Opportunistic Adaptive Non-Orthogonal Multiple Access in Multiuser Wireless Systems: Probabilistic User Scheduling and Performance Analysis

Long Yang, *Member, IEEE*, Hai Jiang, *Senior Member, IEEE*, Qiang Ye, *Member, IEEE*,
Zhiguo Ding, *Fellow, IEEE*, Fang Fang, *Member, IEEE*, Jia Shi,
Jian Chen, *Member, IEEE*, and Xuan Xue, *Member, IEEE*

***Abstract*—This paper designs a novel opportunistic adaptive non-orthogonal multiple access (OA-NOMA) strategy, where a base station (BS) employs NOMA to serve a near user (NU)-far user (FU) pair opportunistically scheduled from M NUs and K FUs. In particular, the NOMA transmission to the scheduled NU-FU pair adaptively operates in one of two modes: *Direct NOMA mode*, in which the BS directly serves the scheduled NU-FU pair with using NOMA; *Cooperative NOMA mode*, in which the scheduled NU receives the messages intended by both scheduled users from the BS, and then forwards the message intended by the scheduled FU. For the OA-NOMA strategy, a scheduling candidate acquisition method and a probabilistic user pair scheduling scheme are proposed to guarantee the transmission reliability and improve the scheduling fairness, respectively. To evaluate the scheduling fairness, we develop a max-min fairness criterion and show that the OA-NOMA strategy approximately achieves max-min fairness. The reliability of the OA-NOMA strategy is also evaluated in terms of outage probability and diversity order. For the outage probability, we derive an approximate expression and numerically verify its tightness. For the diversity order, we show that the proposed OA-NOMA strategy achieves a diversity order of M.**

***Index Terms*—Cooperative non-orthogonal multiple access, outage probability, scheduling fairness, user scheduling.**

Manuscript received November 20, 2019; revised April 8, 2020; accepted May 22, 2020. This work was supported in part by the National Natural Science Foundation of China under Grants 61971320, 61771366, 61901327, 61941105 and 61825104, in part by the National Sciences and Engineering Research Council of Canada under Grants RGPIN-2018-06307 and RGPIN-2017-05853, in part by the China Postdoctoral Science Foundation under Grant 2019M653559, and in part by the International Postdoctoral Exchange Fellowship Program 2017 from the Office of China Postdoctoral Council. The editor coordinating the review of this paper and approving it for publication was Dr. Y. Shi *(Corresponding Author: Jian Chen).*

L. Yang, J. Shi, J. Chen, and X. Xue are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: lyang@xidian.edu.cn, jiashi@xidian.edu.cn, jianchen@mail.xidian.edu.cn, xuanx@xidian.edu.cn).

H. Jiang is with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta T6G 1H9, Canada (e-mail: hai1@ualberta.ca).

Q. Ye is with the Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia B3H 4R2, Canada (e-mail: qye@cs.dal.ca).

Z. Ding is with the School of Electrical and Electronic Engineering, The University of Manchester, M13 9PL, United Kingdom (e-mail: zhiguo.ding@manchester.ac.uk).

F. Fang is with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PR, United Kingdom, and also with the Department of Engineering, Durham University, Durham DH1 3LE, United Kingdom (e-mail: fang.fang@manchester.ac.uk).

I. Introduction

Non-orthogonal multiple access (NOMA) has the potential to resolve the spectral resource scarcity problem for the upcoming fifth generation (5G) mobile networks [1]–[10]. By exploiting superposition coding and successive interference cancellation (SIC), NOMA is able to serve two or more users simultaneously by using power-domain multiplexing. Compared with conventional orthogonal multiple access (OMA), NOMA can achieve higher spectral efficiency as well as better connectivity to *far users* (FUs) which are distant from the source.

Recently, cooperative communications have been introduced into NOMA to further enhance the performance of FUs, which is termed as *cooperative NOMA* [11]–[17]. Following the rationale of conventional cooperative communications [18], cooperative NOMA can be realized by deploying a dedicated relay to help the FUs [11]–[13]. Further, for multiple-relay/antenna NOMA systems, employing the relay/antenna selection can enhance the performance of relay-aided cooperative NOMA [14], [15]. On the other hand, according to the principles of NOMA, *near users* (NUs), which are close to the source, may obtain information of FUs during its SIC-based detection. Thus, the NUs can naturally serve as relays for FUs, which reduces the cost of deploying dedicated relays. Exploiting this feature of NOMA, a cooperative NOMA scheme is proposed in [16] for a two-user NOMA system, where the NU serves as a relay for the FU in addition to decoding its own message. In [17], a sophisticated cooperation scheme is proposed for a multiuser NOMA system that can serve K users simultaneously. In this scheme, each user sequentially forwards other users' messages decoded by itself. However, since $(K-1)$ phases are required for the cooperation, this scheme achieves only $1/K$ spectral efficiency of non-cooperative NOMA, which would become a bottleneck for practical systems with a large amount of users. As indicated in [19], serving all users in a single NOMA transmission may not be practical, because of the strong co-channel interference in a NOMA transmission. Instead, partitioning all the users to a number of groups and then using NOMA within each group is a more promising way to implement NOMA in practical systems.

To enhance the spectral efficiency and reduce the co-channel interference, user scheduling has been integrated into cooperative NOMA in some recent work [20]–[26]. Specifically, only the users that are scheduled can participate in the cooperative NOMA transmissions, which is formally called *opportunistic cooperative NOMA*. When only the distance information is known, three user scheduling schemes are proposed in [20], where the scheduled NU serves as a relay to help the scheduled FU. It is demonstrated in [20] that the scheduled FU always achieves a diversity order of two with all three schemes. On the other hand, if instantaneous channel state information (CSI) is available for user scheduling, the opportunistic cooperative NOMA can provide a full diversity for both scheduled NU and scheduled FU [21]–[23]. For the scenario that one of M NUs and one of K FUs are opportunistically served by cooperative NOMA, a best-near best-far (BNBF) user scheduling scheme is proposed in [21], in which the NU and the FU that own the respectively best downlink channels are scheduled prior to each transmission. With using the BNBF scheme, diversity orders of M and K are achieved by the scheduled NU and the scheduled FU, respectively. Considering the same scenari-o, the work in [22] designs an adaptive OMA/cooperative NOMA strategy with using dynamic user scheduling. It is demonstrated in [22] that by adaptively switching between OMA and cooperative NOMA, the diversity order achieved by the scheduled FU can be further improved to $M+K$. Moreover, for a NOMA-assisted cooperative overlay cognitive radio system, a CSI-based two-stage user scheduling scheme is designed in [23], in order to provide the full diversity for both primary and secondary transmissions. Note that these studies [20]–[23] only focus on wireless unicast applications, where all users require individually different information. When some users demand the same data via wireless multicast, several user scheduling schemes are proposed in [24]–[26]. In these schemes, one of the users that have successfully decoded the multicast messages is selected to forward all its decoded messages.

We have the following observation for the existing s-tudies on opportunistic cooperative NOMA [20]–[26]: With the existing approaches on opportunistic cooperative NOMA, user scheduling is based on the distance information [20] or instantaneous CSI [21]–[26], resulting in a system that favors the users with certain geographic locations/channel conditions. In other words, unfair channel access is provided to the users.

Motivated by this observation, this paper focuses on the design of dynamic direct/cooperative NOMA strategy with novel user scheduling for enhancing fairness and reliability simultaneously. To the best of authors' knowledge, such an issue remains uninvestigated to date. In particular, the main contributions of this paper can be summarized as follows.

- We design a novel opportunistic adaptive NOMA (OA-NOMA) strategy for a downlink NOMA system, where the base station (BS) serves an NU-FU pair opportunisti-cally scheduled from M NUs and K FUs. According to the channel conditions of the scheduled NU-FU pair, the downlink transmission adaptively operates in either the direct NOMA mode or the cooperative NOMA mode, where the power control is employed in each mode to minimize the transmit power consumption.
- To guarantee the reliability of OA-NOMA strategy, we first design a scheduling candidates acquisition (SCA) method to find out the candidate NU-FU pairs for scheduling. Then, with the candidate NU-FU pairs, a probabilistic user pair scheduling (P-UPS) scheme that schedules each candidate NU-FU pair with a certain probability is then proposed to enhance the scheduling fairness of the OA-NOMA strategy.
- Our scheduling scheme can approximately achieve max-min fairness. Further, we theoretically prove that when the system achieves max-min fairness, it also achieves proportional fairness. Thus, our scheduling scheme also approximately achieves proportional fairness.
- The reliability of the OA-NOMA strategy is evaluated in terms of outage probability and diversity order. For the outage probability, we derive an approximate expres-sion and verify its tightness by simulations. Further, it is demonstrated that the proposed OA-NOMA strategy provides a diversity order of M for the scheduled NU-FU pair, showing that the inherent diversity offered by NUs is fully exploited. Thus, our work simultaneously achieves reliability and fairness of dynamic direct/cooperative NO-MA, which remains uninvestigated in the literature so far.

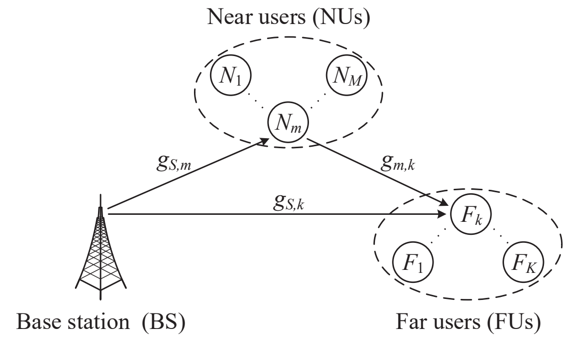


Fig. 1. System model under investigation.

The rest of this paper is organized as follows. Section II describes our system model. Section III proposes the OA-NOMA strategy, in which the user pair scheduling is realized by the proposed SCA method and P-UPS scheme. Section IV presents the max-min fairness criterion for NU-FU pair scheduling. Section V theoretically analyzes the reliability of OA-NOMA strategy in terms of outage probability and diversity order. The simulation results are provided in Section VI, followed by Section VII concluding this paper.

Notations Explanation: The cumulative distribution function (CDF) and probability density function (PDF) of random variable (RV) X are denoted as $F_X(\cdot)$ and $f_X(\cdot)$, respectively. The joint PDF of RVs X and Y is denoted as $f_{X,Y}(\cdot,\cdot)$. $\Pr\{\cdot\}$ means probability of an event $\{\cdot\}$. Boldface uppercase letters are used to represent sets, while blackboard boldface uppercase letters are used to represent events/conditions. When $\mathbf{B} \subseteq \mathbf{A}$, notation $\mathbf{A} \setminus \mathbf{B}$ means the difference of set $\mathbf{A}$ and $\mathbf{B}$, i.e., $\mathbf{A} \setminus \mathbf{B} = \{i | i \in \mathbf{A}, i \notin \mathbf{B}\}$.

II. System Model

As shown in Fig. 1, we consider a downlink wireless system consisting of a BS denoted by S, a number M of NUs denoted by $N_m, m \in \mathbf{M} \triangleq \{1, ..., M\}$ and a number K of FUs denoted by $F_k, k \in \mathbf{K} \triangleq \{1, ..., K\}$. It is assumed that the NUs are relatively close to the BS, whereas the FUs are relatively distant from the BS, as in [20]–[22]. For example, for downlink transmission from a BS to multiple users in a cell, some users are located in the cell-center area, corresponding to the NUs, while other users are located in the cell-edge area, corresponding to the FUs. Prior to each downlink transmission, *user pair scheduling* is carried out to pair an NU and an FU dynamically scheduled from M NUs and K FUs, respectively. For notational convenience, an NU-FU pair consisting of NU N_m and FU F_k is denoted by $[N_m, F_k]$. Then, downlink NOMA transmission is performed to serve the scheduled NU-FU pair with predetermined target rates for both users[1]. More specifically, the target rates for the scheduled NU and the scheduled FU are denoted by r_N and r_F, respectively. The transmit power control is employed to guarantee the reliability of scheduled NU-FU pair with minimized power consumption.

In the considered system, each of the stations has only one antenna, working in a half-duplex mode. The channel coefficients of links $S-N_m$, $S-F_k$ and N_m-F_k are denoted by $g_{S,m}$, $g_{S,k}$ and $g_{m,k}$, respectively. Channels over all the links experience independent but non-identically distributed Rayleigh block-fading. Thus, channel gains $|g_{S,m}|^2$, $|g_{S,k}|^2$ and $|g_{m,k}|^2$ follow exponential distribution with mean values $G_{S,m}$, $G_{S,k}$ and $G_{m,k}$, respectively, and remain unchanged within each transmission block (but may vary independently over different transmission blocks). Channel reciprocity is also assumed for our considered system. The maximal transmit power of the BS is denoted by P_S, while the maximal transmit power of each NU is denoted by P_N. The noise is modeled as additive white Gaussian noise (AWGN) at each receiver with an identical variance σ^2. The transmit signal-to-noise ratio (SNR) of the system is defined as $\rho = P_S/\sigma^2$.

III. Proposed Opportunistic Adaptive NOMA Strategy

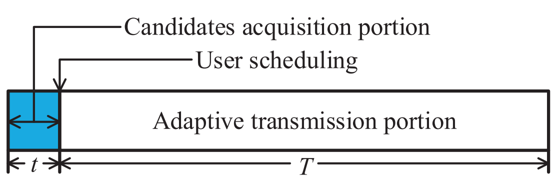


Fig. 2. The structure of a transmission block of OA-NOMA strategy.

This section proposes a novel downlink NOMA transmission strategy for the considered system. In this strategy, each transmission block is partitioned into two portions: the *candidates acquisition portion* with a duration t and the *adaptive transmission portion* with a duration T, as shown in Fig. 2. During the candidates acquisition portion, an SCA method is employed to acquire the necessary instantaneous CSI, with which the power allocation (PA) coefficients of each NU-FU pair are determined for transmit power minimization. Then, with the obtained instantaneous CSI and PA coefficients, the BS can find out the candidate NU-FU pairs for scheduling. Finally, among all candidate NU-FU pairs, a P-UPS scheme is performed to schedule one of them at the end of the candidates acquisition portion. During the adaptive transmission portion, the downlink transmission to the scheduled NU-FU pair adaptively operates in either *the direct NOMA mode* or *the cooperative NOMA mode*. In particular, considering that NU-FU pair $[N_m, F_k]$ is scheduled, the purpose of the two transmission modes can be briefly described as follows.

- Direct NOMA mode: When the direct link to FU F_k is in a good channel condition, i.e., $\rho|g_{S,k}|^2$ is no less than a predetermined threshold Ξ, the BS directly serves both NU N_m and FU F_k by using NOMA, in order to avoid unnecessary cooperation;
- Cooperative NOMA mode: When the direct link to FU F_k is in a bad channel condition, i.e., $\rho|g_{S,k}|^2$ is less than Ξ, the BS sends to NU N_m the messages intended by both NU N_m and FU F_k, and then, recruits NU N_m as a relay to forward message intended by FU F_k, in order to improve the reception quality of FU F_k.

To sum up, in the adaptive transmission portion of each transmission block, the BS *opportunistically* serves an NU-FU pair scheduled from NUs and FUs, and the transmission to the scheduled NU-FU pair *adaptively* switches between the direct NOMA mode and the cooperative NOMA mode. Thus, the proposed strategy is termed as *OA-NOMA strategy*. In the following, for a better understanding of the OA-NOMA strategy, we first detail the two NOMA transmission modes of the adaptive transmission portion, and then present the SCA method and the P-UPS scheme for the candidates acquisition portion.

A. Adaptive Transmission Portion

This subsection presents the transmission procedure, the power control approach and the reliability condition for both the direct NOMA mode and the cooperative NOMA mode. For the clarity of presentation, we assume that NU-FU pair $[N_m, F_k]$ is scheduled in the candidates acquisition portion. The desired messages of NU N_m and FU F_k are denoted by x_N and x_F, respectively.

1) Direct NOMA Mode: When $\rho|g_{S,k}|^2 \geq \Xi$, i.e., the direct link to F_k is in a relatively good condition, the downlink transmission to NU-FU pair $[N_m, F_k]$ operates in the direct NOMA mode. In this mode, the BS sends superimposed signals $\sqrt{P_S\alpha_N}x_N + \sqrt{P_S\alpha_F}x_F$ to both scheduled users, where α_N and α_F are the PA coefficients of messages x_N and x_F, respectively. Since the power control is employed, the BS may not need to fully use its available transmit power,

[1]Theoretically, by using NOMA, even more than two users can be served simultaneously. However, in this work, we consider that only an NU-FU pair is served in each downlink transmission due to the following reasons: 1) As the NUs and the FUs are expected to have very different downlink channel quality, performing NOMA between an NU and an FU can fully utilize the performance gain of NOMA over conventional OMA [19]; 2) Compared with superposition coding over a large number of messages, the superposition coding over two messages can reduce the complexity and error propagation of SIC-based detection [27]. Therefore, two-message NOMA transmission to a scheduled NU-FU pair can exploit the performance gain of NOMA over conventional OMA while avoiding the high complexity and error propagation.

indicating that $\alpha_N + \alpha_F \leq 1$.[2] Once the superimposed signals are received, FU F_k tries to decode its desired message x_F with the signal-to-interference-plus-noise ratio (SINR) given by

$$\gamma^{\text{dir}}_{S\to F_k,x_F} = \frac{\alpha_F \rho |g_{S,k}|^2}{\alpha_N \rho |g_{S,k}|^2 + 1}, \tag{1}$$

where the superscript "dir" indicates the direct NOMA mode. Meanwhile, NU N_m sequentially detects messages x_F and x_N from its received signals. First, NU N_m attempts to decode message x_F with an SINR shown as

$$\gamma^{\text{dir}}_{S\to N_m,x_F} = \frac{\alpha_F \rho |g_{S,m}|^2}{\alpha_N \rho |g_{S,m}|^2 + 1}. \tag{2}$$

If message x_F has been correctly decoded, NU N_m removes this message by using SIC and then tries to decode message x_N with an SNR given by

$$\gamma^{\text{dir}}_{S\to N_m,x_N} = \alpha_N \rho |g_{S,m}|^2. \tag{3}$$

Next, we present the power control approach that aims at guaranteeing the reliability of the scheduled NU-FU pair with minimized power consumption. Since target rates of messages x_N and x_F are r_N and r_F, respectively, NU-FU pair $[N_m, F_k]$ can be reliably served in the direct NOMA mode if the following inequalities hold simultaneously

$$\begin{cases} \min\left\{\log\left(1+\gamma^{\text{dir}}_{S\to F_k,x_F}\right), \log\left(1+\gamma^{\text{dir}}_{S\to N_m,x_F}\right)\right\} \geq r_F, \\ \log\left(1+\gamma^{\text{dir}}_{S\to N_m,x_N}\right) \geq r_N. \end{cases} \tag{4}$$

Here, the first inequality means that message x_F is correctly decoded by both users, while the second inequality means that message x_N is correctly decoded by NU N_m. Using (1), (2) and (3) and with some algebraic manipulations, the inequalities in (4) can be rewritten as

$$\begin{cases} \alpha_F \geq \tau_F \left(\alpha_N + \dfrac{1}{\min\{\rho|g_{S,m}|^2, \rho|g_{S,k}|^2\}}\right), \\ \alpha_N \geq \dfrac{\tau_N}{\rho |g_{S,m}|^2}, \end{cases} \tag{5}$$

where $\tau_N \triangleq 2^{r_N} - 1$ and $\tau_F \triangleq 2^{r_F} - 1$. Apparently, the minimal value for α_N that ensures successful detection of x_N is given by

$$\alpha_N = \varphi_m \triangleq \frac{\tau_N}{\rho |g_{S,m}|^2}. \tag{6}$$

Applying (6) into the first inequality of (5), the minimal value for α_F that ensures successful detection of x_F is obtained as

$$\alpha_F = \xi_{m,k} \triangleq \frac{\tau_F}{\min\{\rho|g_{S,k}|^2, \rho|g_{S,m}|^2\}} + \frac{\tau_F \tau_N}{\rho |g_{S,m}|^2}. \tag{7}$$

Moreover, since the PA coefficients should satisfy $\alpha_N + \alpha_F \leq 1$, the minimized PA coefficients in (6) and (7) are feasible only when $\varphi_m + \xi_{m,k} \leq 1$ holds. Therefore, in the direct NOMA mode, when $\varphi_m + \xi_{m,k} \leq 1$ holds, the BS should employ (6) and (7) to minimize the power consumption with guaranteeing the reliability requirement of NU-FU pair $[N_m, F_k]$. Otherwise, when $\varphi_m + \xi_{m,k} > 1$, no feasible PA is available to guarantee the reliability requirement of NU-FU pair $[N_m, F_k]$.

Recall that when $\rho|g_{S,k}|^2 \geq \Xi$, the downlink transmission to NU-FU pair $[N_m, F_k]$ operates in the direct NOMA mode. Therefore, when NU-FU pair $[N_m, F_k]$ is scheduled, the reliability condition of the direct NOMA mode, denoted by $\mathbb{R}^{\text{dir}}_{m,k}$, can be expressed as

$$\mathbb{R}^{\text{dir}}_{m,k} \triangleq \{\rho|g_{S,k}|^2 \geq \Xi, \varphi_m + \xi_{m,k} \leq 1\}. \tag{8}$$

2) Cooperative NOMA Mode: When $\rho|g_{S,k}|^2 < \Xi$, i.e., the direct link to F_k is in a relatively bad channel condition, the downlink transmission to NU-FU pair $[N_m, F_k]$ operates in the cooperative NOMA mode. In this mode, the adaptive transmission portion is divided into two phases, each with duration $T/2$. During the first phase, the BS sends superimposed signals $\sqrt{P_S \beta_N} x_N + \sqrt{P_S \beta_F} x_F$ to NU N_m, where β_N and β_F represent the PA coefficient of messages x_N and x_F, respectively. Similar to the direct NOMA mode, as the transmit power control is employed, the BS may not fully utilize its transmit power, meaning that PA coefficients β_N and β_F satisfy $\beta_N + \beta_F \leq 1$. After receiving the superimposed signals, NU N_m attempts to successively decode both messages by using SIC. First, NU N_m tries to decode message x_F with an SINR given by

$$\gamma^{\text{coop}}_{S\to N_m,x_F} = \frac{\beta_F \rho |g_{S,m}|^2}{\beta_N \rho |g_{S,m}|^2 + 1}, \tag{9}$$

where superscript "coop" represents cooperative NOMA mode. If message x_F has been correctly decoded, NU N_m performs SIC and tries to decode message x_N with an SNR:

$$\gamma^{\text{coop}}_{S\to N_m,x_N} = \beta_N \rho |g_{S,m}|^2. \tag{10}$$

If NU N_m has successfully decoded x_F, it re-encodes and forwards[3] message x_F to FU F_k by sending a signal $\sqrt{\theta_F P_N} x_F$, where θ_F is the PA coefficient for the forwarded message x_F. Since the power control is employed, the scheduled NU may partially use its transmit power, implying that $\theta_F \leq 1$ holds. Defining $\mu \triangleq P_N/P_S$, the received SNR for FU F_k to decode message x_F is given by[4]

$$\gamma^{\text{coop}}_{N_m\to F_k,x_F} = \theta_F \mu \rho |g_{m,k}|^2. \tag{11}$$

[2]For many existing works, e.g. [14], [19]–[22], it is assumed that $\alpha_N + \alpha_F = 1$, indicating that the available transmit power is always fully used. However, in our system, to meet the target rates of scheduled users, fully using all available transmit power may lead to the waste of transmit power, since it is possible that the available transmit power may be very large in practical wireless systems. Thus, in this paper, we consider a more general setup $\alpha_N + \alpha_F \leq 1$, and then choose α_N and α_F that can meet the target rate requirements with minimal power consumption.

[3]In this work, we consider that the scheduled NU is always willing to act as a relay and forward information to the scheduled FU in the cooperative NOMA mode. The incentive for stimulating cooperation can be handled by a higher-layer control mechanism. In particular, such user cooperation can be encouraged by designing advanced rewarding and pricing strategies. For example, a user which helps other users will be offered rewards, which will be helpful to the user to get help from other users in the future. The design of such rewarding and pricing strategies is an important direction for future research, but will be out of the scope of this paper.

[4]For operational simplicity, signal combining is not considered at the scheduled FU. Thus, the results in this paper can be viewed as a lower bound on the system performance with signal combining. Note that this lower bound becomes tighter when the links from the BS to FUs become weaker. When the direct links from the BS to FUs do not exist (e.g., due to blockage), this lower bound is exactly the performance achieved by the system.

To guarantee the reliability of scheduled NU-FU pair with minimized power consumption, the power control approach for the cooperative NOMA mode is presented as follows. First, with target rates r_N and r_F, guaranteeing the reliability of NU-FU pair $[N_m, F_k]$ requires that the following inequalities hold simultaneously

$$\begin{cases} \frac{1}{2}\log\left(1+\gamma^{\rm coop}_{S\to N_m,x_F}\right)\geq r_F, \\ \frac{1}{2}\log\left(1+\gamma^{\rm coop}_{N_m\to F_k,x_F}\right)\geq r_F, \\ \frac{1}{2}\log\left(1+\gamma_{S\to N_m,x_N}\right)\geq r_N, \end{cases} \tag{12}$$

where the first two inequalities represent that message x_F can be successfully decoded by the scheduled NU and the scheduled FU, respectively, and the third inequality represents that message x_N can be successfully decoded by the scheduled NU. Using (9), (10) and (11) with some algebraic manipulations, the inequalities in (12) can be re-expressed as

$$\begin{cases} \beta_F \geq \lambda_F\left(\beta_N + \frac{1}{\rho|g_{S,m}|^2}\right), \\ \theta_F \geq \frac{\lambda_F}{\mu\rho|g_{m,k}|^2}, \\ \beta_N \geq \frac{\lambda_N}{\rho|g_{S,m}|^2}, \end{cases} \tag{13}$$

where $\lambda_N \triangleq 2^{2r_N}-1$ and $\lambda_F \triangleq 2^{2r_F}-1$. As known from the second and third inequalities of (13), the minimal values for θ_F and β_N are given by

$$\theta_F = \omega_{m,k} \triangleq \frac{\lambda_F}{\mu\rho|g_{m,k}|^2}, \tag{14}$$

$$\beta_N = \psi_m \triangleq \frac{\lambda_N}{\rho|g_{S,m}|^2}. \tag{15}$$

Further, substituting (15) into the first inequality of (13), the minimal value for β_F is obtained as

$$\beta_F = \kappa_m \triangleq \frac{\lambda_F\lambda_N+\lambda_F}{\rho|g_{S,m}|^2}. \tag{16}$$

Recall that the PA coefficients should satisfy $\beta_N+\beta_F \leq 1$ and $\theta_F \leq 1$. Thus, the minimized PA coefficients shown in (14), (15) and (16) are feasible only when $\psi_m+\kappa_m \leq 1$ and $\omega_{m,k}\leq 1$ hold simultaneously. Or in other words, when both $\psi_m+\kappa_m\leq 1$ and $\omega_{m,k}\leq 1$ hold, the minimized PA coefficients in (14), (15) and (16) can guarantee the reliability of NU-FU pair $[N_m, F_k]$ in the cooperative NOMA mode. Otherwise, when either $\psi_m+\kappa_m > 1$ or $\omega_{m,k}>1$ holds, no feasible PA is available to guarantee the reliability of NU-FU pair $[N_m, F_k]$ in the cooperative NOMA mode.

Recall that when $\rho|g_{S,k}|^2 < \Xi$, the downlink transmission to NU-FU pair $[N_m, F_k]$ operates in the cooperative NOMA mode. Therefore, when NU-FU pair $[N_m, F_k]$ is scheduled, the reliability condition of the cooperative NOMA mode, denoted by $\mathbb{R}^{\rm coop}_{m,k}$, can be derived as follows

$$\mathbb{R}^{\rm coop}_{m,k} \triangleq \{\rho|g_{S,k}|^2<\Xi, \psi_m+\kappa_m\leq 1, \omega_{m,k}\leq 1\}. \tag{17}$$

B. Candidates Acquisition Portion

This subsection presents the detailed procedure of the SCA method, and then, proposes the P-UPS scheme for NU-FU pair scheduling.

Algorithm 1 Scheduling Candidates Acquisition Method

Step 1	In mini-slot k ($k\in\mathbf{K}$), FU F_k broadcasts a flag to the BS and all NUs. By reception of this flag, the BS estimates channel gain $|g_{S,k}|^2$, while NU N_m ($m\in\mathbf{M}$) estimates channel gain $|g_{m,k}|^2$.
Step 2	At the end of mini-slot K, each NU, say N_m, computes the values of $\omega_{m,k}$ for $k\in\mathbf{K}$ by applying channel gain $|g_{m,k}|^2$ into (14). Then, NU N_m generates K bits feedback information, where the kth bit is 1 if $\omega_{m,k}\leq 1$ holds, or 0 otherwise.
Step 3	In mini-slot $K+m$ ($m\in\mathbf{M}$), NU N_m sends its feedback information to the BS. Upon receiving this feedback information, the BS measures the reception to estimate channel gain $|g_{S,m}|^2$. Then, by further decoding this information, the BS can know whether $\omega_{m,k}\leq 1$ holds for $k\in\mathbf{K}$.
Step 4	At the end of mini-slot $K+M$, the BS has known channel gains $|g_{S,k}|^2$ for $k\in\mathbf{K}$ (from Step 1) and $|g_{S,m}|^2$ for $m\in\mathbf{M}$ (from Step 3). Applying these channel gains into (6), (7), (15) and (16), the BS computes the values of φ_m, $\xi_{m,k}$, ψ_m and κ_m for $m\in\mathbf{M}$ and $k\in\mathbf{K}$. Then, combining these values with the decoded feedback information from each NU, the BS further obtains $\mathbf{P}$ by using (18).

1) SCA Method: Recall that the downlink transmission intends to serve the scheduled NU-FU pair with predetermined target rates r_N and r_F. Thus, to guarantee the reliability, an NU-FU pair is eligible for being scheduled only when both users can be reliably served in either the direct NOMA mode or the cooperative NOMA mode, referred to as a *candidate NU-FU pair*. Based on reliability conditions in (8) and (17), the indices set of candidate NU-FU pairs can be expressed as

$$\begin{aligned} \mathbf{P} \triangleq & \left\{[m,k] | \mathbb{R}^{\rm dir}_{m,k}\cup\mathbb{R}^{\rm coop}_{m,k}, m\in\mathbf{M}, k\in\mathbf{K}\right\} \\ = & \Big\{[m,k]\Big|\{\rho|g_{S,k}|^2\geq\Xi, \varphi_m+\xi_{m,k}\leq 1\}\cup \\ & \{\rho|g_{S,k}|^2<\Xi, \psi_m+\kappa_m\leq 1, \omega_{m,k}\leq 1\}, \\ & m\in\mathbf{M}, k\in\mathbf{K}\Big\}. \end{aligned} \tag{18}$$

To find out all candidate NU-FU pairs for scheduling, the BS should acquire the information of set $\mathbf{P}$ during the candidate acquisition portion, which can be realized by the SCA method shown in Algorithm 1. In the SCA method, the candidate acquisition portion is divided into $(K+M)$ mini-slots[5], where

[5]As seen in Algorithm 1, the SCA method requires $M+K$ mini-slots to determine the candidate NU-FU pairs for the scheduling over a certain resource block. Thus, the number of mini-slots required by the SCA method increases with the number of users. Note that we consider the NU-FU pair scheduling over *a single resource block*. Actually, a practical system could have *multiple resource blocks*. To implement our proposed scheme in such a practical system, we first should partition all users in the system into multiple groups, and a resource block is assigned to each group (similar to the hybrid NOMA scheme shown in [19]). Then our scheme can work in each group. In this setting, each group has a moderate number of users and thereby the number of mini-slots required by the proposed method can be controlled to an acceptable level.

the duration of each mini-slot is $t/(K+M)$. Since very limited information is exchanged in every mini-slot, strong channel coding can be used to improve the robustness of information exchange. Thus, it is reasonable to assume that the information exchange is error-free in the SCA method.

Note that with the SCA method, the BS can determine the PA coefficients $\alpha_F = \xi_{m,k}$, $\alpha_N = \varphi_m$, $\beta_N = \psi_m$ and $\beta_F = \kappa_m$ (in Step 4) for $m \in \mathbf{M}$ and $k \in \mathbf{K}$, while each NU $N_m (m \in \mathbf{M})$ can determine the PA coefficients $\theta_F = \omega_{m,k}$ for $k \in \mathbf{K}$ (in Step 2). Therefore, the power control can be realized for any candidate NU-FU pair. Moreover, as the BS acquires channel gains $|g_{S,k}|^2$, $k \in \mathbf{K}$ in Step 1, it can adaptively determine the NOMA transmission mode (direct or cooperative) for the scheduled NU-FU pair during the adaptive transmission portion.

Computational complexity and communication overhead: As shown in Algorithm 1, each NU N_m ($m \in \mathbf{M}$) needs to calculate the value of $\omega_{m,k}$ for $k \in \mathbf{K}$ in Step 2, while the BS needs to calculate the values of φ_m, $\xi_{m,k}$, ψ_m and κ_m for $m \in \mathbf{M}$ and $k \in \mathbf{K}$ in Step 4. Therefore, the overall computational complexity is $\mathcal{O}(MK)$. On the other hand, each FU needs to broadcast a flag in Step 1, while each NU sends its feedback information only once in Step 3, indicating that the overall communication overhead is $K + M$.

2) P-UPS Scheme: At the end of the candidates acquisition portion, the user pair scheduling is performed among all candidate NU-FU pairs. In the following, we will propose a P-UPS scheme for condition $\{\mathbf{P} \neq \emptyset\}$, aiming at improving the scheduling fairness.

When $\{\mathbf{P} \neq \emptyset\}$ happens, the BS assigns a *scheduling probability* for each candidate NU-FU pair. More specifically, by introducing a fairness exponent $\eta(> 1)$, the scheduling probability for candidate NU-FU $[N_m, F_k]$ is defined as

$$SP_{m,k}(\mathbf{P}) \triangleq \frac{CP_{m,k}^{-\eta}}{\sum_{[i,j]\in\mathbf{P}} CP_{i,j}^{-\eta}}, \forall [m,k] \in \mathbf{P}, \tag{19}$$

with $CP_{m,k} \triangleq \Pr\{\mathbb{R}_{m,k}^{\text{dir}} \cup \mathbb{R}_{m,k}^{\text{coop}}\}$. Here, $CP_{m,k}$ represents the probability that $[N_m, F_k]$ becomes a candidate NU-FU pair in a long term, called *candidate probability* of NU-FU pair $[N_m, F_k]$. Further, since the global statistical CSI is available at the BS, the candidate probabilities of all NU-FU pairs can be numerically computed by the BS by using the derivations in Appendix A. Then, with the assigned scheduling probability of each candidate NU-FU pair, the BS probabilistically schedules one of candidate NU-FU pairs such that candidate NU-FU pair $\{N_m, F_k\}$ is scheduled with probability $SP_{m,k}(\mathbf{P})$, for $[m, k] \in \mathbf{P}$.

The rationale of the above P-UPS scheme is as follows. As observed from (19), the scheduling probability of an NU-FU pair is monotonically decreasing with its candidate probability. Thus, in a certain transmission block, a candidate NU-FU pair is scheduled with a low probability if it has a high chance to become a candidate for scheduling in a long term, and vice versa. Therefore, the proposed P-UPS scheme can improve the long-term scheduling fairness among all NU-FU pairs.

Additionally, if no candidate NU-FU user pair exists, i.e., $\{\mathbf{P} = \emptyset\}$ happens, no NU-FU pair can be reliably served in either the direct NOMA mode or the cooperative NOMA mode. In this condition, the user pair scheduling will be cancelled, and then, an *outage* will be declared[6].

IV. Max-Min Fairness Criterion

This section investigates the scheduling fairness of proposed OA-NOMA strategy from the perspective of successful serving probability (SSP) of each NU-FU pair. Here, the SSP of NU-FU pair $[N_m, F_k]$ refers to the probability that NU-FU pair $[N_m, F_k]$ is successfully served, denoted by $SSP_{m,k}$. Note that the sum SSP of all NU-FU pairs should satisfy $\sum_{m=1}^{M}\sum_{k=1}^{K} SSP_{m,k} = 1 - P_{\text{out}}$, where $P_{\text{out}} = \Pr\{\mathbf{P} = \emptyset\}$ represents the outage probability of OA-NOMA strategy and will be theoretically evaluated in Section V-A. Further, recall that with the P-UPS scheme, an NU-FU pair has chance to be successfully served only when this NU-FU pair is a candidate NU-FU pair in the current transmission block. Thus, the SSP of NU-FU pair $[N_m, F_k]$ is bounded by its candidate probability, i.e., $SSP_{m,k} \leq CP_{m,k}$.

According to the definition of max-min fairness [31], [32], the max-min fairness for NU-FU pair scheduling in terms of SSP can be defined as follows: Let us treat $1 - P_{\text{out}}$ as the resource and treat the SSPs of NU-FU pairs as a feasible allocation of $1 - P_{\text{out}}$ among all NU-FU pairs. The SSPs of NU-FU pairs are max-min fair, if and only if for every NU-FU pair $[N_i, F_j]$, $SSP_{i,j}$ can only be increased (while remaining $SSP_{i,j} \leq CP_{i,j}$) at the expense of decreasing $SSP_{i',j'} (\leq SSP_{i,j}, [i'j'] \neq [i,j])$.

***Theorem** 1 (Max-min fairness criterion):* When the SSPs of NU-FU pairs are max-min fair, the SSP of NU-FU pair $[N_m, F_k]$ can be expressed as $SSP_{m,k}^{\text{max-min}} = \min\{\zeta, CP_{m,k}\}$ for $m \in \mathbf{M}$ and $k \in \mathbf{K}$, where ζ is a positive constant satisfying $\sum_{m=1}^{M}\sum_{k=1}^{K} \min\{\zeta, CP_{m,k}\} = 1 - P_{\text{out}}$.

Proof: We use the proof by contradiction. Consider a feasible allocation with $SSP_{m,k}^{\text{max-min}} = \min\{\zeta, CP_{m,k}\}$, for $m \in \mathbf{M}, k \in \mathbf{K}$. Assume this allocation is not max-min fair. Then there exist $[i,j]$ and $[i',j']$ ($[i,j] \neq [i',j']$) such that $SSP_{i,j}^{\text{max-min}}$ can be increased to $SSP_{i,j}^{\text{max-min}} + \delta_1$ (while keeping $SSP_{i,j}^{\text{max-min}} + \delta_1 \leq CP_{i,j}$) at the expense of decreasing $SSP_{i',j'}^{\text{max-min}}$ to $SSP_{i',j'}^{\text{max-min}} - \delta_2$, where $SSP_{i',j'}^{\text{max-min}} > SSP_{i,j}^{\text{max-min}}$ holds, δ_1 and δ_2 are sufficiently small positive constants. Since $SSP_{i,j}^{\text{max-min}} = \min\{\zeta, CP_{i,j}\}$ and $SSP_{i,j}^{\text{max-min}} + \delta_1 \leq CP_{i,j}$, it is obtained $SSP_{i,j}^{\text{max-min}} = \zeta$. Further, as $SSP_{i',j'}^{\text{max-min}} = \min\{\zeta, CP_{i',j'}\}$ holds, we know $SSP_{i',j'}^{\text{max-min}} \leq \zeta = SSP_{i,j}^{\text{max-min}}$. This contradicts the fact $SSP_{i',j'}^{\text{max-min}} > SSP_{i,j}^{\text{max-min}}$. This completes the proof. ■

It can be observed that the value of $\sum_{m=1}^{M}\sum_{k=1}^{K} \min\{\zeta, CP_{m,k}\}$ is monotonically non-decreasing with ζ. Thus, based on the fact $\sum_{m=1}^{M}\sum_{k=1}^{K} \min\{\zeta, CP_{m,k}\} = 1 - P_{\text{out}}$, the value of ζ can be determined by a bisection search, with which the max-min fair SSP of each NU-FU pair is obtained. Moreover, the max-min fairness criterion given in Theorem 1

[6]When $\{\mathbf{P} = \emptyset\}$ happens, the system can still schedule one NU or one FU for conventional OMA transmission. However, the user scheduling of OMA systems has been extensively studied in existing literature, such as [28]–[30]. Therefore, in this work, we do not consider the OMA user scheduling and performance analysis for this condition.

will be used as a benchmark for numerically evaluating the scheduling fairness in Section VI.

On the other hand, the proportional fairness is another widely adopted fairness metric for user scheduling. Differing from the max-min fairness that purely focuses on fairness, the proportional fairness can achieve a good tradeoff between the performance and fairness. Fortunately, as demonstrated in [33], the max-min fairness may become proportional fairness under certain conditions. In particular, for the considered NOMA system that uses SSP to evaluate the scheduling fairness, the developed max-min fairness criterion is equivalent to the proportional fairness criterion, as shown in the following Lemma.

***Lemma** 1:* For the considered system, the max-min fair SSPs shown in Theorem 1 are also proportionally fair.

Proof: Please refer to Appendix B. ∎

V. Performance Analysis

In this section, the reliability of the proposed OA-NOMA s-trategy is theoretically evaluated in terms of outage probability and diversity order.

A. Outage Probability

Recall that an outage is declared when no candidate NU-FU pair exists, i.e., $\mathbf{P}=\emptyset$. To proceed the analysis, we introduce an indices set $\mathbf{F}_\Xi \triangleq \{k|\rho|g_{S,k}|^2 \geq \Xi, k\in\mathbf{K}\}$, which is the set of FUs with relatively good channel conditions from the BS. Thus, based on Total Probability Theorem, the outage probability can be expressed as

$$P_{\rm out}=\Pr\{\mathbf{P}=\emptyset\}=\underbrace{\Pr\left\{\bigcap_{k=1}^{K}\mathbb{E}_{1,k},\mathbf{F}_\Xi=\emptyset\right\}}_{\triangleq P_{\rm out}(\emptyset)} \tag{20}$$

$$+\sum_{i=1}^{K}\sum_{\mathbf{A}_i\subseteq\mathbf{K},|\mathbf{A}_i|=i}\underbrace{\Pr\left\{\bigcap_{k\in\mathbf{K}\setminus\mathbf{A}_i}\mathbb{E}_{1,k},\bigcap_{k\in\mathbf{A}_i}\mathbb{E}_{2,k},\mathbf{F}_\Xi=\mathbf{A}_i\right\}}_{\triangleq P_{\rm out}(\mathbf{A}_i)},$$

with $\mathbb{E}_{1,k}$ and $\mathbb{E}_{2,k}$ defined as

$$\begin{aligned}\mathbb{E}_{1,k}&\triangleq\left\{\bigcap_{m=1}^{M}\left[(\psi_m+\kappa_m>1)\cup(\omega_{m,k}>1)\right]\right\} \\ &\overset{\text{(i)}}{=}\left\{\bigcap_{m=1}^{M}\left[\left(\frac{\Lambda}{\rho|g_{S,m}|^2}>1\right)\cup\left(\frac{\lambda_F}{\mu\rho|g_{m,k}|^2}>1\right)\right]\right\},\end{aligned} \tag{21}$$

$$\begin{aligned}\mathbb{E}_{2,k}&\triangleq\left\{\bigcap_{m=1}^{M}\varphi_m+\xi_{m,k}>1\right\} \\ &\overset{\text{(ii)}}{=}\left\{\bigcap_{m=1}^{M}\max\left\{\frac{\Upsilon}{\rho|g_{S,m}|^2},\frac{\tau}{\rho|g_{S,m}|^2}+\frac{\tau_F}{\rho|g_{S,k}|^2}\right\}>1\right\},\end{aligned} \tag{22}$$

where $\mathbf{A}_i$ represents a subset of $\mathbf{K}$ with cardinality $|\mathbf{A}_i|=i$, terms Λ, Υ and τ have been defined in Appendix A. Note that step (i) uses (14), (15) and (16), while step (ii) uses (6) and (7).

In the outage probability expression in (20), the term $P_{\rm out}(\emptyset)$ refers to the case when all FUs have relatively bad channel conditions from the BS, and $\mathbb{E}_{1,k}$ means that FU F_k (which has a relatively bad channel condition) cannot form a candidate NU-FU pair with any NU by using the cooperative NOMA mode. The term $P_{\rm out}(\mathbf{A}_i)$ refers to the case when FUs in set $\mathbf{A}_i$ have relatively good channel conditions from the BS, and other FUs have relatively bad channel conditions. $\mathbb{E}_{2,k}$ means that FU F_k (which has a relatively good channel condition) cannot form a candidate NU-FU pair with any NU by using the direct NOMA mode.

***Lemma** 2:* A closed-form expression for $P_{\rm out}(\emptyset)$ is given by

$$\begin{aligned}P_{\rm out}(\emptyset)=&\prod_{m=1}^{M}\left[\left(1-e^{-\frac{\Lambda}{\rho G_{S,m}}}\right)+e^{-\frac{\Lambda}{\rho G_{S,m}}}\right. \\ &\left.\times\prod_{k=1}^{K}\left(1-e^{-\frac{\lambda_F}{\rho G_{m,k}}}\right)\right]\prod_{k=1}^{K}\left(1-e^{-\frac{\Xi}{\rho G_{S,k}}}\right),\end{aligned} \tag{23}$$

while $P_{\rm out}(\mathbf{A}_i)$ can be expressed as

$$\begin{aligned}&P_{\rm out}(\mathbf{A}_i) \\ &=\Phi_1+\sum_{j=1}^{M}\sum_{\mathbf{B}_j\subseteq\mathbf{M},|\mathbf{B}_j|=j}\sum_{l=0}^{j}\sum_{\mathbf{C}_l\subseteq\mathbf{B}_j,|\mathbf{C}_l|=l}\Phi_2\left(\Phi_3-\Phi_4\cdot\chi\right),\end{aligned} \tag{24}$$

with Φ_1, Φ_2, Φ_3, Φ_4 and χ [7] being given by

$$\begin{aligned}\Phi_1=&e^{-\sum_{k\in\mathbf{A}_i}\frac{\Xi}{\rho G_{S,k}}}\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\left(1-e^{-\frac{\Xi}{\rho G_{S,k}}}\right) \\ &\times\prod_{m=1}^{M}\left(1-e^{-\frac{\Upsilon}{\rho G_{S,m}}}\right),\end{aligned} \tag{25}$$

$$\Phi_2=\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\left[\prod_{m\in\mathbf{C}_l}\left(1-e^{-\frac{\lambda_F}{\mu\rho G_{m,k}}}\right)\right], \tag{26}$$

$$\begin{aligned}\Phi_3=&e^{-\sum_{k\in\mathbf{A}_i}\frac{\Xi}{\rho G_{S,k}}-\sum_{m\in\mathbf{C}_l}\frac{\Lambda}{\rho G_{S,m}}}\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\left(1-e^{-\frac{\Xi}{\rho G_{S,k}}}\right) \\ &\times\left[\prod_{m\in\mathbf{B}_j\setminus\mathbf{C}_l}\left(e^{-\frac{\Upsilon}{\rho G_{S,m}}}-e^{-\frac{\Lambda}{\rho G_{S,m}}}\right)\right. \\ &\left.\times\prod_{m\in\mathbf{M}\setminus\mathbf{B}_j}\left(1-e^{-\frac{\Upsilon}{\rho G_{S,m}}}\right)\right],\end{aligned} \tag{27}$$

$$\Phi_4=\prod_{m\in\mathbf{M}\setminus\mathbf{B}_j}\left(1-e^{-\frac{\Upsilon}{\rho G_{S,m}}}\right)\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\left(1-e^{-\frac{\Xi}{\rho G_{S,k}}}\right), \tag{28}$$

$$\begin{aligned}\chi\triangleq\Pr\Bigg\{&\frac{\tau}{\max\limits_{m\in\mathbf{B}_j}\rho|g_{S,m}|^2}+\frac{\tau_F}{\max\limits_{k\in\mathbf{A}_i}\rho|g_{S,k}|^2}\leq 1, \\ &\bigcap_{k\in\mathbf{A}_i}\rho|g_{S,k}|^2\geq\Xi,\bigcap_{m\in\mathbf{C}_l}\rho|g_{S,m}|^2\geq\Lambda,\end{aligned}$$

[7]For presentation simplicity, for notation χ, we omit its variables i,j,l.

$$\bigcap_{m\in\mathbf{B}_j\backslash\mathbf{C}_l} \Upsilon \le \rho|g_{S,m}|^2 < \Lambda \Bigg\}, \tag{29}$$

where $\mathbf{B}_j$ represents a subset of $\mathbf{M}$ with cardinality $|\mathbf{B}_j| = j$, and $\mathbf{C}_l$ represents a subset of $\mathbf{B}_j$ with cardinality $|\mathbf{C}_l| = l$.

Proof: Please refer to Appendix C. ∎

As seen from (29), the events inside $\Pr\{\cdot\}$ are highly correlated. Thus, it is very challenging to further derive χ in closed form. Alternatively, a tight approximation for χ will be provided as follows.

First, the probability denoted by χ can be re-expressed in the following lemma.

Lemma 3: Considering different combinations of $\{i, j, l\}$, the probability denoted by χ can be re-expressed in eight cases shown in (30) (at the top of the next page), with $\varrho_{j,l}$ being defined as

$$\varrho_{j,l} \triangleq \prod_{m\in\mathbf{B}_j\backslash\mathbf{C}_l} [e^{-\Upsilon/(\rho G_{S,m})} - e^{-\Lambda/(\rho G_{S,m})}], \tag{31}$$

where RVs X_1, X_2, Y_1, Y_2, Z_1 and Z_2 are defined as

$$\begin{cases} X_1 \triangleq \min\limits_{m\in\mathbf{B}_j} \rho|g_{S,m}|^2, & X_2 \triangleq \max\limits_{m\in\mathbf{B}_j} \rho|g_{S,m}|^2, \\ Y_1 \triangleq \min\limits_{k\in\mathbf{A}_i} \rho|g_{S,k}|^2, & Y_2 \triangleq \max\limits_{k\in\mathbf{A}_i} \rho|g_{S,k}|^2, \\ Z_1 \triangleq \min\limits_{m\in\mathbf{C}_l} \rho|g_{S,m}|^2, & Z_2 \triangleq \max\limits_{m\in\mathbf{C}_l} \rho|g_{S,m}|^2, \end{cases} \tag{32}$$

and RVs X_0, Y_0, and Z_0 are defined as follows. If $j = 1$, $\mathbf{B}_j$ only has one element, denoted by m_1, and then we define $X_0 \triangleq \rho|g_{S,m_1}|^2$. If $i = 1$, $\mathbf{A}_i$ only has one element, denoted by k_1, and then we define $Y_0 \triangleq \rho|g_{S,k_1}|^2$. If $l = 1$, $\mathbf{C}_l$ only has one element, denoted by m_1', and then we define $Z_0 \triangleq \rho|g_{S,m_1'}|^2$.

Proof: Please refer to Appendix D. ∎

Next, we focus on deriving the PDFs involved in (30). Recall that channel gains $|g_{S,m}|^2$ and $|g_{S,k}|^2$ are exponentially distributed with mean values $G_{S,m}$ and $G_{S,k}$, respectively. Thus, we have

$$\begin{cases} f_{X_0}(x_0) = \frac{1}{\rho G_{S,m_1}} e^{-\frac{x_0}{\rho G_{S,m_1}}}, \\ f_{Y_0}(y_0) = \frac{1}{\rho G_{S,k_1}} e^{-\frac{y_0}{\rho G_{S,k_1}}}, \\ f_{Z_0}(z_0) = \frac{1}{\rho G_{S,m_1'}} e^{-\frac{z_0}{\rho G_{S,m_1'}}}. \end{cases} \tag{33}$$

On the other hand, for the joint PDFs $f_{X_1,X_2}(x_1,x_2)$, $f_{Y_1,Y_2}(y_1,y_2)$ and $f_{Z_1,Z_2}(z_1,z_2)$, we have the following lemma.

Lemma 4: The joint PDFs of $\{X_1, X_2\}$, $\{Y_1, Y_2\}$ and $\{Z_1, Z_2\}$ can be expressed as follows

$$\begin{aligned} f_{X_1,X_2}(x_1,x_2) = &\sum_{m_1^\star\in\mathbf{B}_j} \sum_{m_2^\star\in\mathbf{B}_j\backslash\{m_1^\star\}} \frac{e^{-\frac{1}{\rho}\left(\frac{x_1}{G_{S,m_1^\star}}+\frac{x_2}{G_{S,m_2^\star}}\right)}}{\rho^2 G_{S,m_1^\star} G_{S,m_2^\star}} \\ &\times \prod_{m\in\mathbf{B}_j\backslash\{m_1^\star,m_2^\star\}} \left(e^{-\frac{x_1}{\rho G_{S,m}}} - e^{-\frac{x_2}{\rho G_{S,m}}}\right), \end{aligned} \tag{34}$$

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) = &\sum_{k_1^\star\in\mathbf{A}_i} \sum_{k_2^\star\in\mathbf{A}_i\backslash\{k_1^\star\}} \frac{e^{-\frac{1}{\rho}\left(\frac{y_1}{G_{S,k_1^\star}}+\frac{y_2}{G_{S,k_2^\star}}\right)}}{\rho^2 G_{S,k_1^\star} G_{S,k_2^\star}} \\ &\times \prod_{k\in\mathbf{A}_i\backslash\{k_1^\star,k_2^\star\}} \left(e^{-\frac{y_1}{\rho G_{S,k}}} - e^{-\frac{y_2}{\rho G_{S,k}}}\right), \end{aligned} \tag{35}$$

$$\begin{aligned} f_{Z_1,Z_2}(z_1,z_2) = &\sum_{m_1^\star\in\mathbf{C}_l} \sum_{m_2^\star\in\mathbf{C}_l\backslash\{m_1^\star\}} \frac{e^{-\frac{1}{\rho}\left(\frac{z_1}{G_{S,m_1^\star}}+\frac{z_2}{G_{S,m_2^\star}}\right)}}{\rho^2 G_{S,m_1^\star} G_{S,m_2^\star}} \\ &\times \prod_{m\in\mathbf{C}_l\backslash\{m_1^\star,m_2^\star\}} \left(e^{-\frac{z_1}{\rho G_{S,m}}} - e^{-\frac{z_2}{\rho G_{S,m}}}\right), \end{aligned} \tag{36}$$

where $x_1 \le x_2$, $y_1 \le y_2$ and $z_1 \le z_2$ should hold for (34), (35) and (36), respectively. Note that for $x_1 > x_2$, $y_1 > y_2$ and $z_1 > z_2$, we have $f_{X_1,X_2}(x_1,x_2) = 0$, $f_{Y_1,Y_2}(y_1,y_2) = 0$ and $f_{Z_1,Z_2}(z_1,z_2) = 0$, respectively.

Proof: Please refer to Appendix E. ∎

Finally, applying (33), (34), (35) and (36) into (30) and then using the mathematical derivations in Appendix F, a closed-form approximation for χ can be obtained as shown in (37) (on the next page), where Q is a positive integer that can be used to control the accuracy of approximation, $B_q \triangleq \min\{1/\Xi, (1/\tau_F)[1-\tau(1/\Lambda+(q/Q)[(1/\Upsilon)-(1/\Lambda)])]\}$, and $D_q \triangleq \min\{1/\Xi, (1/\tau_F)[1-(q\tau/Q\Lambda)]\}$, $q \in \{0, 1, ..., Q\}$. Note that for terms $\varrho_{j,l}$, $I_1(\cdot)$, $J_{1,q}$, $I_2(\cdot)$, $J_{2,q}$, $J_{3,q}$ and $J_{4,q}$ involved in (37), closed-form expressions have been derived in (31), (F.4), (F.5), (F.7), (F.8), (F.9) and (F.10), respectively.

Overall, substituting (37) into (24) and then combining the results with (23), an approximate expression for the outage probability can be obtained. The tightness of obtained approximate expression will be verified by simulations in Section VI.

The computation complexity of evaluating χ given in (37) is $\mathcal{O}(Q)$. Accordingly, the computational complexity for $P_{\text{out}}(\mathbf{A}_i)$ in (24) is $\mathcal{O}(2^{2M}Q)$. Thus, based on (20), the computational complexity of evaluating the outage probability is $\mathcal{O}(2^{2M+K}Q)$.

B. Diversity Order

We know that $e^{-x/\rho} \simeq 1-(x/\rho)$ holds for $\rho \to \infty$, where x represents any positive constant [36, eq. (1.211.1)]. By using this asymptotic expression in (23) while ignoring high-order infinitesimals, we have

$$P_{\text{out}}(\emptyset) \overset{\rho\to\infty}{\simeq} \frac{1}{\rho^{M+K}} \left(\prod_{m=1}^{M} \frac{\Lambda}{G_{S,m}}\right) \left(\prod_{k=1}^{K} \frac{\Xi}{G_{S,k}}\right). \tag{38}$$

However, since no closed-form expression is available for $P_{\text{out}}(\mathbf{A}_i)$, its high-SNR asymptotic expression cannot be derived directly. To facilitate the analysis, we provide an asymptotic upper bound for $P_{\text{out}}(\mathbf{A}_i)$ in the following lemma.

Lemma 5: When $\rho \to \infty$, $P_{\text{out}}(\mathbf{A}_i)$ is asymptotically

$$\chi=\begin{cases}\displaystyle\iiiint\limits_{\substack{y_1\geq\Xi,x_1\geq\Upsilon,x_2<\Lambda,\\ \tau/x_2+\tau_F/y_2\leq1}} f_{X_1,X_2}(x_1,x_2)f_{Y_1,Y_2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}x_1\mathrm{d}y_2\mathrm{d}x_2, & l=0,i\geq2,j\geq2;\\ \displaystyle\iiint\limits_{\substack{y_0\geq\Xi,x_1\geq\Upsilon,x_2<\Lambda\\ \tau/x_2+\tau_F/y_0\leq1}} f_{X_1,X_2}(x_1,x_2)f_{Y_0}(y_0)\mathrm{d}y_0\mathrm{d}x_1\mathrm{d}x_2, & l=0,i=1,j\geq2;\\ \displaystyle\iiint\limits_{\substack{y_1\geq\Xi,\Upsilon\leq x_0<\Lambda\\ \tau/x_0+\tau_F/y_2\leq1}} f_{X_0}(x_0)f_{Y_1,Y_2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}y_2\mathrm{d}x_0, & l=0,i\geq2,j=1;\\ \displaystyle\iint\limits_{\substack{y_0\geq\Xi,\Upsilon\leq x_0<\Lambda\\ \tau/x_0+\tau_F/y_0\leq1}} f_{X_0}(x_0)f_{Y_0}(y_0)\mathrm{d}y_0\mathrm{d}x_0, & l=0,i=1,j=1;\\ \varrho_{j,l}\cdot\displaystyle\iiint\limits_{\substack{y_1\geq\Xi,z_0\geq\Lambda\\ \tau/z_0+\tau_F/y_2\leq1}} f_{Z_0}(z_0)f_{Y_1,Y_2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}y_2\mathrm{d}z_0, & l=1,i\geq2,j\geq l;\\ \varrho_{j,l}\cdot\displaystyle\iint\limits_{\substack{y_0\geq\Xi,z_0\geq\Lambda\\ \tau/z_0+\tau_F/y_0\leq1}} f_{Z_0}(z_0)f_{Y_0}(y_0)\mathrm{d}y_0\mathrm{d}z_0, & l=1,i=1,j\geq l;\\ \varrho_{j,l}\cdot\displaystyle\iiiint\limits_{\substack{y_1\geq\Xi,z_1\geq\Lambda,\\ \tau/z_2+\tau_F/y_2\leq1}} f_{Z_1,Z_2}(z_1,z_2)f_{Y_1,Y_2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}z_1\mathrm{d}y_2\mathrm{d}z_2, & l\geq2,i\geq2;j\geq l;\\ \varrho_{j,l}\cdot\displaystyle\iiint\limits_{\substack{y_0\geq\Xi,z_1\geq\Lambda\\ \tau/z_2+\tau_F/y_0\leq1}} f_{Z_1,Z_2}(z_1,z_2)f_{Y_0}(y_0)\mathrm{d}y_0\mathrm{d}z_1\mathrm{d}z_2, & l\geq2,i=1,j\geq l.\end{cases}\tag{30}$$

$$\chi\approx\begin{cases}\frac{1}{2}\sum_{q=1}^{Q}[I_1(B_q)+I_1(B_{q-1})]J_{1,q}, & l=0,i\geq2,j\geq2;\\ \frac{1}{2}\sum_{q=1}^{Q}[I_2(B_q)+I_2(B_{q-1})]J_{1,q}, & l=0,i=1,j\geq2;\\ \frac{1}{2}\sum_{q=1}^{Q}[I_1(B_q)+I_1(B_{q-1})]J_{2,q}, & l=0,i\geq2,j=1;\\ \frac{1}{2}\sum_{q=1}^{Q}[I_2(B_q)+I_2(B_{q-1})]J_{2,q}, & l=0,i=1,j=1;\\ \frac{\varrho_{j,l}}{2}\sum_{q=1}^{Q}[I_1(D_q)+I_1(D_{q-1})]J_{3,q}, & l=1,i\geq2,j\geq l;\\ \frac{\varrho_{j,l}}{2}\sum_{q=1}^{Q}[I_2(D_q)+I_2(D_{q-1})]J_{3,q}, & l=1,i=1,j\geq l;\\ \frac{\varrho_{j,l}}{2}\sum_{q=1}^{Q}[I_1(D_q)+I_1(D_{q-1})]J_{4,q}, & l\geq2,i\geq2,j\geq l;\\ \frac{\varrho_{j,l}}{2}\sum_{q=1}^{Q}[I_2(D_q)+I_2(D_{q-1})]J_{4,q}, & l\geq2,i=1,j\geq l.\end{cases}\tag{37}$$

---

upper bounded by

$$P_{\mathrm{out}}(\mathbf{A}_i)\overset{\rho\to\infty}{<}\frac{1}{\rho^{M+K-i}}\left(\prod_{m=1}^{M}\frac{\min\{\Lambda,\max(2\tau,\Upsilon)\}}{G_{S,m}}\right)\times\left(\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\frac{\Xi}{G_{S,k}}\right),\forall i\in\mathbf{K}.\tag{39}$$

Proof: Please refer to Appendix G. ■

Applying (38) and (39) into (20), we have

$$\begin{aligned}P_{\mathrm{out}}&\overset{\rho\to\infty}{<}\frac{1}{\rho^{M+K}}\left(\prod_{m=1}^{M}\frac{\Lambda}{G_{S,m}}\right)\left(\prod_{k=1}^{K}\frac{\Xi}{G_{S,k}}\right)\\&+\sum_{i=1}^{K}\sum_{\mathbf{A}_i\subseteq\mathbf{K},|\mathbf{A}_i|=i}\frac{1}{\rho^{M+K-i}}\\&\times\left(\prod_{m=1}^{M}\frac{\min\{\Lambda,2\Upsilon\}}{G_{S,m}}\right)\left(\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\frac{\Xi}{G_{S,k}}\right)\\&\overset{\rho\to\infty}{\simeq}\frac{1}{\rho^{M}}\prod_{m=1}^{M}\frac{\min\{\Lambda,2\Upsilon\}}{G_{S,m}}.\end{aligned}\tag{40}$$

As observed from (40), when $\rho\to\infty$, the upper bound of outage probability is proportional to ρ^{-M}. Therefore, the proposed OA-NOMA strategy can achieve at least a diversity order of M, implying that the inherent diversity offered by all NUs is fully exploited.

VI. Numerical Results

In our simulation, we consider the BS is located at the center of two concentric circles with radius being 80 and 150, respectively. The locations of NUs are randomly generated within the circle with radius of 80, while the locations of FUs are randomly generated between these two concentric circles. We model path loss attenuation from node i to node j as $1/[1+(d_{i,j}/d_0)^{\nu}]$, where $d_{i,j}$ denotes the distance between the two nodes, d_0 denotes the reference distance, and ν denotes the path loss exponent. Following the standard parameters for urban cellular networks [37], we set $d_0 = 5$ and $\nu = 3.2$ in our simulation. The fairness exponent of the proposed P-UPS scheme is set to $\eta = 2$, and the accuracy-controlling parameter for computing the approximation in (37) is set to $Q = 7$. By using a one-dimensional search, we adopt the optimal value of Ξ, which minimizes the derived approximate outage probability. The other parameters of our simulation are given by $\mu = P_N/P_S = 1$ and $\sigma^2 = 1$. All simulated values are averaged over 1×10^9 independent transmission blocks.

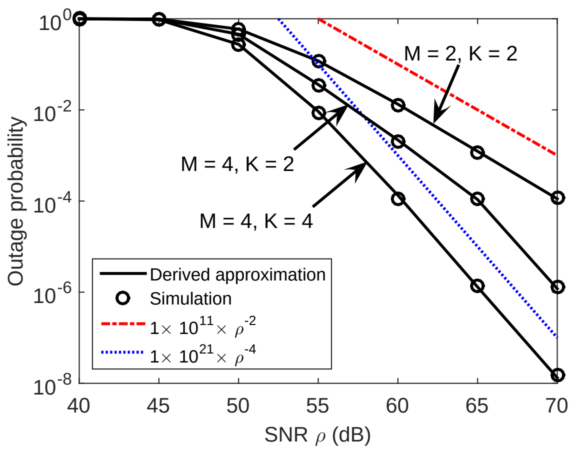


Fig. 3. Approximate and simulated outage probabilities, where $r_N = 5$ bps/Hz, $r_F = 1$ bps/Hz, $\{M, K\} = \{2, 2\}, \{4, 2\}, \{4, 4\}$.

Figs. 3 shows the theoretically approximate and numerically simulated outage probabilities, where $r_N = 5$ bps/Hz, $r_F = 1$ bps/Hz and $\{M, K\} = \{2, 2\}$, $\{4, 2\}, \{4, 4\}$. As shown in Fig. 3, the simulated outage probability coincides with the derived approximate outage probability, which verifies the tightness of the derived approximate outage probability. Further, by comparing the outage probability curves with different combinations of M and K, it can be known that the outage probability decreases with both M and K. This observation can be explained as follows. When more NUs or more FUs are available for NU-FU pair scheduling, the number of candidate NU-FU pairs will increase in average, thereby the probability of no candidate NU-FU pair existing decreases significantly. Furthermore, by comparing the reference lines $1 \times 10^{11} \times \rho^{-2}$ and $1 \times 10^{21} \times \rho^{-4}$ with the curves of outage probability in high-SNR regime, it can be observed that the outage probability with $\{M, K\} = \{2, 2\}$ decays at the same rate with reference line $1 \times 10^{11} \times \rho^{-2}$, whereas the outage probabilities with $\{M, K\} = \{4, 2\}$ and $\{M, K\} = \{4, 4\}$ decay at the same rate with reference line $1 \times 10^{21} \times \rho^{-4}$. Therefore, it can be concluded that the achieved diversity order is equal to the number of NUs, which complies with our theoretical analysis in Section V-B.

Next, we compare the outage probability, the power consumption and the scheduling fairness of the proposed OA-NOMA strategy and its counterpart strategies that use the following scheduling schemes.

1) **Random scheduling (RS)**: If $\mathbf{P} \neq \emptyset$, one of candidate NU-FU pairs will be randomly scheduled. Otherwise, the scheduling and downlink transmission will be cancelled.
2) **Two-stage scheduling (TSS)**: This two-stage scheme follows the idea of the user selection scheme in [22]. In the first stage, the FU owning the highest channel gain from the BS is scheduled out of all FUs, i.e., $k^* = \arg\max_{k \in \mathbf{K}} |g_{S,k}|^2$. In the second stage, the *qualified* NUs which satisfy $\rho|g_{S,m}|^2 \geq \Lambda$ will participate in the scheduling. If $\mathcal{Q} \triangleq \{m|\rho|g_{S,m}|^2 \geq \Lambda\} \neq \emptyset$, the scheduled NU will be $m^* = \arg\max_{m \in \mathcal{Q}} |g_{m,k^*}|^2$; otherwise, no NU will be scheduled and the transmission will be cancelled.
3) **One-stage scheduling (OSS)**: Inspired by the work in [20], we consider three different one-stage scheduling criteria: 1) *Best FU best NU (BFBN)*: the FU with the best channel quality from the BS and the NU with the best channel quality from the BS are scheduled, i.e., $k^* = \arg\max_{k \in \mathbf{K}} |g_{S,k}|^2$, $m^* = \arg\max_{m \in \mathbf{M}} |g_{S,m}|^2$; 2) *Worst FU best NU (WFBN)*: the FU with the worst channel quality from the BS and the NU with the best channel quality from the BS are scheduled, i.e., $k^* = \arg\min_{k \in \mathbf{K}} |g_{S,k}|^2$, $m^* = \arg\max_{m \in \mathbf{M}} |g_{S,m}|^2$; 3) *Random FU and random NU (RFRN)*: an FU and an NU are randomly scheduled from all FUs and all NUs, respectively.

We call those strategies as the *RS strategy*, the *TSS strategy*, and the *OSS strategy*, respectively. Similar to the proposed OA-NOMA strategy, the above three strategies also adaptively serve the scheduled NU-FU pair in either direct NOMA mode or cooperative NOMA mode. Further, the RS strategy employs the same power control approaches as those in the proposed OA-NOMA strategy. On the other hand, since the fixed PA was used in [20] and [22], we consider that the TSS strategy and OSS strategy employ the following approximately optimal fixed PA [22, eq. (18)]:

$$\begin{cases} \text{Direct NOMA} \begin{cases} \alpha_F = \frac{\tau_F \tau_N + \tau_F}{\Upsilon}, \\ \alpha_N = 1 - \alpha_F, \end{cases} \\ \\ \text{Cooperative NOMA} \begin{cases} \beta_F = \frac{\lambda_F \lambda_N + \lambda_F}{\Lambda}, \\ \beta_N = 1 - \beta_F, \\ \theta_F = 1. \end{cases} \end{cases} \tag{41}$$

Fig. 4 shows the simulated outage probability of the proposed OA-NOMA strategy, the RS strategy, the TSS strategy and the OSS strategy, where $M = 3$, $K = 3$, $r_N = 4$ bps/Hz and$r_F = 1$ bps/Hz. Since both the OA-NOMA strategy and the RS strategy perform NU-FU pair scheduling among all candidate NU-FU pairs obtained from the SCA method, they are expected to have the same outage performance, as shown in this figure. It can also be observed that the proposed OA-NOMA strategy achieves better outage performance than the TSS strategy. When the OSS strategy employs the WFBN or RFRN criterion, the achieved outage probability is much higher than the proposed one. Further, when BFBN criterion is employed in the OSS strategy, the outage performance is significantly improved, but still worse than the proposed OA-NOMA strategy. This is because none of these criteria for the

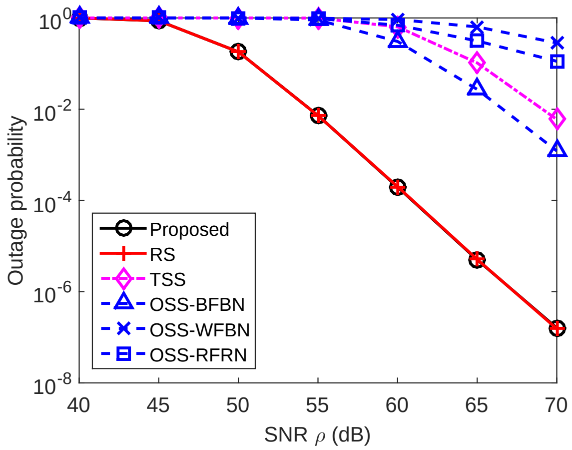


Fig. 4. Simulated outage probability of the proposed strategy, the RS strategy, the TSS strategy and the OSS strategy, where $M = 3$, $K = 3$, $r_N = 4$ bps/Hz and $r_F = 1$ bps/Hz.

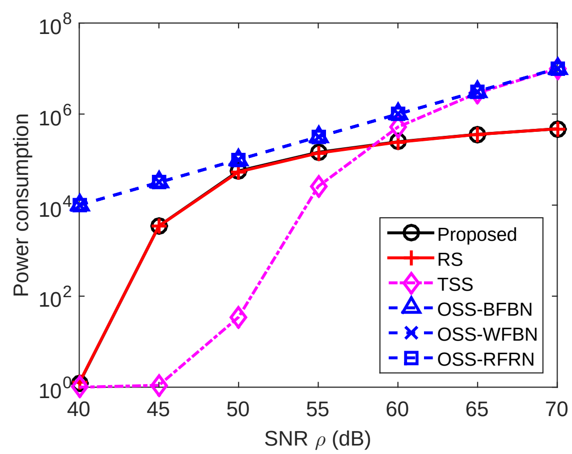


Fig. 5. Simulated power consumption of the proposed strategy, the RS strategy, the TSS strategy and the OSS strategy, where $M = 3$, $K = 3$, $r_N = 4$ bps/Hz and $r_F = 1$ bps/Hz.

OSS strategy considers the quality of links between the FUs and NUs.

Fig. 5 provides the simulated power consumption of the proposed OA-NOMA strategy, the RS strategy, the TSS strategy and the OSS strategy, where $M = 3$, $K = 3$, $r_N = 4$ bps/Hz and $r_F = 1$ bps/Hz. For each strategy, the simulated power consumption is averaged over 1×10^9 independent transmission blocks, where the power consumed in every transmission block is computed as follows: 1) When the transmission to the scheduled NU-FU pair operates in the direct NOMA mode, the power consumption is $(\alpha_F + \alpha_N)P_S$; 2) When the transmission to the scheduled NU-FU pair operates in the cooperative NOMA mode, the power consumption is $\frac{1}{2}(\beta_N + \beta_F)P_S + \frac{1}{2}\theta_F P_N$, where $1/2$ accounts for each phase with $1/2$ duration of the adaptive transmission portion; 3) When no transmission happens due to the cancelled user scheduling, the power consumption is zero. As observed from this figure, the OA-NOMA strategy and the RS strategy consume the same amount of transmit power. This is because the RS strategy uses the same power control approaches as those in the OA-NOMA strategy. On the other hand, since the fixed PA shown in (41) always fully utilizes the transmit power of the BS and the scheduled NU, the OSS strategy with BFBN/WFBN/RFRN criterion incurs much higher power consumption than the proposed OA-NOMA strategy. Interestingly, although the TSS strategy also employs the fixed PA (41), its power consumption is much less than the proposed one when $\rho < 60$ dB. This observation can be explained as follows. Recall that if no qualified NU exists, i.e., $\mathcal{Q} = \{m|\rho|g_{S,m}|^2 \geq \Lambda\} = \emptyset$, the transmission of the TSS strategy will be cancelled. Thus, when $\rho < 60$ dB, some power savings can be achieved by the TSS strategy due to the cancelled transmission. Furthermore, when $\rho \geq 60$ dB, the TSS strategy consumes the similar amount of transmit power to that of the OSS strategy, since $\mathcal{Q}$ is almost always non-empty.

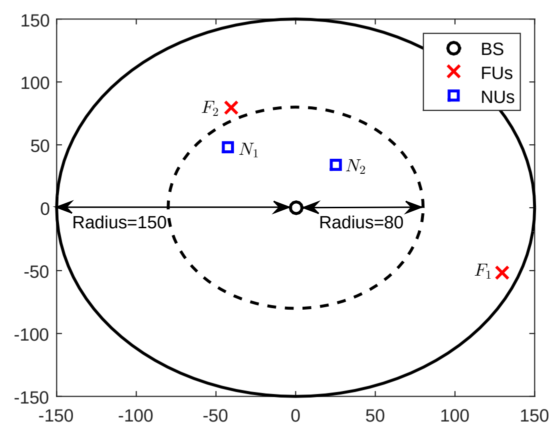


Fig. 6. System topology used for evaluating SSP of each NU-FU pair, where $M = 2$ and $K = 2$.

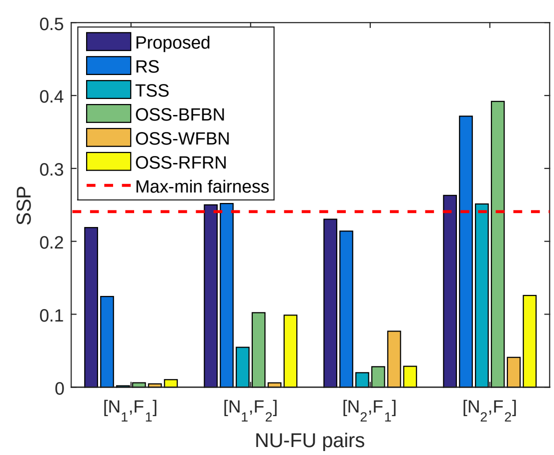


Fig. 7. Simulated SSP of each NU-FU pair, where $M = 2$, $K = 2$, $r_N = 2$ bps/Hz, $r_F = 1$ bps/Hz and $\rho = 50$ dB.

Considering the system topology shown in Fig. 6, Fig. 7 compares the scheduling fairness among the proposed OA-NOMA strategy, the RS strategy, the TSS strategy and the OSS strategy, where $M = 2$, $K = 2$, $r_N = 2$ bps/Hz, $r_F = 1$ bps/Hz and $\rho = 50$ dB. The max-min fair SSPs given in

Theorem 1 are also plotted as dashed line in Fig. 7 to provide a max-min fairness benchmark. As seen from Fig. 7, the SSP achieved by the OA-NOMA strategy is very close to the max-min fairness benchmark at each NU-FU pair, indicating that the OA-NOMA strategy can approximately achieve the max-min fairness. Further, when the RS strategy is employed, the SSP of NU-FU pair $[N_1, F_1]$ is far below the max-min fairness benchmark, whereas the SSP of NU-FU pair $[N_2, F_2]$ is much higher than the max-min fairness benchmark. This fact means that the RS strategy achieves worse scheduling fairness than the proposed OA-NOMA strategy. Moreover, when the TSS strategy or the OSS strategy is employed, the fairness among NU-FU pairs is even worse than its counterpart in the RS strategy.

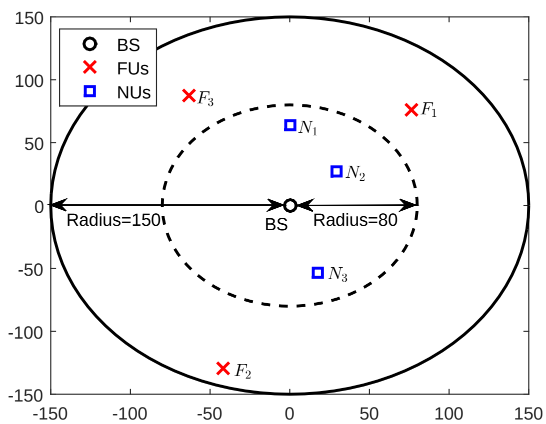


Fig. 8. System topology used for evaluating SSP of each NU-FU pair, where $M = 3$ and $K = 3$.

Moreover, we also demonstrate the scheduling fairness when more NU-FU pairs are available. Considering the system topology shown in Fig. 8, Fig. 9 compares the scheduling fairness among the proposed OA-NOMA strategy, the RS strategy, the TSS strategy and the OSS strategy, where $M = 3$, $K = 3$, $r_N = r_F = 1$ bps/Hz and $\rho = 50$ dB. Note that the max-min fairness benchmark in Fig. 9 also comes from Theorem 1. As seen from Fig. 9, the proposed OA-NOMA strategy still approximately achieves max-min fairness, which is similar to the observation from Fig. 7. When employing the RS strategy, the SSPs of NU-FU pairs $[N_1, F_2]$, $[N_2, F_1]$, $[N_2, F_2]$ and $[N_2, F_3]$ fluctuate around the max-min fairness benchmark. Thus, the proposed OA-NOMA strategy still achieves better fairness than the RS strategy for $M = 3$ and $K = 3$. Furthermore, when employing the TSS strategy or the OSS strategy using BFBN/WFBN/RFRN criterion, the scheduling fairness among NU-FU pairs is even worse than its counterpart in the RS strategy.

VII. Conclusion

In this paper, we have designed an OA-NOMA strategy for a multiuser downlink NOMA system to serve an NU-FU pair opportunistically scheduled from M NUs and K FUs. For this NOMA strategy, an SCA method and a P-UPS scheme have been proposed to improve the reliability and scheduling fairness, respectively. To evaluate the scheduling fairness, we have developed a max-min fairness criterion, with which we have shown that the OA-NOMA strategy achieves the approximate max-min fairness. Further, the reliability of OA-NOMA strategy has also been investigated in terms of outage probability and diversity order. For the outage probability, we have derived an approximate expression and demonstrated its tightness by simulations. Then, by carrying out the high-SNR asymptotic analysis, we have shown that the proposed OA-NOMA strategy can achieve a diversity order of M, indicating that the inherent diversity of NUs can be fully exploited.

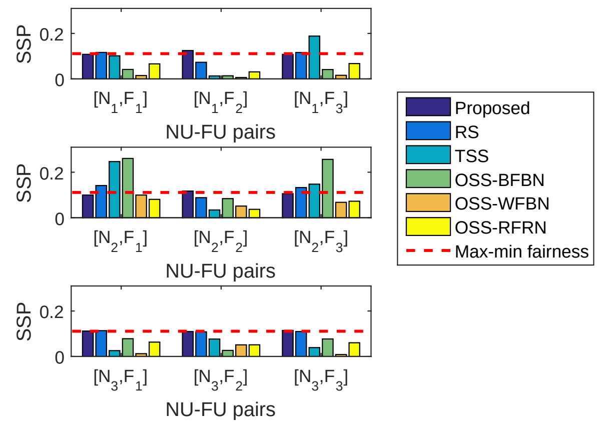


Fig. 9. Simulated SSP of each NU-FU pair, where $M = 3$, $K = 3$, $r_N = r_F = 1$ bps/Hz and $\rho = 50$ dB.

Appendix A
Derivations of Candidate Probability

As observed from (8) and (17), $\mathbb{R}^{\rm dir}_{m,k}$ and $\mathbb{R}^{\rm coop}_{m,k}$ are mutually exclusive. Based on this fact, $CP_{m,k}$ can be further expressed as $CP_{m,k} = \vartheta_{m,k,1} + \vartheta_{m,k,2}$, where $\vartheta_{m,k,1}$ and $\vartheta_{m,k,2}$ are given by

$$\begin{aligned}\vartheta_{m,k,1} &\triangleq \Pr\left\{\mathbb{R}^{\rm dir}_{m,k}\right\} \\ &= \Pr\{\rho|g_{S,k}|^2 \geq \Xi, \varphi_m + \xi_{m,k} \leq 1\},\end{aligned} \tag{A.1}$$

$$\begin{aligned}\vartheta_{m,k,2} &\triangleq \Pr\left\{\mathbb{R}^{\rm coop}_{m,k}\right\} \\ &= \Pr\{\rho|g_{S,k}|^2 < \Xi, \psi_m + \kappa_m \leq 1, \omega_{m,k} \leq 1\}.\end{aligned} \tag{A.2}$$

Applying (6) and (7) into (A.1) with letting $X \triangleq 1/(\rho|g_{S,m}|^2)$ and $Y \triangleq 1/(\rho|g_{S,k}|^2)$, $\vartheta_{m,k,1}$ can be expressed as

$$\begin{aligned}\vartheta_{m,k,1} &= \Pr\left\{1/Y \geq \Xi, \tau X + \tau_F Y \leq 1, \Upsilon X \leq 1\right\} \\ &= \iint\limits_{\substack{0 \leq x \leq 1/\Upsilon, \\ 0 \leq y \leq \min\{1/\Xi, (1-\tau x)/\tau_F\}}} f_X(x) f_Y(y) \mathrm{d}x\mathrm{d}y,\end{aligned} \tag{A.3}$$

where $\tau \triangleq \tau_N + \tau_N\tau_F$ and $\Upsilon \triangleq \tau_N + \tau_F + \tau_N\tau_F$, $f_X(x) = \frac{1}{\rho G_{S,m} x^2} e^{-\frac{1}{\rho G_{S,m} x}}$ and $f_Y(y) = \frac{1}{\rho G_{S,k} y^2} e^{-\frac{1}{\rho G_{S,k} y}}$. Then, following the derivations in [25, Appendix E], a tight approximate expression for $\vartheta_{m,k,1}$ can be obtained as

$$\vartheta_{m,k,1} = \frac{1}{2}\sum_{l=1}^{L}\left(e^{-\frac{1}{\rho G_{S,m} x_l}} - e^{-\frac{1}{\rho G_{S,m} x_{l-1}}}\right)$$

$$\times\left(e^{-\frac{1}{\rho G_{S,k} y_l}}+e^{-\frac{1}{\rho G_{S,k} y_{l-1}}}\right),$$

where $x_l \triangleq \frac{l}{L\Upsilon}$ and $y_l = \min\{\frac{1}{\Xi}, \frac{1}{\tau_F}(1-\frac{\tau l}{L\Upsilon})\}$ for $l = 1, \cdots, L$ with L being a positive integer. As demonstrated in [25], the tightness of the approximation is guaranteed for $L \geq 3$. On the other hand, applying (14), (15) and (16) into (A.2) with some algebraic manipulations, a closed-form expression for $\vartheta_{m,k,2}$ can be obtained as

$$\vartheta_{m,k,2} = \left(1-e^{\frac{\Xi}{\rho G_{S,k}}}\right)e^{-\frac{1}{\rho}\left(\frac{\Lambda}{G_{S,m}}+\frac{\lambda_F}{\mu G_{m,k}}\right)},$$

where $\Lambda \triangleq \lambda_N + \lambda_F + \lambda_N\lambda_F$. Combining the above obtained $\vartheta_{m,k,1}$ and $\vartheta_{m,k,2}$ expressions, a tight approximation for $CP_{m,k}$ is derived.

Appendix B
Proof of Lemma 1

According to the definition of proportional fairness [34], the max-min fair SSPs given in Theorem 1 are proportional fair, if and only if for any other *feasible SSPs of NU-FU pairs*, denoted as $SSP'_{m,k}$ for $m \in \mathbf{M}$ and $K \in \mathbf{K}$, the aggregate of proportional changes is zero or negative:

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{SSP'_{m,k}-SSP^{\text{max-min}}_{m,k}}{SSP^{\text{max-min}}_{m,k}} \leq 0. \quad \text{(B.1)}$$

Here, the term "feasible SSPs of NU-FU pairs" means that 1) $SSP'_{m,k} \leq CP_{m,k}$ holds for $m \in \mathbf{M}$ and $k \in \mathbf{K}$, 2) $\sum_{m=1}^{M}\sum_{k=1}^{K} SSP'_{m,k} = 1 - P_{\text{out}}$ holds.

Next, we will prove that (B.1) always holds. Defining $\delta_{m,k} \triangleq SSP'_{m,k} - SSP^{\text{max-min}}_{m,k}$, the aggregate of proportional changes can be expressed as

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{SSP'_{m,k}-SSP^{\text{max-min}}_{m,k}}{SSP^{\text{max-min}}_{m,k}} = \sum_{m=1}^{M}\sum_{k=1}^{K}\frac{\delta_{m,k}}{SSP^{\text{max-min}}_{m,k}}. \quad \text{(B.2)}$$

Then, by introducing indices sets $\mathbf{D}^+ \triangleq \left\{[m,k]\middle|\delta_{m,k} > 0, m \in \mathbf{M}, k \in \mathbf{K}\right\}$ and $\mathbf{D}^- \triangleq \left\{[m,k]\middle|\delta_{m,k} \leq 0, m \in \mathbf{M}, k \in \mathbf{K}\right\}$, (B.2) can be rewritten as

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{SSP'_{m,k}-SSP^{\text{max-min}}_{m,k}}{SSP^{\text{max-min}}_{m,k}}\\ &= \sum_{[m,k]\in\mathbf{D}^+}\frac{\delta_{m,k}}{SSP^{\text{max-min}}_{m,k}} + \sum_{[m,k]\in\mathbf{D}^-}\frac{\delta_{m,k}}{SSP^{\text{max-min}}_{m,k}}. \quad \text{(B.3)}\end{aligned}$$

Note that for $[m,k] \in \mathbf{D}^+$, we have $\delta_{m,k} = SSP'_{m,k} - SSP^{\text{max-min}}_{m,k} > 0$, meaning that inequality $SSP^{\text{max-min}}_{m,k} = \min\{\zeta, CP_{m,k}\} < SSP'_{m,k}$ holds. Further, since $SSP'_{m,k} \leq CP_{m,k}$ holds for $m \in \mathbf{M}$ and $k \in \mathbf{K}$, it can be concluded that for $[m,k] \in \mathbf{D}^+$, $\min\{\zeta, CP_{m,k}\} < SSP'_{m,k} \leq CP_{m,k}$ always holds, indicating that $\zeta < CP_{m,k}$ should hold for $[m,k] \in \mathbf{D}^+$. Therefore, for $[m,k] \in \mathbf{D}^+$, we have $SSP^{\text{max-min}}_{m,k} = \min\{\zeta, CP_{m,k}\} = \zeta$. Applying this fact into (B.3), the aggregate of proportional changes can be further expressed as

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{SSP'_{m,k}-SSP^{\text{max-min}}_{m,k}}{SSP^{\text{max-min}}_{m,k}}\\ &= \sum_{[m,k]\in\mathbf{D}^+}\frac{\delta_{m,k}}{\zeta} + \sum_{[m,k]\in\mathbf{D}^-}\frac{\delta_{m,k}}{SSP^{\text{max-min}}_{m,k}}. \quad \text{(B.4)}\end{aligned}$$

On the other hand, as $SSP^{\text{max-min}}_{m,k} = \min\{\zeta, CP_{m,k}\} \leq \zeta$, we have $1/SSP^{\text{max-min}}_{m,k} \geq 1/\zeta$ holds. Based on this fact, it can be further known that for $[m,k] \in \mathbf{D}^-$, $\delta_{m,k}/SSP^{\text{max-min}}_{m,k} \leq \delta_{m,k}/\zeta$ holds due to $\delta_{m,k} \leq 0$. Applying this fact into the right-hand side of (B.4), the following results can be obtained

$$\begin{aligned}&\sum_{m=1}^{M}\sum_{k=1}^{K}\frac{SSP'_{m,k}-SSP^{\text{max-min}}_{m,k}}{SSP^{\text{max-min}}_{m,k}}\\ &\leq \sum_{[m,k]\in\mathbf{D}^+}\frac{\delta_{m,k}}{\zeta} + \sum_{[m,k]\in\mathbf{D}^-}\frac{\delta_{m,k}}{\zeta}\\ &\overset{(b.1)}{=} \sum_{m=1}^{M}\sum_{k=1}^{K}\frac{\delta_{m,k}}{\zeta}\\ &\overset{(b.2)}{=} \frac{1}{\zeta}\left(\sum_{m=1}^{M}\sum_{k=1}^{K}SSP'_{m,k} - \sum_{m=1}^{M}\sum_{k=1}^{K}SSP^{\text{max-min}}_{m,k}\right), \quad \text{(B.5)}\end{aligned}$$

where step (b.1) comes from the fact $\mathbf{D}^+ \cup \mathbf{D}^- = \{[m,k] | m \in \mathbf{M}, k \in \mathbf{K}\}$, step (b.2) comes from the fact $\delta_{m,k} = SSP'_{m,k} - SSP^{\text{max-min}}_{m,k}$. Since $\sum_{m=1}^{M}\sum_{k=1}^{K} SSP^{\text{max-min}}_{m,k} = \sum_{m=1}^{M}\sum_{k=1}^{K} SSP'_{m,k} = 1 - P_{\text{out}}$, it can be known that the last line of (B.5) is equal to zero, meaning that (B.1) is obtained. This completes the proof.

Appendix C
Proof of Lemma 2

First, we prove the closed-form expression of $P_{\text{out}}(\emptyset)$ given in (23). When $\mathbf{F}_\Xi = \emptyset$, we have $\rho|g_{S,k}|^2 < \Xi$ holds for $k \in \mathbf{K}$. Moreover, as known from (14), (15) and (16), ψ_m, κ_m and $\omega_{m,k}$ are independent from $|g_{S,k}|^2$ for $k \in \mathbf{K}$. Thus, using (21) with some algebraic manipulations, $P_{\text{out}}(\emptyset)$ can be expressed as

$$\begin{aligned}&P_{\text{out}}(\emptyset)\\ &= \prod_{m=1}^{M}\underbrace{\Pr\left\{\left(\rho|g_{S,m}|^2 < \Lambda\right)\bigcup\left(\bigcap_{k=1}^{K}\mu\rho|g_{m,k}|^2 < \lambda_F\right)\right\}}_{\triangleq\Psi_m}\\ &\quad\times\prod_{k=1}^{K}\Pr\left\{\rho|g_{S,k}|^2 < \Xi\right\}\\ &= \prod_{m=1}^{M}\Psi_m \cdot \prod_{k=1}^{K}\left(1 - e^{-\frac{\Xi}{\rho G_{S,k}}}\right). \quad \text{(C.1)}\end{aligned}$$

Further, since $\Pr\{\mathbb{A} \cup \mathbb{B}\} = \Pr\{\mathbb{A}\} + \Pr\{\mathbb{B}\} - \Pr\{\mathbb{A} \cap \mathbb{B}\}$ holds for any two events $\mathbb{A}$ and $\mathbb{B}$, Ψ_m can be derived as

$$\begin{aligned}\Psi_m =& \Pr\left\{\rho|g_{S,m}|^2 < \Lambda\right\} + \Pr\left\{\bigcap_{k=1}^{K}\mu\rho|g_{m,k}|^2 < \lambda_F\right\}\\ &- \Pr\left\{\rho|g_{S,m}|^2 < \Lambda, \bigcap_{k=1}^{K}\mu\rho|g_{m,k}|^2 < \lambda_F\right\} \quad \text{(C.2)}\\ =& \left(1 - e^{-\frac{\Lambda}{\rho G_{S,m}}}\right) + e^{-\frac{\Lambda}{\rho G_{S,m}}}\prod_{k=1}^{K}\left(1 - e^{-\frac{\lambda_F}{\rho G_{m,k}}}\right).\end{aligned}$$

Substituting (C.2) into (C.1), $P_{\text{out}}(\emptyset)$ is derived into a closed-form expression shown in (23).

Next, we prove the expression for $P_{\text{out}}(\mathbf{A}_i)$ given in (24). To facilitate the proof, we introduce two indices sets $\mathbf{N}_\Upsilon \triangleq \{m|\rho|g_{S,m}|^2 \geq \Upsilon, m \in \mathbf{M}\}$ and $\mathbf{N}_\Lambda \triangleq \{m|\rho|g_{S,m}|^2 \geq \Lambda, m \in \mathbf{M}\}$, where $\mathbf{N}_\Lambda \subseteq \mathbf{N}_\Upsilon$ due to $\Lambda > \Upsilon$. Since $\rho|g_{S,m}|^2 < \Lambda$ always holds for $m \in \mathbf{M} \setminus \mathbf{N}_\Lambda$, it can be known from (21) that $\{\cap_{k\in\mathbf{K}\setminus\mathbf{A}_i}\mathbb{E}_{1,k}\} = \mathbb{H}_1(\mathbf{A}_i,\mathbf{N}_\Lambda) \triangleq \cap_{k\in\mathbf{K}\setminus\mathbf{A}_i}\left\{\cap_{m\in\mathbf{N}_\Lambda}\mu\rho|g_{m,k}|^2 < \lambda_F\right\}$. Similarly, since $\rho|g_{S,m}|^2 < \Upsilon$ always holds for $m \in \mathbf{M}\setminus\mathbf{N}_\Upsilon$, we can know from (22) that $\{\cap_{k\in\mathbf{A}_i}\mathbb{E}_{2,k}\} = \mathbb{H}_2(\mathbf{A}_i,\mathbf{N}_\Upsilon) \triangleq \left\{\cap_{k\in\mathbf{A}_i}\left[\cap_{m\in\mathbf{N}_\Upsilon}\frac{\tau}{\rho|g_{S,m}|^2}+\frac{\tau_F}{\rho|g_{S,k}|^2} > 1\right]\right\}$. Thus, applying these facts into the expression of $P_{\text{out}}(\mathbf{A}_i)$ given in (20), we further have

$$\begin{aligned}
P_{\text{out}}(\mathbf{A}_i) &= \Pr\{\mathbb{H}_1(\mathbf{A}_i,\mathbf{N}_\Lambda),\mathbb{H}_2(\mathbf{A}_i,\mathbf{N}_\Upsilon),\mathbf{F}_\Xi=\mathbf{A}_i\}\\
&\overset{(c.1)}{=}\Pr\{\mathbf{F}_\Xi=\mathbf{A}_i,\mathbf{N}_\Upsilon=\emptyset\}\\
&\quad+\sum_{j=1}^{M}\sum_{\mathbf{B}_j\subseteq\mathbf{M},|\mathbf{B}_j|=j}\sum_{l=0}^{j}\sum_{\mathbf{C}_l\subseteq\mathbf{B}_j,|\mathbf{C}_l|=l}\\
&\quad\Pr\{\mathbb{H}_1(\mathbf{A}_i,\mathbf{C}_l),\mathbb{H}_2(\mathbf{A}_i,\mathbf{B}_j),\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)\}\\
&\overset{(c.2)}{=}\Pr\{\mathbf{F}_\Xi=\mathbf{A}_i,\mathbf{N}_\Upsilon=\emptyset\}\\
&\quad+\sum_{j=1}^{M}\sum_{\mathbf{B}_j\subseteq\mathbf{M},|\mathbf{B}_j|=j}\sum_{l=0}^{j}\sum_{\mathbf{C}_l\subseteq\mathbf{B}_j,|\mathbf{C}_l|=l}\\
&\quad\Pr\{\mathbb{H}_1(\mathbf{A}_i,\mathbf{C}_l)\}\Big(\Pr\{\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)\}\\
&\quad-\Pr\left\{\bar{\mathbb{H}}_2(\mathbf{A}_i,\mathbf{B}_j),\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)\right\}\Big),
\end{aligned}\tag{C.3}$$

with $\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)$ defined as

$$\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)\triangleq\{\mathbf{F}_\Xi=\mathbf{A}_i,\mathbf{N}_\Upsilon=\mathbf{B}_j,\mathbf{N}_\Lambda=\mathbf{C}_l\},\tag{C.4}$$

where step (c.1) comes from the Total Probability Theorem, step (c.2) uses the fact that $\mathbb{H}_1(\mathbf{A}_i,\mathbf{C}_l)$ is independent from $\mathbb{H}_2(\mathbf{A}_i,\mathbf{B}_j)$ and $\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)$, and $\bar{\mathbb{H}}_2(\mathbf{A}_i,\mathbf{B}_j)$ represents the complementary event of $\mathbb{H}_2(\mathbf{A}_i,\mathbf{B}_j)$. With some mathematical derivations, it can be obtained

$$\begin{cases}
\Pr\{\mathbf{F}_\Xi=\mathbf{A}_i,\mathbf{N}_\Upsilon=\emptyset\}=\Phi_1,\\
\Pr\{\mathbb{H}_1(\mathbf{A}_i,\mathbf{C}_l)\}=\Phi_2,\\
\Pr\{\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)\}=\Phi_3,
\end{cases}$$

where closed-form expressions for Φ_1, Φ_2 and Φ_3 have been given in (25), (26) and (27), respectively. On the other hand, as $\mathbb{H}_2(\mathbf{A}_i,\mathbf{B}_j) = \left\{\cap_{k\in\mathbf{A}_i}\left[\cap_{m\in\mathbf{B}_j}\tau/(\rho|g_{S,m}|^2)+\tau_F/(\rho|g_{S,k}|^2) > 1\right]\right\}$, its complementary event can be expressed as

$$\begin{aligned}
\bar{\mathbb{H}}_2(\mathbf{A}_i,\mathbf{B}_j)&=\left\{\bigcup_{k\in\mathbf{A}_i}\left(\bigcup_{m\in\mathbf{B}_j}\frac{\tau}{\rho|g_{S,m}|^2}+\frac{\tau_F}{\rho|g_{S,k}|^2}\leq 1\right)\right\}\\
&=\left\{\frac{\tau}{\max\limits_{m\in\mathbf{B}_j}\rho|g_{S,m}|^2}+\frac{\tau_F}{\max\limits_{k\in\mathbf{A}_i}\rho|g_{S,k}|^2}\leq 1\right\}.
\end{aligned}\tag{C.5}$$

Then, combining (C.4) and (C.5) with some algebraic manipulations, $\Pr\left\{\bar{\mathbb{H}}_2(\mathbf{A}_i,\mathbf{B}_j),\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)\right\}$ can be expressed as $\Pr\left\{\bar{\mathbb{H}}_2(\mathbf{A}_i,\mathbf{B}_j),\mathbb{I}(\mathbf{A}_i,\mathbf{B}_j,\mathbf{C}_l)\right\} = \Phi_4\cdot\chi$, where Φ_4 and χ have been given in (28) and (29). This completes the proof.

Appendix D
Proof of Lemma 3

According to (29), when $l=0$, $i\geq 2$ and $j\geq 2$, χ can be expressed as

$$\chi=\Pr\left\{\frac{\tau}{X_2}+\frac{\tau_F}{Y_2}\leq 1, Y_1\geq\Xi, X_1\geq\Upsilon, X_2<\Lambda\right\}.\tag{D.1}$$

Denoting the joint PDFs of RVs X_1 and X_2 as $f_{X_1,X_2}(x_1,x_2)$ and the joint PDF of RVs Y_1 and Y_2 as $f_{Y_1,Y_2}(y_1,y_2)$, (D.1) can be rewritten as an integral form shown at the first case of (30). With the same rationale, the expressions for cases $\{l=0,i=1,j\geq 2\}$, $\{l=0,i\geq 2,j=1\}$ and $\{l=0,i=1,j=1\}$ can also be obtained as shown at the second, third and fourth cases in (30), respectively.

On the other hand, when $l\geq 2$, $i\geq 2$ and $j\geq l$, χ can be re-expressed as

$$\begin{aligned}
\chi&=\Pr\Bigg\{\frac{\tau}{X_2}+\frac{\tau_F}{Y_2}\leq 1, Y_1\geq\Xi, Z_1\geq\Lambda,\\
&\qquad\bigcap_{m\in\mathbf{B}_j\setminus\mathbf{C}_l}\Upsilon\leq\rho|g_{S,m}|^2<\Lambda\Bigg\}\\
&\overset{(d.1)}{=}\Pr\left\{\frac{\tau}{Z_2}+\frac{\tau_F}{Y_2}\leq 1, Y_1\geq\Xi, Z_1\geq\Lambda\right\}\\
&\qquad\times\prod_{m\in\mathbf{B}_j\setminus\mathbf{C}_l}\Pr\left\{\Upsilon\leq\rho|g_{S,m}|^2<\Lambda\right\},
\end{aligned}\tag{D.2}$$

where step (d.1) uses the fact that $X_2=\max_{m\in\mathbf{B}_j}\rho|g_{S,m}|^2 = \max_{m\in\mathbf{C}_l}\rho|g_{S,m}|^2 = Z_2$. Then, applying the joint PDFs $f_{Y_1,Y_2}(y_1,y_2)$ and $f_{Z_1,Z_2}(z_1,z_2)$ into (D.2) with some algebraic manipulations, an expression for χ is obtained as shown at the seventh case of (30). Following the same rationale, the expressions for $\{l=1,i\geq 2,j\geq l\}$, $\{l=1,i=1,j\geq l\}$ and $\{l\geq 2,i=1,j\geq l\}$ can also be obtained as shown at the fifth, sixth and eighth cases of (30), respectively.

Appendix E
Proof of Lemma 4

Suppose that a number n of independent but non-identically distributed RVs $W_1,W_2,...,W_n$ are ranked into an ascending order as $W_{(1)}<W_{(2)}<\cdots<W_{(n)}$. Then, according to [35, eq.(5.2.8)], the joint PDF of $W_{(1)}=\min_{i=1,...,n}W_i$ and $W_{(n)}=\max_{i=1,...,n}W_i$ can be expressed as

$$\begin{aligned}
f_{W_{(1)},W_{(n)}}(x,y)&=\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}f_{W_i}(x)f_{W_j}(y)\\
&\quad\times\prod_{r=1,r\neq i,j}^{n}\left[F_{W_r}(y)-F_{W_r}(x)\right],
\end{aligned}\tag{E.1}$$

for $x\leq y$ or $f_{W_{(1)},W_{(n)}}(x,y)=0$ otherwise.

Following the above results, for $X_1=\min_{m\in\mathbf{B}_j}\rho|g_{S,m}|^2$ and $X_2=\max_{m\in\mathbf{B}_j}\rho|g_{S,m}|^2$, we have

$$f_{X_1,X_2}(x_1,x_2)=\sum_{m_1^\star\in\mathbf{B}_j}\sum_{m_2^\star\in\mathbf{B}_j\setminus\{m_1^\star\}}f_{\rho|g_{S,m_1^\star}|^2}(x_1)$$

$$\times f_{\rho|g_{S,m_2^\star}|^2}(x_2)\prod_{m\in\mathbf{B}_j\setminus\{m_1^\star,m_2^\star\}}\Big[F_{\rho|g_{S,m}|^2}(x_2)-F_{\rho|g_{S,m}|^2}(x_1)\Big]. \qquad \text{(E.2)}$$

Since $|g_{S,m}|^2$ is exponentially distributed with mean values $G_{S,m}$, the CDF and PDF of $\rho|g_{S,m}|^2$ are given by $F_{\rho|g_{S,m}|^2}(x) = 1 - e^{-x/(\rho G_{S,m})}$ and $f_{\rho|g_{S,m}|^2} = (1/\rho G_{S,m})e^{-x/(\rho G_{S,m})}$, respectively. Applying these results into (E.2), $f_{X_1,X_2}(x_1,x_2)$ can be derived as shown in (34). Following the above derivations, $f_{Y_1,Y_2}(y_1,y_2)$ and $f_{Z_1,Z_2}(z_1,z_2)$ can also be obtained as shown in (35) and (36), respectively.

Appendix F
Derivations of Approximation in (37)

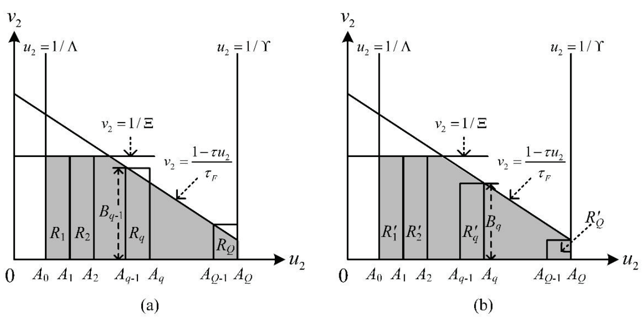


Fig. 10. An upper and lower bound of the integral region $R(u_2,v_2)$.

We first derive an approximation for the case $\{l=0, i\geq 2, j\geq 2\}$. Recall $f_{X_1,X_2}(x_1,x_2)=0$ for $x_1\geq x_2$ and $f_{Y_1,Y_2}(y_1,y_2)=0$ for $y_1\geq y_2$. Thus, for the case $\{l=0,i\geq 2, j\geq 2\}$, the corresponding integral expression in (30) can be equivalently written as

$$\chi = \iiiint\limits_{\substack{y_2>y_1\geq\Xi,\Upsilon\leq x_1<x_2<\Lambda,\\ \tau/x_2+\tau_F/y_2\leq 1}} f_{X_1,X_2}(x_1,x_2) \times f_{Y_1,Y_2}(y_1,y_2)\mathrm{d}y_1\mathrm{d}x_1\mathrm{d}y_2\mathrm{d}x_2. \qquad \text{(F.1)}$$

Then, denoting $u_1=1/x_1$, $u_2=1/x_2$, $v_1=1/y_1$ and $v_2=1/y_2$, we have

$$\chi = \iint\limits_{R(u_2,v_2)}\int_{v_2}^{1/\Xi}\int_{u_2}^{1/\Upsilon}\frac{1}{u_1^2u_2^2v_1^2v_2^2}f_{X_1,X_2}\left(\frac{1}{u_1},\frac{1}{u_2}\right) f_{Y_1,Y_2}\left(\frac{1}{v_1},\frac{1}{v_2}\right)\mathrm{d}u_1\mathrm{d}v_1\mathrm{d}u_2\mathrm{d}v_2, \qquad \text{(F.2)}$$

with integral region $R(u_2,v_2)\triangleq\{1/\Lambda<u_2\leq 1/\Upsilon, 0<v_2\leq\min\{1/\Xi,(1-\tau u_2)/\tau_F\}\}$. As shown in Fig. 10, integral region $R(u_2,v_2)$ (the shadowed area) is included in the combination of Q rectangles denoted by $\{R_1,...,R_Q\}$, and includes the combination of Q rectangles denoted by $\{R'_1,...,R'_Q\}$. All the rectangles have the same bottom edge size, i.e., $A_q-A_{q-1}=\frac{1}{Q}(\frac{1}{\Upsilon}-\frac{1}{\Lambda})$ for $q=1,...,Q$. Since $A_0=1/\Lambda$ and $A_Q=1/\Upsilon$, we have $A_q=\frac{1}{\Lambda}+\frac{q}{Q}(\frac{1}{\Upsilon}-\frac{1}{\Lambda})$ for $q=0,...,Q$. Further, by observing Fig. 10, we have $R_q=\{A_{q-1}<u_2<A_q, 0<v_2\leq B_{q-1}\}$ and $R'_q=\{A_{q-1}<u_2<A_q, 0<v_2\leq B_q\}$, where $B_q\triangleq\min\{\frac{1}{\Xi},\frac{1}{\tau_F}(1-\tau A_q)\}=\min\{\frac{1}{\Xi},\frac{1}{\tau_F}(1-\tau[\frac{1}{\Lambda}+\frac{q}{Q}(\frac{1}{\Upsilon}-\frac{1}{\Lambda})])\}$ for $q=0,...,Q$. Thus, for case $\{l=0,i\geq 2,j\geq 2\}$, we can get a lower bound and an upper bound of χ shown in the following

$$\sum_{q=1}^{Q}\underbrace{\int_0^{B_q}\int_{v_2}^{1/\Xi}\frac{1}{v_1^2v_2^2}f_{Y_1,Y_2}\left(\frac{1}{v_1},\frac{1}{v_2}\right)\mathrm{d}v_1\mathrm{d}v_2}_{\triangleq I_1(B_q)} \times\underbrace{\int_{A_{q-1}}^{A_q}\int_{u_2}^{1/\Upsilon}\frac{1}{u_1^2u_2^2}f_{X_1,X_2}\left(\frac{1}{u_1},\frac{1}{u_2}\right)\mathrm{d}u_1\mathrm{d}u_2}_{\triangleq J_{1,q}} \leq\chi\leq\sum_{q=1}^{Q}I_1(B_{q-1})J_{1,q}. \qquad \text{(F.3)}$$

Using the joint PDFs in (34) and (35) with letting $x_1=1/u_1$, $x_2=1/u_2$, $y_1=1/v_1$ and $y_2=1/v_2$, closed-form expressions for $I_1(a)$ and $J_{1,q}$ can be derived as follows

$$I_1(a)=\sum_{k_1^\star\in\mathbf{A}_i}\sum_{k_2^\star\in\mathbf{A}_i\setminus\{k_1^\star\}}\sum_{r=0}^{i-2}\sum_{\substack{\hat{\mathbf{A}}_r\subseteq\mathbf{A}_i\setminus\{k_1^\star,k_2^\star\}\\|\hat{\mathbf{A}}_r|=r}}(-1)^r \times\left(\frac{e^{-\frac{1}{\rho}\left(\Xi\Omega_1+\frac{\Omega_2}{a}\right)}}{G_{S,k_1^\star}G_{S,k_2^\star}\Omega_1\Omega_2}-\frac{e^{-\frac{\Omega_1+\Omega_2}{\rho a}}}{G_{S,k_1^\star}G_{S,k_2^\star}\Omega_1(\Omega_1+\Omega_2)}\right), \qquad \text{(F.4)}$$

$$J_{1,q}=\sum_{m_1^\star\in\mathbf{B}_j}\sum_{m_2^\star\in\mathbf{B}_j\setminus\{m_1^\star\}}\sum_{s=0}^{j-2}\sum_{\substack{\hat{\mathbf{B}}_s\subseteq\mathbf{B}_j\setminus\{m_1^\star,m_2^\star\}\\|\hat{\mathbf{B}}_s|=s}}(-1)^s \times\left(\frac{e^{-\frac{1}{\rho}\left(\Upsilon\Theta_1+\frac{\Theta_2}{A_q}\right)}-e^{-\frac{1}{\rho}\left(\Upsilon\Theta_1+\frac{\Theta_2}{A_{q-1}}\right)}}{G_{S,m_1^\star}G_{S,m_2^\star}\Theta_1\Theta_2} -\frac{e^{-\frac{\Theta_1+\Theta_2}{\rho A_q}}-e^{-\frac{\Theta_1+\Theta_2}{\rho A_{q-1}}}}{G_{S,m_1^\star}G_{S,m_2^\star}\Theta_1(\Theta_1+\Theta_2)}\right), \qquad \text{(F.5)}$$

where $\Omega_1\triangleq\frac{1}{G_{S,k_1^\star}}+\sum_{k\in\mathbf{A}_i\setminus(\{k_1^\star,k_2^\star\}\cup\hat{\mathbf{A}}_r)}\frac{1}{G_{S,k}}$, $\Omega_2\triangleq\frac{1}{G_{S,k_2^\star}}+\sum_{k\in\hat{\mathbf{A}}_r}\frac{1}{G_{S,k}}$, $\Theta_1\triangleq\frac{1}{G_{S,m_1^\star}}+\sum_{m\in\mathbf{B}_j\setminus(\{m_1^\star,m_2^\star\}\cup\hat{\mathbf{B}}_s)}\frac{1}{G_{S,m}}$ and $\Theta_2\triangleq\frac{1}{G_{S,m_2^\star}}+\sum_{m\in\hat{\mathbf{B}}_s}\frac{1}{G_{S,m}}$. Finally, we approximate χ as the average of the lower and upper bounds shown in (F.3), and we get the approximate expression as shown in the first case of (37).

By following the above rationale with letting $x_0=1/u_0$, $y_0=1/v_0$, $w_n=1/z_n$ $(n=0,1,2)$, the upper and lower bounds of χ for the other seven cases of (30) can be expressed as shown in (F.6) (at the top of the next page), with $I_2(a)$, $J_{2,q}$, $J_{3,q}$ and $J_{4,q}$ being defined as

$$\begin{cases} I_2(a) & \triangleq & \int_0^a\frac{1}{v_0^2}f_{Y_0}\left(\frac{1}{v_0}\right)\mathrm{d}v_0,\\ J_{2,q} & \triangleq & \int_{A_{q-1}}^{A_q}\frac{1}{u_0^2}f_{X_0}\left(\frac{1}{u_0}\right)\mathrm{d}u_0,\\ J_{3,q} & \triangleq & \int_{C_{q-1}}^{C_q}\frac{1}{w_0^2}f_{Z_0}\left(\frac{1}{w_0}\right)\mathrm{d}w_0,\\ J_{4,q} & \triangleq & \int_{C_{q-1}}^{C_q}\int_{w_2}^{1/\Lambda}\frac{1}{w_1^1w_2^2}f_{Z_1,Z_2}\left(\frac{1}{w_1},\frac{1}{w_2}\right)\mathrm{d}w_1\mathrm{d}w_2,\end{cases}$$

$$\begin{cases} \sum_{q=1}^{Q} I_2(B_q)J_{1,q} \le \chi \le \sum_{q=1}^{Q} I_2(B_{q-1})J_{1,q}, & l=0, i=1, j\ge 2;\\ \sum_{q=1}^{Q} I_1(B_q)J_{2,q} \le \chi \le \sum_{q=1}^{Q} I_1(B_{q-1})J_{2,q}, & l=0, i\ge 2, j=1;\\ \sum_{q=1}^{Q} I_2(B_q)J_{2,q} \le \chi \le \sum_{q=1}^{Q} I_2(B_{q-1})J_{2,q}, & l=0, i=1, j=1;\\ \varrho_{j,l}\cdot\sum_{q=1}^{Q} I_1(D_q)J_{3,q} \le \chi \le \varrho_{j,l}\cdot\sum_{q=1}^{Q} I_1(D_{q-1})J_{3,q}, & l=1, i\ge 2, j\ge l;\\ \varrho_{j,l}\cdot\sum_{q=1}^{Q} I_2(D_q)J_{3,q} \le \chi \le \varrho_{j,l}\cdot\sum_{q=1}^{Q} I_2(D_{q-1})J_{3,q}, & l=1, i=1, j\ge l;\\ \varrho_{j,l}\cdot\sum_{q=1}^{Q} I_1(D_q)J_{4,q} \le \chi \le \varrho_{j,l}\cdot\sum_{q=1}^{Q} I_1(D_{q-1})J_{4,q}, & l\ge 2, i\ge 2, j\ge l;\\ \varrho_{j,l}\cdot\sum_{q=1}^{Q} I_2(D_q)J_{4,q} \le \chi \le \varrho_{j,l}\cdot\sum_{q=1}^{Q} I_2(D_{q-1})J_{4,q}, & l\ge 2, i=1, j\ge l. \end{cases} \tag{F.6}$$

where $C_q \triangleq \frac{q}{Q\Lambda}$ and $D_q \triangleq \min\left\{\frac{1}{\Xi}, \frac{1}{\tau_F}\left(1-\frac{q\tau}{Q\Lambda}\right)\right\}$. Using derived PDFs in (33) with letting $x_0 = 1/u_0$, $y_0 = 1/v_0$ and $z_0 = 1/w_0$, $I_2(a)$, $J_{2,q}$ and $J_{3,q}$ can be derived as follows

$$I_2(a) = e^{-\frac{1}{\rho G_{S,k_1} a}}, \tag{F.7}$$

$$J_{2,q} = e^{-\frac{1}{\rho G_{S,m_1} A_q}} - e^{-\frac{1}{\rho G_{S,m_1} A_{q-1}}}, \tag{F.8}$$

$$J_{3,q} = e^{-\frac{1}{\rho G_{S,m_1'} C_q}} - e^{-\frac{1}{\rho G_{S,m_1'} C_{q-1}}}. \tag{F.9}$$

Further, applying the joint PDF given in (36) with letting $z_1 = 1/w_1$ and $z_2 = 1/w_2$, $J_{4,q}$ is obtained as

$$\begin{aligned} J_{4,q} = & \sum_{m_1^\star\in\mathbf{C}_l}\sum_{m_2^\star\in\mathbf{C}_l\setminus\{m_1^\star\}}\sum_{t=0}^{l-2}\sum_{\substack{\hat{\mathbf{C}}_t\subseteq\mathbf{C}_l\setminus\{m_1^\star,m_2^\star\}\\|\hat{\mathbf{C}}_t|=t}}(-1)^t \\ & \times\left(\frac{e^{-\frac{1}{\rho}\left(\Lambda\Delta_1+\frac{\Delta_2}{C_q}\right)} - e^{-\frac{1}{\rho}\left(\Lambda\Delta_1+\frac{\Delta_2}{C_{q-1}}\right)}}{G_{S,m_1^\star}G_{S,m_2^\star}\Delta_1\Delta_2}\right. \\ & \left. - \frac{e^{-\frac{\Delta_1+\Delta_2}{\rho C_q}} - e^{-\frac{\Delta_1+\Delta_2}{\rho C_{q-1}}}}{G_{S,m_1^\star}G_{S,m_2^\star}\Delta_1(\Delta_1+\Delta_2)}\right), \end{aligned} \tag{F.10}$$

where $\Delta_1 \triangleq \frac{1}{G_{S,m_1^\star}} + \sum_{m\in\mathbf{C}_l\setminus(\{m_1^\star,m_2^\star\}\cup\hat{\mathbf{C}}_t)}\frac{1}{G_{S,m}}$ and $\Delta_2 \triangleq \frac{1}{G_{S,m_2^\star}} + \sum_{m\in\hat{\mathbf{C}}_t}\frac{1}{G_{S,m}}$. Overall, for the cases in (F.6), by using the average of the lower and upper bounds in each case, the approximate expressions of χ are derived as shown in the second to the eighth cases of (37).

Appendix G
Proof of Lemma 5

According to (20) and (22), probability $P_{\rm out}(\mathbf{A}_i)$ can be expressed as shown in (G.1) (at the top of the next page), where step (g.1) uses the fact that $\frac{\tau}{\rho|g_{S,m}|^2} + \frac{\tau_F}{\rho|g_{S,k}|^2} < 2\cdot\max\{\frac{\tau}{\rho|g_{S,m}|^2}, \frac{\tau_F}{\rho|g_{S,k}|^2}\}$. Then, with some algebraic manipulations, (G.1) can be further rewritten as shown in (G.2) (on the next page), where step (g.2) uses the fact $\Pr\{\mathbb{A}\cup\mathbb{B}\} = \Pr\{\mathbb{A}\} + \Pr\{\mathbb{B}\} - \Pr\{\mathbb{A}\cap\mathbb{B}\}$ holds for any $\mathbb{A}$ and $\mathbb{B}$.

With some mathematical derivations, we have

$$\begin{aligned} & \Pr\left\{\hat{\mathbb{E}}_1, \hat{\mathbb{E}}_2, \mathbf{F}_\Xi = \mathbf{A}_i\right\} \\ = & \prod_{m=1}^{M}\left[\left(1-e^{-\frac{\min\{\Lambda,\max(2\tau,\Upsilon)\}}{\rho G_{S,m}}}\right) + \left(1-e^{-\frac{\max(2\tau,\Upsilon)}{\rho G_{S,m}}}\right)\right. \\ & \times\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\left(1-e^{-\frac{\tau_F}{\mu\rho G_{m,k}}}\right) - \left(1-e^{-\frac{\min\{\Lambda,\max(2\tau,\Upsilon)\}}{\rho G_{S,m}}}\right) \\ & \left.\times\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\left(1-e^{-\frac{\tau_F}{\mu\rho G_{m,k}}}\right)\right] e^{-\sum_{k\in\mathbf{A}_i}\frac{\Xi}{\rho G_{S,k}}} \\ & \times\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\left(1-e^{-\frac{\Xi}{\rho G_{S,k}}}\right). \end{aligned} \tag{G.3}$$

Then, applying the series representation of exponential function into (G.3) and ignoring the high-order infinitesimals, a high-SNR asymptotic expression is obtained as follows

$$\begin{aligned} \Pr\left\{\hat{\mathbb{E}}_1, \hat{\mathbb{E}}_2, \mathbf{F}_\Xi = \mathbf{A}_i\right\} \overset{\rho\to\infty}{\simeq} & \frac{1}{\rho^{M+K-i}} \\ & \times\left(\prod_{m=1}^{M}\frac{\min\{\Lambda,\max(2\tau,\Upsilon)\}}{G_{S,m}}\right) \\ & \times\left(\prod_{k\in\mathbf{K}\setminus\mathbf{A}_i}\frac{\Xi}{G_{S,k}}\right) \\ \propto & \frac{1}{\rho^{M+K-i}}. \end{aligned} \tag{G.4}$$

Following the above derivations, it can also be shown that if $2\tau_F > \Xi$, both $\Pr\{\hat{\mathbb{E}}_1, \hat{\mathbb{E}}_3, \mathbf{F}_\Xi = \mathbf{A}_i\}$ and $\Pr\{\hat{\mathbb{E}}_1, \hat{\mathbb{E}}_2, \hat{\mathbb{E}}_3, \mathbf{F}_\Xi = \mathbf{A}_i\}$ are proportional to $1/\rho^{M+K}$ ($\ll 1/\rho^{M+K-i}$ for $i\in\mathbf{K}$) in high-SNR regime; otherwise, they both are equal to zero. Applying this fact and (G.4) into (G.2), this lemma is obtained.

References

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in Proc. *IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[2] NTT DOCOMO, "5G radio access: Requirements, concept and technologies," NTT DOCOMO 5G White paper, Tokyo, Japan, Jul. 2014. [Online]. Available: https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/whitepaper_5g/DOCOMO_5G_White_Paper.pdf

[3] P. Popovski, *et al.*, *Proposed Solutions for New Radio Access*, document ICT-317669-METIS/ D2.4, European Commission/Seventh Framework Program 2015, Mobile and Wirless Communications Enabler for the Twenty-Twenty Information Society (METIS), Feb. 2015.

[4] Z. Ding, *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

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

$$
\begin{aligned}
P_{\rm out}(\mathbf{A}_i) &= \Pr\left\{\bigcap_{k\in\mathbf{K}\setminus\mathbf{A}_i}\mathbb{E}_{1,k},\bigcap_{k\in\mathbf{A}_i}\left(\bigcap_{m=1}^{M}\max\left\{\frac{\Upsilon}{\rho|g_{S,m}|^2},\frac{\tau}{\rho|g_{S,m}|^2}+\frac{\tau_F}{\rho|g_{S,k}|^2}\right\}>1\right),\mathbf{F}_{\Xi}=\mathbf{A}_i\right\}\\
&\overset{(g.1)}{<}\Pr\left\{\bigcap_{k\in\mathbf{K}\setminus\mathbf{A}_i}\mathbb{E}_{1,k},\bigcap_{k\in\mathbf{A}_i}\left(\bigcap_{m=1}^{M}\max\left\{\frac{\Upsilon}{\rho|g_{S,m}|^2},2\cdot\max\left\{\frac{\tau}{\rho|g_{S,m}|^2},\frac{\tau_F}{\rho|g_{S,k}|^2}\right\}\right\}>1\right),\mathbf{F}_{\Xi}=\mathbf{A}_i\right\}\\
&=\Pr\left\{\bigcap_{k\in\mathbf{K}\setminus\mathbf{A}_i}\mathbb{E}_{1,k},\bigcap_{k\in\mathbf{A}_i}\left(\bigcap_{m=1}^{M}\max\left\{\frac{\max(2\tau,\Upsilon)}{\rho|g_{S,m}|^2},\frac{2\tau_F}{\rho|g_{S,k}|^2}\right\}>1\right),\mathbf{F}_{\Xi}=\mathbf{A}_i\right\}. \qquad \text{(G.1)}
\end{aligned}
$$

$$
\begin{aligned}
P_{\rm out}(\mathbf{A}_i) &< \Pr\Bigg\{\underbrace{\bigcap_{k\in\mathbf{K}\setminus\mathbf{A}_i}\mathbb{E}_{1,k}}_{\hat{\mathbb{E}}_1},\underbrace{\left(\bigcap_{m=1}^{M}\rho|g_{S,m}|^2<\max(2\tau,\Upsilon)\right)}_{\hat{\mathbb{E}}_2}\bigcup\underbrace{\left(\bigcap_{k\in\mathbf{A}_i}\rho|g_{S,k}|^2<2\tau_F\right)}_{\triangleq\hat{\mathbb{E}}_3},\mathbf{F}_{\Xi}=\mathbf{A}_i\Bigg\}\\
&\overset{(g.2)}{=}\Pr\left\{\hat{\mathbb{E}}_1,\hat{\mathbb{E}}_2,\mathbf{F}_{\Xi}=\mathbf{A}_i\right\}+\Pr\left\{\hat{\mathbb{E}}_1,\hat{\mathbb{E}}_3,\mathbf{F}_{\Xi}=\mathbf{A}_i\right\}-\Pr\left\{\hat{\mathbb{E}}_1,\hat{\mathbb{E}}_2,\hat{\mathbb{E}}_3,\mathbf{F}_{\Xi}=\mathbf{A}_i\right\}. \qquad \text{(G.2)}
\end{aligned}
$$

[6] F. Zhou, Y. Wu, R. Q. Hu, Y. Wang and K. K. Wong, “Energy-efficient NOMA enabled heterogeneous cloud radio access networks,” *IEEE Network*, vol. 32, no. 2, pp. 152–160, Mar. 2018.

[7] F. Zhou, Y. Wu, Y. Liang, Z. Li, Y. Wang and K. Wong, “State of the art, taxonomy, and open issues on cognitive radio networks with NOMA,” *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 100-108, April 2018.

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

[9] Z. Yang, W. Xu, C. Pan, Y. Pan, and M. Chen, “On the optimality of power allocation for NOMA downlinks with individual QoS constraints,” *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1649–1652, July 2017.

[10] Z. Yang, C. Pan, W. Xu, Y. Pan, M. Chen, and M. Elkashlan, “Power control for multi-cell networks with non-orthogonal multiple access,” *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 927–942, Feb. 2018.

[11] R. Jiao, L. Dai, J. Zhang, R. MacKenzie, and M. Hao, “On the performance of NOMA-based cooperative relaying systems over Rician fading channels,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11409–11414, Dec. 2017.

[12] J.-B. Kim and I.-H. Lee, “Non-orthogonal multiple access in coordinated direct and relay transmission,” *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 2037–2040, Nov. 2015.

[13] C. Zhong and Z. Zhang, “Non-orthogonal multiple access with cooperative full-duplex relaying,” *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2478–2481, Dec. 2016.

[14] Z. Ding, H. Dai and H. V. Poor, “Relay selection for cooperative NOMA,” *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416–419, Aug. 2016.

[15] J. Men and J. Ge, “Non-orthogonal multiple access for multiple-antenna relaying networks,” *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1686–1689, Oct. 2015.

[16] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, “Full-duplex device-to-device-aided cooperative nonorthogonal multiple access,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4467–4471, May 2017.

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

[18] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, “Cooperative diversity in wireless networks: Efficient protocols and outage behavior,” *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[19] 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.

[20] 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.

[21] N. T. Do, D. B. da Costa, T. Q. Duong, and B. An, “A BNBF user selection scheme for NOMA-based cooperative relaying systems with SWIPT,” *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 664–667, Mar. 2017.

[22] K. Janghel and S. Prakriya, “Performance of adaptive OMA/cooperative-NOMA scheme with user selection,” *IEEE Comm. Lett.*, vol. 22, no. 10, pp. 2092–2095, Oct. 2018.

[23] L. Lv, L. Yang, H. Jiang, T. H. Luan and J. Chen, “When NOMA meets multiuser cognitive radio: Opportunistic cooperation and user scheduling,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6679–6684, Jul. 2018.

[24] 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, June 2017.

[25] L. Yang, J. Chen, Q. Ni, J. Shi, and X. Xue, “NOMA-enabled cooperative unicast-multicast: Design and outage analysis,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7870–7889, Dec. 2017.

[26] L. Yang, Q. Ni, L. Lv, J. Chen, X. Xue, H. Zhang, and H. Jiang, “Cooperative non-orthogonal layered multicast multiple access for heterogeneous networks,” *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1148–1165, Feb. 2019.

[27] A. Zafar, M. Shaqfeh, M. Alouini and H. Alnuweiri, “On multiple users scheduling using superposition coding over rayleigh fading channels,” *IEEE Commun. Lett.*, vol. 17, no. 4, pp. 733–736, April 2013.

[28] L. Yang, M.-S. Alouini, “Performance analysis of multiuser selection diversity,” *IEEE Trans. Veh. Technol.*, vol. 55, no. 6, pp. 1848–1861, Nov. 2006.

[29] S.-H. Hur and B. D. Rao, “Sum rate analysis of a reduced feedback OFDMA downlink system employing joint scheduling and diversity,” *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 862–876, Feb. 2012.

[30] K. J. Kim and T. A. Tsiftsis, “Performance Analysis of cyclically prefixed single-carrier transmissions with outdated opportunistic user selection,” *IEEE Signal Process. Lett.*, vol. 17, no. 10, pp. 847–850, Oct. 2010.

[31] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed., Englewood Cliffs, NJ, USA: Prentice Hall, 1992.

[32] H. Shi, R. V. Prasad, Ertan Onur, and I.G.M.M. Niemegeers, “Fairness in wireless networks: Issues, measures and challenges,” *IEEE Commun. Surveys Tutorials*, vol. 16, no. 1, pp. 5–24.

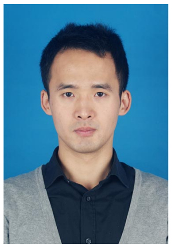
[33] F. Kelly, “Charging and rate control for elastic traffic,” *European Transactions on Telecommunications*, vol. 8, pp. 33–38, Feb. 1997.

[34] F. Kelly, A. K. Maulloo, and D. K. Tan, “Rate control for communication networks: Shadow prices, proportional fairness and stability”, *Journal of the Operational Research Society*, vol. 49, no. 3, pp. 237–252, Mar. 1998.

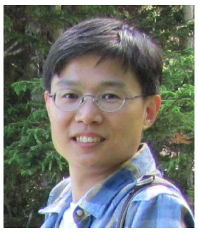
[35] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed., New York, NY, USA: Wiley, 2003.

[36] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed., New York, NY, USA: Academic, 2007.

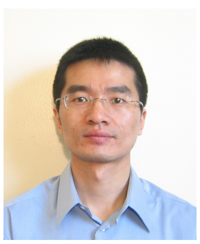
[37] A. Goldsmith, *Wireless Communications*, Cambridge, U.K.: Cambridge University Press, 2005.



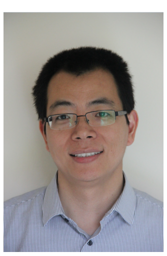
Long Yang (Member, IEEE) received the B.E. and Ph.D. degrees from Xidian University, Xi'an, China, in 2010 and 2015, respectively. Since 2015, he has been a Faculty Member with Xidian University, where he is currently an Associate Professor with the State Key Laboratory of Integrated Services Networks. From 2017 to 2019, he was also a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Alberta, Canada. His current research interests include non-orthogonal multiple access, cooperative communications and wireless physical-layer security.



Hai Jiang (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electronics engineering from Peking University, Beijing, China, in 1995 and 1998, respectively, and the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, ON, Canada, in 2006. Since 2007, he has been a Faculty Member with the University of Alberta, Edmonton, AB, Canada, where he is currently a Professor with the Department of Electrical and Computer Engineering. His research interests include radio resource management, cognitive radio networking, and cooperative communications.



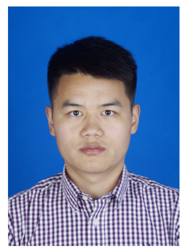
Qiang Ye (Member, IEEE) is an Associate Professor in the Faculty of Computer Sciences at Dalhousie University, Canada. His current research interests lie in the area of communication networks in general. Specifically, he is interested in Wireless Networks, Mobile Computing, Internet of Things (IoT), Network Security, and Data Analytics. He has published a series of papers in top publication venues such as IEEE/ACM Transactions on Networking (TON), IEEE Transactions on Parallel and Distributed Systems (TPDS), and IEEE Transactions on Wireless Communications (TWC). He received a Ph.D. in Computing Science from the University of Alberta in 2007. His M. Engr. and B. Engr. in Computer Science and Technology are from Harbin Institute of Technology, P.R. China. He is a Member of IEEE and ACM.



Zhiguo Ding (Fellow, IEEE) received his B.Eng in Electrical Engineering from the Beijing University of Posts and Telecommunications in 2000, and the Ph.D degree in Electrical Engineering from Imperial College London in 2005. From Jul. 2005 to Apr. 2018, he was working in Queen's University Belfast, Imperial College, Newcastle University and Lancaster University. Since Apr. 2018, he has been with the University of Manchester as a Professor in Communications. From Oct. 2012 to Sept. 2020, he has also been an academic visitor in Princeton University.

Dr. Ding's research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He is serving as an Area Editor for the *IEEE Open Journal of the Communications Society*, an Editor for *IEEE Transactions on Communications*, *IEEE Transactions on Vehicular Technology*, and *Journal of Wireless Communications and Mobile Computing*, and was an Editor for *IEEE Wireless Communication Letters*, *IEEE Communication Letters* from 2013 to 2016. He received the best paper award in IET ICWMC-2009 and IEEE WCSP-2014, the EU Marie Curie Fellowship 2012-2014, the Top IEEE TVT Editor 2017, IEEE Heinrich Hertz Award 2018, IEEE Jack Neubauer Memorial Award 2018, IEEE Best Signal Processing Letter Award 2018 ,and Web of Science Highly Cited Researcher 2019.

Fang Fang (Member, IEEE) received the B.A.Sc. and the M.A.Sc. degrees in electronic engineering from Lanzhou University in 2010 and 2013, respectively, and the Ph.D. degree in electrical engineering from the University of British Columbia (UBC), Kelowna, BC, Canada, in 2018. She is currently a Research Associate with the Department of Electrical and Electronic Engineering, The University of Manchester, UK and an Assistant Professor with the Department of Engineering, Durham University, Durham, UK. Her current research interests include 5G and beyond wireless networks, NOMA, IRS and mobile edge computing. She has served as a TPC Member for IEEE conferences, e.g., GLOBECOM and ICC. She received the Exemplary Reviewer Certificate of the *IEEE Transactions on Communications* in 2017. Currently, she is an Associate Editor of the *IEEE Open Journal of The Communications Society*.



Jia Shi received both his MSc. and Ph.D degrees from University of Southampton, UK, in 2010 and 2015, respectively. He was a research associate with Lancaster University, UK, during 2015-2017. Then, he became a research fellow with 5GIC, University of Surrey, UK, from 2017 to 2018. Since 2018, he joined Xidian University, China, and now is an Associate Professor in the State Key Laboratory of Integrated Services Networks (ISN). His current research interests include mmWave communications, resource allocation in wireless systems, covert communications, physical layer security, etc. He is now serving as an Associate Editor for *Electronic Letters*, and an Editor for *International Journal of Communications System*, and is serving as a Guest Editor for *China Communications*.

Jian Chen (Member, IEEE) received the B.S. degree from Xi'an Jiaotong University, China, in 1989, the M.S. degree from Xi'an Institute of Optics and Precision Mechanics of Chinese Academy of Sciences in 1992, and the Ph.D. degree in Telecommunications Engineering in Xidian University, China, in 2005. He is a Professor with the State Key Laboratory of Integrated Services Networks (ISN), Xidian University. He was a visitor scholar in the University of Manchester from 2007 to 2008. His research interests include cognitive radio, OFDM, wireless sensor networks and non-orthogonal multiple access.

Xuan Xue (Member, IEEE) received the B.E. and Ph.D. degrees from Xidian University, Xi'an, China, in 2010 and 2017, respectively. From 2013 to 2015, she was a Visiting Ph.D. Student with Western University, London, ON, Canada. Since 2018, she has been a Faculty Member with Xidian University, where she is currently a Lecturer with the State Key Laboratory of Integrated Services Networks. Her research interests include massive MIMO, millimeter-wave communications, cooperative communications, and non-orthogonal multiple access.