Optimal Feedback Allocation for Zeroforcing Beamforming Transmission in Downlink NOMA

Kritsada Mamat
Department of Electronic Engineering Technology
College of Industrial Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand 10800
Email: kritsada.m@cit.kmutnb.ac.th

Wiroonsak Santipach
Department of Electrical Engineering
Faculty of Engineering, Kasetsart University
Bangkok, Thailand 10900
Email: wiroonsak.s@ku.ac.th

***Abstract*—We consider non-orthogonal multiple access (NOMA) downlink channels with zeroforcing beamforming transmission. With limited feedback rate, channel direction information (CDI) needs to be quantized and fed back from mobile users to a base station. To increase spectral efficiency, two active users in some clusters share the same beamforming vector and thus, will interfere fully with each other. We derive the approximate achievable rates and outage probability for users in 1-user and 2-user clusters. With the approximate outage probability derived as a function of feedback rate, we find the optimal feedback allocation for both types of cluster that minimizes the maximum outage probability for all users. Numerical results show that significant gain over uniform feedback allocation is achieved in moderate-to-large SNR regimes.**

I. Introduction

Non-orthogonal multiple access (NOMA) has been shown to achieve higher spectral efficiency than orthogonal multiple access (OMA) does [1], [2]. For downlink NOMA, a base station applies superposition coding to transmit users' signals in the same frequency, time, or spatial domain. Thus, multiple users can simultaneously be served by a base station in a single orthogonal channel. However, to decode its own signal successfully, a receiver may have to apply successive interference cancellation (SIC) in conjunction with optimized transmit-power allocation from the base station.

To eliminate interference among users, a base station with multiple transmit antennas may employ zeroforcing beamforming, which requires current channel state information (CSI) [3]. For frequency-division duplex (FDD), base stations obtain quantized CSI via low-rate feedback channels from mobile users. In [4], the authors derive outage probability with quantized CSI as a function of feedback rate. However, all users are assumed to be allocated equal feedback rate for quantizing CSI. In [5], zeroforcing-beamforming transmission with quantized CSI is studied, but only numerical results on the performance are shown with no performance analysis. In [6], 1-bit feedback scheme is proposed for massive MIMO (multiple-input multiple-output) NOMA, which is shown to decompose into multiple SISO (single-input single-output) NOMA's. Transmit power and beams are optimized for 2-user NOMA in millimeter-wave communications [7]. However, a large body of existing work on NOMA with multiple antennas including [7] assumes perfect CSI at the transmitter (CSIT), which is not practical [2], [8, and references therein].

In this work, we assume that some users receive signals coming from similar directions or paths. Thus, their channel direction information (CDI) are approximately the same. Those users are considered to be in the same cluster. Each cluster quantizes and sends back CDI to the base station via feedback channels. We assume that a cluster has either one active user or two active users. In practice, each cluster may consist of more than 2 users. However, to keep receiver's complexity low and sum rate high, some scheduling scheme is used to select either one user or two users in each cluster. It has been shown in [9] that there is a trade-off between sum rate of each cluster and the number of active users in that cluster.

To transmit to different clusters, we assume that the base station applies zeroforcing beamforming. When CDI is perfect or feedback rate is infinite, clusters are orthogonal in spatial domain. Assuming limited feedback, our contribution in this work is as follows.

- With given total-feedback rate, we derive the approximate achievable rate for each user that can be used to find the feedback allocation for each cluster that maximizes the minimum user rate. The approximation derived is shown to be close to corresponding simulation results.
- We also formulate another problem that minimizes the maximum outage probability of active user. The optimal feedback allocation for 1-user cluster is approximated and is shown to increase linearly with total feedback rate when total feedback rate is small.

II. System Model

We consider a discrete-time downlink channel in which a base station with N_t antennas, employs zeroforcing beamforming to transmit data to M clusters of single-antenna users. For tractability of subsequent analysis, we assume that $M = N_t$.

K. Mamat was supported by College of Industrial Technology, King Mongkut's University of Technology North Bangkok while W. Santipach was supported by Kasetsart University Research and Development Institute (KURDI) under FY2018 Kasetsart University research grant.

There are M_1 clusters with single active user and M_2 clusters with 2 active users, and thus, $M_1 + M_2 = M$.

We assume that a channel between each transmit antenna to a receive antenna is Rayleigh flat-fading. Let $\boldsymbol{h}_l$ denote an $N_t \times 1$ channel vector for a single-user cluster where $1 \leq l \leq M_1$, whose entry is the channel gain from each transmit antenna and is independent and complex Gaussian distributed with zero mean and unit variance. Let $\boldsymbol{h}_k$ denote an $N_t \times 1$ channel vector for the user with stronger channel in a 2-user cluster, where $M_1 + 1 \leq k \leq M$. A channel vector for the second user whose channel is weaker in a 2-user cluster, is given by $c_k \boldsymbol{h}_k$ where c_k is a degradation factor and $0 < c_k < 1$. This follows the assumption by [1], [7], [10] that both users' spatial channels are sufficiently correlated that CDI of both users are approximately the same.

To compute zeroforcing beamforming vectors and allocate proper transmit power, the base station requires the current CSI from all clusters. We assume that every user feeds back its channel quality information (CQI) referring to $\|\boldsymbol{h}_l\|$, $\|\boldsymbol{h}_k\|$, and c_k, $\forall l, k$, to the base station *perfectly*. For CDI, each *cluster* quantizes and feeds back its normalized channel vector denoted by $\bar{\boldsymbol{h}}_l \triangleq \boldsymbol{h}_l / \|\boldsymbol{h}_l\|$ and $\bar{\boldsymbol{h}}_k \triangleq \boldsymbol{h}_k / \|\boldsymbol{h}_k\|$ with B bits. We assume channels are independent block fading and the block length is sufficiently long that feeding back CDI is meaningful.

For CDI quantization, a random vector quantization (RVQ) codebook is used. RVQ codebook was shown to perform generally well and can be analyzed to obtain insights on the rate performance with limited feedback [11]. With B bits, an RVQ codebook consists of 2^B independent isotropically distributed $N_t \times 1$ vectors and is denoted by $\mathcal{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{2^B}\}$. Thus, cluster m selects

$$\hat{\boldsymbol{h}}_m = \arg\max_{\boldsymbol{v} \in \mathcal{V}} |\bar{\boldsymbol{h}}_m^\dagger \boldsymbol{v}|^2, \quad 1 \leq m \leq M, \tag{1}$$

and feeds the associated codebook index back to the base station. The base station forms the $N_t \times M$ matrix $\hat{\boldsymbol{H}}$ whose columns are quantized channel vectors $\hat{\boldsymbol{h}}_m$. Zeroforcing beamforming vector for cluster m is denoted by $\boldsymbol{w}_m$, which is the mth column of the $N_t \times M$ matrix given by $\boldsymbol{W} = \hat{\boldsymbol{H}}^\dagger (\hat{\boldsymbol{H}} \hat{\boldsymbol{H}}^\dagger)^{-1}$.

We assume that the average transmit power per user is $P/2$. Thus, clusters with 2 active users will be allocated the combined power of P. The two active users in the same cluster completely interfere with each other in time, frequency, and spatial domains, but will be distinguished in power domain. In 2-user cluster k, we refer to the user with stronger channel $\boldsymbol{h}_k$ as user 1 and the other user with weaker channel $c_k \boldsymbol{h}_k$ as user 2. To maintain performance of user 2, the base station allocates power to the two users inversely proportional to their channel power. Hence, transmit powers for the two users are given by

$$P_{k,1} = \frac{c_k^2}{1 + c_k^2} P \text{ and } P_{k,2} = \frac{1}{1 + c_k^2} P \tag{2}$$

where $P_{k,1} + P_{k,2} = P$, $\forall k$. We note that this power allocation policy may not maximize sum achievable rate; however, it equalizes the received powers of both users in the same cluster.

The base station transmits message signal given by

$$x = \sqrt{\frac{P}{2}} \sum_{l=1}^{M_1} \boldsymbol{w}_l s_l + \sum_{k=M_1+1}^{M} \boldsymbol{w}_k (\sqrt{P_{k,1}} s_{k,1} + \sqrt{P_{k,2}} s_{k,2}) \tag{3}$$

where s is a transmitted symbol with zero mean and unit variance. Since CDI is quantized and thus, not perfect, there is residual interference between different clusters. We can derive instantaneous signal-to-interference plus noise ratio (SINR) of the user in single-user cluster l as follows

$$\gamma_l = \frac{\frac{1}{2} |\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_l|^2}{\frac{1}{2} \sum_{\substack{m=1 \\ m \neq l}}^{M_1} |\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_m|^2 + \sum_{m=M_1+1}^{M} |\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_m|^2 + \frac{\sigma_n^2}{P \|\boldsymbol{h}_l\|^2}} \tag{4}$$

where σ_n^2 is the variance of additive white Gaussian noise at all receivers.

For clusters with 2 active users, SIC is deployed. For the receiver of user 1 (with stronger channel), the signal of the weaker user is decoded first, and then that signal will be reconstructed and subtracted from the received signal. Thus, the signal of user 1 can be decoded without interference from user 2. SINR for user 1 is given by

$$\gamma_{k,1} = \frac{\frac{c_k^2}{1 + c_k^2} |\bar{\boldsymbol{h}}_k^\dagger \boldsymbol{w}_k|^2}{\underbrace{\frac{1}{2} \sum_{m=1}^{M_1} |\bar{\boldsymbol{h}}_k^\dagger \boldsymbol{w}_m|^2 + \sum_{\substack{m=M_1+1 \\ m \neq k}}^{M} |\bar{\boldsymbol{h}}_k^\dagger \boldsymbol{w}_m|^2}_{I_k} + \frac{\sigma_n^2}{P \|\boldsymbol{h}_k\|^2}} \tag{5}$$

where $I_k P$ is the total power of inter-cluster interference. Note that there is no intra-cluster interference caused by the signal of user 2.

For the receiver of user 2 (with weaker channel), the signal of user 2 is decoded directly by treating all interfering signals as noise. Thus, SINR of user 2 is given by

$$\gamma_{k,2} = \frac{\frac{c_k^2}{1 + c_k^2} |\bar{\boldsymbol{h}}_k^\dagger \boldsymbol{w}_k|^2}{c_k^2 I_k + \frac{c_k^4}{1 + c_k^2} |\bar{\boldsymbol{h}}_k^\dagger \boldsymbol{w}_k|^2 + \frac{\sigma_n^2}{P \|\boldsymbol{h}_k\|^2}}. \tag{6}$$

We see in (6) that in addition to inter-cluster interference I_k, user 2 also encounters interference from user 1's signal, which is the second term in the denominator. Also, due to channel-inversion power allocation, the received signal power of user 1 and 2 in a 2-user cluster are the same. The other benefit of the power allocation is the reduction of interference power by a factor of c_k^2. Comparing (5) and (6), we conclude that $\gamma_{k,1} \geq \gamma_{k,2}$ if

$$I_k \leq \frac{c_k^4}{1 - c_k^4} |\bar{\boldsymbol{h}}_k^\dagger \boldsymbol{w}_k|^2. \tag{7}$$

Thus, performance of the user with stronger channel is better than that of the other user when the feedback rate allocated for the cluster is not small or c_k is close to 1.

An achievable rate of each user is given by $R = E[\log_2(1 + \gamma)]$, which depends on the accuracy of CDI or feedback rate. With limited feedback, we would like to maximize the achievable rate over allocation of feedback for each cluster.

III. On Maximizing the Minimum Rate

To analyze the rate of the single user in a 1-user cluster, we first approximate distribution of SINR (4). Since $|\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_l|^2$ is beta distributed [3], its variance can be computed and is given by

$$\sigma^2_{|\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_l|^2} = \frac{1}{N_t^2}. \tag{8}$$

For $m \neq l$, since the probability density function (pdf) of $|\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_m|^2$ is given also by [3], we can compute the variance

$$\sigma^2_{|\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_m|^2} = \frac{4\sum_{i=0}^{2^B} \frac{\binom{2^B}{i}(-1)^i}{(N_t-1)i+2}}{Nt(Nt-1)} - \frac{2^{2B}\beta^2(2^B, \frac{N_t}{N_t-1})}{(N_t-1)^2} \tag{9}$$

where beta function $\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,\mathrm{d}t$. We note that for small to moderate N_t, variance of $|\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_l|^2$ is much larger than that of $|\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_m|^2$. Also, the noise term $\sigma_n^2/(P\|\boldsymbol{h}_l\|^2)$ is very small especially for large SNR. Thus, statistical variation of γ_l in (4) largely depends on $|\bar{\boldsymbol{h}}_l^\dagger \boldsymbol{w}_l|^2$. Hence, γ_l can be well approximated by a scaled beta random variable with the following cumulative distribution function (cdf)

$$F_{\gamma_l}(x) \approx \tilde{F}_{\gamma_l}(x) = 1-(1-\delta(M_1-1, M_2, 1, B)x)^{N_t-1} \tag{10}$$

where $0 \leq x \leq 1/\delta(M_1-1, M_2, 1, B)$ and

$$\delta(x,y,z,B) \triangleq \frac{z+1}{z}\left(\frac{(x+2y)2^B}{2(N_t-1)}\beta(2^B, \frac{N_t}{N_t-1}) + \frac{\sigma_n^2}{PN_t}\right). \tag{11}$$

We note that $1/\delta(M_1-1, M_2, 1, B)$ is an approximate upper bound of γ_l and that δ represents interference power in the SINR expression. As B increases, δ decreases and hence, SINR will increase.

Since the expression of SINR for user 1 in a 2-user cluster is similar to that of γ_l, we can also derive the approximate cdf for $\gamma_{k,1}$ as follows

$$\tilde{F}_{\gamma_{k,1}}(x) = 1-(1-\delta(M_1, M_2-1, c_k^2, B)x)^{N_t-1} \tag{12}$$

where $0 \leq x \leq 1/\delta(M_1, M_2-1, c_k^2, B)$. For user 2, $|\bar{\boldsymbol{h}}_k^\dagger \boldsymbol{w}_k|^2$ appears in both the numerator and the denominator in (6). Thus, the cdf of $\gamma_{k,2}$ is different from the previous cdf's, and can be approximated as follows

$$\tilde{F}_{\gamma_{k,2}}(x) = 1-\left(1-\frac{\delta(c_k^2 M_1, c_k^2(M_2-1), c_k^2, B)x}{1-c_k^2 x}\right)^{N_t-1} \tag{13}$$

where $0 \leq x \leq \frac{1}{c_k^2+\delta(c_k^2 M_1, c_k^2(M_2-1), c_k^2, B)}$.

With the approximate cdf's for SINR, we can derive the approximate ergodic achievable rates, which depend on the number of feedback bits B.

Proposition 1: For $N_t \geq 2$, achievable rates for the user in a 1-user cluster and both users in 2-user cluster are approximated by

$$R_l(B), R_{k,1}(B), R_{k,2}(B) \approx \frac{1}{\ln(2)}\int \frac{1-\tilde{F}(x)}{1+x}\,\mathrm{d}x \tag{14}$$

where $\tilde{F}(x)$ is either the approximate cdf $F_{\gamma_l}(x)$ in (10), $\tilde{F}_{\gamma_{k,1}}(x)$ in (12), or $\tilde{F}_{\gamma_{k,2}}(x)$ in (13) with the stated domains. Limits of each integral correspond to each domain of the approximate cdf.

The above integral can be evaluated by any numerical method. However, for $N_t = 2$, the rates can be expressed in closed forms, which are not stated here due to the page limit.

We expect all rates to increase with feedback rate since residual inter-cluster interference will decrease as quantized CDI becomes more accurate. We note that the rate performance of all single-user clusters are identical since the cdf in (10) is the same. In addition to feedback bits, rates for the two users in a 2-user cluster also depend on the degradation factor c_k, which may be different for different clusters. If c_k is closer to zero, the rate of user 2 will be smaller due to poor channel quality. Assuming that total feedback bits per update is fixed at B_{total}, we would like to optimize feedback-bit allocation for all clusters. Since there are two types of clusters, which could indicate different grade of service, we let B_1 be a number of feedback bits allocated for each single-user cluster and B_2 be a number of feedback bits allocated for each 2-user cluster. Thus,

$$B_{\text{total}} = B_1 M_1 + B_2 M_2. \tag{15}$$

To ensure quality of service, we maximize the minimum rate of all active users over feedback rates as follows

$$\max_{B_1,B_2} \min_{\substack{1\leq l\leq M_1 \\ M_1+1\leq k\leq M}} \{R_l(B_1), R_{k,1}(B_2), R_{k,2}(B_2)\}, \tag{16}$$

subject to total feedback-bit constraint (15). If condition (7) applies, then $R_{k,1} \geq R_{k,2}$. Also, R_l is the same for all l. Thus, we can simplify the optimization problem in (16) as

$$\max_{B_1,B_2} \min_{M_1+1\leq k\leq M} \{R_l(B_1), R_{k,2}(B_2)\}. \tag{17}$$

Finding the optimal feedback allocation B_1^* or B_2^* for this rate maximization requires some numerical method.

Besides individual rate, we can also compute a sum rate of all users in the cell with the proposed NOMA scheme as follows

$$R_{\text{NOMA}} = \sum_{l=1}^{M_1} R_l + \sum_{k=M_1+1}^{M} R_{k,1} + R_{k,2}. \tag{18}$$

If some OMA scheme is applied instead, then the two users in a 2-user cluster will be orthogonal with each other. However, each user in a 2-user OMA cluster utilizes only half of the transmission bandwidth. With OMA, SINR of user 1 remains the same as shown in (5) while SINR of user 2 in (6) will be larger since there is no intra-cluster interference. We can compute the sum rate for an OMA scheme as follows

$$R_{\text{OMA}} = \sum_{l=1}^{M_1} R_l + \sum_{k=M_1+1}^{M} \frac{1}{2}R_{k,1} + \frac{1}{2}E[\log_2(1+\gamma_{\text{OMA};k,2})] \tag{19}$$

where the expression of $\gamma_{\text{OMA};k,2}$ is (6), but with the second term in the denominator removed.

IV. On Minimizing the Maximum Outage Probability

If the objective function is changed from achievable rates in the previous section to outage probability, we can approximate the optimal solutions. An outage occurs when instantaneous achievable rate is smaller than the required transmission rate denoted by R_{req}. The outage probability for the active user in single-user cluster l that is allocated B_1 bits of feedback is given by

$$P_{\text{out};l}(B_1) = F_{\gamma_l}\left(2^{R_{\text{req}}}-1\right) \approx \tilde{F}_{\gamma_l}\left(2^{R_{\text{req}}}-1\right). \quad (20)$$

where the approximate cdf is given by (10). Outage probability for user 2 in 2-user cluster k that is allocated B_2 feedback bits, can be similarly approximated as follows

$$P_{\text{out};k,2}(B_2) \approx \tilde{F}_{\gamma_{k,2}}\left(2^{R_{\text{req}}}-1\right). \quad (21)$$

We obtain the approximations for outage probability

$$P_{\text{out};l}(B_1) \approx (N_t-1)\delta(M_1-1, M_2, 1, B_1)\left(2^{R_{\text{req}}}-1\right) \quad (22)$$

and

$$P_{\text{out};k,2}(B_2) \approx \frac{(N_t-1)\delta(c_k^2 M_1, c_k^2(M_2-1), c_k^2, B_2)}{\frac{1}{2^{R_{\text{req}}}-1}-c_k^2} \quad (23)$$

by binomial expansions of (20) and (21), respectively. The approximations work well when R_{req} is small or SNR is high, and can be improved by adding higher-order terms from the binomial expansions.

In this section, we find feedback allocation that minimizes the maximum outage probability of all active users. Assuming that condition (7) holds, the problem is reduced to

$$\min_{B_1,B_2} \max_{M_1+1\le k\le M} \{P_{\text{out};l}(B_1), P_{\text{out};k,2}(B_2)\}, \quad (24)$$

subject to the same constraint on the total feedback bits (15). We note that $P_{\text{out};l}$ is monotonically decreasing with B_1 while $P_{\text{out};k,2}$ is monotonically increasing with B_1. Thus, we can find B_1 that results in $P_{\text{out};l} = P_{\text{out};k,2}$. By applying the approximations in (22) and (23), we can estimate the feedback allocation for a single-user cluster that gives $P_{\text{out};l} \approx P_{\text{out};k,2}$ and denote that with $\tilde{B}_1(k)$. We can obtain $\tilde{B}_1(k)$ for all $M_1+1 \le k \le M$ with the following Lemma.

Lemma 1: For $R_{\text{req}} < \log_2(1+\frac{1}{c_k^2})$,

$$\tilde{B}_1(k) = \frac{1}{N_t}B_{\text{total}} - (1-\frac{1}{N_t})M_2\log_2(1+\frac{1}{N_t+M_2-2}) - (1-\frac{1}{N_t})M_2\log_2(1-\frac{c_k^2}{1+c_k^2}2^{R_{\text{req}}}). \quad (25)$$

We note that $\tilde{B}_1(k)$ increases with c_k. Thus,

$$\min_{M_1+1\le k\le M} \tilde{B}_1(k) = \tilde{B}_1(k^*) \quad (26)$$

where k^* refers to the 2-user cluster with the weakest user 2 or

$$k^* = \min_{M_1+1\le k\le M} c_k. \quad (27)$$

With Lemma 1, we can derive the approximate optimal feedback allocation for problem (24) as follows.

Proposition 2: If $R_{\text{req}} < \log_2(1+\frac{1}{(c_k^*)^2})$, the optimal feedback allocation to (24) can be approximated by

$$B_1^\star \approx \tilde{B}_1(k^*) \quad (28)$$

$$B_2^\star \approx \frac{1}{M_2}B_{\text{total}} - \frac{M_1}{M_2}B_1^\star. \quad (29)$$

We note that if uniform feedback allocation is applied, $B_1 = B_2 = B_{\text{total}}/N_t$, which is the first term in the feedback allocation for 1-user cluster in (25). How much the optimal allocation must deviate from uniform allocation depends on the second and third terms in (25). For example, if user 2 in the 2-user cluster is also strong (c_k is close to 1), feedback for 1-user clusters (B_1) must increase to combat larger interference from the 2-user cluster. If the number of 2-user clusters (M_2) increases, B_1 may have to decrease. However, if both users in 2-user clusters are strong (c_k is close to 1), B_1 has to increase instead.

However, we remark that the approximation in Proposition 2 is only valid when the required rate R_{req} is less than $\log_2(1+\frac{1}{(c_k^*)^2})$. Also, the approximate feedback allocation is accurate in high-SNR regimes.

V. Numerical Results

In Fig. 1, we compare the approximate cdf of γ_l and $\gamma_{k,2}$ in (10) and (13) with the results obtained from Monte Carlo simulation. In the figure, two sets of system size ($N_t = 4$ and $N_t = 10$) are shown. We see that the derived approximation is close to the simulation results when N_t is large since the variation from interference terms is getting smaller as N_t increases. Furthermore, the derived cdf is more accurate when the argument of cdf is small.

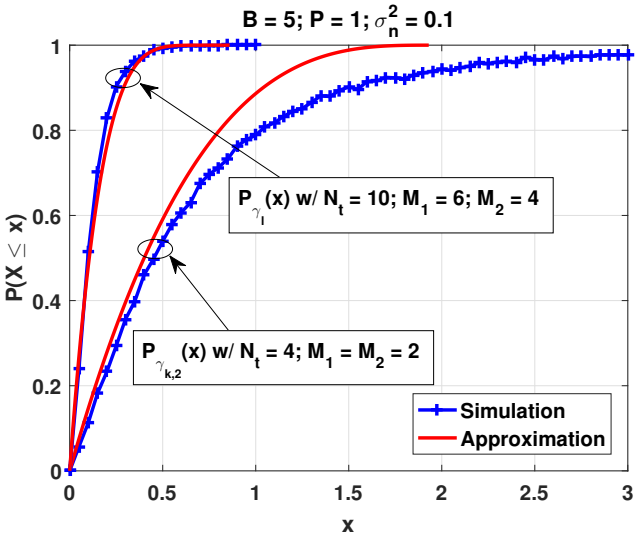

Fig. 1. The derived approximate cdf of γ_l and $\gamma_{k,2}$ are compared with the simulation results for $N_t = 4$ and $N_t = 10$.

Fig. 2 shows the approximate sum rate R_{NOMA} in (18) obtained from Proposition 1 with the sum rate obtained from

Monte Carlo simulation for two system sizes ($N_t = 4$ and $N_t = 10$). For this comparison, uniform feedback allocation ($B_1 = B_2$) is applied in all cases and thus, the total feedback bits $B_{\text{total}} = B_1(M_1 + M_2) = B_1 M$. We see that the NOMA sum rate from our analysis (solid line) approximates that from simulation (solid lines with solid dots) very well. For $N_t = 4$, it takes $B_{\text{total}} = 100$ or 25 bits per 4×1 channel vector or approximately 6 bits per complex channel gain to achieve close to the maximum rate with the optimal beamforming. For $N_t = 10$, larger B_{total} is required to achieve the optimal rate since there is greater interference due to a larger number of clusters. For small feedback rate, the sum rate of the system with $N_t = 4$ outperforms that with $N_t = 10$. This is also due to greater inter-cluster interference from a larger system.

Comparison between a sum rate of the proposed NOMA with that of OMA is also shown in Fig. 2. We see that NOMA outperforms OMA in all feedback rates and the rate gap increases with the total feedback rate. To derive the approximate rate, we assume that RVQ codebook is used to quantize CDI for all clusters. In this figure, besides RVQ, we also show the rate performance with Grassmannian codebook [12] and Discrete Fourier transform (DFT) codebook [13]. Grassmannian codebook is the optimum codebook that maximizes the average inner product squared with normalized channel vectors with independent identically distributed gains, and thus, can outperform RVQ [12]. However, for small feedback rate, all 3 codebooks perform about the same (see the inset figure). DFT codebook is a much simpler codebook and performs much worse when feedback rate is not small. The same performance trends for Grassmannian and DFT codebooks are also observed by [13].

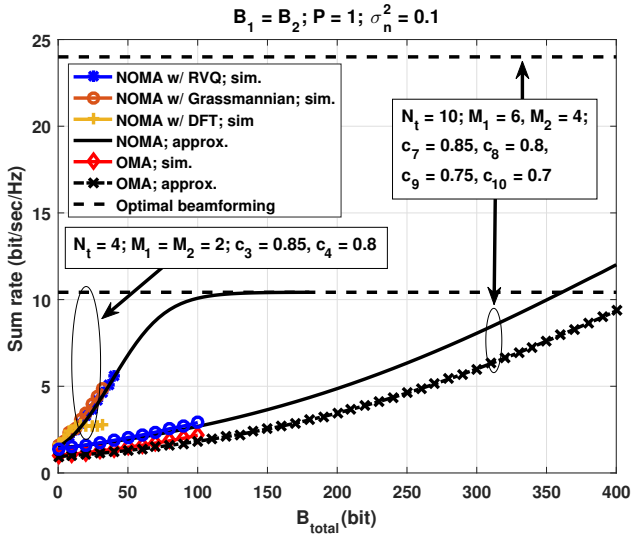

Fig. 2. The derived approximate sum rate of NOMA is compared with the simulation results and also with the sum rate of OMA. Sum rates with other quantization codebooks are shown as well.

We verify the derived approximations for outage probability in (20) - (23) by comparing them with results from numerical simulation in Fig. 3. The solid lines show the derived approximation (20) for 1-user cluster l and (21) for user 2 in 2-user cluster k. Both are derived from the assumption that SINR is approximately scaled beta random variable. We see that those approximations are close to the simulation results shown by solid lines with pluses. Assuming small required rate or high SNR, we obtain approximations (22) and (23) shown in the figure with dashed lines. We note that when $R_{\text{req}} < 0.4$, (22) gives a good approximation to the outage probability for a 1-user cluster and when $R_{\text{req}} < 0.2$, (23) can give a good approximation to that for a 2-user cluster. These approximations (22) and (23) are also used to approximate the optimal feedback allocation in Proposition 2.

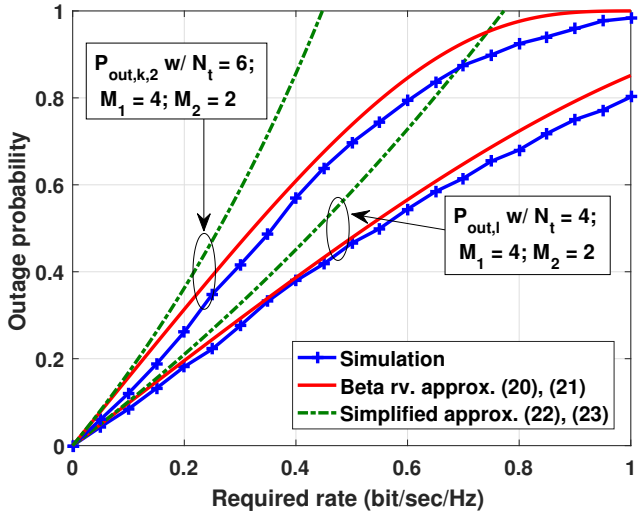

Fig. 3. Various approximations for outage probability for 1-user and 2-user clusters (20) - (23) are compared with simulation results.

In Fig. 4, we compare the approximate optimal feedback allocation $B_1^\star$ in Proposition 2 (solid lines) with that from solving (24) by exhaustive search over all combinations of B_1 and B_2 under (15) (solid lines with pluses). From (25) in Lemma 1, the approximate $B_1^\star$ is a linear function of B_{total} with slope $1/N_t$, which can be observed from the figure as well. We can conclude that in small-feedback regime, the system should only increase B_1 by a factor of $1/N_t$ for every increase of B_{total}. We note that the approximate $B_1^\star$ is accurate up to certain B_{total} and deviates the actual $B_1^\star$ when B_{total} increases. We also see from the numerical-search results that when B_{total} is sufficiently large, $B_1^\star$ is saturated.

In Fig. 5, the outage probability shown is the maximum outage probability among active users for either optimal or uniform feedback allocations. We see that for small-to-moderate SNR regime, outage probability decreases as SNR increases. However, when SNR is increased beyond certain value, there is outage floor. This is due to stronger interference from other users when SNR increases. Since the system considered is fully loaded (the number of clusters equals the degrees of freedom of the system), thus, there is no extra degree of freedom to avoid interference. Thus, the outage probability shown is quite high. In this regime, the optimal feedback

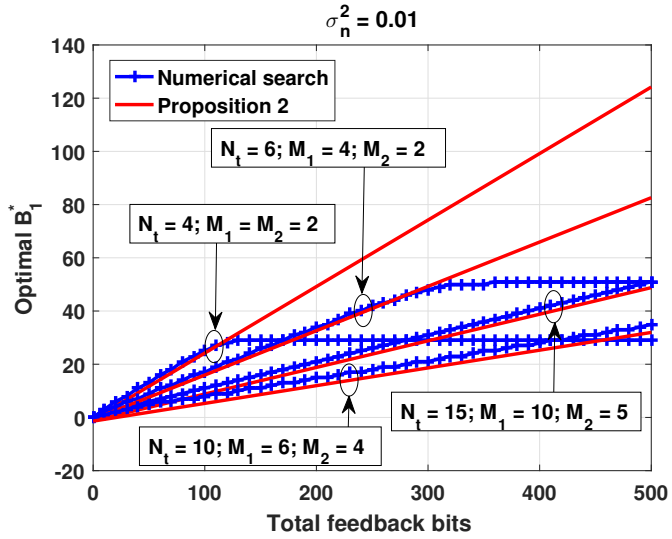

Fig. 4. The approximate optimal feedback allocation for 1-user cluster $B_1^\star$ obtained from Proposition 2 is compared with results from solving (24) numerically.

allocation clearly outperforms uniform allocation significantly. For $N_t = 10$ and the maximum outage probability of 0.15, optimal feedback allocation can reduce transmit power by about 12 dB. However, if the feedback rate is increased or the system load is decreased, we expect the outage probability to be much smaller, but the difference between the optimal and uniform feedback allocation will be small as well.

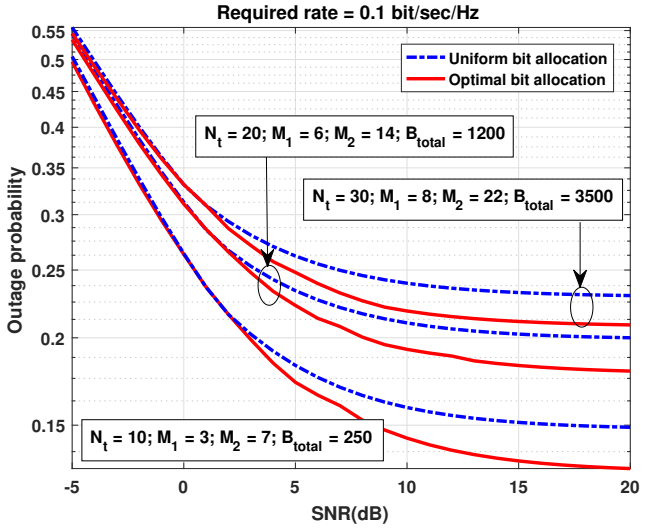

Fig. 5. Maximum outage probability among active users with either optimal or uniform feedback allocations is shown with SNR.

VI. Conclusions

We have analyzed the effect of limited feedback, which is used to relay quantized CDI to the base station, on the achievable rate and outage probability in a single-cell downlink NOMA model. We derive the approximate rates for users in 1-user clusters as well as 2-user clusters and find that the approximation from our analysis can predict the actual rates very well. However, finding the optimal allocation that maximize the minimum rate requires numerical method.

To analyze the optimal feedback allocation for different types of clusters, we consider minimizing the maximum outage probability of all active users. We derive the approximate optimal allocation for feedback rate and find that generally more feedback should be allocated for 2-user clusters, which suffer both intra-cluster and inter-cluster interference. For small feedback-rate regime, the increase in feedback for 1-user cluster should be $1/N_t$ of the total-feedback increase. We also show that NOMA outperforms OMA in all feedback regimes, and a rate gap is significant when feedback rate is low.

References

[1] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[2] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

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

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

[5] S. Liu and C. Zhang, "Downlink non-orthogonal multiple access system with limited feedback channel," in *Proc. Int. Conf. on Wireless Commun. and Signal Processing (WCSP)*, Nanjing, China, Oct. 2015, pp. 1–5.

[6] 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.

[7] Z. Xiao, L. Zhu, J. Choi, P. Xia, and X. Xia, "Joint power allocation and beamforming for non-orthogonal multiple access (NOMA) in 5G millimeter wave communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2961–2974, May 2018.

[8] S. Ali, E. Hossain, and D. I. Kim, "Non-orthogonal multiple access (NOMA) for downlink multiuser MIMO systems: User clustering, beamforming, and power allocation," *IEEE Access*, vol. 5, pp. 565–577, 2017.

[9] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[10] J. Choi, "Minimum transmit power NOMA beamforming for millimeter-wave communications," in *Proc. Vehicular Technol. Conf. (VTC Spring)*, Sydney, Australia, Jun. 2017, pp. 1–5.

[11] W. Santipach and M. L. Honig, "Capacity of a multiple-antenna fading channel with a quantized precoding matrix," *IEEE Trans. Inf. Theory*, vol. 55, no. 3, pp. 1218–1234, Mar. 2009.

[12] D. J. Love, R. W. Heath, and T. Strohmer, "Grassmannian beamforming for multiple-input multiple-output wireless systems," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2735–2747, Oct. 2003.

[13] G. Dietl and G. Bauch, "Linear precoding in the downlink of limited feedback multiuser MIMO systems," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Washington, DC, USA, Nov. 2007, pp. 4359–4364.