{\mathbb{E}\left[\left(|h_n|^2\right)^{\beta_n}\right]}\right),$$

which approaches infinity. This completes the proof that $\lim_{\rho\to\infty}\left(E_{\mathrm{c}}^n - \bar{E}_{\mathrm{c}}^n\right) \to \infty$.

Appendix C: Proof of Lemma 2

To analyze the trends of E_{c}^m and $\bar{E}_{\mathrm{c}}^m$ with respect to ρ, we have

$$\frac{\partial E_{\mathrm{c}}^m}{\partial\rho} = -\frac{1}{\theta_m T_{\mathrm{f}} B}\frac{\left(\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]\right)'}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]}$$

$$= \frac{1-\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m-1} \frac{|h_m|^2}{(\rho\alpha_n|h_m|^2+1)^2}\right]}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]}, \tag{28}$$

where $()'$ is the first derivative with respect to ρ. Apparently, (28) is non-negative. Similarly, for EC in OMA for the m^{th} user, we get

$$\frac{\partial \bar{E}_{\text{c}}^m}{\partial \rho} = \frac{1}{2\ln 2} \frac{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m-1}|h_m|^2\right]}{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m}\right]}, \tag{29}$$

which is non-negative too.

We then start to analyze the trend of $E_{\text{c}}^m - \bar{E}_{\text{c}}^m$ with respect to ρ, as follows.

$$\frac{\partial\left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right)}{\partial \rho} = \frac{\partial E_{\text{c}}^m}{\partial \rho} - \frac{\partial \bar{E}_{\text{c}}^m}{\partial \rho} \tag{30a}$$

$$= \frac{1-\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m-1} \frac{|h_m|^2}{(\rho\alpha_n|h_m|^2+1)^2}\right]}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]} - \frac{1}{2\ln 2} \frac{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m-1}|h_m|^2\right]}{\mathbb{E}\left[\left(1+\rho|h_m|^2\right)^{\beta_m}\right]}. \tag{30b}$$

When $\rho \to 0$, we prove that $\lim_{\rho\to 0} \frac{\partial\left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right)}{\partial \rho} = \left(\frac{1-2\alpha_n}{2\ln 2}\right)\mathbb{E}\left[|h_m|^2\right] \geq 0$, due to the reason that $\alpha_n \in \left(0, \frac{1}{2}\right]$ and $\mathbb{E}\left[|h_m|^2\right] \geq 0$.

When ρ is very large, we can prove that

$$\frac{\partial\left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right)}{\partial \rho} = \frac{\frac{1-\alpha_n}{\alpha_n \ln 2}\mathbb{E}\left[\frac{1}{|h_m|^2}\right] - \frac{1}{2\ln 2}\rho}{\rho^2}. \tag{31a}$$

Since $\mathbb{E}\left[\frac{1}{|h_m|^2}\right]$ is a finite value, unrelated to ρ, therefore when ρ is very large, (31a) can be approximated by $-\frac{1}{2\rho\ln 2}$, which is smaller than 0 and gradually approaches 0 when $\rho \to \infty$. Furthermore, the critical point of the function $E_{\text{c}}^m - \bar{E}_{\text{c}}^m$, i.e., the value of ρ which makes $\frac{\partial\left(E_{\text{c}}^m - \bar{E}_{\text{c}}^m\right)}{\partial \rho} = 0$, can be obtained when (30b) equals to zero.

Appendix D: Proof of Lemma 3

Here, we analyze the trends of E_{c}^n and $\bar{E}_{\text{c}}^n$ versus ρ.

$$\frac{\partial E_{\text{c}}^n}{\partial \rho} = -\frac{1}{\theta_n T_{\text{f}} B} \frac{\left(\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]\right)'}{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]} = \frac{\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]}, \tag{32}$$

which is non-negative. As for the EC in OMA, we can also prove that $\frac{\partial \bar{E}_{\text{c}}^n}{\partial \rho} \geq 0$, which is omitted here due to the page limit. To analyze the trend of $E_{\text{c}}^n - \bar{E}_{\text{c}}^n$ versus ρ, we have that

$$\frac{\partial\left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right)}{\partial \rho} = \frac{\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]} - \frac{1}{2\ln 2} \frac{\mathbb{E}\left[\left(1+\rho|h_n|^2\right)^{\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(1+\rho|h_n|^2\right)^{\beta_n}\right]}. \tag{33}$$

When $\rho \to 0$, we prove that $\lim_{\rho\to 0} \frac{\partial\left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right)}{\partial \rho} = \frac{\alpha_n - \frac{1}{2}}{\ln 2}\mathbb{E}\left[|h_n|^2\right] \leq 0$, due to the fact that $\alpha_n \in \left(0, \frac{1}{2}\right]$, and $\mathbb{E}\left[|h_n|^2\right] \geq 0$.

When ρ is very large, we can prove that

$$\frac{\partial\left(E_{\text{c}}^n - \bar{E}_{\text{c}}^n\right)}{\partial \rho} = \frac{\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(\rho\alpha_n|h_n|^2\right)^{2\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]} - \frac{1}{2\ln 2} \frac{\mathbb{E}\left[\left(\rho|h_n|^2\right)^{\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(\rho|h_n|^2\right)^{\beta_n}\right]} = \frac{1}{2\rho\ln 2}, \tag{34}$$

which is non-negative and approaches 0, when $\rho \to \infty$.

Appendix E: Proof of Lemma 4

From Lemma 1, we note that when $\rho \to 0$, $\mathrm{T}_N = E_{\text{c}}^m + E_{\text{c}}^n \to 0$, and $\lim_{\rho\to\infty} \mathrm{T}_N \to \infty$. For the sum EC in OMA scheme, T_O, we can also get that $\mathrm{T}_O \to 0$ when $\rho \to 0$, and $\lim_{\rho\to\infty} \mathrm{T}_O \to \infty$. In addition, for the sum EC in NOMA scheme, T_N, we can prove that

$$\frac{\partial \mathrm{T}_N}{\partial \rho} = \frac{\partial\left(E_{\text{c}}^m + E_{\text{c}}^n\right)}{\partial \rho} \tag{35a}$$

$$= \frac{1-\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m-1} \frac{|h_m|^2}{(\rho\alpha_n|h_m|^2+1)^2}\right]}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]} + \frac{\alpha_n}{\ln 2} \frac{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n-1}|h_n|^2\right]}{\mathbb{E}\left[\left(1+\rho\alpha_n|h_n|^2\right)^{2\beta_n}\right]}, \tag{35b}$$

which is non-negative because $\frac{\partial E_{\text{c}}^m}{\partial \rho} \geq 0$, and $\frac{\partial E_{\text{c}}^n}{\partial \rho} \geq 0$. When $\rho \to 0$, we have that

$$\lim_{\rho\to 0} \frac{\partial \mathrm{T}_N}{\partial \rho} = \frac{1-\alpha_n}{\ln 2}\mathbb{E}\left[|h_m|^2\right] + \frac{\alpha_n}{\ln 2}\mathbb{E}\left[|h_n|^2\right]. \tag{36}$$

When $\rho \to \infty$, we can prove that

$$\lim_{\rho\to\infty} \frac{\partial \mathrm{T}_N}{\partial \rho} = \lim_{\rho\to\infty} \frac{1-\alpha_n}{\alpha_n \ln 2\rho^2}\mathbb{E}\left[\frac{1}{|h_m|^2}\right] + \frac{1}{\rho\ln 2}, \tag{37}$$

which equals to 0.

By following similar steps, we can also prove that $\frac{\partial \mathrm{T}_O}{\partial \rho} \geq 0$, $\lim_{\rho\to 0}\frac{\partial \mathrm{T}_O}{\partial \rho} = \frac{1}{2\ln 2}\mathbb{E}\left[|h_m|^2\right] + \frac{1}{2\ln 2}\mathbb{E}\left[|h_n|^2\right]$, and $\lim_{\rho\to\infty}\frac{\partial \mathrm{T}_O}{\partial \rho} = 0$. This completes the proof for Lemma 4.

Appendix F: Proof of Lemma 5

Recall that the EC expression in NOMA scheme, for the m^{th} user, is given by

$$E_{\mathrm{c}}^{m} = -\frac{1}{\theta_m T_{\mathrm{f}} B}\ln\left(\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]\right), \quad (38)$$

which gives an indeterminate form $\frac{0}{0}$, when $\theta_m \to 0$.

By applying L'Hopital's rule, $\lim_{\theta_m\to 0} E_{\mathrm{c}}^{m}$ becomes

$$\lim_{\theta_m\to 0} -\frac{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\ln\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)\left(-\frac{1}{\ln 2}\right)\right]}{\mathbb{E}\left[\left(\frac{\rho|h_m|^2+1}{\rho\alpha_n|h_m|^2+1}\right)^{2\beta_m}\right]}$$
$$= \mathbb{E}\left[\log_2\left(1+\frac{\alpha_m|h_m|^2}{\alpha_n|h_m|^2+\frac{1}{\rho}}\right)\right], \quad (39)$$

which equals to $\mathbb{E}\left[R_m\right]$. In other words, when $\theta_m \to 0$, which refers to a user with no delay constraint, the EC in NOMA is equivalent to the ergodic capacity. Similarly, by using L'Hopital's rule, we can also conclude that $\lim_{\theta_m\to 0}\bar{E}_{\mathrm{c}}^{m} = \frac{1}{2}\mathbb{E}\left[\log_2\left(1+\rho|h_m|^2\right)\right]$, which equals to $\mathbb{E}\left[\bar{R}_m\right]$. Hence, when $\theta_m \to 0$, $E_{\mathrm{c}}^{m} - \bar{E}_{\mathrm{c}}^{m} = \mathbb{E}\left[R_m\right] - \mathbb{E}\left[\bar{R}_m\right]$. By following similar steps, we can get the same conclusion for the n^{th} user, i.e., $\lim_{\theta_n\to 0} E_{\mathrm{c}}^{n} = \mathbb{E}\left[R_n\right]$, $\lim_{\theta_n\to 0}\bar{E}_{\mathrm{c}}^{n} = \mathbb{E}\left[\bar{R}_n\right]$, and $\lim_{\theta_n\to 0}\left(E_{\mathrm{c}}^{n} - \bar{E}_{\mathrm{c}}^{n}\right) = \mathbb{E}\left[R_n\right] - \mathbb{E}\left[\bar{R}_n\right]$.

Consider the m^{th} user with no delay constraint, i.e., $\theta_m \to 0$. By inserting $\rho \to \infty$ to (39), we can prove that $\lim_{\substack{\theta_m\to 0\\ \rho\to\infty}} E_{\mathrm{c}}^{m} = \mathbb{E}\left[\log_2\left(\frac{1}{\alpha_n}\right)\right]$. As for the EC in OMA for the m^{th} user, we can get that $\lim_{\substack{\theta_m\to 0\\ \rho\to\infty}}\bar{E}_{\mathrm{c}}^{m} \to \infty$, by inserting $\rho \to \infty$ into $\frac{1}{2}\mathbb{E}\left[\log_2\left(1+\rho|h_m|^2\right)\right]$. Henceforth, we can prove that $\lim_{\substack{\theta_m\to 0\\ \rho\to\infty}}\left(E_{\mathrm{c}}^{m} - \bar{E}_{\mathrm{c}}^{m}\right) \to -\infty$.

Similarly, for the n^{th} user with $\theta_n \to 0$, when the transmit SNR ρ is very large, we can prove that $\lim_{\substack{\theta_n\to 0\\ \rho\to\infty}} E_{\mathrm{c}}^{n} \to \infty$, and $\lim_{\substack{\theta_n\to 0\\ \rho\to\infty}}\bar{E}_{\mathrm{c}}^{n} \to \infty$. As for $\lim_{\substack{\theta_n\to 0\\ \rho\to\infty}}\left(E_{\mathrm{c}}^{n} - \bar{E}_{\mathrm{c}}^{n}\right)$, we have that

$$\lim_{\substack{\theta_n\to 0\\ \rho\to\infty}}\left(E_{\mathrm{c}}^{n} - \bar{E}_{\mathrm{c}}^{n}\right)$$
$$= \lim_{\rho\to\infty}\mathbb{E}\left[\log_2\left(1+\rho\alpha_n|h_n|^2\right)\right] - \frac{1}{2}\mathbb{E}\left[\log_2\left(1+\rho|h_n|^2\right)\right]$$
$$= \lim_{\rho\to\infty}\mathbb{E}\left[\log_2\left(\frac{\frac{1}{\sqrt{\rho}}+\sqrt{\rho}\alpha_n|h_n|^2}{\sqrt{\frac{1}{\rho}+|h_n|^2}}\right)\right]$$
$$= \lim_{\rho\to\infty}\mathbb{E}\left[\log_2\left(\sqrt{\rho}\alpha_n\sqrt{|h_n|^2}\right)\right], \quad (40a)$$

which approaches infinity. Here we complete the proof for Lemma 5.

Appendix G: Proof of Lemma 6

By inserting $\rho \to 0$ into (16a), (17a), (16b), and (17b), we can prove that $E_{\mathrm{c}}^{m_i} - \bar{E}_{\mathrm{c}}^{m_i} \to 0$, and $E_{\mathrm{c}}^{n_i} - \bar{E}_{\mathrm{c}}^{n_i} \to 0$. Therefore, one can easily get that when $\rho \to 0$, $\mathrm{M}_N - \mathrm{M}_O \to 0$, since $\mathrm{M}_N - \mathrm{M}_O = \sum_{i=1}^{M/2}\left(E_{\mathrm{c}}^{m_i} - \bar{E}_{\mathrm{c}}^{m_i} + E_{\mathrm{c}}^{n_i} - \bar{E}_{\mathrm{c}}^{n_i}\right)$. When $\rho \to \infty$, we can prove that

$$\lim_{\rho\to\infty}\left(\mathrm{M}_N - \mathrm{M}_O\right) = \lim_{\rho\to\infty}\sum_{i=1}^{M/2}\left(E_{\mathrm{c}}^{m_i} - \bar{E}_{\mathrm{c}}^{m_i} + E_{\mathrm{c}}^{n_i} - \bar{E}_{\mathrm{c}}^{n_i}\right)$$
$$= \sum_{i=1}^{M/2}\lim_{\rho\to\infty} -\frac{1}{\theta_{m_i}T_{\mathrm{f}}B}\ln\left(\frac{\alpha_{n_i}^{-\frac{4}{M}\beta_{m_i}}}{\mathbb{E}\left[\left(|h_{m_i}|^2\right)^{\frac{2}{M}\beta_{m_i}}\right]}\rho^{-\frac{2}{M}\beta_{m_i}}\right)$$
$$- \frac{1}{\theta_{n_i}T_{\mathrm{f}}B}\ln\left(\frac{\alpha_{n_i}^{\frac{4}{M}\beta_{n_i}}\mathbb{E}\left[\left(|h_{n_i}|^2\right)^{\frac{4}{M}\beta_{n_i}}\right]}{\mathbb{E}\left[\left(|h_{n_i}|^2\right)^{\frac{2}{M}\beta_{n_i}}\right]}\rho^{\frac{2}{M}\beta_{n_i}}\right) \quad (41a)$$
$$= \sum_{i=1}^{M/2} -\frac{1}{\theta_{m_i}T_{\mathrm{f}}B}\ln\left(\frac{\alpha_{n_i}^{-\frac{4}{M}\beta_{m_i}}}{\mathbb{E}\left[\left(|h_{m_i}|^2\right)^{\frac{2}{M}\beta_{m_i}}\right]}\right)$$
$$- \frac{1}{\theta_{n_i}T_{\mathrm{f}}B}\ln\left(\frac{{\alpha_{n_i}}^{\frac{4}{M}\beta_{n_i}}\mathbb{E}\left[\left(|h_{n_i}|^2\right)^{\frac{4}{M}\beta_{n_i}}\right]}{\mathbb{E}\left[\left(|h_{n_i}|^2\right)^{\frac{2}{M}\beta_{n_i}}\right]}\right), \quad (41b)$$

which is a constant with respect to ρ.

Then, we start to consider $\lim_{\rho\to 0}\frac{\partial\left(\mathrm{M}_N - \mathrm{M}_O\right)}{\partial\rho}$ and $\lim_{\rho\to\infty}\frac{\partial\left(\mathrm{M}_N - \mathrm{M}_O\right)}{\partial\rho}$, by analyzing $\frac{\partial \mathrm{M}_N}{\partial\rho}$ and $\frac{\partial \mathrm{M}_O}{\partial\rho}$ separately.

$$\frac{\partial \mathrm{M}_N}{\partial\rho} = \sum_{i=1}^{M/2}\frac{\partial E_{\mathrm{c}}^{m_i}}{\partial\rho} + \frac{\partial E_{\mathrm{c}}^{n_i}}{\partial\rho} \quad (42a)$$
$$= \sum_{i=1}^{M/2}\frac{2(1-\alpha_{n_i})}{M\ln 2}\frac{\mathbb{E}\left[\left(\frac{\rho|h_{m_i}|^2+1}{\rho\alpha_{n_i}|h_{m_i}|^2+1}\right)^{\frac{4}{M}\beta_{m_i}-1}\frac{|h_{m_i}|^2}{\left(\rho\alpha_{n_i}|h_{m_i}|^2+1\right)^2}\right]}{\mathbb{E}\left[\left(\frac{\rho|h_{m_i}|^2+1}{\rho\alpha_{n_i}|h_{m_i}|^2+1}\right)^{\frac{4}{M}\beta_{m_i}}\right]}$$
$$+ \frac{2\alpha_{n_i}}{M\ln 2}\frac{\mathbb{E}\left[\left(1+\rho\alpha_{n_i}|h_{n_i}|^2\right)^{\frac{4}{M}\beta_{n_i}-1}|h_{n_i}|^2\right]}{\mathbb{E}\left[\left(1+\rho\alpha_{n_i}|h_{n_i}|^2\right)^{\frac{4}{M}\beta_{n_i}}\right]}. \quad (42b)$$

By inserting $\rho = 0$ into (42b), we get that $\lim_{\rho\to 0}\frac{\partial \mathrm{M}_N}{\partial\rho} = \sum_{i=1}^{M/2}\frac{2\left(1-\alpha_{n_i}\right)}{M\ln 2}\mathbb{E}\left[|h_{m_i}|^2\right] + \frac{2\alpha_{n_i}}{M\ln 2}\mathbb{E}\left[|h_{n_i}|^2\right]$. On the other

hand, when $\rho \to \infty$, we can prove that $\lim_{\rho\to\infty} \frac{\partial \mathrm{M}_N}{\partial \rho}$ becomes

$$\lim_{\rho\to\infty} \sum_{i=1}^{M/2} \frac{2\left(1-\alpha_{n_i}\right)}{\alpha_{n_i} M \ln 2 \rho^2} \mathbb{E}\left[\frac{1}{|h_{m_i}|^2}\right] + \frac{2}{M \ln 2 \rho}, \tag{43}$$

which equals to 0. Then we start to consider $\frac{\partial \mathrm{M}_O}{\partial \rho}$.

$$\frac{\partial \mathrm{M}_O}{\partial \rho} = \sum_{i=1}^{M/2} \frac{\partial \bar{E}_{\mathrm{c}}^{m_i}}{\partial \rho} + \frac{\partial \bar{E}_{\mathrm{c}}^{n_i}}{\partial \rho} \tag{44a}$$

$$= \sum_{i=1}^{M/2} \frac{1}{M \ln 2} \frac{\mathbb{E}\left[\left(1+\rho|h_{m_i}|^2\right)^{\frac{2}{M}\beta_{m_i}-1}|h_{m_i}|^2\right]}{\mathbb{E}\left[\left(1+\rho|h_{m_i}|^2\right)^{\frac{2}{M}\beta_{m_i}}\right]} + \frac{1}{M \ln 2} \frac{\mathbb{E}\left[\left(1+\rho|h_{n_i}|^2\right)^{\frac{2}{M}\beta_{n_i}-1}|h_{n_i}|^2\right]}{\mathbb{E}\left[\left(1+\rho|h_{n_i}|^2\right)^{\frac{2}{M}\beta_{n_i}}\right]}. \tag{44b}$$

By inserting $\rho = 0$ into (44b), we get that $\lim_{\rho\to 0} \frac{\partial \mathrm{M}_O}{\partial \rho} = \sum_{i=1}^{M/2} \frac{1}{M \ln 2} \mathbb{E}\left[|h_{m_i}|^2\right] + \frac{1}{M \ln 2} \mathbb{E}\left[|h_{n_i}|^2\right]$. When $\rho \to \infty$, we can prove that

$$\lim_{\rho\to\infty} \frac{\partial \mathrm{M}_O}{\partial \rho} = \lim_{\rho\to\infty} \sum_{i=1}^{M/2} \frac{1}{M \ln 2 \rho} + \frac{1}{M \ln 2 \rho}, \tag{45}$$

which equals to 0 apparently. Hence, we can conclude that $\lim_{\rho\to 0} \frac{\partial\left(\mathrm{M}_N - \mathrm{M}_O\right)}{\partial \rho}$ equals to

$$\sum_{i=1}^{M/2} \frac{1-2\alpha_{n_i}}{M \ln 2} \left(\mathbb{E}\left[|h_{m_i}|^2\right] - \mathbb{E}\left[|h_{n_i}|^2\right]\right), \tag{46}$$

which is ≤ 0, because $\alpha_{n_i} - \frac{1}{2} \leq 0$, and $\mathbb{E}\left[|h_{n_i}|^2\right] \geq \mathbb{E}\left[|h_{m_i}|^2\right]$. This is due to the reason that in the i^{th} NOMA pair, the instantaneous channel power gains $|h_{n_i}|^2$ is larger than $|h_{m_i}|^2$. On the other hand, when $\rho \to \infty$, we can get that $\lim_{\rho\to\infty} \frac{\partial\left(\mathrm{M}_N - \mathrm{M}_O\right)}{\partial \rho} = \lim_{\rho\to\infty} \frac{\partial \mathrm{M}_N}{\partial \rho} - \frac{\partial \mathrm{M}_O}{\partial \rho} = 0$.

REFERENCES

[1] GSMA Intelligence. (2014, Dec.) Understanding 5G: Perspectives on future technological advancements in mobile. [Online]. Available: https://gsmaintelligence.com/research/2014/12/understanding-5g/451/

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

[3] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Annu. Symp. Personal, Indoor and Mobile Radio Commun. (PIMRC)*, London, UK, Sep. 2013, pp. 611–615.

[4] Z. Ma, Z. Zhang, Z. Ding, P. Fan, and H. Li, "Key techniques for 5G wireless communications: Network architecture, physical layer, and MAC layer perspectives," *Science China Information Sciences*, vol. 58, no. 4, pp. 1–20, Feb. 2015.

[5] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, no. 99, Oct. 2016.

[6] L. Dai, B. Wang, Y. Yuan, S. Han, C. 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, Sep. 2015.

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

[8] 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.

[9] 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.

[10] L. Lv, J. Chen, Q. Ni, and Z. Ding, "Design of cooperative non-orthogonal multicast cognitive multiple access for 5G systems: User scheduling and performance analysis," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641–2656, Jun. 2017.

[11] F. Fang, H. Zhang, J. Cheng, and V. C. M. Leung, "Energy efficiency of resource scheduling for non-orthogonal multiple access wireless network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5.

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

[13] X. Chen, Z. Zhang, C. Zhong, and D. W. K. Ng, "Exploiting multiple-antenna techniques for non-orthogonal multiple access," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2207–2220, Oct. 2017.

[14] Q. Yang, H. Wang, D. W. K. Ng, and M. H. Lee, "NOMA in downlink SDMA with limited feedback: Performance analysis and optimization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2281–2294, Oct. 2017.

[15] Y. Zhang, H. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.

[16] 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.

[17] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality-of-service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[18] W. Yu, L. Musavian, and Q. Ni, "Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3339–3353, Jan. 2016.

[19] M. Ozmen and M. C. Gursoy, "Wireless throughput and energy efficiency with random arrivals and statistical queuing constraints," *IEEE Trans. Inf. Theory*, vol. 62, no. 3, pp. 1375 – 1395, Mar. 2016.

[20] L. Musavian and T. Le-Ngoc, "Energy-efficient power allocation over nakagmi-m fading channels under delay-outage constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4081–4091, Aug. 2014.

[21] W. Yu, L. Musavian, and Q. Ni, "Statistical delay qos driven energy efficiency and effective capacity tradeoff for uplink multi-user multi-carrier systems," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3494–3508, Aug. 2017.

[22] M. Ozmen and M. C. Gursoy, "Secure transmission of delay-sensitive data over wireless fading channels," *IEEE Trans. Inf. Forensics Security*, no. 99, Apr. 2017.

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

[24] C. S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Autom. Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[25] ——, *Performance Guarantees in Communication Networks*. Springer-Verlag London, 2000.

[26] J. A. Bucklew, *Introduction to Rare Event Simulation*. Springer-Verlag New York Inc., 2004.

[27] 3GPP TD RP-150496, "Study on downlink multiuser superposition transmission," Tech. Rep.

[28] H. A. David and H. N. Nagaraja, *Order Statistics*. John Wiley, New York, 3rd ed., 2003.

[29] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions*. New York: Dover, 1965.

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

Wenjuan Yu received the B.Sc. degree in Electronic and Information Engineering, from Shandong University of Technology, China, in 2010, the M.Sc. degree from School of Telecommunications Engineering, Xidian University, Xi'an, China, in 2013, and the Ph.D. degree from School of Computing and Communications, Lancaster University, UK, in 2018. Her research interests include wireless resource allocation, cross-layer optimization toward green communications, delay QoS provisioning.

Leila Musavian (S'05-M'07) received the Ph.D. degree in telecommunications from the Kings College London, U.K., in 2006. She was a Lecturer with the InfoLab21, Lancaster University from 2012 to 2016. She was a Research Associate with McGill University from 2011 to 2012 and a Post-Doctoral Fellow with INRS-EMT, Canada, from 2006 to 2008. She is currently a Reader with the School of Computer Science and Electronic Engineering, University of Essex. Her research interests lie in radio resource management for next generation wireless networks, CRNs, energy harvesting, green communication, energy-efficient transmission techniques, cross-layer design for delay QoS provisioning, and 5G systems. She is an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, Executive Editor of the Transactions on Emerging Telecommunications Technologies and Associate Editor of Internet Technology Letters.

Qiang Ni (M'04-SM'08) is a Professor and the Head of Communication Systems Group at School of Computing and Communications, Lancaster University, Lancaster, U.K. He received the B.Sc., M.Sc., and Ph.D. degrees from Huazhong University of Science and Technology, China, all in engineering. His main research interests lie in the area of future generation communications and networking, including Green Communications and Networking, Cognitive Radio Network Systems, Heterogeneous Networks, 5G, SDN, Energy Harvesting, Wireless Information and Power Transfer, IoTs and Vehicular Networks in which areas he had already published 200+ papers.