Two-User Downlink Non-Orthogonal Multiple Access with Limited Feedback

Xiaoyi (Leo) Liu, Hamid Jafarkhani
Center for Pervasive Communications and Computing
University of California, Irvine
Email: xiaoyil3@uci.edu

***Abstract*—In this paper, we analyze downlink non-orthogonal multiple access (NOMA) networks with limited feedback. Our goal is to derive appropriate transmission rates for rate adaptation based on distributed channel feedback information from two receivers. We propose an efficient quantizer with variable-length encoding that approaches the best performance of the case where perfect channel state information is available everywhere. We prove that in the typical application with two receivers, the loss in the minimum rate decays at least exponentially with the minimum feedback rate. Numerical simulations are presented to demonstrate the efficiency of our proposed quantizer and the accuracy of the analytical results.**

***Keywords*—NOMA, rate adaptation, minimum rate, limited feedback**

I. Introduction

Non-orthogonal multiple access (NOMA) has received significant attention recently for its superior spectral efficiency [1]. It is a promising candidate for mobile communication networks, and has been included in LTE Release 13 for the scenario of two-user downlink transmission under the name of multi-user superposition transmission [2]. The key idea of NOMA is to multiplex multiple users with superposition coding at different power levels, and utilize successive interference cancellation (SIC) at receivers with better channel conditions. Specifically, for NOMA with two receivers, the messages to be sent are superposed with different power allocation coefficients at the BS side. At the receivers' side, the weaker receiver decodes its intended message by treating the other's as noise, while the stronger receiver first decodes the message of the weaker receiver, and then decodes its own by removing the other message from the received signal. In this way, the weaker receiver benefits from larger power, and the stronger receiver is able to decode its own message with no interference. Hence, the overall performance of NOMA is enhanced, compared with traditional orthogonal multiple access schemes. It is shown in [3] that the rate region of NOMA is the same as the capacity region of Gaussian broadcast channels with two receivers, but with an additional constraint that the stronger receiver is assigned less power than the weaker one.

There has been a lot of work on NOMA. In [1] and [3], the authors evaluated the benefits of downlink NOMA from the system and information theoretic perspectives, respectively. NOMA with multiple antennas was studied in [4]. A lot of effort has been put into the power allocation design in NOMA. For example, the authors in [5] analyzed the necessary conditions for NOMA with two users to beat the performance of time-division-multiple-access (TDMA), and derived closed-form expressions for the expected data rates and outage probabilities. Transmit power minimization subject to rate constraints was discussed in [6].

However, all the mentioned papers have assumed a perfect knowledge of the distributed channel state information (CSI) at the BS and all the geographically-distributed receivers, which is difficult to realize in practice. Therefore, we consider the limited feedback scenario wherein each receiver only has access to its own local CSI, from the BS to itself, and then broadcasts its feedback information to the BS and other receivers [7], [8]. Under such settings, interesting problems arise, for example: How to design a simple but efficient quantizer for NOMA? What are the performance losses compared with the full-CSI case? In [9], the authors proposed a one-bit feedback scheme for ordering users in downlink Massive-MIMO-NOMA systems, and derived the achieved outage probability. In [10], the authors derived the outage probability of NOMA based on one-bit feedback of channel quality from each receiver, and performed power allocation to minimize the outage probability. Additionally, the problems of transmit power minimization and user fairness maximization based on statistical CSI subject to outage constraints were studied in [11]. In [12], the authors derived the outage probability and sum rate with fixed power allocation by assuming imperfect and statistical CSI.

In this paper, we focus on the limited feedback design for the typical scenario of downlink NOMA, where a BS communicates with two receivers simultaneously [2]. Based on distributed feedback and in the interest of user fairness, we wish to have the minimum rate of the receivers be as large as possible. To dynamically adjust the transmission rates for better channel utilization, we propose a uniform quantizer which assigns each value to its left boundary point and employs variable-length encoding (VLE). Then, power allocation is calculated based on the channel feedback. We calculate the transmission rates that can be supported by the current channel states, and analyze the rate loss compared with the full-CSI scenario. The derived upper bound on rate loss shows that it decreases at least exponentially with the minimum feedback rate. The primary goal of this paper is to study

This work was supported in part by the NSF Award ECCS-1642536.

978-1-5090-4096-4/17/$31.00 ©2017 IEEE

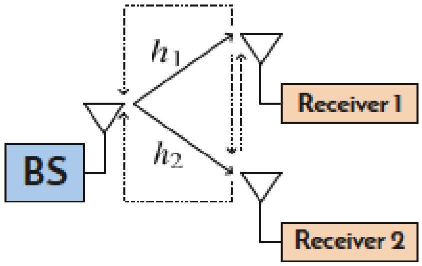

Fig. 1: Downlink NOMA networks. The solid and dashed lines represent the signal and feedback links, respectively.

the impacts of quantization on the performance of NOMA, and provide meaningful insights for practical limited feedback design. Numerical simulations are provided to demonstrate the efficiency of our proposed quantizer and the accuracy of the analytical results.

Notations: The sets of real and natural numbers are represented by $\mathcal{R}$ and $\mathcal{N}$, respectively. For any $x \in \mathcal{R}$, $\lfloor x \rfloor$ is the largest integer that is less than or equal to x. $\Pr\{\cdot\}$ and $\mathbb{E}[\cdot]$ represent the probability and expectation, respectively. For a random variable (r.v.) X, $f_X(\cdot)$ is its probability density function (p.d.f.). $\mathbb{CN}(\mu,\lambda)$ represents a circularly symmetric complex Gaussian r.v. with mean μ and variance λ.

II. Problem Formulation

A. System Model

Consider the downlink transmission in Fig. 1, where a BS is to transmit a superposition of two symbols to two receivers over the same resource block. Both BS and receivers are equipped with only a single antenna. According to the multiuser superposition transmission scheme [2], the transmitted signal is formed as

$$x = \sqrt{P_1}s_1 + \sqrt{P_2}s_2,$$

where s_i is the information bearing symbol for Receiver i with $\mathbb{E}[s_i] = 0$ and $\mathbb{E}\left[|s_i|^2\right] = 1$ for each channel state (the expectation is over all transmitted symbols); P_i is the average transmit power associated with s_i. Let $P = P_1 + P_2$ be the total transmit power, and $\alpha = \frac{P_1}{P}$ be the power allocation coefficient, then, $P_1 = \alpha P$ and $P_2 = (1-\alpha)P$ with $0 \le \alpha \le 1$.

Denote by $h_i \sim \mathbb{CN}(0,\lambda_i)$ the channel coefficient from the BS to Receiver i. Without loss of generality, assume $\lambda_1 \ge \lambda_2$. The received signals at Receivers 1 and 2 are respectively given by

$$y_1 = h_1\sqrt{P_1}s_1 + h_1\sqrt{P_2}s_2 + n_1, \quad y_2 = h_2\sqrt{P_1}s_1 + h_2\sqrt{P_2}s_2 + n_2,$$

where $n_i \sim \mathbb{CN}(0,1)$ represents the background noise. Let $H_i = |h_i|^2$, then, the p.d.f. of H_i is $f_{H_i}(x) = \frac{e^{-\frac{x}{\lambda_i}}}{\lambda_i}$ for $x > 0$.[1] We assume a quasi-static channel model, in which the channels vary independently from one block to another, while remaining constant within each block. Either receiver is assumed to perfectly estimate its local CSI (i.e., H_i), and send the associated quantized local CSI to the other receiver and BS in a broadcast manner via error-free and delay-free feedback links [13], [14]. In some scenario where the two receivers are far away from each other such that they cannot "talk" directly, the BS can play the role of relaying, i.e., forwarding the feedback information received from one receiver to the other.

When $H_1 \ge H_2$, with SIC, Receiver 1 first decodes s_2, and then decodes s_1 after removing s_2 from its received signal y_1; Receiver 2 directly decodes s_2 by treating s_1 as noise [15], [16]. Specifically, the rate for Receiver 2 to decode s_2 by treating s_1 as noise is

$$r_2(\alpha) = \log_2\left(1 + \frac{PH_2(1-\alpha)}{\alpha H_2 P + 1}\right),$$

which is not larger than the rate for Receiver 1 to decode s_2, given as $r_{1\to 2} = \log_2\left(1 + \frac{PH_1(1-\alpha)}{\alpha H_1 P+1}\right)$. If s_2 is transmitted at the rate of $r_2(\alpha)$, Receiver 1 can decode s_2 successfully with an arbitrarily small probability of error [17]. After removing $h_1\sqrt{P_2}s_2$ from y_1, Receiver 1 achieves a data rate for s_1 as

$$r_1(\alpha) = \log_2(1 + \alpha P H_1).$$

On the other hand, when $H_1 < H_2$, Receiver 2 first decodes s_1, removes $h_2\sqrt{P_1}s_1$ from y_2, and then decodes s_2, while Receiver 1 decodes s_1 directly by treating s_2 as noise.

B. Maximum Minimum Rate

Our goal is to maximize the minimum of $r_1(\alpha)$ and $r_2(\alpha)$ to ensure fairness between receivers [8], [18]. When perfect CSI is available at the BS and receivers, the optimal power allocation coefficient $\alpha^\star$ can be found by solving the optimization problem $r_{\max} = \max_{0\le\alpha\le 1} \min\{r_1(\alpha), r_2(\alpha)\}$, the solution of which is given in the following theorem.

Theorem 1. *When $H_1 \ge H_2$, the solution of $\max_{0\le\alpha\le 1} \min\{r_1(\alpha), r_2(\alpha)\}$ is given by*

$$\alpha^\star = \frac{2H_2}{\sqrt{(H_1+H_2)^2 + 4H_1H_2^2P} + (H_1+H_2)}. \quad (1)$$

Proof: Notice that with α increasing from 0 to 1, $r_1(\alpha)$ increases from 0 to $\log_2(1+PH_1)$ and $r_2(\alpha)$ decreases from $\log_2(1+PH_2)$ to 0. Since $\log_2(1+PH_1) \ge \log_2(1+PH_2)$, the maximum minimum rate is reached when $r_1(\alpha^\star) = r_2(\alpha^\star)$, from which $\alpha^\star$ in (1) is derived. ■

The expression of $\alpha^\star$ when $H_1 < H_2$ can be obtained straightforwardly. It is worth noting that both messages attain the same rate at optimality, i.e., $r_1(\alpha^\star) = r_2(\alpha^\star) = r_{\max}$. Moreover, it can be verified that the rate pair $(r_1(\alpha^\star), r_2(\alpha^\star))$ is on the rate region boundaries of both NOMA and Gaussian broadcast channels with two receivers [3].

It is also worth pointing out that $\alpha^\star$ in (1) satisfies the requirement for power allocation considered in [5] and [19]: the achieved individual rate should exceed that in the TDMA scheme, i.e., $r_i(\alpha^\star) \ge \frac{1}{2}\log_2(1+PH_i)$ for $i = 1,2$. Therefore,

[1] The results in this paper can be trivially generalized to other distributions of H_1 and H_2.

the maximum minimum rate we consider in this paper achieves higher rates in addition to better fairness between receivers.

With perfect CSI, the decoding order is determined based on whether $H_1 \geq H_2$ holds. The maximum minimum rate is

$$r_{\max} = \begin{cases} \log_2\left(1+\frac{2H_1H_2P}{\sqrt{(H_1+H_2)^2+4H_1H_2^2P}+(H_1+H_2)}\right), & H_1 \geq H_2, \\ \log_2\left(1+\frac{2H_1H_2P}{\sqrt{(H_1+H_2)^2+4H_1^2H_2P}+(H_1+H_2)}\right), & H_1 < H_2. \end{cases}$$

C. Limited Feedback

In the limited-feedback scenario, for an arbitrary quantizer $q: \mathcal{R} \to \mathcal{R}$, Receiver i maps H_i to $q(H_i)$, and feeds the index of $q(H_i)$ back to the BS and the other receiver, as shown in Fig.1. The index of $q(H_i)$ is decoded and the value of $q(H_i)$ is recovered. The decoding order will be contingent on whether $q(H_1) \geq q(H_2)$. For instance, when $q(H_1) \geq q(H_2)$, Receiver 1 is considered "stronger", while Receiver 2 is "weaker". In this case, the power allocation coefficient is computed based on (1) by treating $q(H_i)$ as H_i, i.e., $\alpha_q = \frac{2q(H_2)}{\sqrt{(q(H_1)+q(H_2))^2+4q(H_1)q^2(H_2)P}+q(H_1)+q(H_2)}$.

For rate adaptation, we shall design appropriate rates $r_{1,q}$ and $r_{2,q}$ for the messages s_1 and s_2 based on limited feedback from the two receivers, such that $r_{1,q}$ and $r_{2,q}$ can be supported and NOMA can be performed. The corresponding rate loss will be

$$r_{\text{loss}} = r_{\max} - \min\{r_{1,q}, r_{2,q}\}.$$

In the subsequent sections, we will propose an efficient quantizer and investigate the performance loss brought by limited feedback.

III. Limited Feedback for Minimum Rate

In this section, we first describe the proposed quantizer when the minimum rate is the concern, then, we show the relationship between the rate loss and the feedback rates.

A. Proposed Quantizer

We consider a uniform quantizer $q_r: \mathcal{R} \to \mathcal{R}$, given by[2]

$$q_r(x) = \begin{cases} \lfloor \frac{x}{\Delta} \rfloor \times \Delta, & x \leq T\Delta, \\ T\Delta, & x > T\Delta, \end{cases}$$

where the bin size Δ and the maximum number of bins $T \in \mathcal{N}$ are adjustable parameters. As shown in Fig. 2, $q_r(x)$ quantizes x to the left boundary of the interval where x is. For any $x \in [n\Delta, (n+1)\Delta)$ when $0 \leq n \leq T-1$, we have $q_r(x) = n\Delta$ and $x - \Delta \leq q_r(x) \leq x$; for any $x \in [T\Delta, \infty)$, $q_r(x) = T\Delta$ and $q_r(x) \leq x$.

B. Rate Adaptation and Loss

When $q_r(\cdot)$ is employed, Receiver 2 is viewed as the "weak" receiver if $q_r(H_1) \geq q_r(H_2)$. Then, according to (1), the power

[2]In q_r, "q" stands for quantizer, and the subscript "r" represents rate.

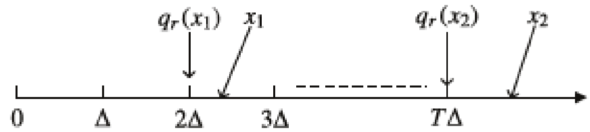

Fig. 2: A uniform quantizer for minimum rate.

allocation coefficient α_{q_r} is calculated as

$$\alpha_{q_r} = \begin{cases} \frac{2q_r(H_2)}{\sqrt{[q_r(H_1)+q_r(H_2)]^2+4q_r(H_1)q_r^2(H_2)P}+[q_r(H_1)+q_r(H_2)]}, \\ \qquad\qquad q_r(H_1) > 0, q_r(H_2) > 0, \\ 0, \qquad\qquad q_r(H_1) = 0 \text{ or } q_r(H_2) = 0. \end{cases}$$

Note that α_{q_r} satisfies $\log_2(1+P\times\alpha_{q_r}\times q_r(H_1)) = \log_2\left(1+\frac{q_r(H_2)\times(1-\alpha_{q_r})}{\alpha_{q_r}\times q_r(H_2)+\frac{1}{P}}\right)$ when $\alpha_{q_r} \neq 0$. To exploit the channels as much as possible, we let the rates for s_1 and s_2 be

$$\begin{aligned} r_{1,q_r} &= \log_2(1+P\times\alpha_{q_r}\times q_r(H_1)), \\ r_{2,q_r} &= \log_2\left(1+\frac{P\times q_r(H_2)(1-\alpha_{q_r})}{P\times q_r(H_2)\alpha_{q_r}+1}\right). \end{aligned} \quad (2)$$

Lemma 1. *When $q_r(H_1) \geq q_r(H_2)$, the rates r_{1,q_r} and r_{2,q_r} in (2) can be achieved.*

Proof: Based on the channel coding theorem [17], if we can show the channel capacities for s_1 and s_2 under the settings of NOMA are no smaller than r_{1,q_r} and r_{2,q_r}, the rates r_{1,q_r} and r_{2,q_r} can be achieved with a probability of error that can be made arbitrarily small.

When $q_r(H_1) = 0$ or $q_r(H_2) = 0$, it is trivial to verify that r_{1,q_r} and r_{2,q_r} can be supported. When $q_r(H_1) \geq q_r(H_2) > 0$, the channel capacity for Receiver 2 by treating s_1 as noise is $r_2 = \log_2\left(1+\frac{H_2(1-\alpha_{q_r})}{\alpha_{q_r}\times H_2+\frac{1}{P}}\right) \geq \log_2\left(1+\frac{q_r(H_2)\times(1-\alpha_{q_r})}{\alpha_{q_r}\times q_r(H_2)+\frac{1}{P}}\right) = r_{2,q_r}$, since $\log_2\left(1+\frac{x(1-\alpha)}{x\alpha+\frac{1}{P}}\right)$ is an increasing function of x and $q_r(H_2) \leq H_2$. At Receiver 1, the channel capacity of s_2 with treating s_1 as noise is $r_{1\to 2} = \log_2\left(1+\frac{H_1(1-\alpha_{q_r})}{\alpha_{q_r}\times H_1+\frac{1}{P}}\right) \geq \log_2\left(1+\frac{q_r(H_1)\times(1-\alpha_{q_r})}{\alpha_{q_r}\times q_r(H_1)+\frac{1}{P}}\right) \geq \log_2\left(1+\frac{q_r(H_2)\times(1-\alpha_{q_r})}{\alpha_{q_r}\times q_r(H_2)+\frac{1}{P}}\right) = r_{2,q_r}$, because $H_1 \geq q_r(H_1) \geq q_r(H_2)$. Hence, s_2 can be decoded at Receiver 1 with an arbitrarily small error and removed from y_1. After that, the channel capacity of s_1 is $r_1 = \log_2(1+P\times\alpha_{q_r}\times H_1) \geq \log_2(1+P\times\alpha_{q_r}\times q_r(H_1)) = r_{1,q_r}$. Therefore, the rates r_{1,q_r} and r_{2,q_r} can be achieved for both s_1 and s_2. ■

To sum up, it is the key fact of $q_r(x) \leq x$ that ensures the rates r_{1,q_r} and r_{2,q_r} can be supported. When $q_r(H_1) \geq q_r(H_2)$, the rate loss is

$$r_{\text{loss}} = r_{\max} - \min\{r_{1,q_r}, r_{2,q_r}\}.$$

Lemma 2. *The average rate loss of the quantizer $q_r(\cdot)$ is upper-bounded by:*

$$\mathrm{E}[r_{\text{loss}}] \leq \log_2\left(1+C_0\times P\times\max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta\right\}\right), \quad (3)$$

where C_0 is a positive constant that is independent of P, T and Δ.

The proof of Lemma 2 is provided in [20]. We mainly focus on showing how the average rate loss changes with the bin size Δ. It is beyond the scope of this paper to find the tightest bounds, i.e., the smallest value for C_0.

It is observed from (3) that when $e^{-\frac{T\Delta}{\lambda_1}} > \Delta$, the maximum number of bins, T, can degrade the rate. To eliminate this effect, we choose T such that $e^{-\frac{T\Delta}{\lambda_1}} = \Delta$, which yields $T = \frac{\lambda_1}{\Delta}\log\frac{1}{\Delta}$.[3] With an appropriate value for T, we can make the rate loss decrease at least linearly with respect to Δ.

Corollary 1. *When $T = \frac{\lambda_1}{\Delta}\log\frac{1}{\Delta}$, the average rate loss of the quantizer $q_r(\cdot)$ is upper-bounded by:*

$$\mathrm{E}[r_{\mathrm{loss}}] \le \log_2(1 + C_0 \times P \times \Delta) \le C_1 \times P \times \Delta, \qquad (4)$$

where C_0 and C_1 are positive constants that are independent of P and Δ.

C. Feedback Rate

Rather than the naive fixed-length encoding (FLE) for feedback that requires $\lceil \log_2(T+1) \rceil$ bits per receiver per channel state, we consider the more efficient variable-length encoding (VLE) [14], [21].[4] An example of VLE that can be applied here is $b_0 = \{0\}$, $b_1 = \{1\}$, $b_2 = \{00\}$, $b_3 = \{01\}$ and so on, sequentially for all codewords in the set $\{0, 1, 00, 01, 10, 11, \ldots\}$, where b_n is the binary string to be fed back when $q_r(x) = n\Delta$. The length of b_n is $\lfloor \log_2(n+2) \rfloor$. The following theorem derives an upper bound on the rate loss with respect to the feedback rate of Receiver i (denoted by $R_{r,\mathrm{VLE},i}$).

Theorem 2. *When variable-length encoding is applied to the quantizer $q_r(\cdot)$, the rate loss decays at least exponentially as:*

$$\begin{aligned}\mathrm{E}[r_{\mathrm{loss}}] &\le \log_2\left(1 + C_2 \times P \times 2^{-\min\{R_{r,\mathrm{VLE},1}, R_{r,\mathrm{VLE},2}\}}\right) \\ &\le C_3 \times P \times 2^{-\min\{R_{r,\mathrm{VLE},1}, R_{r,\mathrm{VLE},2}\}}, \qquad (5)\end{aligned}$$

where C_2 and C_3 are positive constants that are independent of P and $R_{r,\mathrm{VLE},i}$.

Proof: The feedback rate of Receiver i is derived as

$$\begin{aligned}R_{r,\mathrm{VLE},i} &= \sum_{n=0}^{T-1} \lfloor \log_2(n+2) \rfloor \int_{n\Delta}^{(n+1)\Delta} f_{H_i}(H_i)\mathrm{d}H_i \\ &\quad + \lfloor \log_2(T+2) \rfloor \int_{T\Delta}^{\infty} f_{H_i}(H_i)\mathrm{d}H_i \\ &\le \sum_{n=0}^{\infty} \lfloor \log_2(n+2) \rfloor \int_{n\Delta}^{(n+1)\Delta} f_{H_i}(H_i)\mathrm{d}H_i \\ &\le \sum_{n=0}^{\infty} \underbrace{\log_2(n+2)}_{\le \log_2(n+1)+1} \int_{n\Delta}^{(n+1)\Delta} \frac{e^{-\frac{H_i}{\lambda_i}}}{\lambda_i}\mathrm{d}H_i \\ &\le \sum_{n=0}^{\infty} e^{-\frac{n\Delta}{\lambda_i}}\left(1 - e^{-\frac{\Delta}{\lambda_i}}\right) \times \log_2(n+1) \\ &\quad + \underbrace{\sum_{n=0}^{\infty} 1 \times \int_{n\Delta}^{(n+1)\Delta} \frac{e^{-\frac{H_i}{\lambda_i}}}{\lambda_i}\mathrm{d}H_i}_{=1} \\ &= 1 + \left(1 - e^{-\frac{\Delta}{\lambda_i}}\right)\sum_{n=0}^{\infty} e^{-\frac{n\Delta}{\lambda_i}} \times \log_2(n+1) \\ &\le 1 + \frac{\Delta}{\lambda_i}\sum_{n=0}^{\infty} e^{-\frac{n\Delta}{\lambda_i}} \times \log_2(n+1).\end{aligned}$$

With the help of [14, Eq.(22)]: $\sum_{n=1}^{\infty} e^{-\beta n}\log(n) \le \frac{e^{-\beta}}{\beta}\left[2 + \log\left(1 + \frac{1}{\beta}\right)\right]$, by letting $\beta = e^{-\frac{\Delta}{\lambda_i}}$, we have

$$\begin{aligned}\sum_{n=0}^{\infty} e^{-\frac{n\Delta}{\lambda_i}} \times \log_2(n+1) &= \sum_{n=1}^{\infty} e^{-\frac{n\Delta}{\lambda_i}} \times \log_2(n+1) \\ = \frac{e^{\frac{\Delta}{\lambda_i}}}{\log 2}\sum_{n=2}^{\infty} e^{-\frac{n\Delta}{\lambda_i}} \times \log(n) &\le \frac{1}{\frac{\Delta}{\lambda_i}}\left[\frac{2}{\log 2} + \log_2\left(1 + \frac{1}{\frac{\Delta}{\lambda_i}}\right)\right].\end{aligned}$$

Then, $R_{r,\mathrm{VLE},i}$ is upper-bounded by[5]

$$R_{r,\mathrm{VLE},i} \le \frac{2}{\log 2} + 1 + \log_2\left(1 + \frac{1}{\frac{\Delta}{\lambda_i}}\right), \qquad (6)$$

or equivalently (when $R_{r,\mathrm{VLE},i}$ is sufficiently large),

$$\Delta \le \frac{\lambda_i}{2^{R_{r,\mathrm{VLE},i} - 1 - \frac{2}{\log 2}} - 1} \le \frac{\lambda_i}{2^{R_{r,\mathrm{VLE},i} - 2 - \frac{2}{\log 2}}} = C_4 \times 2^{-R_{r,\mathrm{VLE},i}}. \qquad (7)$$

Substituting (7) into (4) proves the theorem. ■

IV. Numerical Simulations and Discussions

In this section, we perform numerical simulations to validate the effectiveness of our proposed quantizer for rate adaptation. In all subsequent simulations for two receivers, we assume the channel variances are $\lambda_1 = 1$ and $\lambda_2 = 0.5$. Results for other values of λ_1 and λ_2 will exhibit similar observations.

In Fig. 3, we simulated the minimum rates of the full-CSI case, $q_r(\cdot)$ and the TDMA scheme (where each receiver occupies half of the time to transmit). We observe that the proposed quantizer with NOMA outperforms the TDMA scheme when $\Delta = 0.01$ and 0.05. The rate loss between the full-CSI case and $q_r(\cdot)$ with $\Delta = 0.01$ is almost negligible. The corresponding values for $T = \frac{\lambda_1}{\Delta}\log\frac{1}{\Delta}$ and the feedback rates for both receivers (bits/per channel state) are listed in Table I. Compared with FLE which costs $\lceil \log_2(T+1) \rceil$ bits per receiver per channel state, VLE can save almost half of the feedback bits.

In Fig. 4, we plot the rate losses of $q_r(\cdot)$ for different values of Δ and the feedback rates $R_{r,\mathrm{VLE},1}$ and $R_{r,\mathrm{VLE},2}$. It shows that the rate loss of $q_r(\cdot)$ decreases at least linearly with respect to Δ and exponentially with $\min\{R_{r,\mathrm{VLE},1}, R_{r,\mathrm{VLE},2}\}$, which

[3] Approaching the performance in the full-CSI case generally requires a small value for Δ. We mainly consider the case where $\Delta \le 1$ in this paper.

[4] For example, when $\Delta = 0.01$ and $\lambda_1 = 1$, $T = \frac{\lambda_1}{\Delta}\log\frac{1}{\Delta} \approx 460.5$. When FLE is adopted, the feedback rate per receiver will be $\lceil \log_2(T+1) \rceil = 9$ bits per channel state. As shown in Section IV, VLE costs far fewer bits.

[5] Although it is intractable to derive a closed-form expression for $R_{r,\mathrm{VLE},i}$, the upper bound in (6) provides a good estimate on how many feedback bits will be consumed.

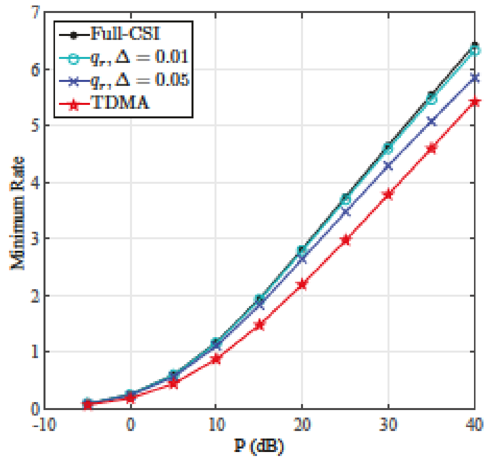

Fig. 3: Simulated minimum rates of NOMA.

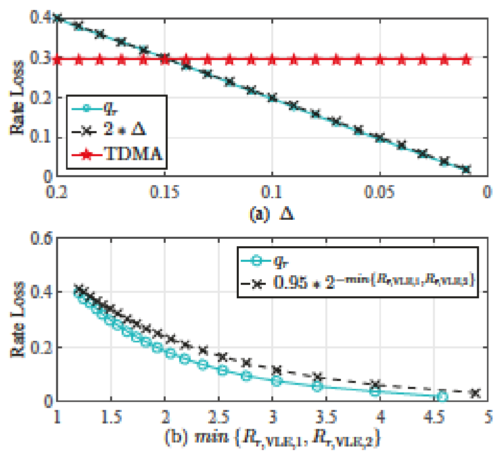

Fig. 4: Simulated rate losses versus (a) Δ and (b) $\min\{R_{r,\mathrm{VLE},1}, R_{r,\mathrm{VLE},2}\}$ for $P = 10$ dB.

TABLE I: Feedback rate for either receiver.

Δ	T	$\lceil \log_2(T+1) \rceil$	Receiver 1	Receiver 2
0.01	461	9	5.3	4.6
0.05	60	6	3.6	2.7

validates the accuracy of our derived upper bounds in (4) and (5). In addition, Fig. 4(a) shows that Δ needs to be less than 0.15 such that $q_r(\cdot)$ can obtain a higher rate compared with the TDMA scheme.

V. Conclusions and Future Work

We have introduced an efficient quantizer for rate adaptation of minimum rate in NOMA with two receivers. We have proved that the loss in rate decreases at least exponentially with the minimum feedback rate. The limited feedback design for the MIMO-NOMA networks will be an interesting future research direction.

References

[1] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sept. 2013, pp. 611–615.

[2] 3rd Generation Partnership Project (3GPP), "Study on downlink multiuser superposition transmission for LTE," Mar. 2015.

[3] 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, 2015.

[4] 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, June 2016.

[5] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov. 2016.

[6] L. Lei, D. Yuan, and P. Varbrand, "On power minimization for non-orthogonal multiple access (NOMA)," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2458–2461, Dec. 2016.

[7] E. Koyuncu and H. Jafarkhani, "Distributed beamforming in wireless multiuser relay-interference networks with quantized feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4538–4576, July 2012.

[8] X. Liu, E. Koyuncu, and H. Jafarkhani, "Cooperative quantization for two-user interference channels," *IEEE Trans. Commun.*, vol. 63, no. 7, pp. 2698–2712, 2015.

[9] 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.

[10] 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.

[11] J. Cui, Z. Ding, and P. Fan, "A novel power allocation scheme under outage constraints in NOMA systems," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1226–1230, Sept. 2016.

[12] 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.

[13] D. J. Love, R. W. Heath, Jr., 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.

[14] X. Liu, E. Koyuncu, and H. Jafarkhani, "Multicast networks with variable-length limited feedback," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 252–264, Jan. 2015.

[15] J. Choi, "On the power allocation for a practical multiuser superposition scheme in NOMA systems," *IEEE Commun. Lett.*, vol. 20, no. 3, pp. 438–441, Mar. 2016.

[16] ——, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, Oct. 2016.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[18] R. Sun, M. Hong, and Z.-Q. Luo, "Joint downlink base station association and power control for max-min fairness: Computation and complexity," *IEEE J. Select. Areas Commun.*, vol. 33, no. 6, pp. 1040–1054, June 2015.

[19] J. A. Oviedo and H. R. Sadjadpour, "A new NOMA approach for fair power allocation," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr. 2016, pp. 843–847.

[20] X. Liu and H. Jafarkhani, "Downlink non-orthogonal multiple access with limited feedback," https://arxiv.org/pdf/1701.05247.pdf, 2017.

[21] E. Koyuncu and H. Jafarkhani, "Variable-length limited feedback beamforming in multiple-antenna fading channels," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7140–7164, Nov. 2014.