NOMA in Downlink SDMA With Limited Feedback: Performance Analysis and Optimization

Qian Yang, *Student Member, IEEE*, Hui-Ming Wang, *Senior Member, IEEE*, Derrick Wing Kwan Ng, *Member, IEEE*, and Moon Ho Lee, *Life Senior Member, IEEE*

***Abstract*— In this paper, the performance of non-orthogonal multiple access (NOMA) is investigated and optimized in a downlink space division multiple access network with a multi-antenna base station and randomly deployed users, under a general channel state information (CSI) limited feedback framework. We first propose a dynamic user scheduling and grouping strategy by leveraging limited feedback. Based on that, an analytical framework is proposed to obtain the outage probability of the network in closed form. The diversity order and the impacts of the number of feedback bits on the outage performance of NOMA are analyzed. Furthermore, the net throughput, which captures the network-wide throughput with the uplink feedback cost considered, is maximized by optimizing the number of feedback bits. Numerical results are demonstrated to verify our analytical findings and show that different from the perfect CSI case, there always exists a performance floor of outage probability in the considered network due to limited feedback. Moreover, the optimal number of feedback bits for net throughput maximization increases as the channel coherence time becomes longer.**

***Index Terms*— Multi-user MIMO, non-orthogonal multiple access (NOMA), space division multiple access (SDMA), limited feedback.**

I. Introduction

NON-ORTHOGONAL multiple access (NOMA), as a promising and enabling technology to enhance the spectral efficiency of wireless communications, has received increasing research interests in the fifth generation (5G) networks [2], [3]. Different from traditional orthogonal multiple access (OMA), NOMA allows multiple users to be simultaneously served at the same resource block (same time, frequency, code, and spatial domain). Therefore, a large number of simultaneous transmissions can be accommodated within limited resources, and spectral efficiency is further improved [4]–[6]. The main idea to realize NOMA lies in the exploitation of the power domain. In NOMA, the signals of multiple users are superimposed with different power allocations according to their different channel conditions, and multi-user signal detection on the receiver side is conducted via successive interference cancellation (SIC) [7]. The concept of NOMA is essentially a special case of superposition coding (SC) developed for broadcast channels with user fairness taken into consideration [8]. Interested readers may refer to [9] for a survey of the recent progress of NOMA in 5G systems.

The NOMA technology has been widely investigated under *single-antenna* systems for spectral efficiency improvement [10]–[13]. In [10], the performance of NOMA is studied under a cellular downlink scenario with randomly deployed users. The power allocation problem in NOMA downlink systems is studied from a fairness and an energy efficiency maximization perspective in [11] and [12], respectively. The impact of user pairing on the performance of NOMA systems is characterized in [13]. In addition, the prevailing physical-layer security issues emerged in NOMA systems are studied in [14] and [15]. As a basic way of multiple access, the concept of NOMA has also been extended to various scenarios, e.g., coordinated multi-point (CoMP) [16], cooperative [17]–[20], cognitive radio [21], [22], and wireless power transfer [23], [24] networks. However, almost all of the above works confine the study of NOMA to perfect channel state information (CSI) cases. Yang *et al.* [25] derive the outage probability and the average sum rate based on imperfect CSI and second order statistics, respectively. In [26], the outage balancing issue is studied in downlink NOMA systems with statistical CSI. Yet the performance of the multi-antenna NOMA case remains unexplored in [25] and [26].

The implementation of NOMA in *multi-antenna* broadcast systems is of great importance, since spectral efficiency can be further improved with additional degrees of freedom (DoFs) [4]. When it comes to multi-antenna NOMA systems, the situation becomes more involved and the framework developed in single-antenna cannot be directly exploited.

Manuscript received January 18, 2017; revised May 15, 2017; accepted May 22, 2017. Date of publication July 11, 2017; date of current version September 15, 2017. The work of Q. Yang and H.-M. Wang was supported in part by the National Natural Science Foundation of China under Grant 61671364, in part by the Foundation for the Author of National Excellent Doctoral Dissertation of China under Grant 201340, and in part by the Young Talent Support Fund of Science and Technology of Shaanxi Province under Grant 2015KJXX-01. The work of D. W. K. Ng was supported through the Australian Research Council's Discovery Early Career Researcher Award funding scheme under Grant DE170100137. The work of M. H. Lee was supported by NRF, South Korea, under Grant MEST 2015R1A2A1A05000977. This paper was partially presented at the IEEE International Conference on Communications, Paris, France, May 2017 [1]. *(Corresponding author: Hui-Ming Wang.)*

Q. Yang and H.-M. Wang are with the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: yangq36@gmail.com; xjbswhm@gmail.com).

D. W. K. Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia (e-mail: w.k.ng@unsw.edu.au).

M. H. Lee is with the Division of Electronics Engineering, Chonbuk National University, Jeonju 561-756, South Korea (e-mail: moonho@jbnu.ac.kr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2017.2725107

The high-dimensional channel gains are coupled with precoding matrices and performing optimal resource allocation is non-trivial [27]. To circumvent this problem, there exist two main branches of research on the multi-antenna NOMA system:

1) In most of the research, the NOMA principle is directly exploited among *all the users* for a pre-assigned SIC ordering determined by the equivalent scalar channel after transmit beamforming [27]–[30]. In these cases, the entire multi-antenna NOMA system is degraded as a single-antenna one. With perfect CSI assumed at the multi-antenna base station (BS), the beamforming vector for NOMA is designed to maximize the sum rate and to minimize the total transmission power in [27] and [28], respectively. In [29], the optimal power allocation for ergodic capacity maximization is obtained under Rayleigh fading multiple-input multiple-output (MIMO) NOMA systems with statistical CSI. In [30], with imperfect CSI at the BS the worst-case achievable sum rate is optimized by designing robust beamforming in NOMA.
2) In some other schemes, the users in the multi-antenna system are first grouped into multiple clusters where inter-cluster users are served by *space division multiple access* (SDMA) to suppress inter-cluster interference, and then NOMA principles are further exploited by *intra-cluster users*. As a result, the original multi-antenna system is decomposed into multiple separate single-antenna NOMA ones [31]–[33]. The performance of MIMO NOMA systems is studied in [31] for randomly divided groups. To relax the demand on the antenna number at the receivers in [31], Ding *et al.* [32] propose a novel MIMO-NOMA framework for downlink and uplink transmission using the concept of signal alignment. By following the similar concept of [31], the performance of NOMA in the massive-MIMO system is investigated in [33].

In the multi-antenna NOMA system with multiple users, the acquisition of CSI is vital at the transmitter for beamforming, power allocation, and interference management. In most of current works such as [27] and [28] perfect CSI is assumed, which is impractical especially when the antenna number at the transmitter is large. In practical multi-user scenarios, collecting CSI at a BS usually relies on the *limited feedback* from its users [34]–[37], especially in frequency-division duplexed (FDD) systems where the channel reciprocity is not guaranteed. Under the limited feedback (also known as digital feedback) framework, each user and the BS maintain a common quantization codebook which is designed off-line and known to the both a priori. While collecting CSI, each user first quantizes their estimated downlink CSI by choosing the optimal channel vector from its codebook, and then the corresponding index of the vector with much less bit overhead is sent back to the BS using a low-rate feedback channel. The BS thereby designs beamforming and power allocation according to the feedback without incurring too much performance loss, when the number of feedback bits is sufficiently large [38]. In [33] and [39], the outage performance of NOMA with only one-bit feedback is investigated in each single-antenna group decomposed from a massive-MIMO system and in a single-antenna system, respectively. In [40], the sum rate of a multi-antenna NOMA system is studied under limited feedback. However, the considered one-bit feedback in [33] and [39] is only applicable to single-antenna NOMA systems and cannot be directly employed in multi-antenna systems. Furthermore, the outage performance remains unexplored in [40], and the work in [40] restricts the number of users in each beam to two. So far, a comprehensive performance analysis concerning outage probability and net throughput (defined as the difference of the downlink transmission and uplink feedback throughputs) under a general limited feedback framework is still missing. Additionally, random or fixed user grouping is usually assumed for the implementation of NOMA in multi-antenna NOMA systems [31]–[33], and how to fulfill dynamic user scheduling and clustering/grouping by leveraging limited feedback remains uncovered in the existing literature. These motivate our work.

In this paper, we consider a downlink single-cell cellular network with a multi-antenna BS and randomly deployed single-antenna users for delay-sensitive applications. The outage probability subject to constant rate traffic and net throughput are used as the relevant performance metrics. The NOMA technology is incorporated with SDMA to simultaneously serve multiple users under a general limited feedback framework. By leveraging the feedback framework, a dynamic user scheduling and grouping strategy is proposed. The users in different groups exploit different SDMA beams and those within a group are concurrently served by NOMA. In this way, both the spatial and power domains are fully exploited. Besides, we provide a comprehensive study on the performance of NOMA and the impacts of the number of feedback bits under the considered system. The novelties and main contributions of this paper can be summarized as follows:

1) We provide a comprehensive analysis on both the outage probability and net throughput under SDMA-NOMA systems with a general limited feedback framework, which has not been well studied before.
2) In the existing works, random or fixed user grouping is usually assumed for the implementation of NOMA in multi-antenna systems. In our paper, by leveraging the general limited feedback framework, a dynamic user scheduling and grouping strategy is proposed based on the feedback information available at the BS.
3) We derive a closed-form approximation of outage probability, and it is shown that the outage probability floor always exists due to limited feedback. Furthermore, we find that the diversity order is one for all the users when the number of feedback bits is sufficiently large.
4) We devise a low-complexity algorithm to find the optimal feedback rate for the maximization of net throughput. In addition, the impacts of the channel coherence time and group number on the optimum are analytically investigated.

The rest of this paper is organized as follows: In Section II, we present the system model and performance metrics of the considered SDMA-NOMA system. In Section III, we focus on characterizing the outage performance. In Section IV, we maximize the net throughput by finding the optimal number

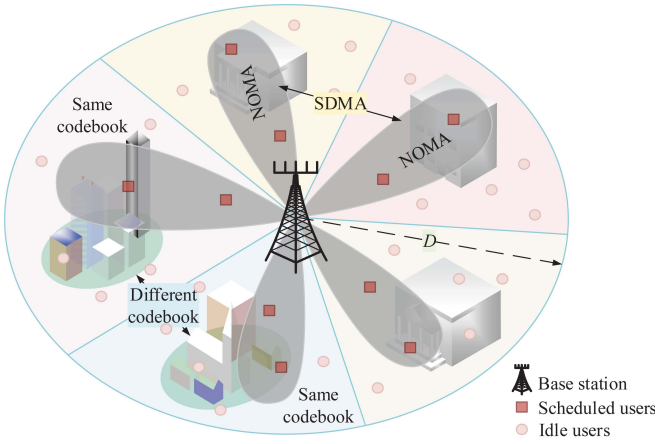


Fig. 1. The considered downlink cellular network with a multi-antenna BS and multiple single-antenna users. The NOMA technology is exploited to simultaneously serve multiple users under each spatial beam of SDMA.

of feedback bits. Numerical results are presented in Section V before the conclusions drawn in Section VI.

Notations: $\mathbf{A}^T$ and $\mathbf{A}^H$ represent the transpose and Hermitian transpose of a matrix $\mathbf{A}$, respectively. The factorial of a non-negative integer n is denoted by $n!$, and $\binom{n}{k} = \frac{n!}{(n-k)!k!}$. $X \sim \text{Exp}(\lambda)$ denotes the exponential distributed random variable with rate λ, $X \sim \text{Gamma}(M, \alpha)$ denotes the Gamma-distributed random variable with shape M and scale α, and $X \sim \text{Beta}(a, b)$ denotes the Beta-distributed random variable with parameters a and b. $\Gamma(x)$ is the Gamma function [41, eq. (8.310)], and $\gamma(\alpha, x)$ is the lower incomplete gamma function [41, eq. (8.350.1)]. $X \overset{d}{=} Y$ means that the two random variables X and Y are equal in distribution.

II. Models and Metrics

In this section, we will describe the models and performance metrics in the considered SDMA-NOMA system under a general limited feedback framework. In the conventional SDMA scheme [35], [36], only one user in each spatial beam is served and the multiple access is implemented over spatial domain while power domain is generally neglected. By following the line of [31]–[33], in this paper, the users are divided into multiple groups served by multiple spatial beams and power domain is further exploited by NOMA within each group. The details will be presented in the following subsections.

A. System Model

Consider a downlink single-cell cellular network with M transmit antennas equipped at the BS and $W \geq M$ single-antenna users. The BS aims to simultaneously serve multiple scheduled users through broadcast. We assume that the users are uniformly distributed within the round disc with radius D as in [10], [15], [21], and [25], and the BS locates at its center as shown in Fig. 1. Therefore, the distances between the BS and the users follow the independent and identically distributed (i.i.d.) distribution with the cumulative distribution function (CDF) given by

$$F_d(x) = \frac{x^2}{D^2}, \quad 0 \leq x \leq D. \tag{1}$$

In the considered cellular network, the cell is assumed to be divided into G disjoint sectors where the scheduled users in each sector form a group and are served by a common beam of SDMA as shown in Fig. 1. The detailed user scheduling and grouping strategy will be introduced in Section II-C. Over the conventional SDMA framework, the NOMA technology is concurrently exploited to simultaneously serve the users within each group for the further improvement of spectral efficiency.

B. Channel and Feedback Models

In FDD systems, before data transmission each user is assumed to have perfect CSI through downlink channel estimation, and then each user feeds the CSI back to the BS through an error-free but limited-rate feedback channel [35].

All the wireless channels are assumed to experience frequency flat Rayleigh fading together with a large-scale path loss. Denote the channel from the BS to the k-th user in the n-th sector ($1 \leq n \leq G$) as $\mathbf{h}_{n,k} = \mathbf{g}_{n,k} d_{n,k}^{-\alpha/2}$, where $\mathbf{g}_{n,k} \in \mathbb{C}^{M \times 1}$ is the small-scale fading vector with zero-mean unit-variance i.i.d. complex Gaussian entries, α is the path-loss factor, and $d_{n,k}$ denotes the distance from the BS to the user whose CDF is given in (1). In a general limited feedback framework, the quantization of channel direction information (CDI), i.e., $\tilde{\mathbf{g}}_{n,k} = \mathbf{g}_{n,k}/\|\mathbf{g}_{n,k}\|$ is critical for beamforming design of SDMA. Compared with small-scale fading, the user locations remain static and are easier to track. Therefore, the BS simply uses the distances (large-scale fading) to its users, i.e., $\{d_{n,k}\}$, as the channel quality information (CQI), based on which the order of SIC will be determined. Furthermore, we propose a low-complexity codebook distribution scheme. Specifically, the same codebook for CDI quantization is shared by the users within the same sector, while the users in different sectors use different and independent codebooks[1] as shown in Fig. 1. Once a user in a sector receives its codebook from the BS, the codebook can be shared within the sector via decentralized ways such as device-to-device (D2D) transmission and thus the overhead of the BS is reduced.

Assuming that the feedback rate is B bits per channel coherence time T_c which denotes the number of downlink symbols experiencing the same channel fading, then the size of the codebook consisting of M-dimensional unit-norm vectors is $N = 2^B$. Denoting the codebook in the n-th sector as

$$\mathcal{C}_n = \{\mathbf{c}_{n,1}, \mathbf{c}_{n,2}, \ldots, \mathbf{c}_{n,N}\}, \quad 1 \leq n \leq G, \tag{2}$$

the channel quantization for the k-th user in the n-th sector ($1 \leq n \leq G$) is to find the nearest vector in terms of maximum inner product [35], [36] from codebook $\mathcal{C}_n$

[1]The assumption that the users in different sectors employ different and independent codebooks originates from the fact that independent codebooks are usually used by different users in conventional SDMA systems as in [35] and [36].

satisfying

$$j_{n,k}^{\star} = \arg \max_{1 \le j \le N} |\tilde{\mathbf{g}}_{n,k}^{H} \mathbf{c}_{n,j}|. \tag{3}$$

After channel quantization, each user feeds the selected index $j_{n,k}^{\star}$ back to the BS.

In the subsequent performance analysis, we will leverage the well-known quantization cell approximation (QCA) and the codebook design based on random vector quantization (RVQ) as in [35] and [36]. The main idea of the QCA is that each quantization cell can be approximated by a Voronoi region of a spherical cap with the area 2^{-B} in a unit sphere [36]. In fact, we can decompose the actual channel direction of the k-th user in the n-th sector ($1 \le n \le G$) as

$$\tilde{\mathbf{g}}_{n,k} = \cos\theta_{n,k} \cdot \mathbf{c}_{n,j_{n,k}^{\star}} + \sin\theta_{n,k} \cdot \mathbf{e}_{n,k}, \tag{4}$$

where $\theta_{n,k}$ denotes the angle between the actual CDI $\tilde{\mathbf{g}}_{n,k}$ and the quantized CDI $\mathbf{c}_{n,j_{n,k}^{\star}}$, $\mathbf{e}_{n,k}$ denotes the unit-norm error vector isotropically distributed in the nullspace of $\mathbf{c}_{n,j_{n,k}^{\star}}$, and the CDF of $\sin^2\theta_{n,k}$ resulted from QCA is given by [35], [36]

$$F_{\sin^2\theta_{n,k}}(x) = \begin{cases} 2^B x^{M-1}, & 0 \le x \le \delta, \\ 1, & x \ge \delta, \end{cases} \tag{5}$$

with $\delta = 2^{-\frac{B}{M-1}}$. It has been pointed out in [36] that the QCA actually provides a performance upper bound, and the performance gap between the QCA and RVQ, which gives a performance lower bound, is small. Therefore, the results derived in this work by leveraging the QCA will capture the accurate performance for any well-designed codebook.

C. Scheduling and Grouping Model

To simultaneously serve multiple users in different sectors using SDMA, the BS schedules a subset of the users in each sector which have the similar channel direction as a group. In this way, the scheduled users in each group exploit a common beam provided by SDMA, while the intra-group users are concurrently served via NOMA during data transmission.

We remark that the design of scheduling and grouping faces the following two problems. On the one hand, we have to select the users in each sector which have the similar channel direction. On the other hand, the user grouping should be dynamically changed according to the fluctuation of the channel. In the considered network, the proposed general limited feedback framework enables us to solve the both problems concurrently. The basic idea to realize this is to fulfill user scheduling and grouping by leveraging the feedback information $\{j_{n,k}^{\star}\}$ at the BS. For analytical simplicity, in the n-th sector ($1 \le n \le G$) we propose to randomly schedule K users, who feed back *the same quantization index*, to form the n-th NOMA group. The same quantization index is chosen randomly to ensure the fairness among the users with various channel directions in each sector, and the index in the n-th sector ($1 \le n \le G$) is denoted by $j_n^{\star}$. Under this random scheduling scheme,[2] the distances from the BS to the scheduled users still follow the i.i.d. distribution with the CDF given in (1).

Since the BS has the full knowledge of the codebooks used in all the sectors, the common beamformer shared by the K scheduled users in the n-th group ($1 \le n \le G$) can be designed by viewing these users' channel directions as the same $\hat{\mathbf{g}}_n = \mathbf{c}_{n,j_n^{\star}}$. According to (4), the actual channel direction of the k-th scheduled user in the n-th group ($1 \le k \le K$, $1 \le n \le G$) is now decomposed as

$$\tilde{\mathbf{g}}_{n,k} = \cos\theta_{n,k} \cdot \hat{\mathbf{g}}_n + \sin\theta_{n,k} \cdot \mathbf{e}_{n,k}. \tag{6}$$

D. Data Transmission Model

The main idea of our transmission scheme is to exploit NOMA to simultaneously serve the K scheduled users in each group over the conventional SDMA framework.

In the NOMA scheme, SIC is exploited to suppress the intra-group interference, which allows the users with strong channel condition to be served by allocating little power for ensuring fairness. The decoding order of SIC for single-antenna systems is well-known, but the optimal one for multi-antenna systems is difficult to obtain [27]. Recalling that the distances (large-scale fading) of the users $\{d_{n,k}\}$ as CQI is assumed to be available at the BS in Section II-B, here we sort the users within a group for performing SIC according to $\{d_{n,k}\}$.[3] Without loss of generality, the distances between the users in the n-th group and the BS are sorted as $d_{n,1} \ge d_{n,2} \ge \ldots \ge d_{n,K}$ for $1 \le n \le G$. Based on the NOMA and SDMA protocol, the BS simultaneously serves all the scheduled users in G groups and its transmit signal is characterized by

$$\begin{aligned} \mathbf{x} &= \sqrt{P}\mathbf{W}\bar{\mathbf{s}} \\ &= \sqrt{P}\sum_{n=1}^{G} \mathbf{w}_n \bar{s}_n \\ &= \sqrt{P}\sum_{n=1}^{G} \left(\mathbf{w}_n \left(\sum_{k=1}^{K} \sqrt{\beta_{n,k}} s_{n,k} \right) \right), \end{aligned} \tag{7}$$

where P is the total transmit power; $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_G] \in \mathbb{C}^{M \times G}$ and $\bar{\mathbf{s}} = [\bar{s}_1, \bar{s}_2, \ldots, \bar{s}_G]^T \in \mathbb{C}^{G \times 1}$ are the collections of the unit-norm beamforming vectors and the information bearing signals for all the G groups, respectively; $\beta_{n,k}$ and $s_{n,k}$ are the power allocation factor and unit-power information bearing signal for the k-th user in the n-th group, respectively. According to the principle of NOMA, we usually have $\beta_{n,1} \ge \beta_{n,2} \ge \ldots \ge \beta_{n,K}$ to ensure fairness among users within a group. Moreover, we assume that the transmit power is equally

[2]The considered random scheduling scheme here also takes care of the fairness among the users with different large-scale path losses in each sector. Additionally, the proposed analytical framework in this paper can be easily extended to some other greedy scheduling schemes by only changing the distance distribution, as will be pointed out in Remark 2 of Section IV.

[3]Note that determining the decoding order of SIC according to the distances from the BS to the users has also been adopted in [25], and the motivation is justified therein.

allocated among different groups,[4] i.e., $\sum_{k=1}^{K}\beta_{n,k}=1/G$ for $1\le n\le G$.

The received signal at the k-th user in the n-th group ($1\le k\le K,\ 1\le n\le G$) is given by

$$y_{n,k}=\sqrt{P\beta_{n,k}}\mathbf{h}_{n,k}^{H}\mathbf{w}_n s_{n,k}+\underbrace{\mathbf{h}_{n,k}^{H}\mathbf{w}_n\sum_{j=1,j\neq k}^{K}\sqrt{P\beta_{n,j}}s_{n,j}}_{\text{intra-group interference}}+\underbrace{\sqrt{P}\sum_{m=1,m\neq n}^{G}\left(\mathbf{h}_{n,k}^{H}\mathbf{w}_m\bar{s}_m\right)}_{\text{inter-group interference}}+n_{n,k},\qquad(8)$$

where $n_{n,k}$ is zero-mean additive white Gaussian noise (AWGN) with variance σ^2. To suppress the intra-group interference in (8), SIC is employed at each user. To be specific, in the n-th group the k-th user first detects the i-th user's message ($i<k$) and then removes the decoded message from its observation in a successive manner for $i=1,2,\ldots,k-1$. Then the k-th user detects its own message by regarding the i-th user's message ($i>k$) as noise. Based on the above SIC procedure, the signal-to-interference-plus-noise ratio (SINR) for detecting the i-th user's message ($1\le i\le k$) by the k-th user is denoted by

$$\mathrm{SINR}_{n,k}^{i}=\frac{|\mathbf{h}_{n,k}^{H}\mathbf{w}_n|^2\beta_{n,i}}{|\mathbf{h}_{n,k}^{H}\mathbf{w}_n|^2\sum_{j=i+1}^{K}\beta_{n,j}+\Lambda\sum_{m=1,m\neq n}^{G}|\mathbf{h}_{n,k}^{H}\mathbf{w}_m|^2+1/\rho},\qquad(9)$$

where $\Lambda=1/G$ and $\rho=P/\sigma^2$ denotes the transmit signal-to-noise ratio (SNR).

For the specific design of precoding matrix $\mathbf{W}$ in SDMA, we assume that zero-forcing (ZF) beamforming is adopted to suppress the inter-group interference in (8) for analytical tractability. To fully exploit the spatial DoFs provided by multiple antennas equipped at the BS, we only focus on the case where $G=M$ groups are simultaneously served by SDMA. The BS uses the quantized CSI $\{\hat{\mathbf{g}}_m\}_{m=1}^{G}$ to design the beamformer for each group. Accordingly, the beamformer $\mathbf{w}_n$ for group n ($1\le n\le G$) is designed to satisfy

$$\hat{\mathbf{g}}_m^{H}\mathbf{w}_n=0,\quad\forall m\neq n,\ 1\le m\le G.\qquad(10)$$

E. Performance Metrics

We will use the following two main metrics in the performance analysis of the considered SDMA-NOMA system:

1) Outage Probability: In the subsequent analysis, we focus on the outage performance of the users in one group and omit the group subscript n to lighten the notations. According to (9), the corresponding rate for the k-th user to decode the i-th user's message is denoted as $C_k^i=\log_2(1+\mathrm{SINR}_k^i)$ for $1\le k\le K$ and $1\le i\le k$. Note that the rates for the users to decode their own messages are given by $\{C_k^k\}_{k=1}^{K}$. However, we remark that different from the case in [10] the multi-user rates $\{C_k^k\}_{k=1}^{K}$ here may not be simultaneously achievable, since failure can occur during SIC due to the partial CSI and the considered multi-antenna system here.[5] Therefore, the ergodic sum rate adopted in [10] will not capture the actual performance of the considered system in this paper. We resort to the outage probability in delay-sensitive applications as a more suitable performance metric in this paper. By denoting the target rates, namely quality of service (QoS), for the K users in a group as $\{\tilde{R}_k\}_{k=1}^{K}$, an outage event occurs at the k-th user ($1\le k\le K$) when either one of the following two constraints are not satisfied:

i) the SIC is successful, i.e., $C_k^i\ge\tilde{R}_i$ for all $i<k$, which is actually a worst-case SIC with error propagation taken into account [42];
ii) the user can successfully decode the message targeted for itself, i.e., $C_k^k\ge\tilde{R}_k$.

Accordingly, the outage probability of the k-th user in a group is expressed as

$$\mathrm{P}_k^{\mathrm{out}}=1-\Pr\left\{\bigcap_{i=1}^{k}C_k^i\ge\tilde{R}_i\right\}.\qquad(11)$$

When no outage occurs at the users in all the G groups, the sum rate of the system is simply $G\sum_{j=1}^{K}\tilde{R}_j$.

2) Net Throughput: To ensure fairness for inter-group users, we assume that the power allocation and transmit rates in NOMA are the same among all the G groups. As in [43], the net throughput in this paper is defined as the throughput difference of the total scheduled users' downlink information transmission and uplink CSI feedback. When each user's feedback rate is B bits per channel coherence time T_c, the net throughput is given by

$$\Gamma_{\mathrm{net}}(B)=G\left(\sum_{j=1}^{K}\left(1-\mathrm{P}_j^{\mathrm{out}}(B)\right)\tilde{R}_j-B\cdot K/T_c\right),\qquad(12)$$

where $\mathrm{P}_j^{\mathrm{out}}(B)$ denotes the outage probability of the j-th user. The intuition behind this metric is that the downlink data transmission comes from the cost of uplink CSI feedback. Therefore, the net throughput envisions this fundamental trade-off and serves as an indicator to the overall network throughput gain. The detailed analysis and the optimal feedback rate for the maximization of net throughput will be introduced in Section IV.

III. Outage Performance

In this section, we analyze the outage performance, which is an important metric in delay-sensitive communications where the transmitter transmits the message at a fixed data rate.

[4]This assumption is made to lighten the notations and ensure the fairness among different groups by recalling that each group has the same number of users who are randomly scheduled from each sector. Note that the performance analysis framework developed in our paper can be easily extended to the scenarios with some other power allocation policies.

[5]In the single-antenna NOMA system with perfect CSI, the success of SIC is guaranteed based on the fact that once the channel is ordered, the user with the strong channel condition can always decode the messages for the users with weaker channel conditions [10]. However, this fact does not hold for the cases of either partial CSI [25] or multi-antenna [27].

According to (11), the outage probability of the k-th user in a group is calculated as

$$\mathrm{P}_k^{\text{out}} = 1 - \Pr\left\{\bigcap_{i=1}^{k} \mathrm{SINR}_{n,k}^{i} \geq \phi_i\right\}, \quad 1 \leq k \leq K, \tag{13}$$

where $\phi_i = 2^{\tilde{R}_i} - 1$. Since the obtained CSI is imperfect under limited feedback, the inter-group interference is still present. By leveraging (6), (9), and (10), (13) can be recast as

$$\begin{aligned}\mathrm{P}_k^{\text{out}} &= 1 - \Pr\left\{\bigcap_{i=1}^{k} \frac{\beta_i X}{\left(\sum_{j=i+1}^{K}\beta_j\right) X + \Lambda Y + Z/\rho} \geq \phi_i\right\} \\ &\overset{(a)}{=} 1 - \Pr\left\{\bigcap_{i=1}^{k} \frac{X}{\Lambda Y + Z/\rho} \geq c_i\right\} \\ &= \Pr\left\{\frac{X}{\Lambda Y + Z/\rho} < \eta_k\right\}, \end{aligned} \tag{14}$$

where $Y = \|\mathbf{g}_{n,k}\|^2 \sin^2\theta_{n,k} \sum_{m=1,m\neq n}^{G} |\mathbf{e}_{n,k}^H \mathbf{w}_m|^2$, $X = |\mathbf{g}_{n,k}^H \mathbf{w}_n|^2$, $Z = d_{n,k}^{\alpha}$, $c_i = \frac{\phi_i}{\beta_i - \phi_i \sum_{j=i+1}^{K}\beta_j}$, and $\eta_k = \max_{1\leq i\leq k} c_i$. Note that the equality (a) in (14) is obtained by assuming $\beta_i > \phi_i \sum_{j=i+1}^{K}\beta_j$ for $\forall 1 \leq i \leq k$, otherwise the k-th user's outage probability is always one [10].

The computation of the probability in (14) is challenging due to the following two reasons. On the one hand, the norm of the channel $\|\mathbf{g}_{n,k}\|^2$ appears in both the signal term X and the inter-group interference term Y, making the numerator and denominator in (14) coupled with the random variable and hard to deal with. On the other hand, the term $\sum_{m=1,m\neq n}^{G} |\mathbf{e}_{n,k}^H \mathbf{w}_m|^2$ in Y involves the sum of $G-1$ random variables, of which the CDF is difficult to figure out. Here we adopt the similar method in [43] and [44] by imposing the independence on the signal and inter-group interference terms, i.e., X and Y, to obtain an approximate result, which is shown to be reasonable in [44]. The following theorem gives an approximate closed-form expression of the outage probability in (14).

Theorem 1: The outage probability of the k-th user ($1 \leq k \leq K$) in a group in (14) can be approximated as

$$\begin{aligned}\mathrm{P}_k^{\text{out}} = 1 - 2k\tau_1 \binom{K}{k} \sum_{i=0}^{k-1} \binom{k-1}{i} \frac{(-1)^i}{\alpha D^{2(K+i-k+1)}} \\ \times \left(\frac{\rho}{\eta_k}\right)^{\frac{2}{\alpha}(K+i-k+1)} \gamma\left(\frac{2}{\alpha}(K+i-k+1), \frac{\eta_k}{\rho} D^{\alpha}\right),\end{aligned} \tag{15}$$

where $\tau_1 = a_1 \sum_{n=0}^{\infty} \frac{b_n}{(1+\eta_k \Lambda a_2)^{G+n-1}}$, $a_1 = \frac{\sqrt{M-1}-1}{G-2+\sqrt{M-1}}$, $a_2 = \delta\left(1 - \frac{1}{\sqrt{M-1}}\right)$, $b_n = \frac{1}{n}\sum_{i=1}^{n}(1-a_1)^i b_{n-i}$ for $n = 1, 2, \ldots$ with $b_0 = 1$, and $G = M$ from the full-loaded ZF-SDMA assumption.

Proof: The proof is given in Appendix A. ∎

The accuracy of approximation (15) will be checked by the simulations in Section V. The outage probability given in (15) is in a closed form, but it is hard to find some insights for the performance of the system due to its sophisticated expression. We will consider the two limiting situations, namely the *high SNR regime* where the transmit SNR ρ is large and the *high resolution regime* where the number of feedback bits B is large in the next two subsections.

A. High SNR Regime

In the high SNR regime, the diversity order of the users in the system is of much importance, which reflects the limiting performance of a system about how outage reduces as SNR increases. Before deriving the diversity order, we first find an asymptotic expression of (15) under high SNR.

By rewriting the lower incomplete gamma function in the series form [41, eq. (8.354.1)], (15) can be equivalently written as

$$\begin{aligned}\mathrm{P}_k^{\text{out}} = 1 - k\tau_1 \binom{K}{k} \sum_{n=0}^{\infty} \sum_{i=0}^{k-1} \binom{k-1}{i} \\ \times \frac{(-1)^{i+n} D^{n\alpha} \left(\frac{\eta_k}{\rho}\right)^n}{n!\,(K+i-k+1+n\alpha/2)}.\end{aligned} \tag{16}$$

By neglecting the $O\left(\frac{1}{\rho}\right)$ terms in (16), we obtain an approximation of the outage probability for the k-th user in the high SNR regime as

$$\begin{aligned}&\mathrm{P}_{k,\text{out}}^{\rho\to\infty} \\ &= 1 - k\tau_1 \binom{K}{k} \sum_{i=0}^{k-1} \binom{k-1}{i} \frac{(-1)^i}{K+i-k+1} \\ &\quad + k\tau_1 \binom{K}{k} \sum_{i=0}^{k-1} \binom{k-1}{i} \frac{(-1)^i D^{\alpha} \eta_k}{(K+i-k+1+\alpha/2)\,\rho} \\ &= 1 - \tau_1 + \tau_1 \tau_2 D^{\alpha} \eta_k \frac{1}{\rho}, \end{aligned} \tag{17}$$

where the last equality follows from the fact that the CDF $F_{d_{n,k}}(D) = k\binom{K}{k}\sum_{i=0}^{k-1}\binom{k-1}{i}\frac{(-1)^i}{K+i-k+1} = 1$ in (27), and $\tau_2 = k\binom{K}{k}\sum_{i=0}^{k-1}\binom{k-1}{i}\frac{(-1)^i}{K+i-k+1+\alpha/2}$.

From (17), we know that under this situation there exists an outage probability floor. The outage probability remains as a constant, i.e., $1-\tau_1$ when the transmit power approaches infinity. In this case, there is no diversity. From Theorem 1, we see that τ_1 increases with an increase in B, meaning that the outage floor is reduced when we increase the number of feedback bits. Additionally, η_k is an increasing function of the QoS constraint $\tilde{R}_i$ and thereby can be viewed as an indicator for the QoS. It can be easily seen from (17) that the outage floor increases when the QoS constraint becomes more stringent.

B. High Resolution Regime

As the number of feedback bits goes to infinity, we can find that $\tau_1 \to 1$ since $B \to \infty$ leads to $\delta \to 0$, $a_2 \to 0$, and finally $\tau_1 \to a_1 \sum_{n=0}^{\infty} b_n = 1$, where the last equality follows from the fact that the integral of PDF (30) in Appendix A is equal to one. By substituting $\tau_1 = 1$ into (15), the outage

probability for the k-th user in the high resolution regime becomes

$$\mathrm{P}_{k,\mathrm{out}}^{B\to\infty} = 1 - 2k\binom{K}{k}\sum_{i=0}^{k-1}\binom{k-1}{i}\frac{(-1)^i}{\alpha D^{2(K+i-k+1)}} \times \left(\frac{\rho}{\eta_k}\right)^{\frac{2}{\alpha}(K+i-k+1)} \gamma\left(\frac{2}{\alpha}(K+i-k+1), \frac{\eta_k}{\rho}D^{\alpha}\right). \tag{18}$$

When we investigate the diversity order and let the SNR approach infinity now, by following the similar method in Section III-A, the outage probability in (18) reduces to

$$\mathrm{P}_{k,\mathrm{out}}^{\rho,B\to\infty} = \tau_2 D^{\alpha}\eta_k\frac{1}{\rho}. \tag{19}$$

From (19), the outage probability floor vanishes now and all the users achieve the same diversity order equal to one. Furthermore, it is straightforward to see that the outage probability increases if either the QoS constraint at the user becomes more stringent or the cellular size becomes larger.

It is interesting to see that even though a significantly more numbers of users in multiple groups are concurrently served in our scheme, the diversity gain here still maintains the same as that achieved by the single-antenna NOMA system under the second order statistical CSI in [25] where only one user group is considered. The reasons behind can be summarized as follows:

1) Since the BS perfectly knows the direction of the channel in the high resolution regime, the inter-group interference in (8) is completely eliminated using ZF precoding. Therefore, the considered SDMA-NOMA system now degenerates to multiple pure NOMA systems separated by different groups.
2) We have assumed the full-loaded ZF-SDMA case where $G = M$ in this paper, meaning that the available DoF at each user is only one and the system under each SDMA beam is actually equivalent to a single-antenna one.
3) As in [25] under the second order statistical CSI case, the decoding order of SIC in NOMA is determined by the distances from the BS to the users in this paper. Both the diversity orders here and in [25] for all the users served by NOMA turn out to be one. This is because the amplitude of small scale fading is absent at the BS during the user ordering in SIC, while it can severely deteriorate the outage performance of NOMA.

Remark 1: The SDMA-NOMA system investigated here in the high resolution regime is superior to the full-loaded ZF-SDMA system where only M users are concurrently scheduled in terms of spectral efficiency. The reason behind is that far more users (larger than M) can be concurrently served in our SDMA-NOMA framework while the diversity orders remain the same.

IV. Net Throughput Maximization

In this section, we evaluate the net throughput of the considered network, which serves as a direct indicator to the overall network throughput gain. Based on the derived expression, we obtain the optimal feedback rate that maximizes the net throughput by proposing a low-complexity algorithm.

By substituting (15) into (12), we obtain the net throughput given by

$$\Gamma_{\mathrm{net}}(B) = G\left(\sum_{k=1}^{K}\tau_1(k,B)\cdot f_k\tilde{R}_k - B\cdot K/T_c\right), \tag{20}$$

where the term

$$f_k \triangleq 2k\binom{K}{k}\sum_{i=0}^{k-1}\binom{k-1}{i}\frac{(-1)^i}{\alpha D^{2(K+i-k+1)}} \times \left(\frac{\rho}{\eta_k}\right)^{\frac{2}{\alpha}(K+i-k+1)} \gamma\left(\frac{2}{\alpha}(K+i-k+1), \frac{\eta_k}{\rho}D^{\alpha}\right)$$

is irrelevant to the number of feedback bits B, and τ_1 is a function of both k and B. From (20), it is straightforward to see that an exceedingly large of B incurs huge cost of uplink CSI feedback while an insufficient of B also compromises the outage performance of downlink data transmission. Both cases limit the performance of the net throughput. Therefore, there exists an optimal number of feedback bits $B^{\star}$ that maximizes the net throughput. The net throughput maximization problem is then expressed as

$$\underset{B\in\mathbb{N}\cup 0}{\text{maximize}} \quad \Gamma_{\mathrm{net}}(B). \tag{21}$$

Since problem (21) is a univariate integer optimization problem and the number of feedback bits B cannot be very large in practice, problem (21) can be easily solved optimally via the one-dimensional search method. However, the exhaustive search method cannot provide any insight on the guidance of the system design. Thus, in the sequel of this section we aim to solve problem (21) analytically and devise a relatively low-complexity algorithm for net throughput maximization.

Problem (21) is generally nonconvex and solving it analytically is very challenging, one of the main obstacles lies in the fact that the exact τ_1 derived in Theorem 1 is a cumbersome function of B and it is also a function of k. To facilitate the following analysis, here we first approximate the exact $\tau_1(k,B)$ using the similar method in [43] and [44] via assuming that all inter-group interference terms in Y in (14) are independent of one another. This assumption is shown to be reasonable in [44], and we will verify the accuracy of this approximation by the simulations in Section V. Accordingly, the inter-group interference Y in (14) now follows the Gamma distribution with shape $G-1$ and scale δ. From (29) in Appendix A, we obtain the approximated $\tilde{\tau}_1(k,B)$ given by

$$\tilde{\tau}_1(k,B) = \int_0^{\infty}\frac{y^{G-2}e^{-y/\delta}}{\Gamma(G-1)\delta^{G-1}}e^{-\eta_k\Lambda y}dy = \frac{1}{(1+\eta_k\Lambda\delta)^{G-1}}, \tag{22}$$

where the last equality follows from [41, eq. (3.326.2)]. Furthermore, another obstacle in solving problem (21) is that with $\tau_1(k,B)$ in (20) replaced by (22) the objective function in problem (21) includes a summation of K high-order fraction terms related to B. This summation renders

the subsequent optimization intractable. To obtain a low-complexity result which gives insight on the system design but is still representative, we only consider a simplified scenario where each user imposes the same SINR constraint $\bar{\phi}$ and the power allocation factors in each group satisfy $\beta_{k+1} \leq \beta_k \leq (1+\bar{\phi})\beta_{k+1}$ for $1 \leq k \leq K-1$ in this section. The general case is left for future work. Under this scenario, we have the same $\eta_k = \bar{\eta} = c_1$ for all $1 \leq k \leq K$ since $c_1 \geq c_2 \geq \cdots \geq c_K$ is satisfied at this time. By substituting (22) into (20), the approximated net throughput now becomes

$$\tilde{\Gamma}_{\text{net}}(B) = G\left(\frac{1}{(1+\bar{\eta}\Lambda\delta)^{G-1}}Q - B\cdot K/T_c\right), \quad (23)$$

where $Q \triangleq \sum_{k=1}^{K} f_k \bar{R}$ with $\bar{R} \triangleq \log_2(1+\bar{\phi})$. The optimal number of feedback bits that maximizes the net throughput in (23) is derived in the following theorem.

Theorem 2: The optimal number of feedback bits for the maximization of net throughput in (23) is given by the integer that is closest to

$$B^\star = \begin{cases} 0, & \frac{K}{QT_c\ln 2} \geq \frac{(G-1)^{G-1}}{G^G}, \\ \arg\max\limits_{B\in\{0,B_1\}\cap[0,\infty)} \tilde{\Gamma}_{\text{net}}(B), & \frac{K}{QT_c\ln 2} < \frac{(G-1)^{G-1}}{G^G}. \end{cases} \quad (24)$$

where $B_1 = -(M-1)\log_2\delta_1$ and δ_1 is obtained by solving the equality

$$g(\delta) \triangleq \frac{\delta}{(1+\bar{\eta}\Lambda\delta)^G} = \frac{KG}{Q\bar{\eta}T_c\ln 2} \quad (25)$$

in the interval $\delta \in (0, \delta_0)$ using the bisection method with $\delta_0 \triangleq \frac{1}{(G-1)\bar{\eta}\Lambda}$.

Proof: The proof is given in Appendix B. ∎

Note that the optimal number of feedback bits $B^\star$ obtained from Theorem 2 can be zero, which can be resulted from a small channel coherence time T_c. In fact, if the cost or namely penalty factor of uplink CSI feedback K/T_c is large, the optimal $B^\star$ cannot be large and the net throughput actually becomes a decreasing function of B. On the contrary, when the channel coherence time T_c increases, we know intuitively that the optimal feedback rate $B^\star$ increases. The reason behind this is that under this situation more accurate CSI should be acquired at the BS since it will be used for a longer channel coherence time, and at the same time $B^\star$ can be large since the penalty factor K/T_c for the downlink throughput in a group is small. The following proposition analytically proves this intuition and also unveils the impact of T_c on the optimal net throughput $\tilde{\Gamma}_{\text{net}}(B^\star)$, which relies on the elegant result given in Theorem 2.

Proposition 1: The optimal number of feedback bits $B^\star$ and the corresponding optimal net throughput $\tilde{\Gamma}_{\text{net}}(B^\star)$ are both increasing functions of the channel coherence time T_c.

Proof: The proof is given in Appendix C. ∎

Furthermore, under the scenario of $B^\star \neq 0$, we note that the growth rate of the optimal number of feedback bits $B^\star$ due to the increase of the channel coherence time T_c becomes higher, when more antennas (and thereby more groups since $G = M$) are employed at the BS. Intuitively, this higher growth stems from the needs to concurrently quantize the higher-dimensional channel space and cope with the increased inter-group interference. By leveraging Theorem 2, this result can be analytically proved in the following proposition.

Proposition 2: When $B^\star \neq 0$, the growth rate of the optimal number of feedback bits $B^\star$ due to the increase of the channel coherence time T_c is an increasing function of the group and antenna number $G = M$.

Proof: The proof is given in Appendix D. ∎

Remark 2: The proposed analytical framework including outage performance and net throughput maximization can be easily extended to some other scheduling schemes with large-scale path loss taken into account. The only difference lies in the distribution of ordered distances from the BS to the scheduled users shown in (27), and its effect on net throughput is fully captured in parameter Q.

V. Simulation Results

Numerical results are presented in this section to verify the derived analytical expressions in the considered system. In addition, a comparison between the proposed NOMA scheme and the conventional OMA scheme is given. To be specific, we consider a time division multiple access (TDMA) counterpart implemented under each spatial beam, where the BS serves one user in each group per time slot. To make a fair comparison, in each simulation trial the same K users in each sector are scheduled as in the NOMA scheme. Note that since only one user is served in each group per time slot for OMA, all the power in each group is allocated to the user. The outage probability of the k-th user in each group under the OMA scheme is then given by $\mathrm{P}_k^\circ = \Pr\{C_k^\circ < \tilde{R}_k\}$ for $1 \leq k \leq K$, where

$$C_k^\circ = \frac{1}{K}\log_2\left(1 + \frac{\Lambda|\mathbf{h}_{n,k}^H\mathbf{w}_n|^2}{\Lambda\sum_{m=1,m\neq n}^{G}|\mathbf{h}_{n,k}^H\mathbf{w}_m|^2 + 1/\rho}\right). \quad (26)$$

In each trial of the Monte Carlo simulation, the channels and RVQ-based codebooks are randomly generated according to the models described in Section II-B. The simulation settings are as follows, unless otherwise specified: The cell is modeled by a disc with the normalized disc radius $D = 10$ and the pass-loss exponent $\alpha = 2$. The sector (group) number in the cell and the antenna number at the BS are $G = M = 3$. The number of users in each group is $K = 2$. The transmit SNR is $\rho = 30$ dB, and the target rates for the users in each group are identical $\bar{R} = 0.5$ bits/s/Hz. The power allocation factors in each group are $\beta_i = \frac{1}{\mu}(1+\bar{\phi})^{K-i}$ for $1 \leq i \leq K$, where $\bar{\phi} = 2^{\bar{R}} - 1$ and μ is set to ensure $\sum_{i=1}^{K}\beta_i = \Lambda$. The normalized channel coherence time is $T_c = 200$. All the simulation results to be shown are averaged over 10^4 trials.

In Fig. 2, the outage performance of NOMA and OMA is compared. One can see that for both the two schemes there always exists a performance floor due to the limited feedback. It is shown that under the given power allocation, whether NOMA is superior to conventional OMA for a specific user in a group depends on the distance from the BS to this user. For the user with the farther distance from the BS (User 1), the outage probability achieved by NOMA is lower than that

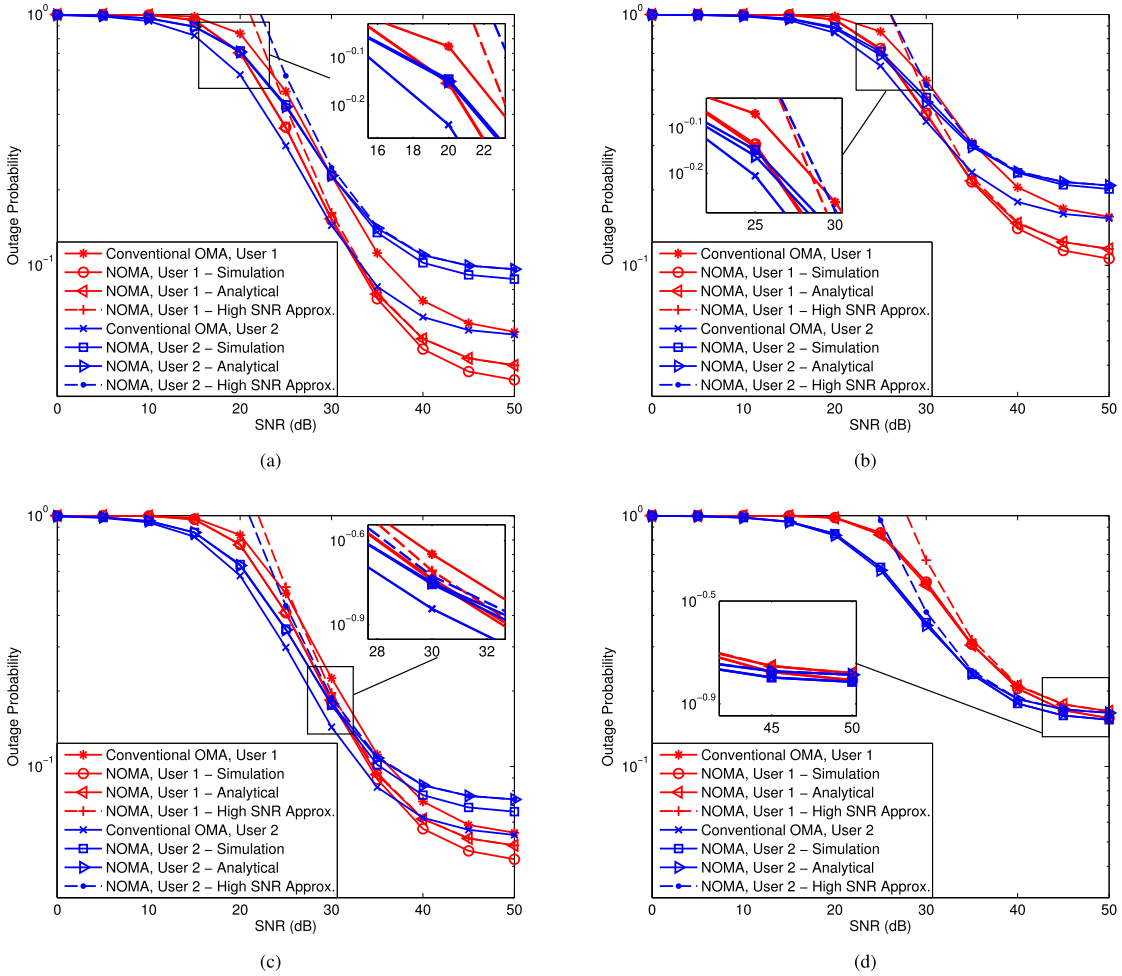


Fig. 2. The outage probability versus transmit SNR ρ under different multiple access technologies. The number of feedback bits is $B = 10$ bits. The target rates are (a)(c) $\tilde{R}_1 = \tilde{R}_2 = 0.5$ bits/s/Hz and (b)(d) $\tilde{R}_1 = \tilde{R}_2 = 1$ bits/s/Hz, and the power allocation factors in each group satisfy (a)(b) $\beta_1/\beta_2 = 3$ and (c)(d) $\beta_1/\beta_2 = 2$.

achieved by conventional OMA, whereas the situation is opposite for the user with the nearer distance (User 2). The reason behind this is two-fold. On the one hand, User 1 benefits from a more fair power allocation policy in NOMA which allows the user with weak channel condition to be allocated with more power. On the other hand, User 2 with less allocated power actually has more stringent constraints compared to User 1 since the additional SIC constraint in NOMA has to be satisfied, which makes NOMA inferior to OMA under this scenario. Furthermore, it is also worth noting that different from the conventional NOMA case with perfect CSI as in [10], here in Fig. 2(a) the user with a better channel condition (User 2) has a better outage performance than that achieved by the other user (User 1) only in the low SNR regime (from 0 to 20 dB). This is because performance floors exist due to limited feedback, and the outage floor of User 2 is higher than that of User 1 since User 2 has less allocated power and the additional SIC constraint. As pointed out in [10], the outage performance of NOMA largely depends on the choices of the users' targeted data rates and allocated power. Indeed, as observed from Fig. 2(c)(d), the outage performance of User 2 gets improved while that of User 1 becomes degraded when less power is allocated to User 1 under the power allocation $\beta_1/\beta_2 = 2$. Additionally, from Fig. 2 we know that the crossing point of the two users' outage probability curves, which reflects user fairness in NOMA, can be adjusted by changing different targeted rates and power allocation schemes. Moreover, it can be observed from Fig. 2(b)(d) that the outage performance of NOMA deteriorates when a stricter QoS is demanded as expected.

Fig. 3 shows the impacts of the number of feedback bits on the outage performance of NOMA. We can see that the outage probability largely depends on the number of feedback bits. With a larger B the outage probability floor approaches vanishing as predicted in Section III-A. Under a large number of feedback bits ($B = 16$ bits), it can be observed from Fig. 3 that the two users approximately have the same diversity order equal to one as expected. Moreover, according to Fig. 3 the derived analytical expression is shown to coincide well with the simulation results especially in the low-to-moderate SNR range (from 0 to 35 dB).

Fig. 4 depicts the effects of the number of feedback bits on the net throughput of NOMA under limited feedback. It can be clearly seen that there exists an optimal number of feedback bits for the maximization of net throughput. The analytical result in (20) and its approximation in (23) are found

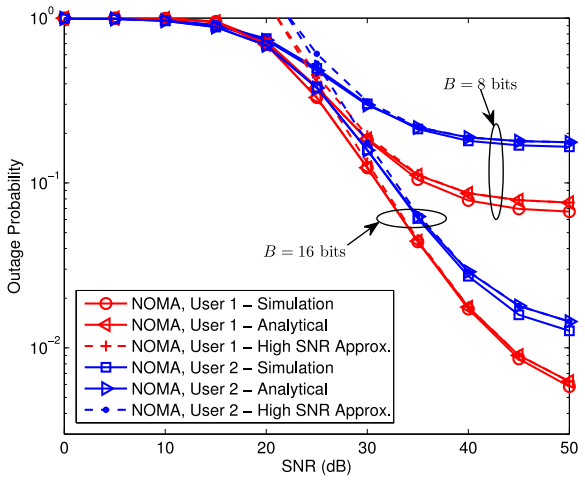


Fig. 3. The outage probability versus transmit SNR ρ under different numbers of feedback bits. The target rates are $\tilde{R}_1 = \tilde{R}_2 = 0.5$ bits/s/Hz, and the power allocation factors in each group satisfy $\beta_1/\beta_2 = 3$.

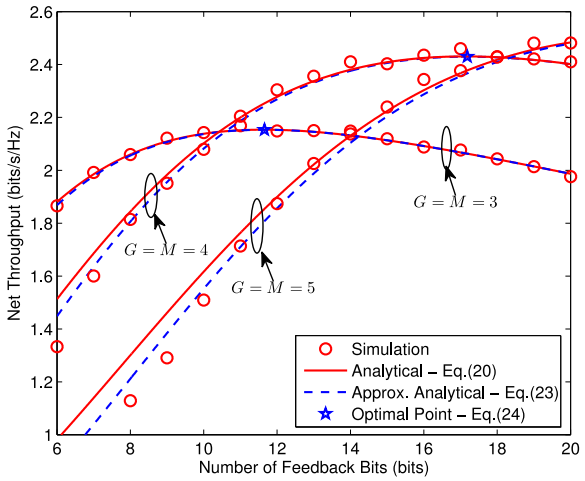
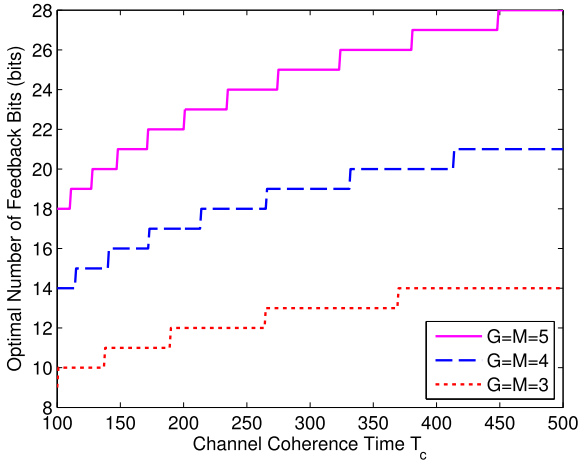
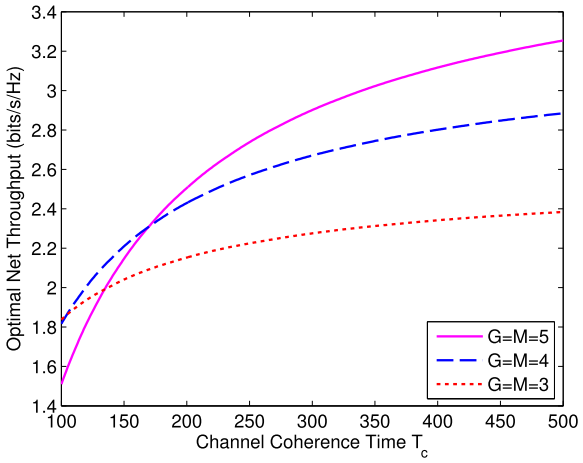


Fig. 4. The net throughput versus the number of feedback bits B under different group and antenna numbers $G = M$.



(a)



(b)

Fig. 5. (a) The optimal number of feedback bits and (b) the corresponding optimal net throughput versus the channel coherence time T_c under different group and antenna numbers $G = M$.

to be well coincident with the simulations, which validates our theoretical derivations. Furthermore, from Fig. 4 when the group (antenna) number increases, the optimal net throughput is improved due to the fact that more scheduled users can be served by more SDMA beams. The optimal number of feedback bits also increases with an increase in G and M. The reason behind is that more feedback bits are needed to quantify the higher-dimensional channel with more antennas, such that a better ZF beamforming can be designed to suppress the inter-group interference from more groups. It is also worth noting that from Fig. 4 under a larger group (antenna) number, the network performance can be even worse when the number of feedback bits is insufficient (less than 10 bits) due to unmanageable interference.

Fig. 5 plots the optimal number of feedback bits $B^\star$ and the corresponding optimal net throughput $\tilde{\Gamma}_{\text{net}}(B^\star)$ versus the channel coherence time T_c. It can be seen that both $B^\star$ and $\tilde{\Gamma}_{\text{net}}(B^\star)$ increase with an increasing T_c, which validates our analytical findings in Proposition 1. This outcome is expected since the penalty of uplink CSI feedback becomes smaller as T_c increases. Moreover, the growth rate of $B^\star$ due to the increase of T_c becomes higher under a larger group and antenna number $G = M$. Indeed, throughout the presented range of T_c in Fig. 5(a), the increment of $B^\star$ is 5 bits under $G = M = 3$ while this range increases to 10 bits under $G = M = 5$. This observation validates the conclusion presented in Proposition 2. The reason behind stems from the needs to concurrently quantize the higher-dimensional channel space and cope with the increased inter-group interference under larger group and antenna numbers. However, as G and M increase, the corresponding optimal net throughput can be smaller when the channel coherence time is short from Fig. 5(b). This is because the channel fading is fast under this scenario and the overhead for uplink CSI feedback is large, while more feedback bits are required when G and M are large.

Fig. 6 illustrates the impacts of the user number in each group K on the optimal net throughput under different group and antenna numbers. It is shown that the user number in each group has a large influence on the optimal net throughput. In particular, when the user number per group is small, the optimal net throughput increases as more users are concurrently scheduled in each sector. However, as more and more users (larger than four when $G = M = 3$)

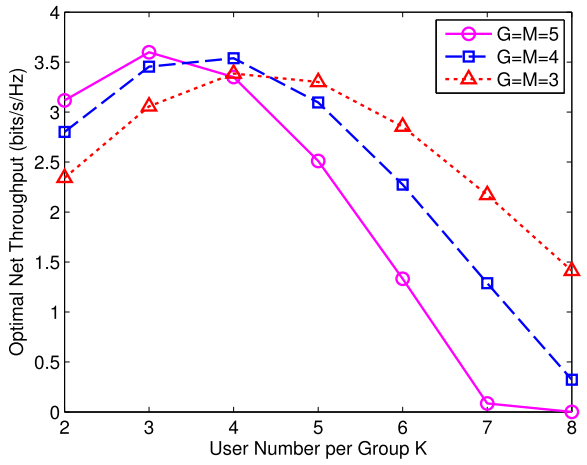


Fig. 6. The optimal net throughput versus user number in each group K under different group and antenna numbers $G = M$. The normalized channel coherence time is $T_c = 400$.

are scheduled, the throughput begins to decrease rapidly as observed from Fig. 6. This is because when K is large, the increases of feedback overhead and outage probability incurred by the more intra-group interference dominate that of network-wide throughput owing to the more served users. In addition, this phenomenon becomes more prominent as the group and antenna numbers are larger, since the inter-group interference can be relatively larger due to the fixed total transmit power at the BS.

VI. Conclusion

In this paper, we have investigated the outage performance in a downlink SDMA-NOMA cellular network with general limited feedback for the first time. A closed-form expression of the outage probability in the considered network has been derived. A low-complexity algorithm has been proposed to find the optimal number of feedback bits for net throughput maximization. It has been shown that there always exists an outage probability floor due to limited feedback, and the optimal number of feedback bits for net throughput maximization is an increasing function of the channel coherence time.

Appendix A
Proof of Theorem 1

According to [36], since $\tilde{\mathbf{g}}_{n,k}$ and $\mathbf{w}_n$ are independent and isotropically distributed in $\mathbb{C}^{M\times 1}$, we have $|\tilde{\mathbf{g}}_{n,k}^H \mathbf{w}_n|^2 \sim \text{Beta}(1, M-1)$ and thus $X \sim \text{Exp}(1)$. Since the CDF of the unordered distance is given in (1), the probability distribution function (PDF) of the ordered distance from the BS to its k-th farthest user is obtained by order statistics [45] as

$$\begin{aligned} f_{d_{n,k}}(x) &= k\binom{K}{k} F_d(x)^{K-k}\left(1-F_d(x)\right)^{k-1} f_d(x) \\ &= 2k\binom{K}{k}\frac{x^{2(K-k)+1}}{D^{2(K-k+1)}}\left(1-\frac{x^2}{D^2}\right)^{k-1} \\ &= 2k\binom{K}{k}\sum_{i=0}^{k-1}\binom{k-1}{i}(-1)^i\frac{x^{2(K+i-k)+1}}{D^{2(K+i-k+1)}}, \\ &\quad 0 \le x \le D, \end{aligned} \tag{27}$$

where $f_d(x)$ is the corresponding PDF of the CDF in (1). Since we have assumed that X is independent of Y, once we know PDF $f_Y(y)$ the outage probability in (14) reduces to

$$\begin{aligned} &\mathrm{P}_k^{\text{out}} \\ &= \Pr\{X - \eta_k \Lambda Y < (\eta_k/\rho)Z\} \\ &= \int_0^D \left(1 - e^{-\frac{\eta_k}{\rho}z^\alpha}\int_0^\infty f_Y(y)e^{-\eta_k \Lambda y}dy\right) f_{d_{n,k}}(z)dz \\ &= 1 - \tau_1 \int_0^D f_{d_{n,k}}(z)e^{-\frac{\eta_k}{\rho}z^\alpha}dz \\ &= 1 - 2k\tau_1\binom{K}{k}\sum_{i=0}^{k-1}\binom{k-1}{i}\frac{(-1)^i}{\alpha D^{2(K+i-k+1)}} \\ &\quad \times \left(\frac{\rho}{\eta_k}\right)^{\frac{2}{\alpha}(K+i-k+1)} \gamma\left(\frac{2}{\alpha}(K+i-k+1), \frac{\eta_k}{\rho}D^\alpha\right), \end{aligned} \tag{28}$$

where

$$\tau_1 = \int_0^\infty f_Y(y)e^{-\eta_k \Lambda y}dy, \tag{29}$$

and the last equality in (28) follows from [41, eq. (3.381.8)].

Next, we aim to find the PDF of random variable Y. From [36], we know that $\mathbf{e}_{n,k}$ and $\mathbf{w}_m$ are independent and isotropically distributed in $\mathbb{C}^{(M-1)\times 1}$ whose hyperplane is orthogonal to $\hat{\mathbf{g}}_n$. Therefore, we have $|\mathbf{e}_{n,k}^H \mathbf{w}_m|^2 \sim \text{Beta}(1, M-2)$. Additionally, $\|\mathbf{g}_{n,k}\|^2 \sin^2\theta_{n,k} \sim \text{Gamma}(M-1, \delta)$ is satisfied from [36, Lemma 1]. Thus, the random variable Y can be represented as $Y \overset{d}{=} \text{Gamma}(M-1,\delta)\sum_{i=1}^{G-1}\text{Beta}(1, M-2)$. Since we know that $\text{Gamma}(M-1,\delta)\text{Beta}(1,M-2) \overset{d}{=} \text{Gamma}(1,\delta)$ [44], Y is actually the sum of $G-1$ $\text{Gamma}(1,\delta)$ random variables correlated by a common $\text{Gamma}(M-1,\delta)$ factor. It is not hard to see that the correlation coefficient between any two addends in Y is the same and given by $\frac{1}{M-1}$. By applying the method proposed in [46], we obtain the exact PDF of Y given by

$$f_Y(y) = a_1 \sum_{n=0}^{\infty} \frac{b_n y^{G+n-2}e^{-y/a_2}}{a_2^{G+n-1}\Gamma(G+n-1)}, \quad y \ge 0, \tag{30}$$

where $a_1 = \frac{\sqrt{M-1}-1}{G-2+\sqrt{M-1}}$, $a_2 = \delta\left(1-\frac{1}{\sqrt{M-1}}\right)$, and b_n is recursively obtained by

$$b_n = \begin{cases} 1, & n = 0, \\ \frac{1}{n}\sum_{i=1}^{n}(1-a_1)^i b_{n-i}, & n = 1, 2, \cdots. \end{cases} \tag{31}$$

Substituting (30) into (29) yields

$$\begin{aligned} \tau_1 &= a_1 \sum_{n=0}^{\infty} \frac{b_n}{a_2^{G+n-1}\Gamma(G+n-1)} \\ &\quad \times \int_0^\infty y^{G+n-2}e^{-\left(\frac{1}{a_2}+\eta_k \Lambda\right)y}dy \\ &= a_1 \sum_{n=0}^{\infty} \frac{b_n}{(1+\eta_k \Lambda a_2)^{G+n-1}}, \end{aligned} \tag{32}$$

where the last equality follows from [41, eq. (3.326.2)].

By substituting (32) into (28), the proof is completed.

APPENDIX B
PROOF OF THEOREM 2

Since finding the optimal integer B maximizing (23) is NP-hard, we first relax the feasible field of B as the non-negative continuous real number, i.e., $B \geq 0$. Then we find the optimal δ instead of B that maximizes (23) by noting that $\delta = 2^{-\frac{B}{M-1}}$ is a monotonically decreasing function of B. Therefore, (23) is recast as

$$\tilde{\Gamma}_{\text{net}}(\delta) = G\left(\frac{1}{(1+\bar{\eta}\Lambda\delta)^{G-1}}Q + (M-1)K/T_c \log_2 \delta\right), \quad 0 < \delta \leq 1. \tag{33}$$

The first derivative of $\tilde{\Gamma}_{\text{net}}(\delta)$ is calculated as

$$f(\delta) = \frac{d\tilde{\Gamma}_{\text{net}}(\delta)}{d\delta} = \frac{H_1}{\delta}(H_2 - g(\delta)), \tag{34}$$

where $H_1 = Q(G-1)\bar{\eta}$ and $H_2 = \frac{(M-1)KG}{H_1 T_c \ln 2}$ are two positive terms, and $g(\delta) = \frac{\delta}{(1+\bar{\eta}\Lambda\delta)^G}$. To find the optimal δ, it is necessary to study the property of function $g(\delta)$. The first derivative of $g(\delta)$ is given by

$$\frac{dg(\delta)}{d\delta} = \frac{1-(G-1)\bar{\eta}\Lambda\delta}{(1+\bar{\eta}\Lambda\delta)^{G+1}}. \tag{35}$$

It is not hard to see that when δ changes from 0 to infinity, $g(\delta)$ first increases from 0 and then decreases to 0. By setting $\frac{dg(\delta)}{d\delta} = 0$, we find that the maximum of $g(\delta)$ is achieved when $\delta = \delta_0 = \frac{1}{(G-1)\bar{\eta}\Lambda}$.

The optimal $\delta^\star$ maximizing $\tilde{\Gamma}_{\text{net}}(\delta)$ is discussed as follows:

1) When $H_2 \geq g(\delta_0)$, we have $f(\delta_0) \geq 0$. $\tilde{\Gamma}_{\text{net}}(\delta)$ is an increasing function of δ and thus the optimal $\delta^\star = 1$;

2) When $H_2 < g(\delta_0)$, the function $y = g(\delta)$ has two crossing points with the line $y = H_2$ for $\delta > 0$. Denote the first crossing point as δ_1. Note that δ_1 can be easily obtained through a bisectional search, since $g(\delta)$ is an increasing function for $0 < \delta < \delta_0$ and δ_1 is uniquely determined. Based on the relative position of 1 and the two crossing points, the optimal $\delta^\star$ is further discussed as follows under this situation:

a) when $H_2 \geq g(1)$ and $\delta_0 \geq 1$, $\tilde{\Gamma}_{\text{net}}(\delta)$ is monotonically increasing in $\delta \in (0, 1]$ and thus the optimal $\delta^\star = 1$;
b) when $H_2 < g(1)$, $\tilde{\Gamma}_{\text{net}}(\delta)$ first increases and then decreases in $\delta \in (0, 1]$, and thus the optimal $\delta^\star = \delta_1$;
c) when $H_2 \geq g(1)$ and $\delta_0 < 1$, $\tilde{\Gamma}_{\text{net}}(\delta)$ first increases, then decreases and at last increases in $\delta \in (0, 1]$. The optimum under this situation is given by $\delta^\star = \arg\max_{\delta \in \{1, \delta_1\}} \tilde{\Gamma}_{\text{net}}(\delta)$.

Based on the discussion above, the optimal $\delta^\star$ is given in the following compact from

$$\delta^\star = \begin{cases} 1, & H_2 \geq g(\delta_0), \\ \arg\max_{\delta \in \{1, \delta_1\} \cap (0,1]} \tilde{\Gamma}_{\text{net}}(\delta), & H_2 < g(\delta_0). \end{cases} \tag{36}$$

From (36), the optimal number of feedback bits can be obtained as in Theorem 2 using the relation $\delta = 2^{-\frac{B}{M-1}}$. The proof is completed.

APPENDIX C
PROOF OF PROPOSITION 1

To prove Proposition 1, we resort to show that the optimal $\delta^\star$ in (36) is a decreasing function of T_c and the corresponding optimal net throughput $\tilde{\Gamma}_{\text{net}}(\delta^\star)$ is an increasing function of T_c based on the relation $\delta = 2^{-\frac{B}{M-1}}$.

From the analysis in Appendix B, we know that the optimal $\delta^\star \in (0, 1]$ maximizing $\tilde{\Gamma}_{\text{net}}(\delta)$ has to choose from $\{1, \delta_1\}$ where $\delta_1 \in (0, 1]$ exists only when $H_2 < g(\delta_0)$. Since $\tilde{\Gamma}_{\text{net}}(1) = \frac{GQ}{(1+\bar{\eta}\Lambda)^{G-1}}$ is fixed with respect to (w.r.t.) T_c, we put our emphasis on the behavior of $\tilde{\Gamma}_{\text{net}}(\delta_1)$ w.r.t. T_c when $H_2 < g(\delta_0)$. Recalling that δ_1 is the first crossing point of the functions $y = g(\delta)$ and $y = H_2$, the derivative of δ_1 w.r.t. T_c can be found as

$$\frac{d\delta_1}{dT_c} = \frac{\delta_1(1+\bar{\eta}\Lambda\delta_1)}{T_c((G-1)\bar{\eta}\Lambda\delta_1 - 1)}. \tag{37}$$

It is not hard to see that $\frac{d\delta_1}{dT_c} < 0$ since $\delta_1 < \delta_0 = \frac{1}{(G-1)\bar{\eta}\Lambda}$ according to Appendix B. In addition, the derivative of $\tilde{\Gamma}_{\text{net}}(\delta_1)$ w.r.t. T_c can be calculated as

$$\begin{aligned} \frac{d\tilde{\Gamma}_{\text{net}}(\delta_1)}{dT_c} &= \frac{d\tilde{\Gamma}_{\text{net}}(\delta_1)}{d\delta_1}\frac{d\delta_1}{dT_c} \\ &= \left(f(\delta_1) - G(M-1)K/T_c^2 \log_2 \delta_1 \frac{dT_c}{d\delta_1}\right)\frac{d\delta_1}{dT_c} \\ &\overset{(b)}{=} -KG(M-1)/T_c^2 \log_2 \delta_1 \geq 0, \end{aligned} \tag{38}$$

where the function $f(\delta)$ is defined in (34), and equality (b) holds because of $f(\delta_1) = 0$ from the analysis in Appendix B.

Based on the optimal structure of $\delta^\star$ in (36) and the monotonic properties of δ_1 and $\tilde{\Gamma}_{\text{net}}(\delta_1)$ in (37) and (38), respectively, we finally obtain the conclusion that the optimal $\delta^\star$ is a decreasing function of T_c and the corresponding optimal net throughput $\tilde{\Gamma}_{\text{net}}(\delta^\star)$ is an increasing function of T_c. The proof is completed.

APPENDIX D
PROOF OF PROPOSITION 2

To prove Proposition 2, we aim to prove that the mixed partial derivative $\frac{\partial^2 B_1}{\partial T_c \partial G} > 0$. According to the relation $B_1 = -(M-1)\log_2 \delta_1$ in Theorem 2, the first partial derivative of B_1 w.r.t. G can be found as

$$\frac{\partial B_1}{\partial G} = -\log_2 \delta_1 - \frac{M-1}{\delta_1 \ln 2}\frac{\partial \delta_1}{\partial G}. \tag{39}$$

By rewriting the equality of δ_1 in (25) of Theorem 2, we obtain a implicit equation given by

$$\Xi(\delta_1, G, T_c) \triangleq \frac{K}{Q\bar{\eta}\ln 2}\frac{G(1+\bar{\eta}\Lambda\delta_1)^G}{\delta_1} - T_c = 0. \tag{40}$$

By leveraging the derivative formula of implicit function, we find

$$\frac{\partial \delta_1}{\partial G} = -\frac{\Xi_G}{\Xi_{\delta_1}}, \tag{41}$$

where

$$\Xi_G \triangleq \frac{\partial \Xi}{\partial G} = \frac{T_c(G + \bar{\eta}\delta_1 - \bar{\eta}\delta_1 G)}{G(G+\bar{\eta}\delta_1)} \tag{42}$$

and

$$\Xi_{\delta_1} \triangleq \frac{\partial \Xi}{\partial \delta_1} = \frac{T_c(\bar{\eta}\delta_1 G - G - \bar{\eta}\delta_1)}{\delta_1(G + \bar{\eta}\delta_1)}, \tag{43}$$

respectively. Substituting (42) and (43) into (41) yields

$$\frac{\partial \delta_1}{\partial G} = \frac{\delta_1}{G}. \tag{44}$$

By inserting (44) into (39), we finally obtain the first partial derivative of B_1 w.r.t. G as

$$\frac{\partial B_1}{\partial G} = -\log_2 \delta_1 - \frac{M-1}{G\ln 2}. \tag{45}$$

Based on (45), we have the mixed partial derivative

$$\frac{\partial^2 B_1}{\partial T_c \partial G} = -\frac{1}{\delta_1 \ln 2}\frac{\partial \delta_1}{\partial T_c}. \tag{46}$$

By applying the similar method as in (41)–(44), we obtain

$$\frac{\partial \delta_1}{\partial T_c} = \frac{1}{\Xi_{\delta_1}}. \tag{47}$$

Substituting (47) into (46) results in

$$\begin{aligned}\frac{\partial^2 B_1}{\partial T_c \partial G} &= \frac{G + \bar{\eta}\delta_1}{T_c \ln 2}\frac{\Lambda}{1-(G-1)\bar{\eta}\Lambda\delta_1} \\ &\overset{(c)}{=} \frac{G + \bar{\eta}\delta_1}{T_c \ln 2}\frac{\Lambda}{1-\frac{\delta_1}{\delta_0}} \overset{(d)}{>} 0,\end{aligned} \tag{48}$$

where equality (c) and inequality (d) follow from the relation $0 < \delta_1 < \delta_0 = \frac{1}{(G-1)\bar{\eta}\Lambda}$ according to Theorem 2. The proof is completed.

REFERENCES

[1] Q. Yang *et al.*, "Outage performance of NOMA in downlink SDMA systems with limited feedback," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.

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

[3] V. W. S. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

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

[5] Y. Yuan *et al.*, "Non-orthogonal transmission technology in LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 68–74, Jul. 2016.

[6] 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.

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

[8] P. Xu, Z. Ding, X. Dai, and H. V. Poor, "A new evaluation criterion for non-orthogonal multiple access in 5G software defined networks," *IEEE Access*, vol. 3, pp. 1633–1639, Sep. 2015.

[9] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[10] 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.

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

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

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

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

[15] Z. Qin, Y. Liu, Z. Ding, Y. Gao, and M. Elkashlan, "Physical layer security for 5G non-orthogonal multiple access in large-scale networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

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

[17] J.-B. Kim and I.-H. Lee, "Capacity analysis of cooperative relaying systems using non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1949–1952, Nov. 2015.

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

[19] 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.

[20] 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.

[21] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Nonorthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10152–10157, Dec. 2016.

[22] Y. Zhang, Q. Yang, T.-X. Zheng, H.-M. Wang, Y. Ju, and Y. Meng, "Energy efficiency optimization in cognitive radio inspired non-orthogonal multiple access," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Valencia, Spain, Sep. 2016, pp. 1–6.

[23] 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.

[24] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-powered communications with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8422–8436, Dec. 2016.

[25] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 654–667, Feb. 2016.

[26] S. Shi, L. Yang, and H. Zhu, "Outage balancing in downlink nonorthogonal multiple access with statistical channel state information," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4718–4731, Jul. 2016.

[27] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[28] J. Choi, "Minimum power multicast beamforming with superposition coding for multiresolution broadcast and application to NOMA systems," *IEEE Trans. Commun.*, vol. 63, no. 3, pp. 791–800, Mar. 2015.

[29] Q. Sun, S. Han, C.-L. I., and Z. Pan, "On the ergodic capacity of MIMO NOMA systems," *IEEE Wireless Commun. Lett.*, vol. 4, no. 4, pp. 405–408, Aug. 2015.

[30] Q. Zhang, Q. Li, and J. Qin, "Robust beamforming for nonorthogonal multiple-access systems in MISO channels," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10231–10236, Dec. 2016.

[31] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[32] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[33] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[34] K. K. Mukkavilli, A. Sabharwal, E. Erkip, and B. Aazhang, "On beamforming with finite rate feedback in multiple-antenna systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2562–2579, Oct. 2003.

[35] N. Jindal, "MIMO broadcast channels with finite-rate feedback," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5045–5060, Nov. 2006.

[36] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, pp. 1478–1491, Sep. 2007.
[37] C. K. Au-Yeung and D. J. Love, "On the performance of random vector quantization limited feedback beamforming in a MISO system," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 458–462, Feb. 2007.
[38] D. J. Love, R. W. Heath, Jr., V. K. N. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.
[39] P. Xu, Y. Yuan, Z. Ding, X. Dai, and R. Schober, "On the outage performance of non-orthogonal multiple access with 1-bit feedback," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6716–6730, Oct. 2016.
[40] S. Liu and C. Zhang, "Downlink non-orthogonal multiple access system with limited feedback channel," in *Proc. Int. Wireless Commun. Signal Process. (WCSP) Conf.*, Nanjing, China, Oct. 2015, pp. 1–5.
[41] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. New York, NY, USA: Academic, 2007.
[42] H. Tabassum, E. Hossain, and M. J. Hossain, "Modeling and analysis of uplink non-orthogonal multiple access (NOMA) in large-scale cellular networks using Poisson cluster processes," *IEEE Trans. Commun.*, to be published, doi: 10.1109/TCOMM.2017.2699180.
[43] J. Park, N. Lee, J. G. Andrews, and R. W. Heath, Jr., "On the optimal feedback rate in interference-limited multi-antenna cellular systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5748–5762, Aug. 2016.
[44] M. Kountouris and J. G. Andrews, "Downlink SDMA with limited feedback in interference-limited wireless networks," *IEEE Trans. Wireless Commun.*, vol. 11, no. 8, pp. 2730–2741, Aug. 2012.
[45] H. A. David and H. N. Nagaraja, *Order Statistics*, 3rd ed. Hoboken, NJ, USA: Wiley, 2003.
[46] M. S. Alouini, A. Abdi, and M. Kaveh, "Sum of gamma variates and performance of wireless communication systems over Nakagami-fading channels," *IEEE Trans. Veh. Technol.*, vol. 50, no. 6, pp. 1471–1480, Nov. 2001.

Qian Yang (S'16) received the B.S. degree in information engineering from Xi'an Jiaotong University, Xi'an, China, in 2014, where he is currently pursuing the Ph.D. degree in information and communication engineering. His research interests include 5G networks, MIMO systems, convex optimization, and statistical signal processing.

Hui-Ming Wang (S'07–M'10–SM'16) received the B.S. and Ph.D. degrees (Hons.) in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2010, respectively. From 2007 to 2010, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, University of Delaware, USA. He is currently a Full Professor with the Department of Information and Communications Engineering, Xi'an Jiaotong University, and also with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, China. His research interests include cooperative communication systems, physical-layer security of wireless communications, MIMO, and space-timing coding.

He has co-authored the book *Physical Layer Security in Random Cellular Networks* (Springer, 2016), and has authored or co-authored over 100 IEEE journal and conference papers. Ten of his papers are the ESI Highly Cited Papers. He received the National Excellent Doctoral Dissertation Award in China in 2012, the Best Paper Award from the International Conference on Wireless Communications and Signal Processing in 2011, and the Best Paper Award from the IEEE/CIC International Conference on Communications in China in 2014. He has served as a TPC Member of various IEEE sponsored conferences, including the IEEE Globecom, ICC, WCNC, VTC, and PIMRC. He also served as a Symposium Chair of the Wireless Communications and Networking in ChinaCom 2015 and a Technical Program Committee Chair of the Workshop on Physical Layer Security in the IEEE Globecom 2016. He is currently an Associate Editor of the IEEE ACCESS.

Derrick Wing Kwan Ng (S'06–M'12) received the bachelor's degree (Hons.) and the M.Phil. degree in electronic engineering from The Hong Kong University of Science and Technology (HKUST) in 2006 and 2008, respectively, and the Ph.D. degree from The University of British Columbia (UBC) in 2012. In 2011 and in 2012, he was a Visiting Scholar with the Centre Tecnològic de Telecomunicacions de Catalunya-Hong Kong. He was a Senior Post-Doctoral Fellow with the Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nürnberg, Germany. He is currently a Senior Lecturer and an ARC DECRA Research Fellow with the University of New South Wales, Sydney, Australia. His research interests include convex and non-convex optimization, physical layer security, wireless information and power transfer, and green (energy-efficient) wireless communications.

Dr. Ng received the Best Paper Award at the IEEE International Conference on Computing, Networking and Communications 2016, the IEEE Wireless Communications and Networking Conference (WCNC) 2012, the IEEE Global Telecommunication Conference (Globecom) 2011, and the IEEE Third International Conference on Communications and Networking in China 2008. He received the IEEE Student Travel Grants for attending the IEEE WCNC 2010, the IEEE International Conference on Communications (ICC) 2011, and the IEEE Globecom 2011. He was also a recipient of the 2009 Four Year Doctoral Fellowship from UBC, Sumida&Ichiro Yawata Foundation Scholarship in 2008, and the R&D Excellence Scholarship from the Center for Wireless Information Technology, HKUST, in 2006. He was a Co-Chair of the Wireless Access Track of the 2014 IEEE 80th Vehicular Technology Conference, the Globecom 2016, the 2017 International Workshop on Wireless Energy Harvesting Communication Networks, and the Globecom International Workshop on Sub-6 GHz Spectrum for 5G Progress. He has been a TPC Member of various conferences, including the Globecom, WCNC, ICC, VTC, and PIMRC. He was honored as an Exemplary Reviewer of the IEEE TRANSACTIONS ON COMMUNICATIONS in 2015, the Top Reviewer of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in 2014 and 2016, and an Exemplary Reviewer of the IEEE WIRELESS COMMUNICATIONS LETTERS in 2012, 2014, and 2015. He has been serving as an Editorial Assistant to the Editor-in-Chief of the IEEE TRANSACTIONS ON COMMUNICATIONS since 2012. He is currently an Editor of the IEEE COMMUNICATIONS LETTERS, the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING.

Moon Ho Lee (S'81–M'85–SM'86–LSM'15) received the Ph.D. degree in electrical engineering from Chonnam National University, South Korea, in 1984, and the Ph.D. degree in electrical engineering from The University of Tokyo, Japan, in 1990. He held a post-doctoral position at the University of Minnesota, USA, from 1985 to 1986. He was a Chief Engineer with Namyang MBC Broadcasting from 1970 to 1980. He joined Chonbuk National University, South Korea, as a Professor, where he is currently a Professor and a former Chair of the Department of Electronics Engineering. He has made significant original contributions in the areas of mobile communication code design, channel coding, and multidimensional source and channel coding. He holds 116 Patents. He was the Inventor of Jacket Matrix and was cited over 95 559 times, in 2014, in Wikipedia. He is a member of the National Academy of Engineering, South Korea, and a Foreign Fellow of the Bulgaria Academy of Sciences.