Low-Latency Driven Performance Analysis for Single-Cluster NOMA Networks

Zhengyu Song[1],Wenjuan Yu[2], Lixia Xiao[3], Leila Musavian[4], Qiang Ni[2], and Xin Sun[1]
[1]School of Electronic and Information Engineering, Beijing Jiaotong University, China
[2]School of Computing and Communications, InfoLab21, Lancaster University, UK
[3] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, China
[4] School of Computer Science and Electronic Engineering, University of Essex, UK
Emails: {songzy, xsun}@bjtu.edu.cn, {w.yu8, q.ni}@lancaster.ac.uk, lixiaxiao@hust.edu.cn, leila.musavian@essex.ac.uk

***Abstract*—In this paper, we study the total effective capacity (EC) of single-cluster non-orthogonal multiple access (NOMA) networks and demonstrate the performance gain of single-cluster NOMA over user-paired NOMA and orthogonal multiple access (OMA). Specifically, the exact closed-form expression and an approximate closed-form expression at high signal-to-noise ratios (SNRs), in terms of the total EC, are derived for single-cluster NOMA networks. The derivations reveal that the total EC at high SNRs only relies on the statistical delay requirement of the strongest user and is independent of the other users' delay requirements. Further, we theoretically analyze the total EC differences between single-cluster NOMA and user-paired NOMA/OMA communications and explore the impact of transmit SNR. Simulation results verify the accuracy of analytical results and further reveal that the single-cluster NOMA network achieves a greater gain in terms of the total EC, compared to the conventional OMA, when the number of users increases.**

***Index Terms*—Low latency, effective capacity, NOMA, OMA.**

I. Introduction

Non-orthogonal multiple access (NOMA) technology can enhance connectivity, boost spectral efficiency and realize low transmission latency, which has been viewed as one of the most promising techniques for the next generation of cellular networks [1], [2]. Meanwhile, driven by the emerging verticals in beyond 5G era requiring diversified Quality of Service (QoS) requirements, it is of crucial importance for future networks to provide flexible scheduling of resources to support differentiated services such as massive connectivity and low-latency transmissions. It is shown that NOMA can be a potential solution [3], [4]. In this paper, we focus on the single-class traffic in terms of delay QoS and aim to explore the performance of NOMA in supporting delay-sensitive services.

There have been various investigations on multi-user NOMA systems [5]–[7]. For example, the authors in [5] analyze the outage probability and outage capacity for a downlink NOMA network, while it only reflects the ability of satisfying the target data rate instead of the delay requirements of users. In [6], the joint optimization of user clustering and power allocation is considered to maximize the sum rate of NOMA systems. Moreover, the effect of imperfect successive interference cancellation (SIC) in the downlink transmission of a NOMA network supporting massive access is studied in [7], where the transmit beams and powers are jointly optimized. However, the above-mentioned studies on multi-user NOMA networks do not consider the delay requirements of users.

Considering delay-sensitive applications, the author in [8] investigates the power control problem in a downlink NOMA network by maximizing the sum effective capacity (EC). Similarly, in [9], the delay constraint is characterized by the delay QoS exponent, where the power allocation problem is studied based on the max-min EC criterion. However, [8] and [9] focus on designing the power control policy, and the analysis in terms of delay-constrained achievable rate is not involved. Analytical expressions of EC for downlink NOMA transmissions are provided in [10] and [11], and the EC for uplink NOMA is considered in [12], but all studies consider *user-paired NOMA*, where all N users are separated into $\frac{N}{2}$ clusters each with two users. To the best of our knowledge, the performance of the EC for *single-cluster NOMA*, where all users are in one cluster and share the same resource via NOMA, has not been explored in the existing literature.

Motivated by such observations, considering users' delay provisioning, we focus on the EC analysis for the networks of single-cluster NOMA, and try to provide the quantitative performance comparison in terms of the total EC among single-cluster NOMA, user-paired NOMA and orthogonal multiple access (OMA). Specifically, we derive the exact closed-form expression for the total EC of single-cluster NOMA and its approximate closed-form at high signal-to-noise ratios (SNRs). It is found for the first time that the total EC of single-cluster NOMA at high SNRs is independent of power coefficients and only depends on the delay requirement of the user with the best channel quality. Accordingly, the total EC differences between single-cluster NOMA and user-paired NOMA/OMA are analyzed theoretically and validated via simulation results. We find that the performance gain of single-cluster NOMA over user-paired NOMA/OMA remains stable when the SNR is high. Numerical results indicate that for both loose and strict latency scenarios, single-cluster NOMA outperforms the other two models at high SNRs, and the performance gain is independent of the power coefficients of all users. Simulation results further reveal that increasing the number of users can

improve the total EC of single-cluster NOMA.

II. System Model

We consider the cellular downlink transmission with one base station (BS) and N single-antenna users. Suppose that wireless channels between the BS and users undergo block fading, which means that the channel gain in one fading-block is constant and independent with that in another fading-block. Let h_n represent the channel gain between the BS and the n-th user, and assume that $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_N|^2$. The length of each fading-block is denoted by $T_{\rm f}$, and it is assumed that the frame size equals to $T_{\rm f}$. Besides, let W represent the bandwidth. Fixed power allocation is considered, where each user is allocated a fixed fraction of total transmit power at the BS [5], [13]. In this paper, we focus on the single-cluster NOMA where all N users share the same radio resource such as time slots via NOMA. For comparison purposes, we also introduce the user-paired NOMA and OMA in the following.

In the network of single-cluster NOMA, with successive interference cancelation (SIC) technique, the transmission data rate of the n-th user can be expressed as

$$R_n^{\rm M}=\begin{cases} \log_2\left(1+\frac{p|h_n|^2a_n}{p|h_n|^2A_{n+1}+1}\right), & n=1,2,\cdots,N-1,\\ \log_2\left(1+p|h_n|^2a_n\right), & n=N, \end{cases} \tag{1}$$

where p denotes the transmit SNR, a_n denotes the power coefficient of the n-th user, and $A_{n+1}=\sum_{l=n+1}^{N}a_l$. Note that $\sum_{n=1}^{N}a_n=1$.

In the network of user-paired NOMA, all N users are divided into $N/2$ NOMA clusters [10], and each cluster occupies $2/N$ of orthogonal resources. Time division multiple access (TDMA) is applied for the inter-cluster multiple access. Thus, the transmission data rates of two users in the k-th NOMA cluster, $k=1,2,\cdots,N/2$, are given by $R_{u_k}^{\rm P}=\frac{2}{N}\log_2\left(1+\frac{p|h_{u_k}|^2a_{u_k}}{p|h_{u_k}|^2a_{v_k}+1}\right)$ and $R_{v_k}^{\rm P}=\frac{2}{N}\log_2\left(1+p|h_{v_k}|^2a_{v_k}\right)$, respectively.

In the OMA network, time-division multiple access is considered, and each user occupies $1/N$ of orthogonal resources. The transmission data rate of the n-th user in the OMA network is expressed as $R_n^{\rm O}=\frac{1}{N}\log_2\left(1+p|h_n|^2\right)$.

III. Effective Capacity Analysis

In this section, we first briefly introduce the theory of EC and analyze the total EC for single-cluster NOMA and compare it with user-paired NOMA and OMA. Assume that there exists a buffer (with an infinite buffer size) for the n-th user at the BS. Then, its delay violation probability can be estimated as

$$P_{\rm delay}^{\rm out}=\Pr\{D_n(t)>D_{\max}^n\}\approx\Pr\{Q(t)>0\}{\rm e}^{-\theta_n\mu D_{\max}^n}, \tag{2}$$

where $D_n(t)$ is the delay experienced by a packet arriving at time t, $D_{\max}^n$ is the given delay bound in the unit of symbol duration, $Q(t)$ indicates the number of packets in the queue at time t, $\Pr\{Q(t)>0\}$ is the probability of a non-empty buffer, and $\theta_n(\theta_n>0)$ is the n-th user's delay QoS exponent representing the exponential decay rate. It was proved that the constant arrival rate has to be limited to the value of EC, so that a target delay violation probability limit can be met. Assume that the service process of wireless channel satisfies Gärtner-Ellis theorem. Then, the EC for the n-th user on a block-fading channel in the single-cluster NOMA network is defined as

$$E_n^{\rm C_M}=\begin{cases} \eta_n\ln\left(\mathbb{E}\left[\left(1+\frac{p|h_n|^2a_n}{p|h_n|^2A_{n+1}+1}\right)^{\lambda_n}\right]\right), & n=1,2,\cdots,N-1,\\ \eta_n\ln\left(\mathbb{E}\left[\left(1+p|h_n|^2a_n\right)^{\lambda_n}\right]\right), & n=N, \end{cases} \tag{3}$$

where $\eta_n=-\frac{1}{\theta_nT_{\rm f}W}$, and $\lambda_n=-\frac{\theta_nT_{\rm f}W}{\ln 2}$, and the expectation $\mathbb{E}[\cdot]$ is over h_n. Since h_n is the ordered channel gain, its probability density function (PDF) is given by [14]

$$f_n(g_n)=\frac{\beta_n}{p}e^{-\frac{g_n(N-n+1)}{p}}\left(1-e^{-\frac{g_n}{p}}\right)^{n-1}, \tag{4}$$

where $g_n=p|h_n|^2$, and $\beta_n=\frac{1}{{\rm B}(n,N-n+1)}$. Here, ${\rm B}(\cdot,\cdot)$ is the beta function. Similarly, for the user-paired NOMA network, the ECs of users in the k-th cluster can be expressed as

$$E_{u_k}^{\rm C_P}=\eta_{u_k}\ln\left(\mathbb{E}\left[\left(\frac{p|h_{u_k}|^2+1}{p|h_{u_k}|^2a_{v_k}+1}\right)^{\frac{2\lambda_{u_k}}{N}}\right]\right), \tag{5a}$$

$$E_{v_k}^{\rm C_P}=\eta_{v_k}\ln\left(\mathbb{E}\left[\left(p|h_{v_k}|^2a_{v_k}+1\right)^{\frac{2\lambda_{v_k}}{N}}\right]\right). \tag{5b}$$

For the OMA network, the EC of the n-th user is given by

$$E_n^{\rm C_O}=\eta_n\ln\left(\mathbb{E}\left[\left(p|h_n|^2+1\right)^{\frac{\lambda_n}{N}}\right]\right). \tag{6}$$

A. Effective Capacity Closed-Forms for Single-Cluster NOMA

Theorem 1: The exact closed-form expression of the total EC for single-cluster NOMA is calculated as $\Phi_{\rm M}=\sum_{n=1}^{N}E_n^{\rm C_M}$, where $E_n^{\rm C_M}, n=\{1,2\cdots N\}$, is given in (7) on the top of next page. For brevity, the total EC is further approximated as $\Phi'_{\rm M}=\sum_{n=1}^{N-1}\tilde{E}_n^{\rm C_M}+E_N^{\rm C_M}$, where $\tilde{E}_n^{\rm C_M}$ is given in (8). At high SNRs, the total EC approximates to

$$\widetilde{\Phi}_{\rm M}=\eta_N\ln\left(\frac{\beta_N}{p}\sum_{i=0}^{N-1}\binom{N-1}{i}(-1)^i\left(\frac{i+1}{p}\right)^{-\lambda_N-1}\Gamma(\lambda_N+1)\right). \tag{9}$$

Proof: See Appendix A.

Remark 1: The accuracy of the above exact and approximate closed-form expressions will be validated in Section IV by comparing with Monte Carlo results. From (9), we can find that when the SNR is high, the approximate total EC of single-cluster NOMA, i.e., $\widetilde{\Phi}_{\rm M}$, is independent of the power coefficients of all users and only relies on the delay QoS exponent of the N-th user, i.e., the statistical delay requirement of the user with the best channel condition. In other words, when the transmit SNR is high, the other users' delay requirements, i.e.

$$E_n^{\mathrm{C_M}} = \eta_n \ln\left(\frac{\beta_n}{p}\left(\frac{A_n}{A_{n+1}}\right)^{\lambda_n}\left(\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\left(\frac{1}{c_n}+\frac{\lambda_n a_n e^{\frac{c_n}{A_n}}}{A_n A_{n+1}}E_i\left(-\frac{c_n}{A_{n+1}}\right)\right)+\sum_{j=2}^{\infty}\binom{\lambda_n}{j}\left(\frac{-a_n}{A_n A_{n+1}}\right)^j\frac{1}{(j-1)!}\right.\right.$$
$$\left.\left.\times\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i(-c_n)^{j-1}\left(\sum_{l=1}^{j-1}(l-1)!\left(-\frac{c_n}{A_{n+1}}\right)^{-l}-e^{\frac{c_n}{A_{n+1}}}E_i\left(-\frac{c_n}{A_{n+1}}\right)\right)\right)\right), \text{for } n=1,2,\cdots,N-1, \tag{7a}$$

$$E_N^{\mathrm{C_M}} = \eta_N \ln\left(\frac{\beta_n}{pa_N}\sum_{i=0}^{N-1}\binom{N-1}{i}(-1)^i U\left(1,\lambda_N+2,\frac{i+1}{pa_N}\right)\right). \tag{7b}$$

$$\tilde{E}_n^{\mathrm{C_M}} \approx \eta_n \ln\left(\frac{\beta_n}{p}\left(\frac{A_n}{A_{n+1}}\right)^{\lambda_n}\left(\sum_{i=0}^{n-1}\binom{n-1}{i}(-1)^i\left(\frac{1}{c_n}+\frac{\lambda_n a_n}{A_{n+1}A_n}e^{\frac{c_n}{A_n}}E_i\left(-\frac{c_n}{A_n}\right)\right)\right)\right), \text{ for } n=1,2,\cdots,N-1, \tag{8}$$

the n-th user ($n \neq N$), have no impact on the total achievable delay-constrained rate.

B. Comparisons of the Total EC among Three Models

Define $\Delta_1 = \Phi_{\mathrm{M}} - \Phi_{\mathrm{P}}$ to represent the total EC difference between single-cluster NOMA and user-paired NOMA, where Φ_{P} is the total EC for the user-paired NOMA network derived in [10]. Similarly, $\Delta_2 = \Phi_{\mathrm{M}} - \Phi_{\mathrm{O}}$ denotes the total EC difference between single-cluster NOMA and OMA networks, where Φ_{O} is the total EC for the OMA network [10].

Theorem 2: Aiming to analyze the total EC differences among three system models, i.e., Δ_1 and Δ_2, we prove that

(a) When $p \to 0$, $\Delta_1 \to 0$, $\Delta_2 \to 0$, and

$$\lim_{p\to 0}\frac{\partial\Delta_1}{\partial p} = \sum_{n=1}^{N}\frac{a_n}{\ln 2}\mathbb{E}\left[|h_n|^2\right] - \frac{2}{N\ln 2}\sum_{k=1}^{N/2}\left(a_{u_k}\mathbb{E}\left[|h_{u_k}|^2\right] + a_{v_k}\mathbb{E}\left[|h_{v_k}|^2\right]\right), \tag{10a}$$

$$\lim_{p\to 0}\frac{\partial\Delta_2}{\partial p} = \sum_{n=1}^{N}\frac{\mathbb{E}\left[|h_n|^2\right]}{\ln 2}\left(a_n - \frac{1}{N}\right). \tag{10b}$$

(b) When $p \to \infty$, $\frac{\partial\Delta_1}{\partial p} \to 0$, $\frac{\partial\Delta_2}{\partial p} \to 0$, and

$$\lim_{p\to\infty}\Delta_1 = \eta_N \ln\left(\mathbb{E}\left[\left(|h_N|^2\right)^{\lambda_N}\right]\right) - \sum_{k=1}^{N/2}\eta_{v_k}\ln\left(\mathbb{E}\left[\left(|h_{v_k}|^2\right)^{\frac{2\lambda_{v_k}}{N}}\right]\right), \tag{11a}$$

$$\lim_{p\to\infty}\Delta_2 = \eta_N \ln\left(\mathbb{E}\left[\left(|h_N|^2\right)^{\lambda_N}\right]\right) - \sum_{n=1}^{N}\eta_n\ln\left(\mathbb{E}\left[\left(|h_n|^2\right)^{\frac{\lambda_n}{N}}\right]\right). \tag{11b}$$

Proof: See Appendix B.

Remark 2: Theorem 2 indicates that, with the increase of the SNR p, the total EC differences Δ_1 and Δ_2 both start at the value of 0 (when $p \to 0$), and finally remain stable (when $p \to \infty$). When SNR is very high, (11a) implies that the performance gap between single-cluster NOMA and user-paired NOMA is not affected by the power coefficients of users. This gap only relies on the delay requirements and user pairing setting of user-paired NOMA. Similarly, the performance gap between single-cluster NOMA and OMA only depends on the delay requirements of all users, which is independent of the power coefficients of all users, as shown in (11b).

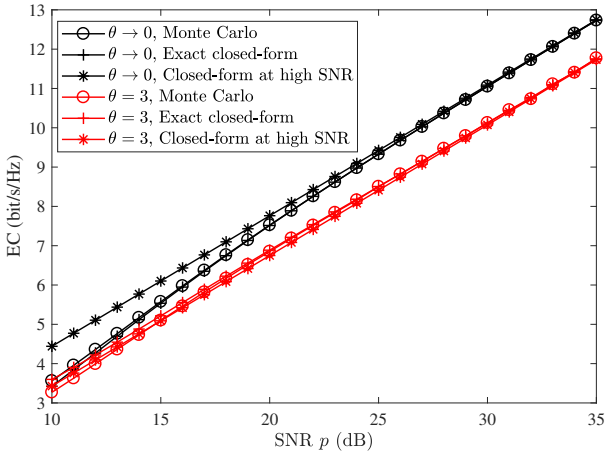

Fig. 1. The total EC for single-cluster NOMA *vs.* the SNR p.

The conclusions of Theorem 2 will be further demonstrated numerically in the next section. In addition, for the considered power model in the simulation, it can be derived that $\lim_{p\to 0}\frac{\partial\Delta_2}{\partial p} \leq 0$, which means that Δ_2 is definitely negative when p is sufficiently small. In other words, OMA outperforms single-cluster NOMA at small SNRs. On the other hand, the closed-form of $\lim_{p\to 0}\frac{\partial\Delta_1}{\partial p}$ is complicated, and its sign not only depends on the power allocation model but also on the user pairing scheme for user-paired NOMA. Also, it is worth noting (in the next section) that the values of Δ_1 and Δ_2 are positive at high SNRs for $\theta \to 0$ and $\theta = 3$, with the consideration of the best and the worst user pairing schemes. This means that in this case, single-cluster NOMA outperforms user-paired NOMA at high SNRs.

IV. Simulation Results

In the simulations, it is assumed that there are 6 users, i.e., $N = 6$, and all users have the same delay exponents, denoted by θ. The unordered channel gains between the BS and users follow Rayleigh fading distribution with unit variance, which are then sorted to satisfy $|h_1|^2 \leq |h_2|^2 \leq \cdots \leq |h_N|^2$. For the single-cluster NOMA network, the power coefficient of user

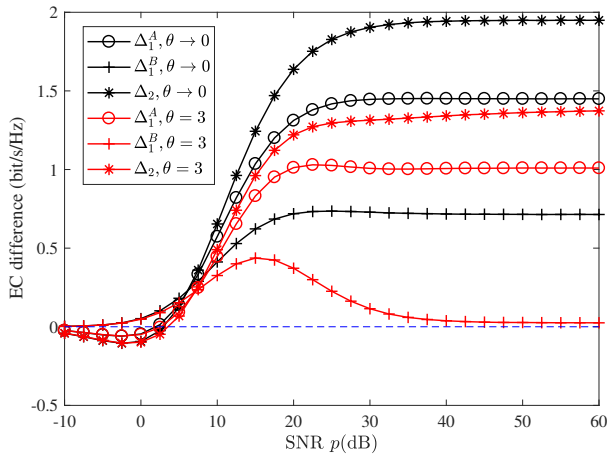

Fig. 2. The total EC difference *vs.* the SNR p.

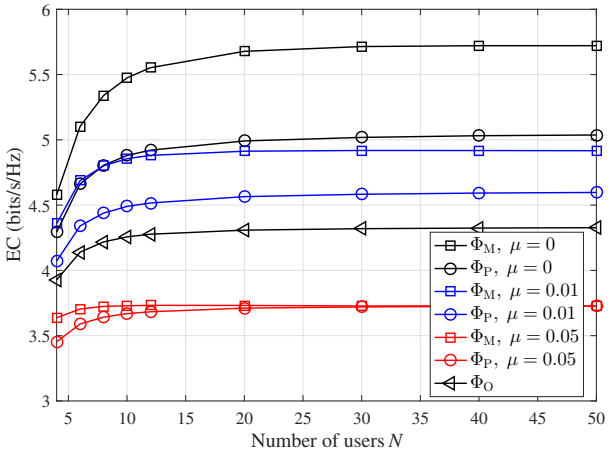

Fig. 4. The total ECs with perfect and imperfect SIC *vs.* the number of users.

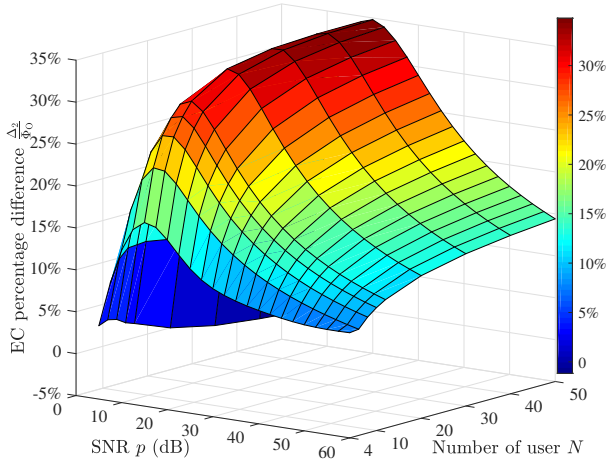

Fig. 3. The total EC percentage difference $\frac{\Delta_2}{\Phi_{\mathrm{O}}}$ *vs.* the number of users N and the SNR p.

n is $a_n = \frac{N-n+1}{C}$ [5], where C is the normalized coefficient to guarantee $\sum_{n=1}^{N} a_n = 1$. Without loss of generality, for the user-paired NOMA network, the power coefficients of two users in the k-th cluster are set to $a_{\mu_k} = 0.8$ and $a_{v_k} = 0.2$, respectively, which follows the settings in [5], [13]. Besides, we assume $B = 100$ KHz and $T_{\mathrm{f}} = 0.01$ ms.

We first verify the accuracy of the exact closed-form expression of the total EC for single-cluster NOMA and its approximate closed-form at high SNRs given in Theorem 1. In Fig. 1, the curves of Φ_{M} and $\widetilde{\Phi}_{\mathrm{M}}$ are plotted versus the SNR p for different θ. It can be found from Fig. 1 that the derived exact closed-form Φ_{M} is accurate, which matches with the Monte Carlo curve. Furthermore, when the SNR p is high, the approximate closed-form $\widetilde{\Phi}_{\mathrm{M}}$ coincides with the exact values. Besides, as shown in Fig. 1, the total EC increases with the SNR p but decreases with the delay exponent θ.

Then, to validate Theorem 2, the total EC differences Δ_1 and Δ_2 are plotted against the SNR p in Fig. 2. To provide a comprehensive comparison, we conduct exhaustive search to find the best and the worst user pairing schemes for the user-paired NOMA with the aim of maximizing the total EC. It has been found that the pairing scheme B = $\{(1,6)\,(2,5)\,(3,4)\}$ is the best choice which supports the maximum total EC, while A = $\{(1,2)\,(3,4)\,(5,6)\}$ is the worst choice which achieves the minimum total EC [10]. Therefore, $\Delta_1^A = \Phi_M - \Phi_P^A$/ $\Delta_1^B = \Phi_M - \Phi_P^B$ provides the largest or the smallest performance gap between single-cluster NOMA and user-paired NOMA. For the other pairing options, the absolute performance differences can only be within $|\Delta_1^A|$ and $|\Delta_1^B|$.

As shown in Fig. 2, all curves begin at the value of 0 and finally remain stable at high SNRs, which confirms Theorem 2. Observing two curves of Δ_1^{B}, we can find that single-cluster NOMA always achieves higher total EC than user-paired NOMA with the user pairing setting B, under the given simulation settings. In contrast, Δ_1^{A} and Δ_2 first decrease under 0 and then turn to go up with the increase of the SNR p, which means that OMA and user-paired NOMA with the user pairing setting A outperform single-cluster NOMA when the SNR p is relatively small but fall behind single-cluster NOMA when p is larger than 3dB. Besides, with the variation of θ from 0 to 3, since the total ECs of all the three networks become small, all the total EC differences decline.

The total EC percentage difference $\frac{\Delta_2}{\Phi_{\mathrm{O}}}$ is presented versus N and p in Fig. 3. It can be observed in Fig. 3, that $\frac{\Delta_2}{\Phi_{\mathrm{O}}}$ reaches its peak when the SNR p is about 15 dB. As shown in Fig. 3, for a given SNR, when the number of users increases, the total EC percentage difference first goes up, and then remains stable, which indicates that increasing the number of clustered users can improve the total EC, but this improvement will vanish when the number of users is large enough.

Finally, the total ECs with perfect and imperfect SIC versus the number of users N are demonstrated in Fig. 4, where μ is the linear coefficient of imperfect SIC and $\mu = 0$ represents the perfect SIC [7]. As shown in this figure, the ECs of NOMA and OMA all grow with the increase of N, while for the same N, the ECs of single-cluster and user-paired NOMA decline with the increase of μ since larger μ results in more residual interference for SIC. Moreover, regardless of μ, single-cluster

NOMA always outperforms user-paired NOMA. It is also observed that when μ is smaller, e.g., $\mu = 0.01$, the ECs of single-cluster and user-paired NOMA are both larger than that of OMA. Contrarily, when $\mu = 0.05$, the ECs of NOMA are less than that of OMA. This is because when performing SIC, more residual interference significantly decreases the ECs of NOMA, while OMA is free from the impact of imperfect SIC.

V. Conclusion

In this paper, we derived the closed-form expressions of total EC for single-cluster NOMA and analyzed the total EC differences among single-cluster NOMA, user-paired NOMA and OMA. Simulation results verified the accuracy of derived closed-forms and showed that single-cluster NOMA outperforms user-paired NOMA and OMA in terms of the total EC when the SNR is larger than 3dB. Further, it was found that when the number of users increases, single-cluster NOMA achieves greater gain on the total EC, compared to OMA.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61901027, Grant 62001179, in part by the Fundamental Research Funds for the Central Universities under Grant 2020kfyXJJS111, in part by the European Union Horizon 2020, RISE 2018 scheme (H2020-MSCA-RISE-2018) under the Marie Skłodowska-Curie grant agreement No. 823903 (RECENT), and in part by the UK EPSRC under grant number EP/K011693/1 and the EU FP7 CROWN project under grant number PIRSES-GA-2013-610524.

Appendix A

Proof of Theorem 1: With the PDF given in (4), the EC of the n-th user, $n = 1, 2, \cdots, N-1$ can be re-expressed as

$$E_n^{\mathrm{C_M}} = \eta_n \ln\left(\frac{\beta_n A_n^{\lambda_n}}{p A_{n+1}^{\lambda_n}} \int_0^\infty \left(1+\frac{a_n/A_{n+1}}{g_n A_n+1}\right)^{-\lambda_n} \times e^{-\frac{g_n(N-n+1)}{p}} \left(1-e^{-\frac{g_n}{p}}\right)^{n-1} dg_n\right). \tag{12}$$

Then, since $\left|\frac{a_n/A_n}{g_n A_{n+1}+1}\right| = \left|\frac{a_n/A_{n+1}}{g_n A_n+1+a_n/A_{n+1}}\right| \leq 1$, the generalized binomial theorem can be applied, which gives

$$\left(1+\frac{a_n/A_{n+1}}{g_n A_n+1}\right)^{-\lambda_n} = \left(1-\frac{a_n/A_n}{g_n A_{n+1}+1}\right)^{\lambda_n} = \sum_{j=0}^{\infty} \binom{\lambda_n}{j} \left(\frac{-a_n/A_n}{g_n A_{n+1}+1}\right)^j \tag{13}$$

We note that the expansion of binomial $(1+x)^a$ is given by $(1+x)^a = \sum_{n=0}^{\infty} \binom{a}{n} x^n$, which holds for a being any real number, i.e., $a \in \mathbb{R}$, and $|x| < 1$. Here, $\binom{a}{n}$ is the binomial coefficient, defined as [15]

$$\binom{a}{n} = \begin{cases} \dfrac{a(a-1)\ldots(a-n+1)}{n!} & \text{if } n \geq 1, \\ 1 & \text{if } n = 0. \end{cases} \tag{14}$$

Then, from the binomial expansion

$$\left(1-e^{-\frac{g_n}{p}}\right)^{n-1} = \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i e^{-\frac{i g_n}{p}}, \tag{15}$$

and by defining $c_n = \frac{N-n+1+i}{p}$, (12) can be approximated as

$$E_n^{\mathrm{C_M}} = \eta_n \ln\left(\frac{\beta_n A_n^{\lambda_n}}{p A_{n+1}^{\lambda_n}} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \left(\int_0^\infty e^{-c_n g_n} dg_n - \frac{\lambda_n a_n}{A_n} \int_0^\infty \frac{e^{-c_n g_n}}{g_n A_{n+1}+1} dg_n + \int_0^\infty \sum_{j=2}^{\infty} \binom{\lambda_n}{j} \left(\frac{-a_n/A_n}{g_n A_{n+1}+1}\right)^j \times e^{-c_n g_n} dg_n\right)\right) \tag{16}$$

Finally, by applying (3.352.4) in [16],

$$\int_0^\infty \frac{e^{-ax}}{(x+b)^m} dx = \frac{\sum_{i=1}^{m-1} (i-1)!(-a)^{m-i-1} b^{-i} - (-a)^{m-1} e^{ab} \mathrm{E}_i(-ab)}{(m-1)!}, \quad |\arg \mathrm{b}| < \pi, \mathrm{Re\ a} > 0, \tag{17}$$

(7a) is obtained. To further truncate the infinite series in (7a), the approximation

$$\left(1+\frac{a_n/A_{n+1}}{g_n A_n+1}\right)^{-\lambda_n} = \left(1-\frac{a_n/A_n}{g_n A_{n+1}+1}\right)^{\lambda_n} \approx 1-\frac{\lambda_n a_n/A_n}{g_n A_{n+1}+1}, \tag{18}$$

can be employed on (12), and (8) is derived by following the above steps.

For the N-th user, its EC can be re-expressed as

$$E_N^{\mathrm{C_M}} = \eta_N \ln\left(\frac{\beta_N}{p} \sum_{i=0}^{N-1} \binom{N-1}{i} (-1)^i \int_0^\infty (1+g_N a_N)^{\lambda_N} \times e^{-\frac{g_N(i+1)}{p}} dg_N\right), \tag{19}$$

by replacing $\left(1-e^{-\frac{g_N}{p}}\right)^{N-1}$ with its binomial expansion. Then, by applying the confluent hypergeometric function,

$$U(a,b,c) = \frac{1}{\Gamma(a)} \int_0^\infty e^{-ct} t^{a-1} (1+t)^{b-a-1} dt, \quad \text{for Re } a, \mathrm{Re}\ c > 0, \tag{20}$$

(7b) can be finally obtained.

When the SNR is high, the EC for the first $N-1$ users, i.e., $n = 1, 2, \cdots, N-1$, can be approximated as

$$\widetilde{E}_n^{\mathrm{C_M}} = \lim_{p\to\infty} \eta_n \ln\left(\left(1+\frac{a_n}{A_{n+1}}\right)^{\lambda_n}\right) = \log_2\left(1+\frac{a_n}{A_{n+1}}\right), \tag{21}$$

and the EC for the strongest user at high SNRs becomes

$$\widetilde{E}_N^{\mathrm{C_M}} = \lim_{p\to\infty} \eta_N \ln\left(\mathbb{E}\left[\left(p a_N |h_N|^2\right)^{\lambda_N}\right]\right), \tag{22}$$

which can be expressed as

$$\begin{aligned}\widetilde{E}_N^{\mathrm{C_M}} &\overset{(\vartheta)}{=} \lim_{p\to\infty} \eta_N \ln\Bigg(\frac{\beta_N a_N^{\lambda_N}}{p}\sum_{i=0}^{N-1}\binom{N-1}{i}(-1)^i \\ &\qquad \times \int_0^\infty g_N^{\lambda_N} e^{-\frac{g_N(i+1)}{p}} dg_N\Bigg) \\ &\overset{(\rho)}{=} \lim_{p\to\infty} \log_2(a_N) + \eta_N \ln\Bigg(\frac{\beta_N}{p}\sum_{i=0}^{N-1}\binom{N-1}{i}(-1)^i \\ &\qquad \times\Big(\frac{i+1}{p}\Big)^{-\lambda_N-1}\Gamma(\lambda_N+1)\Bigg),\end{aligned} \tag{23}$$

where the equality (ϑ) is derived according to the binomial expansion (15), and (ρ) is obtained from (3.382.4) [16]

$$\int_0^\infty (x+a)^b e^{-cx}dx = c^{-b-1}e^{ac}\Gamma(b+1, ac), \mathrm{Re}\ c>0.$$

Thus, by inserting (21) and (23) into $\widetilde{\Phi}_{\mathrm{M}}$, (9) can be derived.

Appendix B

Proof of Theorem 2: When $p=0$, we have $\Phi_{\mathrm{M}}=\Phi_{\mathrm{P}}=\Phi_{\mathrm{O}}=0$. Hence, $\lim_{p\to 0}\Delta_1=\lim_{p\to 0}\Delta_2=0$ holds. Considering $\lim_{p\to 0}\frac{\partial\Delta_1}{\partial p}$ and $\lim_{p\to 0}\frac{\partial\Delta_2}{\partial p}$, the first-derivative of Φ_{M} is first given by

$$\begin{aligned}\frac{\partial\Phi_{\mathrm{M}}}{\partial p} &= \frac{1}{\ln 2}\sum_{n=1}^{N-1}\frac{\mathbb{E}\left[\left(1+\frac{p|h_n|^2a_n}{p|h_n|^2A_{n+1}+1}\right)^{\lambda_{n-1}}\frac{|h_n|^2a_n}{(p|h_n|^2A_{n+1}+1)^2}\right]}{\mathbb{E}\left[\left(1+\frac{p|h_n|^2a_n}{p|h_n|^2A_{n+1}+1}\right)^{\lambda_n}\right]} \\ &\quad + \frac{a_N}{\ln 2}\frac{\mathbb{E}\left[\left(1+p|h_N|^2a_N\right)^{\lambda_N-1}|h_N|^2\right]}{\mathbb{E}\left[\left(1+p|h_N|^2a_N\right)^{\lambda_N}\right]},\end{aligned} \tag{24}$$

and the first-derivatives of Φ_{P} and Φ_{O} are given in (42) and (44) in [10]. Thus, by inserting $p=0$ into $\frac{\partial\Delta_1}{\partial p}=\frac{\partial\Phi_{\mathrm{M}}}{\partial p}-\frac{\partial\Phi_{\mathrm{P}}}{\partial p}$ and $\frac{\partial\Delta_2}{\partial p}=\frac{\partial\Phi_{\mathrm{M}}}{\partial p}-\frac{\partial\Phi_{\mathrm{O}}}{\partial p}$, (10a) and (10b) are respectively proved.

When $p\to\infty$, $\lim_{p\to\infty}\Delta_1$ can be calculated by

$$\begin{aligned}&\lim_{p\to\infty}\Delta_1 = \lim_{p\to\infty}\left(\sum_{n=1}^{N-1}E_n^{\mathrm{C_M}}+E_N^{\mathrm{C_M}}-\sum_{k=1}^{N/2}E_{u_k}^{\mathrm{C_P}}-\sum_{k=1}^{N/2}E_{v_k}^{\mathrm{C_P}}\right) \\ &= \lim_{p\to\infty}\Bigg(\sum_{n=1}^{N-1}\log_2\left(1+\frac{a_n}{A_{n+1}}\right)+\eta_N\ln\Big(\mathbb{E}\left[\left(pa_N|h_N|^2\right)^{\lambda_N}\right]\Big) \\ &\quad -\frac{2}{N}\sum_{k=1}^{N/2}\log_2\Big(\frac{1}{a_{v_k}}\Big)-\sum_{k=1}^{N/2}\eta_{v_k}\ln\left(\mathbb{E}\left[\left(pa_{v_k}|h_{v_k}|^2\right)^{\frac{2\lambda_{v_k}}{N}}\right]\right)\Bigg),\end{aligned} \tag{25}$$

which simplifies to (11a). Similarly, (11b) can also be validated. On the other hand, it can be obtained that

$$\lim_{p\to\infty}\frac{\partial\Phi_{\mathrm{M}}}{\partial p}=\lim_{p\to\infty}\frac{1}{\ln 2}\left(\frac{1}{p^2}\sum_{n=1}^{N-1}\frac{a_n}{A_nA_{n+1}}\mathbb{E}\left[\frac{1}{|h_n|^2}\right]+\frac{1}{p}\right)=0, \tag{26}$$

and similarly, we can prove that $\lim_{p\to\infty}\frac{\partial\Phi_{\mathrm{P}}}{\partial p}=\lim_{p\to\infty}\frac{\partial\Phi_{\mathrm{O}}}{\partial p}=0$. Therefore, $\lim_{p\to\infty}\frac{\partial\Delta_1}{\partial p}=\lim_{p\to\infty}\frac{\partial\Delta_2}{\partial p}=0$ is demonstrated.

References

[1] Y. Liu, Z. Qin *et al.*, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[2] M. Vaezi and H. V. Poor, *NOMA: An Information-Theoretic Perspective*, Jan. 2019, pp. 167–193.

[3] W. Shin, M. Vaezi *et al.*, "Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges," *IEEE Commun. Mag.*, vol. 55, no. 10, pp. 176–183, Oct. 2017.

[4] P. Popovski *et al.*, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, Oct. 2018.

[5] Z. Ding, Z. Yang *et al.*, "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.

[6] M. S. Ali, H. Tabassum *et al.*, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, Oct. 2016.

[7] X. Chen, R. Jia *et al.*, "On the design of massive non-orthogonal multiple access with imperfect successive interference cancellation," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2539–2551, Mar. 2018.

[8] J. Choi, "Effective capacity of NOMA and a suboptimal power control policy with delay QoS," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1849–1858, Jan. 2017.

[9] G. Liu, Z. Ma *et al.*, "Cross-layer power allocation in nonorthogonal multiple access systems for statistical QoS provisioning," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11 388–11 393, Dec. 2017.

[10] W. Yu, L. Musavian, and Q. Ni, "Link-layer capacity of NOMA under statistical delay QoS guarantees," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4907–4922, Oct. 2018.

[11] C. Xiao, J. Zeng *et al.*, "Delay guarantee and effective capacity of downlink NOMA fading channels," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 508–523, Jun. 2019.

[12] M. Bello, A. Chorti *et al.*, "Asymptotic performance analysis of NOMA uplink networks under statistical qos delay constraints," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1691–1706, Oct. 2020.

[13] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[14] H. A. David and H. N. Nagaraja, *Order Statistics*. JohnWiley, New York, 3rded., 2003.

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

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