Liu and Zhang *EURASIP Journal on Wireless Communications and Networking* (2016) 2016:239
DOI 10.1186/s13638-016-0735-9

EURASIP Journal on Wireless Communications and Networking

RESEARCH Open Access

Non-orthogonal multiple access in a downlink multiuser beamforming system with limited CSI feedback

Shimei Liu[1] and Chao Zhang[1,2*]

Abstract

Non-orthogonal multiple access (NOMA) has been recognized as a promising multiple access technology for fifth generation (5G) mobile communication system. However, the advantage of NOMA is only verified under the ideal condition that the transmitter has the perfect knowledge of channel state information (CSI). In this paper, NOMA downlink multiuser system, in which the transmitter acquires the CSI through limited feedback channel, is studied. Two traditional beamforming technologies, zero-forcing beamforming and random beamforming, are investigated in the NOMA downlink multiuser system. Making use of the imperfect CSI feedback, channel direction information (CDI), and channel quality indicator (CQI), we propose the user selection scheme to reduce the interference between the NOMA users. Furthermore, a power allocation scheme is proposed to improve the sum-rate of NOMA system. Numerical results show that NOMA system with limited feedback channel still gains larger system rate than traditional orthogonal multiple access system. It is also shown that random beamforming is more suitable for the NOMA system with limited CSI feedback.

1 Introduction

As a promising multiple access technology for fifth generation (5G) wireless systems, non-orthogonal multiple access (NOMA) has received many attentions recently, as it can improve the system capacity and spectrum efficiency [1–3]. Besides, compared to orthogonal multiple access (OMA) systems, it has the exciting advantage of facilitating ubiquitous accessing of wireless nodes from heterogeneous networks [2]. Therefore, NOMA has been treated as a candidate technology of air interface for future communication systems.

There have been some early work of analyzing and optimizing the NOMA systems. In [4] and [5], the system-level performances of NOMA system have been proven to be superior than that of the OMA system. Ding et al. in [6] demonstrated that NOMA with appropriate transmit power allocation can achieve superior system sum-rates and better outage performances than the OMA system. In [7], a cooperative NOMA transmission scheme which makes use of users with better channel conditions as relays to transmit messages for users with poor channel conditions was proposed. In [8], the authors proposed a NOMA-based downlink cooperative cellar system, where the transmitter transmits information to two paired mobile users simultaneously with the help of a dedicated relay station. A NOMA scheme to minimize spectrum usage was proposed in [9]. In [10], a NOMA with coordinated relay transmission was introduced, in which a base station (BS) directly communicates with NOMA user 1 while communicating with NOMA user 2 via a relay. The impact of user pairing on 5G NOMA system was investigated in [11]. For coordinated two-point systems, [12] showed that NOMA can bring cell-edge users to obtain a reasonable transmission rate without suffering loss of the near users' rates. By [13] and [14], it was shown that multi-antenna NOMA system outperforms the traditional time division multiple access (TDMA) system. To further enhance the performance of NOMA, the impact of applying the successive interference cancelation (SIC) receiver in downlink NOMA was addressed in [15]. User clustering scheme and precoding matrix optimization for the NOMA system were studied in [16]. In [17],

*Correspondence: chaozhang@mail.xjtu.edu.cn
[1]School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2]National Mobile Communications Research Laboratory, Southeast University, Nanjing, China

the user selection and power schedule scheme for downlink NOMA system based on zero-forcing beamforming (NOMA-ZFBF) were proposed to improve the NOMA system capacity. An uplink NOMA system was proposed in [18]. All the above work is proposed to improve the overall performance of the NOMA system. On the other hand, an important issue for NOMA system, the fairness between NOMA users, is investigated in [19–21].

Most of the aforementioned results are based on the assumption that the perfect channel state information (CSI) is known at the transmitter. In practice, however, it is difficult or impossible to let the transmitter have the perfect knowledge of CSI [22]. Thus, whether the advantage of NOMA still appears should be checked and answered. In [23], the performance of downlink NOMA system with channel estimation error was analyzed and it was shown that NOMA still obtains system rate gains compared to traditional OMA systems. Bedsides, a channel correlation matrix-based CSI feedback scheme was proposed in massive-antenna NOMA system [24]. One promising approach to provide the transmitter with imperfect CSI is limited CSI feedback, where each user quantizes its CSI and feeds the corresponding quantized index to the transmitter [25]. To the best of our knowledge, there is little research on the NOMA system with limited CSI feedback.

In this paper, we intend to investigate the performance of NOMA system in downlink multiuser beamforming system with limited CSI feedback.[1] Two beamforming technologies in downlink NOMA system, zero-forcing beamforming (NOMA-ZFBF) and random beamforming (NOMA-RBF), are employed in our system model. Via the limited feedback channel, the channel direction information (CDI) and channel quality indicator (CQI) are fed back to the BS. With the imperfect CSI, we propose the user selection scheme to reduce the interference between the NOMA users and the power allocation scheme to improve the sum-rate of NOMA system. Finally, numerical results show that the NOMA-BF system with our proposed schemes can improve the sum-rate performance under condition of the limited CSI feedback. We also compare the performances of NOMA-ZFBF and NOMA-RBF, and it is shown that NOMA-RBF has better performance than NOMA-ZFBF in the limited CSI feedback case.

This paper is organized as follows. We provide the system model of the downlink NOMA and introduce the SIC detection mechanisms in Section 2. Section 3 presents the limited feedback model of the NOMA system in terms of RBF and ZFBF and introduces the proposed user selection schemes for both beamforming schemes. Power optimization for NOMA system is presented in Section 4. Simulation results are investigated in Section 5, and we conclude this paper in Section 6.

We adopt the following notations. Uppercase boldface letters denote matrices, and lowercase boldface letters denote vectors. Sets are indicated by uppercase calligraphic letters. $|\mathcal{C}|$ denotes the size of a set $\mathcal{C}$. $\mathbf{A}^{\mathrm{T}}(\mathbf{a}^{\mathrm{T}})$ denotes the transpose of a matrix $\mathbf{A}$ (vector $\mathbf{a}$), $\mathbf{A}^{\mathrm{H}}(\mathbf{a}^{\mathrm{H}})$ denotes the conjugate transpose of a matrix $\mathbf{A}$ (vector $\mathbf{a}$), and $\mathbf{A}^{\dagger} = \mathbf{A}^{\mathrm{H}}(\mathbf{A}\mathbf{A}^{\mathrm{H}})^{-1}$ stands for the pseudo-inverse of $\mathbf{A}$. $\|\bullet\|^2$ is the two-norm of a vector. $\mathbb{E}\{\bullet\}$ represents the expectation operator.

2 System model

We consider a downlink multiuser BF transmission system consisting of a single base station (BS) with K antennas and M candidate users with one antenna ($M \geq 2K$), as depicted in Fig. 1. Define the candidate user set as $\mathcal{U}$, i.e., $|\mathcal{U}| = M$. The BS simultaneously transmits K beams to provide multiuser downlink transmission, and each beam serves one user cluster, which includes two or more users to perform NOMA. For simplicity of analyzing, we assume the user cluster contains two users like [7] and [17]. To exploit the advantage of NOMA, users in one cluster have different target data rates. The user, who requires a higher data rate, is named the high-rate (HR) user. The other one with a lower transmission rate is named the low-rate (LR) user. Denote the HR user set as $\mathcal{U}_h$ and the LR user set as $\mathcal{U}_l$. $\mathcal{U}_h(k)$ and $\mathcal{U}_l(k)$ stand for the index numbers of the HR user and LR user in the kth cluster, respectively. Note that in the kth cluster, we always denote the HR user with "1" and the LR user with "2" to distinguish the related signals, i.e., $k1 = \mathcal{U}_h(k)$ and $k2 = \mathcal{U}_l(k)$.

The BS tries to send $\sum_{k=1}^{K} \mathbf{w}_k x_k$ with power $p_k = \frac{P}{K\|\mathbf{w}_k\|^2}$, where $\mathbf{w}_k$ is the $K \times 1$ BF vector for the kth cluster, P is the total transmitted power, and x_k is the superposition symbol of the transmitted symbols for the kth cluster, i.e., $x_k = \sqrt{\alpha_k} s_{k1} + \sqrt{1-\alpha_k} s_{k2}$. Here, s_{k1} and s_{k2} represent the transmitted symbol for the HR and LR users in the kth cluster, respectively. We assume that the information symbols $s_{ki}, i = 1, 2$ have unit energy, i.e., $\mathbb{E}\{\|s_{ki}\|^2\} = 1$. Additionally, $\alpha_k \in (0, 1)$ is the power allocation coefficient between the HR user and the LR user in the kth user cluster. Denote the $1 \times K$ vectors $\mathbf{h}_{k1}$ and $\mathbf{h}_{k2}$ as the channel coefficient vectors of the HR user and the LR user in the kth cluster, respectively. Herein, $\mathbf{h}_{ki}$ has zero mean and unit variance i.i.d complex Gaussian entries. Therefore, the received signal y_{ki} at the two users in the kth cluster can be given as

$$y_{ki} = \sqrt{p_k}\mathbf{h}_{ki}\mathbf{w}_k x_k + \sqrt{p_k}\mathbf{h}_{ki} \sum_{j=1, j\neq k}^{K} \mathbf{w}_j x_j + n_{ki} \quad \text{for } i = 1, 2 \tag{1}$$

where n_{ki} is the receiver noise and follows complex Gaussian with zero mean and unit variance, i.e., $n_{ki} \sim \mathcal{CN}(0, 1)$. Observing (1), the first term on the right-hand

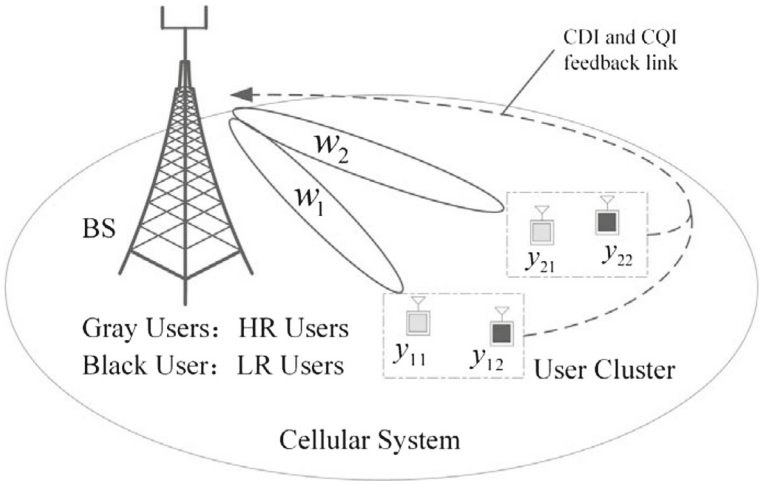

Fig. 1 System model

side of the equality sign is a useful signal for the kth cluster and the second term is the interference from other $K-1$ beams. Following the results in [17], we define the indicator function for HR and LR users as

$$g_{ki} = \frac{p_k |\mathbf{h}_{ki}\mathbf{w}_k|^2}{1 + \frac{p}{K}\sum_{j=1,j\neq k}^{K} |\mathbf{h}_{ki}\mathbf{w}_j|^2} \quad \text{for } i = 1,2; k = 1,\ldots,K \tag{2}$$

Note that g_{k1} and g_{k2} can be treated as the received signal-to-interference-plus-noise ratio (SINR) with respect to x_k at the HR user and LR user, respectively. Furthermore, x_k contains both s_{k1} and s_{k2}. Thus, g_{ki} is not the SINR with respect to s_{ki}. We define g_{ki} herein to determine the detection order at the receivers.

2.1 Successive interference cancelation receiver

As both y_{k1} and y_{k2} include the intra-beam interference and users in one cluster cannot exchange its received information, SIC is implemented to remove the intra-beam interference in NOMA system [3–17]. As the system capacity is up to the detecting order of the SIC receiver, we consider two cases:

2.1.1 Case 1. *Detect s_{k2} firstly*

In this case, s_{k2} is firstly detected by both users. Then, the LR user directly detects its desired information from y_{k2}. So, the maximum achievable rate of the LR user is

$$R_{k2} = \log\left(1 + \frac{1-\alpha_k}{\alpha_k + \frac{1}{g_{k2}}}\right) \tag{3}$$

Yet, the HR user needs to remove $\mathbf{h}_{k1}\mathbf{w}_k\sqrt{1-\alpha_k}s_{k2}$ from y_{k1} and detect s_{k1} subsequently. To avoid error propagation, s_{k2} should be detected correctly by the HR user. As R_{k2} is directly determined by the LR user, the HR should provide large enough channel capacity to ensure detecting s_{k2} without error. It means $R_{k,2\to1} \geq R_{k2}$, where

$$R_{k,2\to1} = \log\left(1 + \frac{1-\alpha_k}{\alpha_k + \frac{1}{g_{k1}}}\right) \tag{4}$$

stands for the maximum achievable rate for the HR user to detect s_{k2}. Thus, after that, the maximum achievable rate of the HR user is

$$R_{k1} = \log\left(1 + \alpha_k g_{k1}\right) \tag{5}$$

To ensure $R_{k,2\to1} \geq R_{k2}$, by (2), there must be

$$g_{k1} \geq g_{k2} \tag{6}$$

2.1.2 Case 2. *Detect s_{k1} firstly*

In this case, s_{k1} is firstly detected by both users. Hence, the LR user needs to remove $\mathbf{h}_{k2}\mathbf{w}_k\sqrt{\alpha_k}s_{k1}$ from y_{k2} and detect s_{k2} from the remaining signal. Therefore, to prevent error propagation at the LR user, similar to Case 1, the condition $R_{k,1\to2} \geq R_{k1}$ is required, where $R_{k,1\to2}$ denotes the maximum achievable rate for the LR user to detect s_{k1}. Herein, R_{k1}, $R_{k,1\to2}$, and R_{k2} are

$$R_{k1} = \log\left(1 + \frac{\alpha_k}{(1-\alpha_k) + \frac{1}{g_{k1}}}\right), \tag{7}$$

$$R_{k,1\to2} = \log\left(1 + \frac{\alpha_k}{(1-\alpha_k) + \frac{1}{g_{k2}}}\right), \tag{8}$$

and

$$R_{k2} = \log\left(1 + (1-\alpha_k)g_{k2}\right). \tag{9}$$

Due to (7) and (8), $R_{k,1\to2} \geq R_{k1}$ is equivalent to

$$g_{k1} \leq g_{k2} \tag{10}$$

Like Case 1, (10) is the necessary condition to detect s_{k1} firstly without error propagation. According to (6) and

(10), we know that the decoding order of the SIC receivers in each cluster is determined by the values of g_{k1} and g_{k2}. In summary, if $g_{k1} \geq g_{k2}$, detect s_{k2} firstly and vice versa. Specially, if $g_{k1} = g_{k2}$, both orders work well. We herein choose Case 1 as default. To let the users in a cluster know the detection order, the BS can use the flag bit to inform HR and LR users.

In the above analysis, we assume that the BS has the perfect CSI to determine the beamforming vectors and the detection order. In practice, it is difficult to acquire the perfect CSI at the transmitter. One promising method to obtain g_{k1} and g_{k2} is to let the users feed back their related channel information by limited feedback channels.

3 NOMA system with limited CSI feedback

In this paper, we assume that all M candidate users can perfectly estimate their channel state information and have to send CSI via a finite-rate, zero-delay, and error-free feedback channel, which is usually modeled in multiuser multiple-input and multiple-output (MIMO) systems [26]. In the limited feedback channel, CDI and CQI of candidate users are fed back to the BS. After that, the BS uses its obtained CDI and CQI to design beamforming vectors $\mathbf{w}_k, k = 1, 2, \ldots, K$ and determines the detection orders in these K clusters. To be compatible with existing mobile communication systems, we consider two classic beamforming technologies: zero-forcing beamforming (ZFBF), which is a linear precoding and easy to be implemented, and random beamforming (RBF), which is specially designed for the transmitter with limited CSI feedback. Since different beamforming technologies may require different CDI and CQI, we next separately discuss the feedback models for ZFBF and RBF. Additionally, the imperfect feedback scenario, e.g., latency and noise in the limited feedback channel, is left for our further study.

3.1 Zero-forcing beamforming

ZFBF technology is widely studied in multi-antenna-based wireless systems, as it incurs low computing complexity of precoding and decoding [26]. So, we first address the NOMA downlink multiuser system with ZFBF (NOMA-ZFBF).

3.1.1 CDI feedback model

All M candidate users quantize the direction of their channel vector $\tilde{\mathbf{h}}_m = \mathbf{h}_m / \|\mathbf{h}_m\|, 1 \leq m \leq M$ to a unit norm vector $\hat{\mathbf{h}}_m$ which is chosen from a codebook. Each user has its own codebook formed by $N = 2^B$ unit norm row vectors which are generated by a random vector quantization (RVQ) algorithm in [26]. For the mth user, the codebook stored by both the BS and mth user is

$$\mathcal{C}_m = \left\{ \mathbf{c}_{m,1}^H, \ldots, \mathbf{c}_{m,N}^H \right\} \tag{11}$$

and $\hat{\mathbf{h}}_m = \mathbf{c}_{m,n}$ is based on the minimum distance criterion:

$$n = \arg \max_{1 \leq j \leq N} \left| \tilde{\mathbf{h}}_m \mathbf{c}_{m,j}^H \right| \tag{12}$$

Thus, each user just needs to feed back the quantization index n with B bits to the BS .

In this section, we assume that K HR users and K LR users have been selected from M candidate users. How to select these $2K$ users in K user clusters is illustrated later. After obtaining K HR users, we define $\hat{\mathbf{H}} = \left[\hat{\mathbf{h}}_{11}^T, \ldots, \hat{\mathbf{h}}_{K1}^T \right]^T$ as the HR user CDI vector and $\mathbf{W}$ as the beamforming matrix. By the ZFBF method in [26], the beamforming matrix is

$$\mathbf{W} = \hat{\mathbf{H}}^{\dagger} = \hat{\mathbf{H}}^{\mathrm{H}} \left(\hat{\mathbf{H}} \hat{\mathbf{H}}^{\mathrm{H}} \right)^{-1} \tag{13}$$

where $\hat{\mathbf{H}}^{\dagger}$ indicates the pseudo-inverse of $\hat{\mathbf{H}}$ and $\hat{\mathbf{H}}^{\mathrm{H}}$ is the conjugate transpose of matrix $\hat{\mathbf{H}}$. Thus, the beamforming vector of the kth beam, $\mathbf{w}_k$, is the kth column of $\mathbf{W}$ and satisfies

$$\hat{\mathbf{h}}_{k1} \mathbf{w}_j = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases} \tag{14}$$

3.1.2 CQI feedback model

To decide the SIC detecting order and optimize system capacity, the BS needs to know g_{k1} and g_{k2}. As using CDI to calculate g_{k1} and g_{k2} incurs great distortion, each user has to provide the CQI information for the BS to estimate g_{k1} and g_{k2}. For the mth user , the angle $\theta_m \in [0, \pi/2]$ between vectors $\tilde{\mathbf{h}}_m$ and $\hat{\mathbf{h}}_m$ is given by $\cos \theta_m = |\tilde{\mathbf{h}}_m \hat{\mathbf{h}}_m^{\mathrm{H}}|$. Following the decomposition method in [26], there is

$$\tilde{\mathbf{h}}_m = \left(\tilde{\mathbf{h}}_m \hat{\mathbf{h}}_m^{\mathrm{H}} \right) \hat{\mathbf{h}}_m + \mathbf{e}_m = \cos \theta_m \hat{\mathbf{h}}_m + \sin \theta_m \tilde{\mathbf{e}}_m \tag{15}$$

where $\mathbf{e}_m$ indicates the quantization error vector and $\tilde{\mathbf{e}}_m = \frac{\mathbf{e}_m}{\|\mathbf{e}_m\|}$.

First, we start from g_{k1}. Define $\tilde{\mathbf{w}}_k = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|}$. So, according to (14), there are $|\hat{\mathbf{h}}_{k1} \tilde{\mathbf{w}}_k| = \frac{1}{\|\mathbf{w}_k\|}$ and $\mathbf{e}_{k1} \tilde{\mathbf{w}}_k \approx 0$. Therefore, we have

$$\begin{aligned} g_{k1} &= \frac{p_k |\mathbf{h}_{k1} \mathbf{w}_k|^2}{1 + \sum_{j=1, j \neq k}^{K} p_j |\mathbf{h}_{k1} \mathbf{w}_j|^2} \\ &= \frac{\frac{P}{K} \|\mathbf{h}_{k1}\|^2 \left| \left(\tilde{\mathbf{h}}_{k1} \hat{\mathbf{h}}_{k1}^H \right) \hat{\mathbf{h}}_{k1} \tilde{\mathbf{w}}_k + \mathbf{e}_{k1} \tilde{\mathbf{w}}_k \right|^2}{1 + \|\mathbf{h}_{k1}\|^2 \sum_{j=1, j \neq k}^{K} p_j \left| \left(\tilde{\mathbf{h}}_{k1} \hat{\mathbf{h}}_{k1}^H \right) \hat{\mathbf{h}}_{k1} \mathbf{w}_j + \mathbf{e}_{k1} \mathbf{w}_j \right|^2} \\ &\approx \frac{\frac{1}{\|\mathbf{w}_k\|^2} \frac{P}{K} \|\mathbf{h}_{k1}\|^2 \cos^2 \theta_{k1}}{1 + \frac{P}{K} \|\mathbf{h}_{k1}\|^2 \sin^2 \theta_{k1} \sum_{j=1, j \neq k}^{K} |\tilde{\mathbf{e}}_{k1} \tilde{\mathbf{w}}_j|^2} \end{aligned} \tag{16}$$

Since the quantization error vectors are random variables, we should get the expectation of (16) with respect to $\tilde{\mathbf{e}}_{k1}, k = 1, 2, \ldots, K$. Note that both $\tilde{\mathbf{e}}_{k1}$ and $\tilde{\mathbf{w}}_j$ are isotropically distributed on the $K - 1$ dimensional hyperplane orthogonal to $\hat{\mathbf{h}}_{k1}$. Also, as $\tilde{\mathbf{w}}_j$ is determined by $\hat{\mathbf{H}}$, we

infer that $\tilde{\mathbf{w}}_j$ is independent of $\tilde{\mathbf{e}}_{k1}, k1 \neq j$. Therefore, $|\tilde{\mathbf{e}}_{k1}\tilde{\mathbf{w}}_j|^2$ is a beta-distributed random variable with parameters $(1, K-2)$, and its mean value is $\frac{1}{K-1}$. Furthermore, there is $\mathbb{E}(\sum_{j=1,j\neq k}^{K}|\tilde{\mathbf{e}}_{k1}\tilde{\mathbf{w}}_j|^2) = 1$. As a result, we have

$$\begin{aligned}\mathbb{E}\left(g_{k1}\right) &\overset{(a)}{\geq} \frac{\frac{1}{\|\mathbf{w}_k\|^2}\frac{P}{K}\|\mathbf{h}_{k1}\|^2\cos^2\theta_{k1}}{1+\frac{P}{K}\|\mathbf{h}_{k1}\|^2\sin^2\theta_{k1}\mathbb{E}\left(\sum_{j=1,j\neq k}^{K}|\tilde{\mathbf{e}}_{k1}\tilde{\mathbf{w}}_j|^2\right)} \\ &= \frac{\frac{1}{\|\mathbf{w}_k\|^2}\frac{P}{K}\|\mathbf{h}_{k1}\|^2\cos^2\theta_{k1}}{1+\frac{P}{K}\|\mathbf{h}_{k1}\|^2\sin^2\theta_{k1}} \end{aligned} \tag{17}$$

where Jensen's inequality is applied in step (a). Because these M candidate users have no idea on the beamforming matrix $\mathbf{W}$ and selected HR and LR users before CSI feedback, according to (17), all candidate users have to transmit their CQI as

$$f_1(\mathbf{h}_m) = \frac{\frac{P}{K}\|\mathbf{h}_m\|^2\cos^2\theta_m}{1+\frac{P}{K}\|\mathbf{h}_m\|^2\sin^2\theta_m} \tag{18}$$

For g_{k2}, similarly, we also define $\tilde{\mathbf{h}}_{k2} = \mathbf{h}_{k2}/\|\mathbf{h}_{k2}\|$, which can be composed into $\tilde{\mathbf{h}}_{k2} = \left(\tilde{\mathbf{h}}_{k2}\hat{\mathbf{h}}_{k1}^H\right)\hat{\mathbf{h}}_{k1} + \mathbf{e}_{k2}$. Here, define $\phi_k \in [0, \pi/2]$ as the angle between $\tilde{\mathbf{h}}_{k2}$ and $\hat{\mathbf{h}}_{k1}$, i.e., $\cos\phi_k = |\tilde{\mathbf{h}}_{k2}\hat{\mathbf{h}}_{k1}^H|$. Thus, there is

$$\begin{aligned} g_{k2} &= \frac{p_k|\mathbf{h}_{k2}\mathbf{w}_k|^2}{1+\sum_{j=1,j\neq k}^{K}p_j|\mathbf{h}_{k2}\mathbf{w}_j|^2} \\ &= \frac{p_k\|\mathbf{h}_{k2}\|^2|\left(\tilde{\mathbf{h}}_{k2}\hat{\mathbf{h}}_{k1}^H\right)\hat{\mathbf{h}}_{k1}\mathbf{w}_k + \mathbf{e}_{k2}\mathbf{w}_k|^2}{1+\frac{P}{K}\|\mathbf{h}_{k2}\|^2\sum_{j=1,j\neq k}^{K}|\left(\tilde{\mathbf{h}}_{k2}\hat{\mathbf{h}}_{k1}^H\right)\hat{\mathbf{h}}_{k1}\tilde{\mathbf{w}}_j + \mathbf{e}_{k2}\tilde{\mathbf{w}}_j|^2} \\ &\overset{(b)}{\approx} \frac{p_k\|\mathbf{h}_{k2}\|^2\cos^2\phi_k}{1+\frac{P}{K}\|\mathbf{h}_{k2}\|^2\sin^2\phi_k\sum_{j=1,j\neq k}^{K}|\tilde{\mathbf{e}}_{k2}\tilde{\mathbf{w}}_j|^2} \end{aligned} \tag{19}$$

Note that key step (b) is based on the assumption that $\mathbf{e}_{k2}\tilde{\mathbf{w}}_k \approx 0$, which is guaranteed by the user selection algorithm in the next subsection. Like (17), we can obtain

$$\begin{aligned}\mathbb{E}[g_{k2}] &\geq \frac{p_k\|\mathbf{h}_{k2}\|^2\cos^2\phi_k}{1+\frac{P}{K}\|\mathbf{h}_{k2}\|^2\sin^2\phi_k\mathbb{E}\left(\sum_{j=1,j\neq k}^{K}|\tilde{\mathbf{e}}_{k2}\tilde{\mathbf{w}}_j|^2\right)} \\ &= \frac{p_k\|\mathbf{h}_{k2}\|^2\cos^2\phi_k}{1+\frac{P}{K}\|\mathbf{h}_{k2}\|^2\sin^2\phi_k} \\ &\overset{(c)}{\geq} \frac{p_k\|\mathbf{h}_{k2}\|^2\cos^2(\theta_{k2}+\eta_k)}{1+\frac{P}{K}\|\mathbf{h}_{k2}\|^2\sin^2(\theta_{k2}+\eta_k)} \\ &\overset{(d)}{\approx} \frac{p_k\|\mathbf{h}_{k2}\|^2\cos^2\theta_{k2}}{1+\frac{P}{K}\|\mathbf{h}_{k2}\|^2\sin^2\theta_{k2}} \end{aligned} \tag{20}$$

where $\eta_k \in [0, \pi/2]$ is the angle between $\hat{\mathbf{h}}_{k2}$ and $\hat{\mathbf{h}}_{k1}$ and step (c) is due to the inequality $\phi_k \leq \theta_{k2} + \eta_k$. In step (d), in order to reduce the intra-beam interference, the BS should try to select the LR user to let η_k as small as possible. By (20), each candidate user needs to send CQI as

$$f_2(\mathbf{h}_m) = \frac{\frac{P}{K}\|\mathbf{h}_m\|^2\cos^2\theta_m}{1+\frac{P}{K}\|\mathbf{h}_m\|^2\sin^2\theta_m} \tag{21}$$

Since each candidate cannot determine whether it is a HR user or LR user, observing (18) and (21), the CQI of candidate user m can be written as

$$f(\mathbf{h}_m) = \frac{\frac{P}{K}\|\mathbf{h}_m\|^2\cos^2\theta_m}{1+\frac{P}{K}\|\mathbf{h}_m\|^2\sin^2\theta_m} \tag{22}$$

Summarily, each candidate user feeds back its CDI $\hat{\mathbf{h}}_m$ and CQI $f(\mathbf{h}_m)$ to the BS in a predefined manner, e.g., time division order. After that, the BS can obtain the ZFBF matrix $\mathbf{W}$ and select M HR users and M LR users to perform the NOMA-ZFBF downlink multiuser system. As said, to make the CQI approach the real g_{k1} and g_{k2} as closely as possible, we should select users in a cluster to ensure $\eta_k \approx 0$.

3.1.3 User selection algorithm

To be compatible with existing MIMO systems and meet the high-rate requirement of HR users, K HR users are suggested to be selected first by the so-called semi-orthogonal user selection (SUS) algorithm proposed in [26]. The purpose of using the SUS algorithm is to let the HR users experience as small as possible inter-beam interferences to meet the high-rate quality of service (QoS) requirement. Then, to select K LR users from the residual $M-K$ users, we propose a LR user selection algorithm herein. After K HR users have been selected, ZFBF vectors $\{\mathbf{w}_k, 1 \leq k \leq K\}$ are obtained by the transmitter. From (20), the user meeting $\eta_k \approx 0$ should be matched for the existing HR user in a cluster, i.e.,

$$\mathcal{U}_l(k) = \arg\max_{m\in\mathcal{U}-\mathcal{U}_h}|\hat{\mathbf{h}}_{k1}\hat{\mathbf{h}}_m^H| \tag{23}$$

In summary, we give the LR user selection algorithm in Algorithm 1.

3.2 Random beamforming

In the NOMA-ZFBF system, the quantization codebook is independently generated by each candidate user, so there is no coordination among these quantizers. Nevertheless, the beamforming vectors are designed through the quantized CDI of the selected HR users. Thus, the quantization errors could be so large that the inter-beam interferences may degrade the system performances. On the one hand, to reduce the quantization error of the whole system, we have to coordinate the codebooks of these users. On the other hand, a low-complexity beamforming suitable for limited feedback condition is expected. In this section, we introduce a random beamforming scheme for

Algorithm 1 LR user selection algorithm in the NOMA-ZFBF system

1: The BS receives all CDI $\{\hat{\mathbf{h}}_m, m = 1, 2, \ldots, M\}$ and CQI $\{f(\mathbf{h}_m), m = 1, 2, \ldots, M\}$ fed back by all candidate users in $\mathcal{U}$ via a limited feedback channel.

2: By SUS algorithm proposed in [26], the HR user set $\mathcal{U}_h$ is formed. Correspondingly, beamforming vectors $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K\}$ are also generated based on (13).

3: Initialize the LR user set as $\mathcal{U}_l = \varnothing$ and let $k = 1$.

4: For the kth cluster,

$$\mathcal{U}_l(k) = \arg \max_{m \in \mathcal{U} - \mathcal{U}_h - \mathcal{U}_l} |\hat{\mathbf{h}}_{k1} \hat{\mathbf{h}}_m^H|$$

$$\mathcal{U}_l \leftarrow \mathcal{U}_l \cup \{\mathcal{U}_l(k)\}$$

$$k \leftarrow k + 1$$

5: If $k \leq K$, then go to 4. Otherwise, the algorithm is finished.

the downlink NOMA multiuser system with limited CSI feedback.

3.2.1 CDI feedback

The BS and all candidate users share a common codebook $\mathcal{I}$ which is used to quantize the direction of user channel vectors. The codebook $\mathcal{I}$ consists of L subcodebooks, i.e., $\mathcal{I} = \bigcup_{l=1}^{L} \mathcal{I}_l$, and each subcodebook is comprised of K unit norm mutually orthogonal vectors with size $1 \times K$, i.e., $\mathcal{I}_l = \{\mathbf{v}_{l,1}, \mathbf{v}_{l,2}, \ldots, \mathbf{v}_{l,K} | \mathbf{v}_{l,k} \mathbf{v}_{l,k}^H = 1, \forall k \neq j, \mathbf{v}_{l,k} \mathbf{v}_{l,j}^H = 0\}$. Moreover, the K vectors in the subcodebook $\mathcal{I}_l$ are independently selected from the uniform distribution on the complex unit sphere [27]. To reduce the quantization error, L should be set as large as possible. Of course, the lager L is, the more memory space is needed for both users and BS.

For the mth user, quantized CDI $\hat{\mathbf{h}}_m$ is selected based on the following guidelines:

$$\hat{\mathbf{h}}_m = \arg \max_{\mathbf{v}_{l,k} \in \mathcal{I}} |\tilde{\mathbf{h}}_m \mathbf{v}_{l,k}^H| \tag{24}$$

Accordingly, the mth user just feeds back the CDI, two-tuple (l, k), to the BS. To be fair, each user feeds back $B = \log N$ bits CQI like the one with the ZFBF scheme. Thus, there is $N = L \times K$. In the RBF scheme, the BS chooses one of the L subcodebooks to perform beamforming. For example, if the lth subcodebook $\mathcal{I}_l$ is selected, then the kth beamforming vector is $\tilde{\mathbf{w}}_k = \mathbf{v}_{l,k}^H$ and the users quantized into $\mathcal{I}_l$ are treated as candidate users. Assume M is large enough, K HR users can be selected from these candidate users. How to select both these HR users and the subcodebook are illustrated in the next subsection. Therefore, we have

$$\hat{\mathbf{h}}_{k1} \mathbf{w}_j = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases} \tag{25}$$

3.2.2 CQI feedback

Due to (25), taking into consideration the same requirement $\eta_k \approx 0$, we can also get the same expressions of (18) and (21), respectively. It means that NOMA-RBF has the same CQI $f(\mathbf{h}_m)$ as NOMA-ZFBF.

3.2.3 User selection algorithm

After obtaining CDI and CQI from the users, the BS first needs to determine K HR users and K beamforming vectors. Then, K LR users can be scheduled according to the K beamforming vectors. Since there may be more than one users quantized into a vector in one subcodebook, we have to select the candidate HR users for a subcodebook. For example, assume there are more than one user whose CDI information meets $\hat{\mathbf{h}}_m = \mathbf{v}_{l,k}$ for the subcodebook $\mathcal{I}_l$. By (24), it means we should select the user with maximum $\cos\theta_m = |\tilde{\mathbf{h}}_m \mathbf{v}_{l,k}^H|$ as the HR user for the vector $\mathbf{v}_{l,k}$. As the BS knows nothing about θ_m, we can use the CQI $f(\mathbf{h}_m)$ fed back by the mth user instead. Define $\mathcal{U}_{l,k}$ as the candidate HR user for $\mathbf{v}_{l,k}$. Then, there is

$$\begin{aligned} \mathcal{U}_{l,k} &= \arg \max_{m \in \left\{j | \hat{\mathbf{h}}_j = \mathbf{v}_{l,k}\right\}} f(\mathbf{h}_m) \\ &= \arg \max_{m \in \left\{j | \hat{\mathbf{h}}_j = \mathbf{v}_{l,k}\right\}} \frac{\frac{P}{K} \|\mathbf{h}_m\|^2 \cos^2 \theta_m}{1 + \frac{P}{K} \|\mathbf{h}_m\|^2 \sin^2 \theta_m} \end{aligned} \tag{26}$$

Then, the optimal subcodebook is given by

$$l^* = \arg \max_{1 \leq l \leq L} \sum_{m \in \{\mathcal{U}_{l,k}, k = 1, 2, \ldots, K\}} f(\mathbf{h}_m) \tag{27}$$

As a result, $\mathbf{v}_{l^*,k}^H \in \mathcal{I}_{l^*}, k = 1, 2, \ldots, K$ are the RBF vectors. Meanwhile, the users with index $\mathcal{U}_{l^*,k}, k = 1, 2, \ldots, K$ are the HR user in the kth beam, i.e., $\mathcal{U}_h(k) = \mathcal{U}_{l^*,k}$. For the LR user, as the angle η_k between $\hat{\mathbf{h}}_{k2}$ and $\hat{\mathbf{h}}_{k1}$ should meet $\eta_k \approx 0$, the LR user in the kth cluster is

$$\mathcal{U}_l(k) = \arg \max_{m \in \mathcal{U} - \mathcal{U}_h} |\hat{\mathbf{h}}_m \mathbf{v}_{l^*,k}^H| \tag{28}$$

In summary, we give the user selection algorithm in Algorithm 2.

4 Power allocation for NOMA users

Since α_k affects the data rate greatly, we need to carefully choose a suitable power allocation ratio to maximize the system rate. Note that, if the NOMA multiuser system has a worse performance than the traditional OMA multiuser

Algorithm 2 User selection algorithm for the NOMA-RBF system

1: The BS receives all CDI $\{\hat{\mathbf{h}}_m, m = 1,2,\ldots,M\}$ and CQI $\{f(\mathbf{h}_m), m = 1,2,\ldots,M\}$ fed back by all candidate users in $\mathcal{U}$ via a limited feedback channel.

2: By (26), select a HR user for each vector $\mathbf{v}_{l,k} \in \mathcal{I}$.

3: Use (27), seek the optimal subcodebook $\mathcal{I}_{l^*}$. Then, the beamforming vectors are $\{\mathbf{v}_{l^*,k}^H, k = 1,2,\ldots,K\}$ and the HR user in the kth beam is $\mathcal{U}_{l^*,k}$. $\mathcal{U}_h = \{\mathcal{U}_{l^*,k}, k = 1,2,\ldots,K\}$.

4: Initialize the LR user set as $\mathcal{U}_l = \varnothing$ and let $k = 1$.

5: For kth cluster,

$$\mathcal{U}_l(k) = \arg \max_{m \in \{\mathcal{U} - \mathcal{U}_h - \mathcal{U}_l\}} |\hat{\mathbf{h}}_m \mathbf{v}_{l^*,k}^H|$$
$$\mathcal{U}_l \leftarrow \mathcal{U}_l \cup \{\mathcal{U}_l(k)\}$$
$$k \leftarrow k + 1$$

6: If $k \le M$, then go to 2. Otherwise, the user algorithm is finished.

system, there is no need to perform NOMA which incurs some extra decoding complexity. As a baseline of our system, we herein consider a traditional downlink multiuser system, where during the same resource block as the NOMA system, each beam supports two users in a time division multiple access (TDMA) way, denoted as TDMA-BF. In other words, the transmit duration is divided into two equal slots for two users. For the kth beam, denote the sum-rate of TDMA-BF system as $R_{k,\mathrm{TDMA}}$. Then, the system sum-rate of the kth cluster in the proposed downlink NOMA multiuser system is

$$\max\{R_{k,\mathrm{TDMA}}, R_{k1} + R_{k2}\}. \tag{29}$$

For the given CSI, $R_{k,\mathrm{TDMA}}$ is determined. Next, we just need to optimize $R_{k1} + R_{k2}$ through adjusting the power allocation ratio α_k.

Recalling (14) and (25), even though the beamforming vectors are dependent on the quantized CDI from the users, the inter-beam interference is very slight in the scenario with massive candidate users. It means the relativity between beams can be insignificant. Therefore, it is dispensable to seek for the optimal power allocation to maximize the object (29). Instead, we can maximize the sum-rate of a user cluster individually. Moreover, to demonstrate the different QoS requirements of the HR user and LR user, we set the minimum user rate constraints $R_{k1} \ge R_h > 0$ and $R_{k2} \ge R_l > 0$. That is to say, the NOMA downlink multiuser system provides two transmission rate services in a cluster. For the kth cluster, the optimization problem can be expressed as

$$\begin{aligned} \max_{0<\alpha_k<1} \ & \{R_{k1} + R_{k2}\} \\ \text{s.t. } & R_{k1} \ge R_h > 0 \\ & R_{k2} \ge R_l > 0 \end{aligned} \tag{30}$$

As the BS only can use CDI and CQI to calculate g_{k1} and g_{k2}, in the following, we denote $\hat{g}_{k1}$ and $\hat{g}_{k2}$ as the obtained g_{k1} and g_{k2}, respectively. According to the relationship between $\hat{g}_{k1}$ and $\hat{g}_{k2}$, we consider two cases to solve (30).

4.1 Case 1: $\hat{g}_{k1} \ge \hat{g}_{k2}$

If $\hat{g}_{k1} \ge \hat{g}_{k2}$, substituting (3) and (5) with (30), the optimal problem becomes

$$\max_{\alpha_k} \left\{ \log(1 + \hat{g}_{k1}\alpha_k) + \log\left(1 + \frac{(1-\alpha_k)}{\alpha_k + \frac{1}{\hat{g}_{k2}}}\right) \right\} \tag{31}$$

$$\text{s.t. } \frac{2^{R_h} - 1}{\hat{g}_{k1}} \le \alpha_k \le \frac{1 - \frac{2^{R_l}-1}{\hat{g}_{k2}}}{2^{R_l}} \tag{32}$$

To ensure that there exists feasible α_k for (31), there is

$$\frac{1 - \frac{2^{R_l}}{\hat{g}_{k2}}}{2^{R_l}} \ge \frac{2^{R_h} - 1}{\hat{g}_{k1}} \tag{33}$$

Then, we have

$$R_l \le \log\left(\frac{1}{\frac{2^{R_h}-1}{\hat{g}_{k1}} + \frac{1}{\hat{g}_{k2}}}\right) \tag{34}$$

As $R_l > 0$, due to (34), R_h has a constraint

$$R_h < \log\left(1 + \hat{g}_{k1}\left(1 - \frac{1}{\hat{g}_{k2}}\right)\right) \tag{35}$$

If the given R_h and R_l make (34) and (35) hold, there exists feasible α_k for (31). Next, we assume that both (34) and (35) hold.

To optimize the objective function in (31) is equivalent to maximize

$$F_1(\alpha_k) = \left(1 + \hat{g}_{k1}\alpha_k\right)\left(1 + \frac{(1-\alpha_k)}{\alpha_k + \frac{1}{\hat{g}_{k2}}}\right) \tag{36}$$

Thus, the first-order derivative of $F_1(\alpha_k)$ with respect to α_k is

$$F_1'(\alpha_k) = \frac{\left(1 + g_{k2}\right)\left(g_{k1} - g_{k2}\right)}{\left(1 + g_{k2}\alpha_k\right)^2} \tag{37}$$

As $\hat{g}_{k1} \ge \hat{g}_{k2}$ holds in Case 1, there is $F_1'(\alpha_k) \ge 0$. It means $F_1(\alpha_k)$ is a monotonic increasing function of α_k. Hence, the optimal solution is

$$\alpha_k^* = \frac{1 - \frac{2^{R_l}-1}{\hat{g}_{k2}}}{2^{R_l}} \tag{38}$$

It is easy to confirm that $0 < \alpha_k^* < 1$. So, α_k^* is the expected solution of problem (31) under the condition of (34) and (35).

4.2 Case 2: $\hat{g}_{k1} < \hat{g}_{k2}$

In this case, due to (7) and (9), similarly, we consider the following optimization problem

$$\max_{\alpha_k} \left\{ \log\left(\frac{\frac{1}{\hat{g}_{k1}} + 1}{\frac{1}{\hat{g}_{k2}}} \right) + \log\left(1 + \frac{\frac{1}{\hat{g}_{k1}} - \frac{1}{\hat{g}_{k2}}}{\alpha_k - \left(\frac{1}{\hat{g}_{k1}} + 1 \right)} \right) \right\} \tag{39}$$

$$\text{s.t.} \frac{\left(1 + \frac{1}{\hat{g}_{k1}}\right)\left(2^{R_h} - 1\right)}{2^{R_h}} \le \alpha_k \le 1 - \frac{2^{R_l} - 1}{\hat{g}_{k2}} \tag{40}$$

Considering the condition

$$1 - \frac{2^{R_l} - 1}{\hat{g}_{k2}} \ge \frac{\left(1 + \frac{1}{\hat{g}_{k1}}\right)\left(2^{R_h} - 1\right)}{2^{R_h}} \tag{41}$$

we have

$$R_l \le \log\left(\frac{1 - \frac{2^{R_h} - 1}{\hat{g}_{k1}}}{2^{R_h}} \hat{g}_{k2} + 1 \right) \tag{42}$$

From (42), there is

$$R_h < \log(\hat{g}_{k1} + 1) \tag{43}$$

Assume R_h and R_l meet (42) and (43), we define the objective function as

$$F_2(\alpha_k) = \log\left(\frac{\frac{1}{\hat{g}_{k1}} + 1}{\frac{1}{\hat{g}_{k2}}} \right) + \log\left(1 + \frac{\frac{1}{\hat{g}_{k1}} - \frac{1}{\hat{g}_{k2}}}{\alpha_k - \left(\frac{1}{\hat{g}_{k1}} + 1 \right)} \right) \tag{44}$$

It is easy to see that $F_2(\alpha_k)$ is a monotonic decreasing function of α_k because of $\hat{g}_{k1} < \hat{g}_{k2}$ in Case 2. Therefore, the optimal solution is

$$\alpha_k^* = \frac{\left(1 + \frac{1}{\hat{g}_{k1}}\right)\left(2^{R_h} - 1\right)}{2^{R_h}} \tag{45}$$

Since (43) holds, there is $0 < \alpha_k^* < 1$. Therefore, α_k^* is the expected solution of problem (39) under the condition of (42) and (43).

4.3 Discussion

In the above analysis, we assume R_h and R_l make (34) and (35) hold in Case 1 or (42) and (43) in Case 2. If both conditions in a case cannot hold, it means the NOMA system cannot offer sufficient rates to meet the minimum rate constraints R_h and R_l. There could be two possible methods to deal with this situation:

1. **Adjusting R_h and R_l:** Since $\hat{g}_{k1}$ and $\hat{g}_{k2}$ are dependent on instantaneous CSI from the BS to these users, we can dynamically adjust R_h and R_l to meet (34) and (35) or (42) and (43) as much as possible. For example, the BS can estimate the possible values of $\hat{g}_{k1}$ and $\hat{g}_{k2}$ via some training signals. Then, the BS could set the R_h and R_l which can make the rate conditions hold with a very large probability.
2. **Employing traditional OMA system:** If the NOMA system cannot offer the desired rate, the most direct method is to employ traditional multiuser access schemes with relatively lower complexity. Note that given a total power P, these traditional OMA schemes, such as TDMA, could not provide the required minimum rates either. In this case, the minimum rate constraints have to be ignored.

5 Simulation results

In this section, we show the performances of the proposed NOMA-ZFBF and NOMA-RBF systems. The number of BS antennas is $K = 2$. Channel and noise parameters are described in the system model. We set $R_l = 1$ bps/Hz and $R_h = 3$ bps/Hz. In the following simulations, if the NOMA system cannot offer the minimum required rates, we employ the traditional TDMA scheme to transmit information for HR and LR users.

5.1 NOMA-ZFBF versus NOMA-RBF

Figure 2 shows the user rates of the NOMA-ZFBF and NOMA-RBF systems versus the total transmit power. From Fig. 2a, we can see that NOMA systems with perfect CSI at the transmitter (CSIT) achieve larger system sum-rates than NOMA systems with limited CSI feed back. Thereinto, NOMA-ZFBF system with perfect CSIT has the best performances. That is to say, the performances of NOMA system with perfect CSIT are the upper bound of NOMA system with limited CSI feedback. For the same N in limited CSI feedback, NOMA-RBF outperforms NOMA-ZFBF. It means RBF is more suitable for NOMA downlink multiuser transmission with limited CSI feedback, which is different from the results with perfect CSIT. In Fig. 2b, as P increases, the average rates of the LR user in all systems increase slightly. Meanwhile, observing Fig. 2c, the average rates of the HR user become larger and larger as P increases. The reason is that the HR user is always selected from the candidate users in priority. Thus, the HR users could gain the most of the advantages of increasing transmit power. For a given beamforming scheme, more bits fed back incurs more rates. For example, the NOMA-RBF system with $N = 64$ has about 3-bps sum-rate gain than the NOMA-RBF system with $N = 16$ at the situation $P = 35$ dB.

In addition, we also show the user rates of the NOMA-ZFBF and NOMA-RBF systems versus the candidate user number in Fig. 3. We also can see that NOMA-ZFBF with perfect CSIT achieves the largest sum-rate among

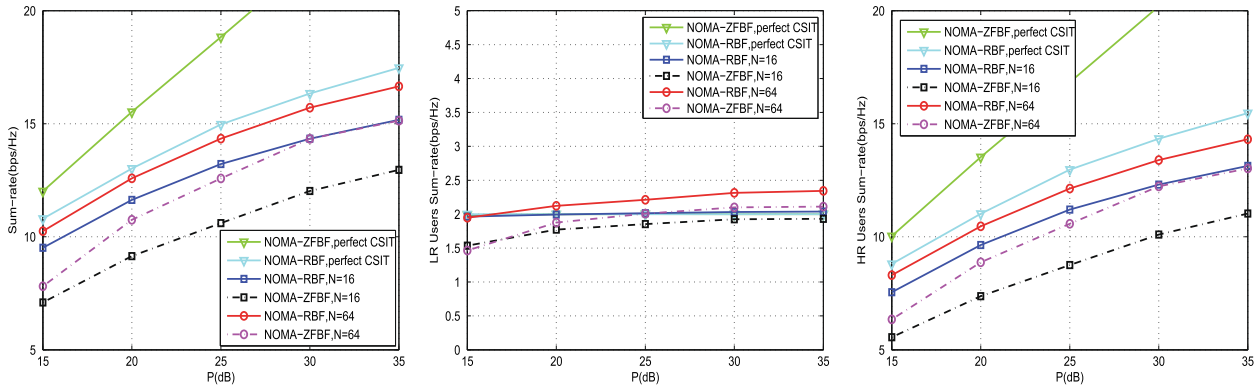

Fig. 2 The sum-rates versus total transmit power $P, M = 50$. **a** The average system sum-rate. **b** The average sum-rate of LR users. **c** The average sum-rate of HR users

all NOMA systems. Moreover, NOMA-RBF is superior to NOMA-ZFBF with limited CSI feedback. Through Fig. 3, as the number of candidate users M increases, all average user rates increase, especially for HR users. This is because more candidate users could provide more opportunities to select a user with little quantization error. By Figs. 2 and 3, we can conclude that RBF has better performances than ZFBF in the NOMA downlink multiuser system with limited CSI feedback and the feedback link with higher capacity improves the rates of NOMA users.

5.2 NOMA-RBF versus TDMA-RBF

In Fig. 4, we compare the sum-rate performances of NOMA-RBF and TDMA-RBF. Obviously, NOMA-RBF with perfect CSIT has the best performance. In NOMA-RBF system with limited CSI feedback, more feedback bits incur larger average sum-rates. For the same N, NOMA-RBF system outperforms TDMA-RBF system. For example, with $N = 64$ and $M = 100$, NOMA-RBF has about 4-bps sum-rate gain than TDMA-RBF system. When $N = 16$ and $M = 100$, the rate gain between NOMA-RBF and TDMA-RBF becomes larger. Therefore, it is shown that NOMA system still gains rate advantage than the traditional TDMA system with limited CSI feedback.

5.3 Optimal power allocation

We herein consider two power allocation schemes to be compared with the proposed optimal power allocation scheme. Define

$$\tilde{\alpha}_k = \begin{cases} \frac{2^{R_h}-1}{\hat{g}_{k1}} + \frac{1-\frac{2^{R_l}-1}{\hat{g}_{k2}}}{2^{R_l}}, & \text{If } \hat{g}_{k1} \geq \hat{g}_{k2} \text{ and (34), (35) hold} \\ \frac{\left(1+\frac{1}{\hat{g}_{k1}}\right)\left(2^{R_h}-1\right)}{2^{R_h}} + 1 - \frac{2^{R_l}-1}{\hat{g}_{k2}}, & \text{If } \hat{g}_{k1} < \hat{g}_{k2} \text{ and (42), (43) hold} \end{cases} \quad (46)$$

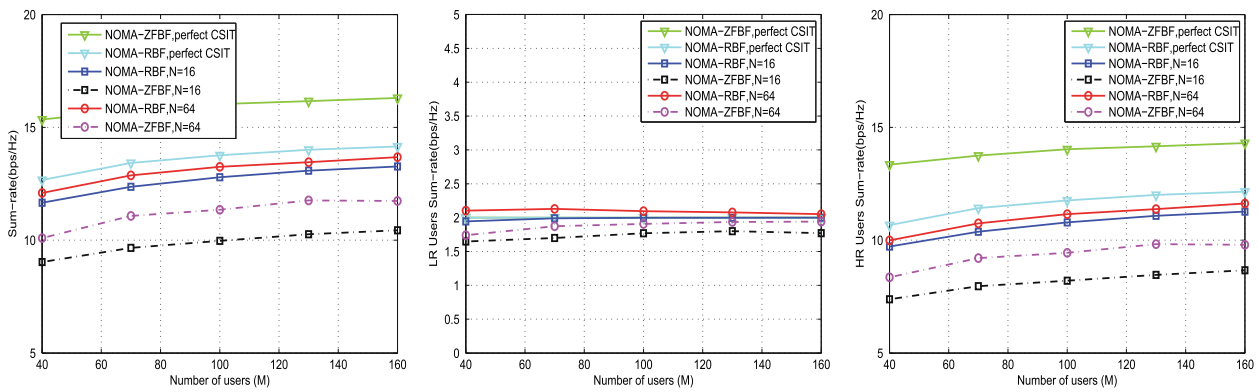

Fig. 3 The sum-rates versus number of candidate users M, $P = 20$ dB. **a** The average system sum-rate. **b** The average sum-rate of LR users. **c** The average sum-rate of HR users

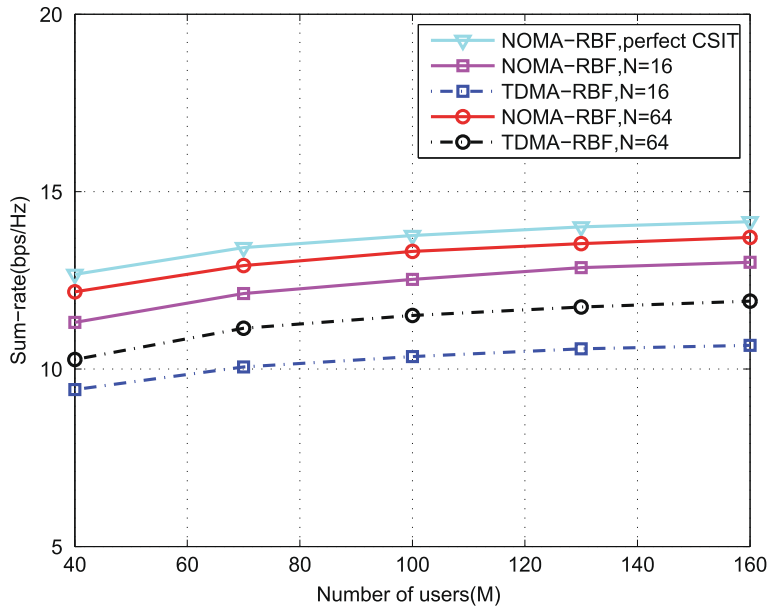

Fig. 4 Average system sum-rate of NOMA-RBF and TDMA-RBF, $P = 20$ dB

which is the sum of maximum and minimum values in the feasible region of α_k in both cases. Then, we have four power allocation schemes: $\alpha_k^{\mathrm{I}} = \frac{1}{2}\tilde{\alpha}_k$, $\alpha_k^{\mathrm{II}} = \frac{1}{3}\tilde{\alpha}_k$, $\alpha_k^{\mathrm{III}} = \frac{1}{4}\tilde{\alpha}_k$, and $\alpha_k^{\mathrm{IV}} = \frac{1}{10}\tilde{\alpha}_k$. In Fig. 5, as P increases, the average sum-rates of all five power allocation schemes increase. The proposed optimal power allocation scheme achieves the best sum-rate among all presented schemes.

6 Conclusions

In this paper, we investigate NOMA downlink multiuser beamforming system with limited CSI feedback. We introduce the zero-forcing beamforming and random beamforming into NOMA downlink multiuser system and give the feedback forms of channel direction information and channel quality indicator, which are transmitted by the candidate users via a limited rate channel. In addition, we

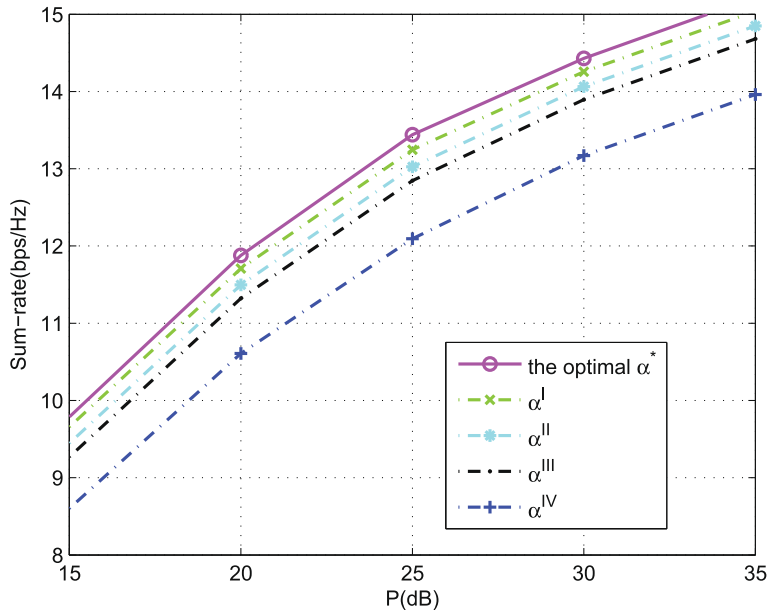

Fig. 5 Average sum-rate of NOMA-RBF with power allocation, $M = 50$, $N = 16$

also propose the user selection algorithms to determine the users sharing a beam in one user cluster. To improve the performances of NOMA system further, we also provide an optimal power allocation scheme. Finally, simulation results show that NOMA system still can provide more user rates than the traditional TDMA system and the RBF is more suitable for NOMA downlink multiuser system with limited CSI feedback.

Endnote

[1] We focus our attention on the overall system performance of NOMA system with limited CSI feedback. Different from [19–21], fairness issue is not considered in this work, which can be an important direction in future work.

Acknowledgements
The authors want to thank the editor and anonymous reviewers for the suggestions and comments to improve the quality of this paper.
This work was supported by the Research Fund of National Mobile Communications Research Laboratory, Southeast University (No.2011D14), the National Natural Science Foundation of China (No.61102082), the Natural Science Basic Research Plan in Shaanxi Province of China under Program No.2015JQ6234, and the Fundamental Research Funds for the Central Universities.

Competing interests
The authors declare that they have no competing interests.

Received: 14 March 2016 Accepted: 19 September 2016
Published online: 04 October 2016

References

1. C Wang, F Haider, X Gao, *et al*, Cellular architecture and key technologies for 5G wireless communication networks. IEEE Commun. Mag. **52**(2), 122–130 (2014)
2. Q Li, H Niu, A Papathanassiou, *et al*, 5G network capacity: key elements and technologies. IEEE Vehic. Technol. Mag. **9**(1), 71–78 (2014)
3. L Dai, B Wang, Y Yuan, S Han, *et al*, Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends. IEEE Commun. Mag. **53**(9), 74–81 (2015)
4. Y Saito, A Benjebbour, Y Kishiyama, *et al*, in *Int. Conf. Personal, Indoor, and Mobile Radio Communications (PIMRC)*. System-level performance evaluation of downlink non-orthogonal multiple access (NOMA), (London, 2013), pp. 611–615
5. Y Saito, Y Kishiyama, A Benjebbour, et al., in *Int. Conf. Vehicular Technology Conference (VTC Spring)*. Non-orthogonal multiple access (NOMA) for cellular future radio access, (Dresden, German, 2013)
6. Z Ding, Z Yang, P Fan, et al., On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users. IEEE Signal Process. Lett. **21**(12), 1501–1505 (2014)
7. Z Ding, M Peng, et al., Cooperative non-orthogonal multiple access in 5G systems. IEEE Commun. Lett. **19**(8), 1462–1465 (2015)
8. J Men, J Ge, Performance analysis of non-orthogonal multiple access in downlink cooperative network. IET Commun. **9**(18), 2267–2273 (2015)
9. M Hojeij, J Farah, C Nour, *et al*, New optimal and suboptimal resource allocation techniques for downlink non-orthogonal multiple access. Wireless Pers. Commun. **87**(3), 837–867 (2016)
10. J Kim, I Lee, Non-orthogonal multiple access in coordinated direct and relay transmission. IEEE Commun. Lett. **19**(11), 2037–2040 (2015)
11. Z Ding, P Fan, V Poor, Impact of user pairing on 5g non-orthogonal multiple access downlink transmissions. IEEE Trans. Vehic. Technol. **65**(8), 6010–6023 (2015)
12. J Choi, Non-orthogonal multiple access in downlink coordinated two-point systems. IEEE Commun. Lett. **18**(2), 313–316 (2014)
13. Q Sun, S Han, I Chin-Lin, *et al*, On the ergodic capacity of MIMO NOMA systems. IEEE Wireless Commun. Lett. **4**(4), 405–408 (2015)
14. Q Sun, S Han, Z Xu, S Wang, et al., in *IEEE Wireless Communications and Networking Conference (WCNC)*. Sum rate optimization for MIMO non-orthogonal multiple access systems, (New Orleans, USA, 2015), pp. 747–752
15. X Chen, A Beiijebbour, A Li, et al., in *Int. Conf. on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. Consideration on successive interference canceller (SIC) receiver at cell-edge users for non-orthogonal multiple access (NOMA) with SU-MIMO, (Hong Kong, 2015), pp. 522–526
16. J Kim, J Koh, J Kang, *et al*, in *Int. Conf. on Military Communications*. Design of user clustering and precoding for downlink non-orthogonal multiple access (NOMA), (Florida, 2015), pp. 1170–1175
17. S Liu, C Zhang, G Lyu, in *IEEE International Conference on Communication Workshop Heterogeneous Converged Networks*. User selection and power schedule for downlink non-orthogonal multiple access (NOMA) system, (London, 2015), pp. 2561–2565
18. B Kim, W Chung, in *IEEE VTC 2015-Spring*. Uplink NOMA with multi-antenna, (Scotland, UK, 2015), pp. 1–5
19. S Timotheou, I Krikidis, Fairness for non-orthogonal multiple access in 5G systems. IEEE Signal Process. Lett. **22**(10), 1647–1651 (2015)
20. Y Liu, Z Ding, M Elkashlan, et al., Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer. IEEE J. Selected Areas Commun. **34**(4), 938–953 (2016)
21. Y Liu, M Elkashlan, Z Ding, GK Karagiannidis, Fairness of user clustering in MIMO non-orthogonal multiple access systems. IEEE Commun. Lett. **20**(7), 1465–1468 (2016)
22. A Tukmanov, S Boussakta, Z Ding, *et al*, Outage performance analysis of imperfect-CSI-based selection cooperation in random networks. IEEE Trans. Commun. **62**(8), 2747–2757 (2014)
23. Z Yang, Z Ding, P Fan, et al., On the performance of non-orthogonal multiple access systems with partial channel information. IEEE Trans. Commun. **64**(2), 654–667 (2015)
24. Z Ding, H Poor, Design of massive-MIMO-NOMA with limited feedback. IEEE Signal Process. Lett. **23**(5), 629–633 (2016)
25. D Love, R Heath, V Lau, et al., An overview of limited feedback in wireless communication systems. IEEE J. Sel. Areas Commun. **26**(8), 1341–1365 (2008)
26. T Yoo, N Jindal, A Goldsmith, Multi-antenna downlink channels with limited feedback and user selection. IEEE J.Selected Areas Commun. **25**(7), 1478–1491 (2007)
27. K Huang, R Heath, Performance of orthogonal beamforming for SDMA with limited feedback. IEEE Trans. Vehic. Technol. **58**(1), 152–164 (2009)