Title
Downlink Non-Orthogonal Multiple Access With Limited Feedback.

Authors
Liu, Xiaoyi
Jafarkhani, Hamid

Publication Date
2017

DOI
10.1109/TWC.2017.2720169

Peer reviewed

# Downlink Non-Orthogonal Multiple Access with Limited Feedback

Xiaoyi (Leo) Liu, *Student Member, IEEE*, Hamid Jafarkhani, *Fellow, IEEE*

### 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 and minimize outage probability of minimum rate for the constant-rate data service, based on distributed channel feedback information from 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 losses in the minimum rate and outage probability decay at least exponentially with the minimum feedback rate. We analyze the diversity gain and provide a sufficient condition for the quantizer to achieve the maximum diversity order. For NOMA with $K$ receivers where $K > 2$, we solve the minimum rate maximization problem within an accuracy of $\varepsilon$ in time complexity of $O\left(K \log \frac{1}{\varepsilon}\right)$, then, we apply the previously proposed quantizers for $K = 2$ to the case of $K > 2$. Numerical simulations are presented to demonstrate the efficiency of our proposed quantizers and the accuracy of the analytical results.

### Index Terms

NOMA, rate adaptation, outage probability, minimum rate, limited feedback

This work was supported in part by the NSF Award ECCS-1642536. Part of this work will be presented at IEEE International Symposium on Information Theory (ISIT), 2017 [1]. X. Liu, H. Jafarkhani are with Center for Pervasive Communications & Computing, University of California, Irvine, Irvine, CA 92697–2625 USA (Email: {xiaoyil3, hamidj}@uci.edu).

# I. Introduction

Non-orthogonal multiple access (NOMA) has received significant attention recently for its superior spectral efficiency [2]. 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 [3]. 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 [4]. 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 [5] 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 [2] and [5], the authors evaluated the benefits of downlink NOMA from the system and information theoretic perspectives, respectively. The performance of NOMA with randomly deployed users was investigated in [6]. A lot of effort has been put into the power allocation design in NOMA. For example, the authors in [7] and [8] 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. In [9], power allocation based on proportional fairness scheduling was investigated for downlink NOMA. Transmit power minimization subject to rate constraints was discussed in [10]. A joint consideration of dynamic user clustering and power allocation was studied in [11].

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 [12]. Under such settings, interesting problems arise, for example: How to design simple but efficient quantizers for NOMA? What are the performance losses compared with the full-CSI case? A user-selection scheme based on limited feedback was studied in [13]. In [14], 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 [15]. In [16], the authors derived the outage probability and sum rate with fixed power allocation by assuming imperfect and statistical CSI. In [17], the authors solved the sum rate maximization problem for downlink NOMA networks using a minorization-maximization algorithm in statistics. In [18], several antenna selection schemes were proposed for the NOMA systems, and the user fairness was evaluated using the Jain's fairness index.

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 [3]. 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. Like [19], we also use the minimum achieved rate of all receivers as the performance measure, but moreover, the main focus of our work is to design efficient quantizers for downlink NOMA and analyze the achieved performance. With this goal, 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 of the feedback rates. The feedback rate in this paper refers to the number of feedback bits each receiver sends for each channel state. where the target data rate needs to be supported and outage probability is the main concern, we conversely propose a uniform quantizer which quantizes each value to its right boundary point.[1] Through the developed upper bound, we show the outage probability loss also decays at least exponentially with the minimum of feedback rate. Additionally, we analyze the achieved diversity gain and provide a sufficient condition on the proposed quantizer in order to achieve the full-CSI diversity order. For the general scenario with $K$ receivers, we solve the minimum rate maximization problem within an accuracy of $\varepsilon$ in time complexity of $O\left(K \log \frac{1}{\varepsilon}\right)$, and apply the previously proposed quantizers for the two-user case here by treating the quantized channels as the perfect ones. We perform Monte Carlo numerical simulations to verify the superiority of our proposed quantizers and the accuracy of the theoretical analysis.

The primary goal of this paper is to study the impacts of quantization on the performance of NOMA, and provide meaningful insights for practical limited feedback design. To summarize, the main contributions of this paper are three-fold:

(1) We propose efficient quantizers to maximize the minimum rate in NOMA. The ideas of our proposed quantizers and VLE as well as the designs for rate adaptation and outage probability based on distributed feedback can be generalized to many other scenarios, e.g., NOMA with other performance measures, the more general interference channels, and so on.

(2) Our theoretical analysis serves as a general framework to analyze the performances of such quantizers in NOMA and other scenarios. For instance, it can be easily applied to study the performances of other power allocation schemes in NOMA based on limited feedback, i.e., [7], [8].

(3) We solve the minimum rate maximization problem for any number of receivers with linear

[1] For example, in some real-time multimedia service applications, the minimum data rate needs to be supported as often as possible, such that the chance of service outage can be greatly reduced.

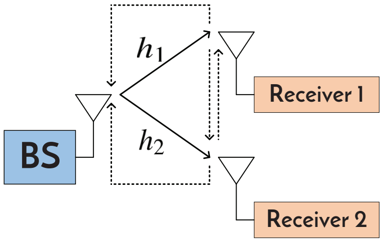

Fig. 1: Downlink NOMA networks. The solid and dashed lines represent the signal and feedback links, respectively.

time complexity.

The remainder of this paper is organized as follows: In Section II, we provide a brief description of the system model and formulate the problem of limited feedback. In Sections III and IV, we propose efficient quantizers for rate adaptation and outage probability, and analyze the performance loss. We extend our proposed quantizers to the general case with any number of receivers in Section V. Numerical simulations are provided in Section VI. We draw the main conclusions and summarize future work in Section VII. Technical proofs are presented in the appendices.

***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, and $\lceil x \rceil$ is the smallest integer that is larger than or equal to $x$. $\Pr\{\cdot\}$ and $\mathtt{E}[\cdot]$ 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 $\mu$ and variance $\lambda$. For a logical statement $\mathtt{ST}$, we let $\mathbf{1}_{\mathtt{ST}} = 1$ when $\mathtt{ST}$ is true, and $\mathbf{1}_{\mathtt{ST}} = 0$ otherwise. Finally, the expression $X \sim_Y Z$ means $0 < \lim_{Y\to\infty} \frac{X}{Z} < \infty$.

## 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.[2] Both BS and receivers are equipped

[2] We assume the two receivers have been pre-selected for the NOMA transmission based on user scheduling algorithms [2], [8]. In this paper, we mainly focus on the physical-layer performance of NOMA with limited feedback, and the study of user scheduling algorithms is beyond our scope of discussions.

with only a single antenna. According to the multiuser superposition transmission scheme [3], 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 $\mathrm{E}[s_i] = 0$ and $\mathrm{E}\left[|s_i|^2\right] = 1$ 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 $0 \leq \alpha \leq 1$.

Denote by $h_i \sim \mathbb{CN}(0, \lambda_i)$ the channel coefficient from the BS to Receiver $i$. Without loss of generality, assume $\lambda_1 \geq \lambda_2$. 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 $x > 0$.[3] 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 the BS in a broadcast manner via error-free and delay-free feedback links [20], [21]. 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.

With SIC, the stronger receiver with better channel condition (i.e., larger $H_i$) first decodes the message for the weaker receiver, and then decodes its own after removing the message of the weaker one from its received signal; the weaker receiver with poorer channel condition directly decodes its own message by treating the message of the stronger one as noise [9], [22]. Specifically, when $H_1 \geq H_2$, the rate for Receiver 2 (i.e., the weaker one) to decode $s_2$ by treating

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

$s_1$ 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 [23]. Afterwards, Receiver 1 can remove $h_1\sqrt{P_2}s_2$ from $y_1$, and achieve a data rate for $s_1$ as

$$r_1(\alpha) = \log_2(1 + \alpha PH_1).$$

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 [12], [24]. 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 $r_{\max} = \max\limits_{0\le\alpha\le 1} \min\{r_1(\alpha), r_2(\alpha)\}$, the solution of which is given in the following theorem.

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

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

*Proof:* Notice that with $\alpha$ 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 found from (1) that: (i) 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 [5]. (ii) When $P \to 0$, $\alpha^\star \to \frac{H_2}{H_1+H_2}$, in which case the power assigned to the stronger receiver is in proportion to the channel quality of the weaker one; when $P \to \infty$, $\alpha^\star \to 0$, then, BS should allocate almost all the power to the weaker one. [4] (iii) $\alpha^\star \geq \frac{1}{2}$. Generally, NOMA steers more power towards the weaker receiver to balance their transmissions.

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

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

and the outage probability of minimum rate is

$$\mathtt{out}_{\min} = \Pr\{r_{\max} < r_{\text{th}}\}, \tag{3}$$

where $r_{\text{th}}$ is the data rate at which the BS will transmit $s_1$ and $s_2$ for every channel state.

### 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 $r_{\text{loss}} = r_{\max} - \min\{r_{1,q}, r_{2,q}\}$, where $r_{\max}$ is given in (2).

For a constant-rate service, we care more about whether the current channels are strong enough to support target data rate with the power allocation coefficient computed based on

[4] Note that $r_1(\alpha^\star) = r_2(\alpha^\star)$ holds for any $P$. When $P \to \infty$, $\alpha^\star \to 0$, and $r_1(\alpha^\star) = r_2(\alpha^\star) = \log_2\left(1 + \frac{2PH_1H_2}{\sqrt{(H_1+H_2)^2+4H_1H_2^2P}+(H_1+H_2)}\right)$ will approach infinity.

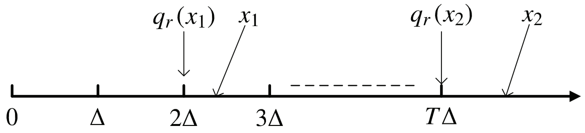

Fig. 2: A uniform quantizer for minimum rate.

limited feedback. The achieved outage probability is $\mathtt{out}_q = \Pr\left\{r_q < r_{\text{th}}\right\}$, where

$$r_q = \min\left\{r_1\left(\alpha_q\right), r_2\left(\alpha_q\right)\right\} = \begin{cases} \min\left\{\log_2\left(1+P\times\alpha_q\times H_1\right), \log_2\left(1+\frac{PH_2\left(1-\alpha_q\right)}{PH_2\alpha_q+1}\right)\right\}, & q(H_1)\geq q(H_2), \\ \min\left\{\log_2\left(1+\frac{PH_1\left(1-\alpha_q\right)}{PH_1\alpha_q+1}\right), \log_2\left(1+P\times\alpha_q\times H_2\right)\right\}, & q(H_1)< q(H_2), \end{cases}$$

The outage probability loss is given as

$$\mathtt{out}_{\text{loss},q} = \mathtt{out}_q - \mathtt{out}_{\min}, \tag{4}$$

where $\mathtt{out}_{\min}$ is given in (3). In the subsequent sections, we will propose efficient quantizers and investigate the performance losses 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[5]

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

where $x$ can be any non-negative real number, and the bin size $\Delta$ 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 $q_r(x) \leq x$.

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

### *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 allocation coefficient $\alpha_{q_r}$ is calculated as

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

which satisfies $\log_2\left(1 + P \times \alpha_{q_r} \times q_r(H_1)\right) = \log_2\left(1 + \frac{q_r(H_2)\times\left(1-\alpha_{q_r}\right)}{\alpha_{q_r}\times q_r(H_2)+\frac{1}{P}}\right)$ when $\alpha_{q_r} \neq 0$. To exploit the channels as much as possible, we let the BS send messages $s_1$ and $s_2$ at rates of

$$r_{1,q_r} = \log_2\left(1 + P \times \alpha_{q_r} \times q_r(H_1)\right), r_{2,q_r} = \log_2\left(1 + \frac{P \times q_r(H_2)\left(1-\alpha_{q_r}\right)}{P \times q_r(H_2)\alpha_{q_r} + 1}\right). \tag{5}$$

**Lemma 1.** *When $q_r(H_1) \geq q_r(H_2)$, the rates $r_{1,q_r}$ and $r_{2,q_r}$ in* (5) *can be achieved.*

*Proof:* Based on the channel coding theorem [23], 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 the side of 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 $r_1 = \log_2\left(1 + P \times \alpha_{q_r} \times H_1\right) \geq \log_2\left(1 + P \times \alpha_{q_r} \times q_r(H_1)\right) = r_{1,q_r}$. 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 $q_r(x) \geq x$ that ensures the rates $r_{1,q_r}$ and $r_{2,q_r}$ in (5) 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:*

$$\mathtt{E}\left[r_{\text{loss}}\right] \leq \log_2\left(1+C_0 \times P \times \max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta\right\}\right), \tag{6}$$

*where $C_0$ is a positive constant that is independent of $P, T$ and $\Delta$.*

*Proof:* See Appendix A. ■

We mainly focus on showing how the average rate loss changes with the bin size $\Delta$. It is beyond the scope of this paper to find the tightest bounds, i.e., the smallest value for $C_0$. A value for $C_0$ which is derived from the proof in Appendix A is $C_0 = \max\left\{4+\frac{\lambda_1}{\lambda_2}, \lambda_2\right\}$.

It is observed from (6) 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 $T = \frac{\lambda_1}{\Delta}\log\frac{1}{\Delta}$.[6] With an appropriate value for $T$, we can make the rate loss decrease at least linearly with $\Delta$.

**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:*

$$\mathtt{E}\left[r_{\text{loss}}\right] \leq \log_2\left(1+C_0 \times P \times \Delta\right) \leq C_1 \times P \times \Delta, \tag{7}$$

*where $C_0$ and $C_1$ are positive constants that are independent of $P$ and $\Delta$.*

### C. Feedback Rate

Rather than the naive fixed-length encoding (FLE) for feedback information which requires $\lceil \log_2(T+1) \rceil$ bits per receiver per channel state, we consider the more efficient variable-length encoding (VLE) [21], [25].[7] 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 $\lfloor \log_2(n+2) \rfloor$.

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

[7] 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 by the theoretical analysis and numerical simulations later, VLE will cost far fewer bits.

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:*

$$\mathtt{E}\left[r_{\mathrm{loss}}\right] \le \log_2\left(1 + C_2 \times P \times 2^{-\min\left\{R_{r,\mathrm{VLE},1}, R_{r,\mathrm{VLE},2}\right\}}\right) \le C_3 \times P \times 2^{-\min\left\{R_{r,\mathrm{VLE},1}, R_{r,\mathrm{VLE},2}\right\}}, \qquad (8)$$

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

*Proof:* The feedback rate of Receiver $i$ is derived as

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

With the help of [21, 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

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

Then, $R_{r,\mathrm{VLE},i}$ is upper-bounded by[8]

$$R_{r,\mathrm{VLE},i} \le \frac{2}{\log 2} + 1 + \log_2\left(1 + \frac{1}{\frac{\Delta}{\lambda_i}}\right), \qquad (9)$$

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

$$\Delta \le \frac{\lambda_i}{2^{R_{r,\mathrm{VLE},i} - 1 - \frac{2}{\log 2}} - 1} \le \frac{\lambda_i}{2^{R_{r,\mathrm{VLE},i} - 2 - \frac{2}{\log 2}}} = C_4 \times 2^{-R_{r,\mathrm{VLE},i}}. \qquad (10)$$

[8] Although it is intractable to derive a closed-form expression for $R_{r,\mathrm{VLE},i}$, the upper bound in (9) provides a good estimate on how many feedback bits will be consumed.

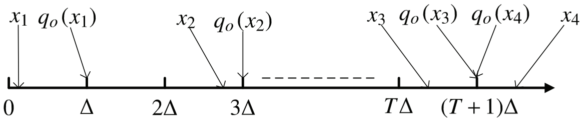

Fig. 3: A uniform quantizer for outage probability.

Substituting (10) into (7) proves the theorem. ■

Therefore, we can see that appropriate values for $T$ and the use of VLE enable the rate loss to decrease at least exponentially with the feedback rate.

## IV. Limited Feedback for Outage Probability

Outage probability is an important performance metric that evaluates the chance that the channels are not strong enough to support the constant-rate data service [26]. An ideal quantizer for outage probability should have at least the following properties: (i) The outage probability loss should decrease toward zero when the feedback rate increases toward infinity. (ii) The outage probability loss should approach zero whenever $P \to 0$ or $P \to \infty$. The intuition of (ii) comes from the fact that when $P$ is adequately small, the outage probabilities of both the full-CSI case and the quantizer should be close to one; when $P$ is significantly large, both outage probabilities should be almost zero. Then, the outage probability losses in both scenarios go to zero.

### A. Proposed Quantizer

As portrayed in Fig. 3, the uniform quantizer proposed for outage probability is given by

$$q_o(x) = \begin{cases} \left\lceil \frac{x}{\Delta} \right\rceil \times \Delta, & x \leq T\Delta, \\ (T+1)\Delta, & x > T\Delta. \end{cases} \tag{11}$$

The only difference between $q_o(\cdot)$ and $q_r(\cdot)$ lies in whether the left or right boundary of the interval is used as the reconstruction point. The quantizer proposed for rate adaptation cannot be directly inherited because when the channel is very weak (i.e., $H_i < \Delta$), it will be quantized as zero (i.e., $q_r(H_i) = 0$), which will result in a zero-value power allocation coefficient, i.e.,$\alpha_{q_r} = 0$, and a minimum rate of zero, i.e., $r_1\left(\alpha_{q_r}\right) = 0$ or $r_2\left(\alpha_{q_r}\right)$. In this case, the transmission will surely encounter an outage. However, even a weak channel reserves the possibility of non-outage, so long as the transmit power $P$ is large enough. Therefore, an appropriate quantizer for outage probability should not quantize any value to zero. The quantizer in (11) fulfills this requirement.

### *B. Outage Probability Loss*

**Lemma 3.** *The outage probability loss of the quantizer $q_o(\cdot)$ is upper-bounded by:*

$$\mathtt{out}_{\mathrm{loss},q_o} \leq C_5 \times e^{-\frac{C_6}{P}} \times \frac{1+\sqrt{P}}{P} \times \max\left\{\Delta^{\frac{1}{2}}, \Delta^{\frac{3}{2}}, e^{-\frac{T\Delta}{\lambda_1}}\right\}, \tag{12}$$

*where $C_5$ and $C_6$ are positive constants that are independent of $P$ and $\Delta$.*

*Proof:* See Appendix B. ■

Different from the rate loss which increases linearly in terms of $P$, because of the term $e^{-\frac{C_6}{P}} \times \frac{1+\sqrt{P}}{P}$, the upper bound on $\mathtt{out}_{\mathrm{loss},q_o}$ in (12) converges to zero either when $P \to 0$ or $P \to \infty$.

To have good performance, we mainly focus on the quantizers with small granularities. When $\Delta \leq 1$, we have $\Delta^{\frac{3}{2}} \leq \Delta^{\frac{1}{2}}$, and the upper bound in (12) is restricted by $\max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta^{\frac{1}{2}}\right\}$. For fixed $\Delta$, the optimal choice for $T$ should satisfy $e^{-\frac{T\Delta}{\lambda_1}} = \Delta^{\frac{1}{2}}$, given by $T = \frac{\lambda_1}{2\Delta}\log\frac{1}{\Delta}$.

**Corollary 2.** *When $0 < \Delta \leq 1$ and $T = \frac{\lambda_1}{2\Delta}\log\frac{1}{\Delta}$, the average rate loss of the quantizer $q_o(\cdot)$ is upper-bounded by:*

$$\mathtt{out}_{\mathrm{loss},q_o} \leq C_5 \times e^{-\frac{C_6}{P}} \times \frac{1+\sqrt{P}}{P} \times \Delta^{\frac{1}{2}}, \tag{13}$$

*where $C_5$ and $C_6$ are positive constants independent of $P$ and $\Delta$.*

### *C. Feedback Rate*

The same VLQ for rate adaptation can be applied to $q_o(\cdot)$ for a better utilization of the feedback resource. From (9) and (10), we obtain $R_{o,\mathrm{VLE},i} \leq \frac{2}{\log 2} + 1 + \log_2\left(1 + \frac{1}{\frac{\Delta}{\lambda_i}}\right)$ and $\Delta \leq C_4 \times 2^{-R_{o,\mathrm{VLE},i}}$. Thus, $\Delta^{\frac{1}{2}} \leq \sqrt{C_4 \times 2^{-R_{o,\mathrm{VLE},i}}} = C_7 \times 2^{-\frac{R_{o,\mathrm{VLE},i}}{2}} \leq C_7 \times 2^{-\frac{\min\left\{R_{o,\mathrm{VLE},1}, R_{o,\mathrm{VLE},2}\right\}}{2}}$. The following theorem states the relationship between the outage probability loss of $q_o(\cdot)$ and the feedback rates.

**Theorem 3.** *When variable-length encoding is applied to the quantizer $q_o(\cdot)$, the rate loss decays at least exponentially as:*

$$\mathtt{out}_{\text{loss},q_o} \leq C_8 \times e^{-\frac{C_6}{P}} \times \frac{1+\sqrt{P}}{P} \times 2^{-\frac{\min\left\{R_{o,\text{VLE},1},R_{o,\text{VLE},2}\right\}}{2}}, \tag{14}$$

*where $C_6$ and $C_8$ are positive constants independent of $P$ and $R_{o,\text{VLE},i}$.*

### *D. Diversity Order*

With an outage probability $\mathtt{out}$, the achieved diversity order is given as $d = \lim_{P\to\infty} \frac{\log \mathtt{out}}{\log P}$ [26, Section 2.3]. The following lemma shows the achievable diversity order of $q_o(\cdot)$ and a sufficient condition to achieve the maximum diversity order in the full-CSI scenario.

**Lemma 4.** (1) *With $q_o(\cdot)$ and fixed $\Delta$, the diversity orders of $\frac{1}{2}$ and 1 are achievable for Receivers 1 and 2, respectively.*

(2) *A sufficient condition for both receivers to achieve the maximum diversity order of 1 is $\Delta \sim_P P^{-\frac{1}{3}}$.*

*Proof:* See Appendix C. ■

In the full-CSI case, both receivers can achieve the same diversity order of 1 as in the case when no interference exists. In the limited feedback case, it can be found from the proofs in Appendices B and C that the cause of this insufficient diversity order for Receiver 1 comes from the marginal region when $0 < H_1, H_2 \leq \Delta$. Therefore, an adequately small $\Delta$ that scales at least in proportion to $P^{-\frac{1}{3}}$ in the high-$P$ region is desired to diminish the probability that $H_i$ falls into that region so as to obtain the maximum diversity gain.

## V. Extension to More than Two Receivers

### *A. Full-CSI Performance*

In this section, we consider NOMA with more than two downlink receivers. Assuming perfect CSI universally available and $H_1 \geq H_2 \geq \cdots \geq H_K$, the maximum minimum rate can be obtained by solving the optimization problem:

$$r_{\max} = \max_{\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_K]} \min_{k=1,\ldots,K} r_k(\boldsymbol{\alpha}), \text{ subject to } 0 \le \alpha_k \le 1, \sum_{k=1}^{K} \alpha_k = 1, \tag{15}$$

where $K$ is the number of receivers, and $r_k(\boldsymbol{\alpha}) = \log_2\left(1 + \frac{\alpha_k}{\sum_{i=1}^{k-1}\alpha_i + \frac{1}{PH_k}}\right)$ is the achieved rate for Receiver $k$ under superposition coding and SIC. To the best of our knowledge, no closed-form solution for $r_{\max}$ is available in the literature. We present the following lemma that helps solving the above optimization problem numerically.

**Lemma 5.** *There exists $\boldsymbol{\alpha}^\star = [\alpha_1^\star, \alpha_2^\star, \ldots, \alpha_K^\star]$, such that all receivers achieve the same rate at optimality, i.e., $r_{\max} = r_1(\boldsymbol{\alpha}^\star) = r_2(\boldsymbol{\alpha}^\star) = \cdots = r_K(\boldsymbol{\alpha}^\star)$.*

The proof of Lemma 5 is given in Appendix D. Since $r_{\max} = r_k(\boldsymbol{\alpha}^\star) = \log_2\left(1 + \frac{\alpha_k^\star}{\sum_{i=1}^{k-1}\alpha_i^\star + \frac{1}{PH_k}}\right)$ for $k = 1, \ldots, K$, we have $\alpha_k^\star = (2^{r_{\max}} - 1) \times \left(\sum_{i=1}^{k-1}\alpha_i^\star + \frac{1}{PH_k}\right)$, which leads to[9]

$$\alpha_k^\star = (2^{r_{\max}} - 1)\left[\frac{1}{PH_k} + (2^{r_{\max}} - 1)\sum_{i=1}^{k-1}\frac{2^{(k-1-i)r_{\max}}}{PH_i}\right]. \tag{16}$$

To find $\alpha_k^\star$, we need to solve for $r_{\max}$ first. Summing both sides from $k = 1, \ldots, K$ and after trivial calculations, we obtain

$$\sum_{k=1}^{K}\alpha_k^\star = 1 = \underbrace{(2^{r_{\max}} - 1)\sum_{i=1}^{K}\frac{2^{(K-i)r_{\max}}}{PH_i}}_{=\varpi(r_{\max})}. \tag{17}$$

In other words, $r_{\max}$ satisfies $\varpi(r_{\max}) = 1$.[10]

Let $r_{\text{ub}} = \log_2\left(1 + \min_{k=1,\ldots,K} PH_k\right) = \log_2(1 + PH_K)$. Since $\varpi(x)$ is an increasing function of $x$ as well as $\varpi(0) < 1$ and $\varpi(r_{\text{ub}}) \ge 1$, we could use the bisection method to find the root of $\varpi(x) = 1$ in the interval $(0, r_{\text{ub}}]$. The calculation of $\varpi(x)$ costs $O(K)$, thus, the time complexity of finding $r_{\max}$ within an accuracy of $\varepsilon$ is $O\left(K \log \frac{1}{\varepsilon}\right)$.

[9] Note that [19] also derives (16), but using the tools of convex optimization.

[10] Note that [27] has solved a different optimization problem, i.e. maximizing the sum rate subject to a minimum rate constraint, which satisfies $\sum_{k=1}^{K}\alpha_k^\star = 1$ but results in different $\alpha_k^\star$s.

### *B. Limited Feedback*

Under limited feedback, the previously proposed quantizers $q_r(\cdot)$ and $q_o(\cdot)$ in Figs. 2 and 3 can still be applied here for rate adaptation and outage probability, respectively. The maximum minimum rate can be calculated using the bisection method by treating $q_r(H_k)$ or $q_o(H_k)$ as $H_k$, and the corresponding power allocation coefficients can be computed. Although it is non-trivial to derive upper bounds on the losses in rate or outage probability for $K > 2$ theoretically, numerical simulations in Section VI show that the relationships between the performance loss and the feedback rate are similar to the case of $K = 2$.

## VI. Numerical Simulations and Discussions

In this section, we perform numerical simulations to validate the effectiveness of our proposed quantizers for rate adaptation and outage probability. In all subsequent simulations for $K$ receivers, we use the channel variances in Table I.

TABLE I: Channel variances for numerical simulations.

| | |
|---|---|
| $K = 2$ | $\lambda_1 = 1, \lambda_2 = 0.5$ |
| $K > 2$ | $\lambda_k = \frac{1}{k}, k = 1, \ldots, K$ |

Results for other values of channel variances will exhibit similar observations. For outage probability, sufficiently large number of channel realizations are generated to observe at least 10000 outage events.

In Fig. 4, 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 II. 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. 5, we plot the rate losses of $q_r(\cdot)$ for different values of $\Delta$ and the feedback rates $R_{r,\text{VLE},1}$ and $R_{r,\text{VLE},2}$. It shows that the rate loss of $q_r(\cdot)$ decreases at least linearly with respect

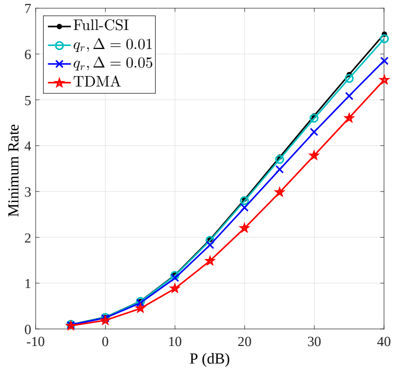

Fig. 4: Simulated minimum rates of NOMA for $K = 2$.

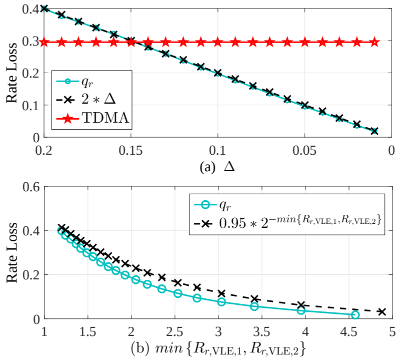

Fig. 5: Simulated rate losses versus (a) $\Delta$ and (b) $\min\{R_{r,\text{VLE},1}, R_{r,\text{VLE},2}\}$ for $K = 2$ and $P = 10$ dB.

TABLE II: Feedback rate for either receiver.

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

to $\Delta$ and exponentially with $\min\{R_{r,\text{VLE},1}, R_{r,\text{VLE},2}\}$, which validates the accuracy of our derived upper bounds in (7) and (8). In addition, Fig. 5(a) shows that $\Delta$ needs to be less than 0.15 such that $q_r(\cdot)$ can obtain a higher rate compared with the TDMA scheme.

In Fig. 6, we compare the outage probabilities of the full-CSI case, $q_o(\cdot)$ under various values of $\Delta$ and the TDMA scheme. It can be seen that: (i) The curve for $q_o(\cdot)$ with $\Delta = 0.01$ almost coincides with that of the full-CSI case. (ii) When $P$ is large, $q_o(\cdot)$ with $\Delta = 0.2$ suffers from an

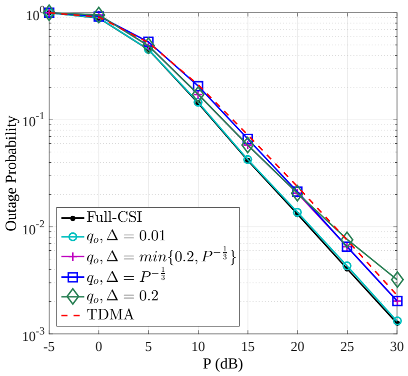

Fig. 6: Simulated outage probabilities of NOMA for $K = 2$.

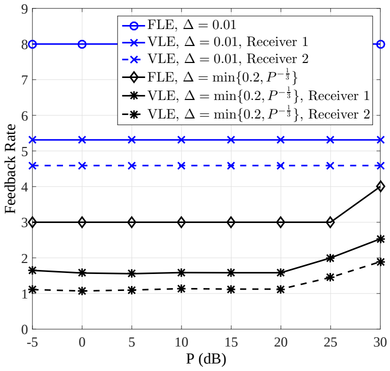

Fig. 7: Simulated feedback rates versus $P$ for $K = 2$.

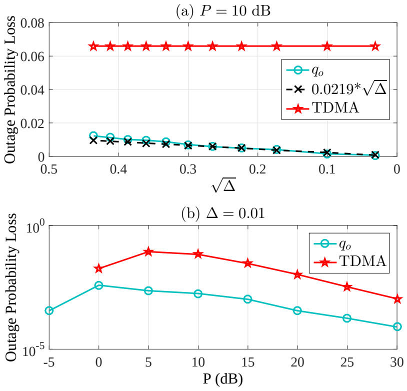

Fig. 8: Simulated outage probability losses versus $\sqrt{\Delta}$ and $P$ for $K = 2$.

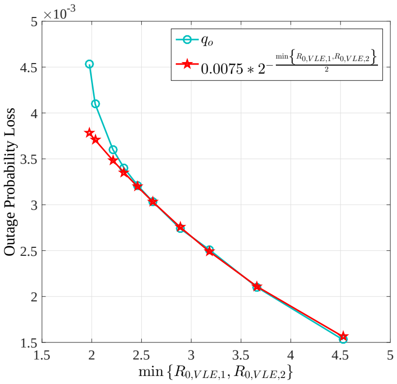

Fig. 9: Simulated outage probability losses versus $\min\{R_{0,\mathrm{VLE},1}, R_{0,\mathrm{VLE},2}\}$ for $K = 2$.

insufficient diversity gain in the high-$P$ region. According to our analysis in Lemma 4, $\Delta = 0.2$ is large enough not to scale with $P^{-\frac{1}{3}}$.[11] (iii) Although the maximum diversity order is achieved when $\Delta = P^{-\frac{1}{3}}$, much less array gain is obtained in the lower and medium-$P$ regions (where $\Delta$ is large). Alternatively, $\Delta = \min\left\{0.2, P^{-\frac{1}{3}}\right\}$ will reserve both benefits of the maximum diversity order brought by $P^{-\frac{1}{3}}$ and the higher array gain of $\Delta = 0.2$.[12] The comparison of feedback rates for VLE and FLE (which requires $\lceil \log_2(T+2) \rceil = \left\lceil \log_2\left(\frac{\lambda_1}{2\Delta} \log \frac{1}{\Delta} + 2\right) \right\rceil$ bits per channel state) under different values of $\Delta$ and $P$ is shown in Fig. 7, which verifies the superiority of VLE. It can be seen that the feedback rates for $\Delta = \min\left\{0.2, P^{-\frac{1}{3}}\right\}$ stay flat in the low and medium-$P$ regions (since $0.2 \leq P^{-\frac{1}{3}}$). When $P^{-\frac{1}{3}} \leq 0.2$ where $P \geq 20.9$ dB, the feedback rates start to increase as $\Delta$ gets smaller.

In Fig. 8(a), the outage probability loss decays at least linearly with respect to $\Delta$; in Fig. 8(b), the outage probability loss approaches zero whenever $P \to 0$ or $P \to \infty$; in Fig. 9, the outage probability loss decays at least exponentially with $\frac{\min\left\{R_{0,\mathrm{VLE},1}, R_{0,\mathrm{VLE},2}\right\}}{2}$. All these observations validate our theoretical analysis.

In Figs. 10 and 11, we simulated the rate and outage probability losses for more than two receivers. For Receiver $k$, the channel variance is set to be $\lambda_k = \frac{1}{k}$, the maximum number of

[11] The value 0.01 for $\Delta$ will also exhibit an insufficient diversity order as long as $P$ is large enough, although we might not be able to observe this in the region of $P \leq 30$ dB in Fig. 6.

[12] We also observe a similar effect of $\Delta$ on the achieved minimum rates, but we mainly elaborate it on outage probability.

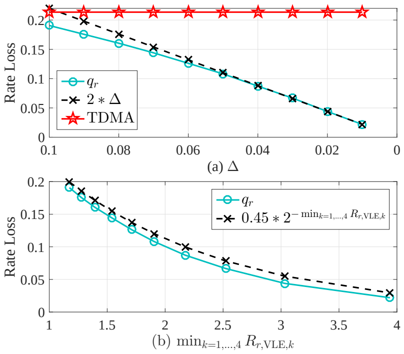

Fig. 10: Simulated rate losses versus (a) $\Delta$ and (b) $\min_{k=1,\dots,K} R_{r,\mathrm{VLE},k}$ for $K=4$ and $P=10$ dB.

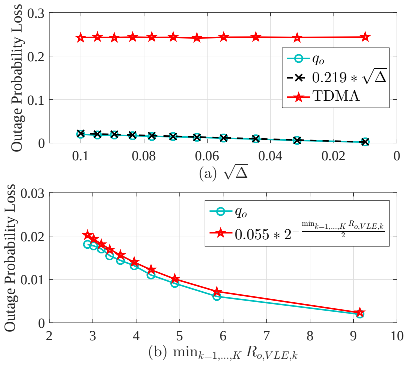

Fig. 11: Simulated outage probability losses versus (a) $\sqrt{\Delta}$ and (b) $\min_{k=1,\dots,K} R_{o,\mathrm{VLE},k}$ for $K=4$ and $P=10$ dB.

bins $T$ for $q_r(\cdot)$ and $q_o(\cdot)$ is $T = \frac{\lambda_k}{\Delta}\log\frac{1}{\Delta}$, and the accuracy used by the bisection method is $\varepsilon = 10^{-4}$. We simply treat the result of bisection method based on perfect CSI as the "full-CSI" performance. Compared with Figs. 5, 8 and 9 for $K=2$, Figs. 10 and 11 exhibit very similar relationships between the losses and $\Delta$ or the feedback rates.

## VII. Conclusions and Future Work

We have introduced efficient quantizers for rate adaptation and outage probability of minimum rate in NOMA with two receivers. We have proved that the losses in rate and outage probability both decrease at least exponentially with the minimum of the feedback rates. Furthermore, we generalized the proposed quantizers to NOMA with any number of receivers. The performance of NOMA with noisy quantized feedback and the user scheduling under limited feedback will be interesting future research directions.

## Appendix A: Proof of Lemma 2

To clarify, the notation $D_i$ for $i \in \mathbb{N}$ represents a positive constant independent of $P, T$ and $\Delta$. The average rate loss of $q_r(\cdot)$ can be expressed as

$$\mathrm{E}\left[r_{\mathrm{loss}}\right] = \underbrace{\int_{\mathcal{H}_{0,\geq}} r_{\mathrm{loss}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i}_{=\mathrm{E}_{\geq}[r_{\mathrm{loss}}]} + \underbrace{\int_{\mathcal{H}_{0,<}} r_{\mathrm{loss}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i}_{=\mathrm{E}_{<}[r_{\mathrm{loss}}]},$$

where $\mathcal{H}_{0,\geq} = \{(H_1,H_2) : q_r(H_1) \geq q_r(H_2)\}$ and $\mathcal{H}_{0,<} = \{(H_1,H_2) : q_r(H_1) < q_r(H_2)\}$. We will only show $\mathtt{E}_{\geq}[r_{\text{loss}}] \leq \log_2\left(1 + D_0 \times P \times \max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta\right\}\right)$, and skip the proof for $\mathtt{E}_{<}[r_{\text{loss}}]$ due to similarity. Note that $q_r(H_1) \geq q_r(H_2)$ does not necessarily mean $H_1 \geq H_2$, since it is possible that $q_r(H_1) = q_r(H_2)$ and $H_1 < H_2$. When $q_r(H_1) \geq q_r(H_2)$, define

$$\mathsf{snr}_{\max} = \begin{cases} \alpha^{\star} H_1 = g_{\geq}(H_1,H_2), & \text{if } H_1 \geq H_2, \\ \alpha^{\star} H_2 = g_{<}(H_1,H_2), & \text{if } H_1 < H_2, \end{cases} \tag{18}$$

$$\mathsf{snr}_{q_r} = \alpha_{q_r} \times q_r(H_1) = g_{\geq}\left(q_r(H_1), q_r(H_2)\right), \mathsf{snr}_{\text{loss}} = \mathsf{snr}_{\max} - \mathsf{snr}_{q_r}.$$

where $g_{\geq}(x,y) = \frac{2xy}{\sqrt{(x+y)^2+4xy^2P}+x+y}$ and $g_{<}(x,y) = \frac{2xy}{\sqrt{(x+y)^2+4x^2yP}+x+y}$. Then, we have $r_{\text{loss}} = \log_2(1+P\times\mathsf{snr}_{\max}) - \log_2\left(1+P\times\mathsf{snr}_{q_r}\right) = \log_2\left(1+P\frac{\mathsf{snr}_{\text{loss}}}{1+P\times\mathsf{snr}_{q_r}}\right) \leq \log_2(1+P\times\mathsf{snr}_{\text{loss}})$. Grounded on this, the main steps of the proof are listed as follows:

(1) Partition $\mathcal{H}_{0,\geq}$ into the following mutually disjoint sub-regions $\mathcal{H}_1,\ldots,\mathcal{H}_4$:

$$\mathcal{H}_1 = \{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), H_1 < T\Delta, H_2 < T\Delta, H_1 < \Delta \text{ or } H_2 < \Delta\},$$
$$\mathcal{H}_2 = \{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), H_1 \geq H_2, \Delta \leq H_1 < T\Delta, \Delta \leq H_2 < T\Delta\}$$
$$\mathcal{H}_3 = \{(H_1,H_2) : q_r(H_1) = q_r(H_2), H_1 < H_2, \Delta \leq H_1 < T\Delta, \Delta \leq H_2 < T\Delta\}$$
$$\mathcal{H}_4 = \{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), H_1 \geq T\Delta \text{ or } H_2 \geq T\Delta\}.$$

Here, $\mathcal{H}_1$ and $\mathcal{H}_4$ are edge regions where $H_i < \Delta$ or $H_i \geq T\Delta$; $\mathcal{H}_2$ and $\mathcal{H}_3$ are the dominant regions where $\Delta \leq H_i < T\Delta$. It can be verified that $\mathcal{H}_i \cap \mathcal{H}_j = \emptyset$ for $i \neq j$, and $\mathcal{H}_{0,\geq} = \bigcup_{i=1}^{4} \mathcal{H}_i$.

(2) Let $\mathcal{E}_i = \int_{\mathcal{H}_i} \mathsf{snr}_{\text{loss}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i$. Then, $\mathtt{E}_{\geq}[\mathsf{snr}_{\text{loss}}] = \sum_{i=1}^{4} \mathcal{E}_i$. Prove $\mathcal{E}_i \leq D_i \times \max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta\right\}$ for $i = 1,\ldots,4$.

(3) After Steps (1) and (2), we obtain $\mathtt{E}_{\geq}[\mathsf{snr}_{\text{loss}}] \leq D_0 \times \max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta\right\}$. Based on Jensen's inequality, we have

$$\mathtt{E}_{\geq}[r_{\text{loss}}] \leq \mathtt{E}_{\geq}[\log_2(1+P\times\mathsf{snr}_{\text{loss}})] \leq \log_2(1+P\times\mathtt{E}_{\geq}[\mathsf{snr}_{\text{loss}}]) \leq \log_2\left(1+D_0\times P\times\max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta\right\}\right).$$

Now, we only need to show the upper bound on $\mathcal{E}_i$ in Step (2).

For $\mathcal{E}_1$, since $\mathcal{H}_1 \subseteq \{(H_1,H_2) : H_2 \leq \Delta\}$ and $\mathsf{snr}_{\text{loss}} \leq \mathsf{snr}_{\max} \leq H_1$, we obtain

$$\mathcal{E}_1 \leq \int_0^{\infty} H_1 \frac{e^{-\frac{H_1}{\lambda_1}}}{\lambda_1}\mathrm{d}H_1 \int_0^{\Delta} \frac{e^{-\frac{H_2}{\lambda_2}}}{\lambda_2}\mathrm{d}H_2 = \lambda_1\left(1-e^{-\frac{\Delta}{\lambda_2}}\right) \leq \lambda_1 \times \frac{\Delta}{\lambda_2} = D_1 \times \Delta,$$

where the last inequality follows since $1-e^{-x} \leq x$ for $x \geq 0$.

For $\mathscr{E}_2$, since $H_1 \geq H_2$ and $q_r(H_i) \leq H_i \leq q_r(H_i) + \Delta$ for $H_i \leq T\Delta$, we upper-bound $\mathsf{snr}_{\text{loss}}$ by

$$\begin{aligned}
\mathsf{snr}_{\text{loss}} = & \frac{2H_1H_2}{\underbrace{\sqrt{(H_1+H_2)^2+4H_1H_2^2P}}_{=\Upsilon}+(H_1+H_2)} \\
& - \frac{2q_r(H_1)\,q_r(H_2)}{\underbrace{\sqrt{[q_r(H_1)+q_r(H_2)]^2+4q_r(H_1)\,q_r^2(H_2)\,P}+[q_r(H_1)+q_r(H_2)]}_{\leq \Upsilon+H_1+H_2}} \\
& \leq 2\frac{H_1H_2-q_r(H_1)\,q_r(H_2)}{\Upsilon+H_1+H_2} \leq 2\frac{H_1H_2-(H_1-\Delta)(H_2-\Delta)}{\Upsilon+H_1+H_2} = 2\Delta\frac{H_1+H_2-\Delta}{\Upsilon+H_1+H_2} \leq 2\Delta. \qquad (19)
\end{aligned}$$

Then, an upper bound on $\mathscr{E}_2$ can be $\mathscr{E}_2 \leq 2\Delta\int_{\mathscr{H}_2}\prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i \leq 2\Delta = D_2 \times \Delta$.

For $\mathscr{E}_3$, we have $q_r(H_1) = q_r(H_2) \leq H_1 < H_2$ and $q_r(H_i) \leq H_i \leq q_r(H_i) + \Delta$ hold for $(H_1, H_2) \in \mathscr{H}_3$. Similar to (19), we can also obtain $\mathsf{snr}_{\text{loss}} \leq 2\Delta$ and $\mathscr{E}_3 \leq D_3 \times \Delta$.

For $\mathscr{E}_4$, since $\mathscr{H}_4 \subseteq \{(H_1,H_2) : H_1 > T\Delta\}$ and $\mathsf{snr}_{\text{loss}} \leq \mathsf{snr}_{\max} \leq H_2$, the upper-bound on $\mathscr{E}_4$ can be $\mathscr{E}_4 \leq \int_{T\Delta}^{\infty} f_{H_1}(H_1)\mathrm{d}H_1 \int_0^{\infty} H_2 f_{H_2}(H_2)\mathrm{d}H_2 = \int_{T\Delta}^{\infty} \frac{e^{-\frac{H_1}{\lambda_1}}}{\lambda_1}\mathrm{d}H_1 \int_0^{\infty} H_2 \frac{e^{-\frac{H_2}{\lambda_2}}}{\lambda_2}\mathrm{d}H_2 = \lambda_2 e^{-\frac{T\Delta}{\lambda_1}} = D_4 \times e^{-\frac{T\Delta}{\lambda_1}}$. We have accomplished Step (2) and the proof of (6) is complete. ■

## Appendix B: Proof of Lemma 3

When the uniform quantizer $q_o(\cdot)$ is applied, the outage probability loss in (4) is rewritten as

$$\mathtt{out}_{\text{loss},q_o} = \underbrace{\int_{I_{0,\geq}} \mathbf{1}_{\min\{r_1(\alpha_{q_o}),r_2(\alpha_{q_o})\}<r_{\text{th}}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i}_{=\mathtt{out}_{\geq,\text{loss},q_o}} + \underbrace{\int_{I_{0,<}} \mathbf{1}_{\min\{r_1(\alpha_{q_o}),r_2(\alpha_{q_o})\}<r_{\text{th}}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i}_{=\mathtt{out}_{<,\text{loss},q_o}}.$$

where

$$\begin{aligned}
I_{0,\geq} &= \{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), r_{\max} = \log_2(1+P\times\mathsf{snr}_{\max}) \geq r_{\text{th}}\} \\
&= \left\{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), \mathsf{snr}_{\max} \geq \tfrac{\beta}{P} = \tfrac{2^{r_{\text{th}}}-1}{P}\right\}, \\
I_{0,<} &= \left\{(H_1,H_2) : q_r(H_1) < q_r(H_2), \mathsf{snr}_{\max} < \tfrac{\beta}{P}\right\}.
\end{aligned}$$

and $\mathsf{snr}_{\max}$ is defined in (18). We show $\mathtt{out}_{\geq,\text{loss},q_o} \leq D_5 \times e^{-\frac{D_6}{P}} \times \frac{1+\sqrt{P}}{P} \times \max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta^{\frac{1}{2}}, \Delta^{\frac{3}{2}}\right\}$ and skip the proof for $\mathtt{out}_{<,\text{loss},q_o}$ due to similarity. The main steps of the proof are:

(1) Partition $\mathcal{I}_{0,\geq}$ into the following mutually disjoint sub-regions:

$$\begin{aligned}
\mathcal{I}_1 &= \left\{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), \mathsf{snr}_{\max} \geq \tfrac{\beta}{P}, H_1 \leq \Delta, H_2 \leq \Delta\right\},\\
\mathcal{I}_2 &= \left\{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), \mathsf{snr}_{\max} = g_{\geq}(H_1,H_2) \geq \tfrac{\beta}{P}, \Delta < H_1 \leq T\Delta, H_2 \leq \Delta\right\},\\
\mathcal{I}_3 &= \left\{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), H_1 \geq H_2, g_{\geq}(H_1,H_2) \geq \tfrac{\beta}{P}, \Delta < H_1 \leq T\Delta, \Delta < H_2 \leq T\Delta\right\},\\
\mathcal{I}_4 &= \left\{(H_1,H_2) : q_r(H_1) = q_r(H_2), H_1 < H_2, g_{<}(H_1,H_2) \geq \tfrac{\beta}{P}, \Delta < H_1 \leq T\Delta, \Delta < H_2 \leq T\Delta\right\},\\
\mathcal{I}_5 &= \left\{(H_1,H_2) : q_r(H_1) \geq q_r(H_2), \mathsf{snr}_{\max} \geq \tfrac{\beta}{P}, H_1 > T\Delta \text{ or } H_2 > T\Delta\right\}.
\end{aligned}$$

Here, $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_5$ are the marginal regions where $H_i \leq \Delta$ or $H_i > T\Delta$; $\mathcal{I}_3$ and $\mathcal{I}_4$ are the main regions where $\Delta < H_i \leq T\Delta$. It can be verified that $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset$ for $i \neq j$, and $\mathcal{I}_{0,\geq} = \bigcup_{i=1}^{5} \mathcal{I}_i$.

(2) Let $\mathscr{F}_i = \int_{\mathcal{I}_i} \mathbf{1}_{\min\{r_1(\alpha_{q_o}), r_2(\alpha_{q_o})\} < r_{\text{th}}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i$. Then, $\texttt{out}_{\geq,\text{loss},q_o} = \sum_{i=1}^{5} \mathscr{F}_i$. Prove $\mathscr{F}_i \leq D_{2i+5} \times e^{-\frac{D_{2i+6}}{P}} \times \frac{1+\sqrt{P}}{P} \times \max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta^{\frac{1}{2}}, \Delta^{\frac{3}{2}}\right\}$ for $i = 1, \ldots, 5$.

Now, we need to show the upper bound on $\mathscr{F}_i$ in Step (2).

For $\mathscr{F}_1$, we have $q_o(H_1) = q_o(H_2) = \Delta \geq H_2$, and thus, $\alpha_{q_o} = \frac{1}{\sqrt{P\Delta+1}+1} \leq \frac{1}{\sqrt{PH_2+1}+1}$. For any $(H_1,H_2) \in \mathcal{I}_1$, since $g_{\geq}(x,y) \leq \min\{x,y\}$ and $g_{<}(x,y) \leq \min\{x,y\}$, it must have $\frac{\beta}{P} \leq \mathsf{snr}_{\max} \leq \min\{H_1,H_2\}$. Moreover, we obtain $\mathbf{1}_{\min\{r_1(\alpha_{q_o}), r_2(\alpha_{q_o})\} < r_{\text{th}}} \leq \mathbf{1}_{r_1(\alpha_{q_o}) < r_{\text{th}}} + \mathbf{1}_{r_2(\alpha_{q_o}) < r_{\text{th}}}$, and

$$\begin{aligned}
\mathbf{1}_{r_1(\alpha_{q_o}) < r_{\text{th}}} &= \mathbf{1}_{H_1 \times \alpha_{q_o} < \frac{\beta}{P}} = \mathbf{1}_{H_1 < \beta\frac{\sqrt{P\Delta+1}+1}{P}},\\
\mathbf{1}_{r_2(\alpha_{q_o}) < r_{\text{th}}} &= \mathbf{1}_{\frac{H_2(1-\alpha_{q_o})}{PH_2\alpha_{q_o}+1} < \frac{\beta}{P}} \leq \mathbf{1}_{\frac{H_2\left(1-\frac{1}{\sqrt{PH_2+1}+1}\right)}{PH_2 \times \frac{1}{\sqrt{PH_2+1}+1}+1} < \frac{\beta}{P}} = \mathbf{1}_{H_2 < \frac{\beta^2+2\beta}{P}}.
\end{aligned}$$

Thus, an upper bound on $\mathscr{F}_1$ is

$$\begin{aligned}
\mathscr{F}_1 &\leq \int_{\mathcal{I}_1} \mathbf{1}_{H_1 < \beta\frac{\sqrt{P\Delta+1}+1}{P}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i + \int_{\mathcal{I}_1} \mathbf{1}_{H_2 < \frac{\beta^2+2\beta}{P}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i\\
&\leq \int_{\frac{\beta}{P}}^{\beta\frac{\sqrt{P\Delta+1}+1}{P}} \frac{e^{-\frac{H_1}{\lambda_1}}}{\lambda_1} \int_{\frac{\beta}{P}}^{\Delta} \frac{e^{-\frac{H_2}{\lambda_2}}}{\lambda_2} \mathrm{d}H_1\mathrm{d}H_2 + \int_{\frac{\beta}{P}}^{\Delta} \frac{e^{-\frac{H_1}{\lambda_1}}}{\lambda_1} \int_{\frac{\beta}{P}}^{\frac{\beta^2+2\beta}{P}} \frac{e^{-\frac{H_2}{\lambda_2}}}{\lambda_2} \mathrm{d}H_1\mathrm{d}H_2\\
&\leq \frac{e^{-\frac{\frac{\beta}{P}}{\lambda_1}}}{\lambda_1} \times \left[\beta\frac{\sqrt{P\Delta+1}+1}{P} - \frac{\beta}{P}\right] \times \frac{1}{\lambda_2} \times \left[\Delta - \frac{\beta}{P}\right] + \frac{1}{\lambda_1} \times \left[\Delta - \frac{\beta}{P}\right] \times \frac{e^{-\frac{\frac{\beta}{P}}{\lambda_2}}}{\lambda_2} \times \left[\frac{\beta^2+2\beta}{P} - \frac{\beta}{P}\right]\\
&\leq \frac{e^{-\frac{\frac{\beta}{P}}{\lambda_1}}}{\lambda_1} \times \beta \times \frac{\overbrace{\sqrt{P\Delta+1}}^{\leq\sqrt{P\Delta}+1}}{P} \times \frac{1}{\lambda_2} \times \Delta + \frac{1}{\lambda_1} \times \Delta \times \frac{e^{-\frac{\frac{\beta}{P}}{\lambda_2}}}{\lambda_2} \times \frac{\beta^2+\beta}{P}
\end{aligned}$$

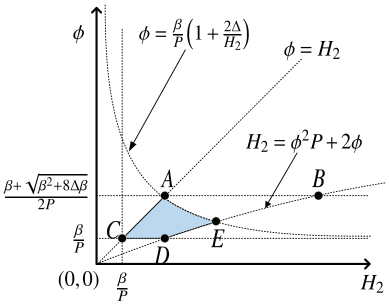
$\phi$
$\phi = \frac{\beta}{P}\left(1+\frac{2\Delta}{H_2}\right)$
$\phi = H_2$
$H_2 = \phi^2 P + 2\phi$
$A$
$B$
$\frac{\beta+\sqrt{\beta^2+8\Delta\beta}}{2P}$
$\frac{\beta}{P}$
$C$
$D$
$E$
$(0,0)$
$\frac{\beta}{P}$
$H_2$

Fig. 12: The integration region $I_2^{''}$.

$$\leq D_{17} \times e^{-\frac{D_{18}}{P}} \times \frac{\sqrt{P\Delta}+1}{P} \times \Delta + D_{19} \times e^{-\frac{D_{20}}{P}} \times \frac{\Delta}{P} \leq D_7 \times e^{-\frac{D_8}{P}} \times \frac{1+\sqrt{P}}{P} \times \max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta^{\frac{1}{2}}, \Delta^{\frac{3}{2}}\right\}. \tag{20}$$

For $\mathscr{F}_2$, let $\mathscr{F}_{2,i} = \int_{I_2} \mathbf{1}_{r_i(\alpha_{q_o})<r_{\text{th}}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i$ for $i = 1,2$. Then, $\mathscr{F}_2 \leq \mathscr{F}_{2,1} + \mathscr{F}_{2,2}$. For $\mathscr{F}_{2,1}$, since $H_1 > H_2$ for $(H_1,H_2) \in I_2$ and $g_{\geq}(x,y)$ is increasing on $x$ and $y$, we have

$$\mathbf{1}_{r_1(\alpha_{q_o})<r_{\text{th}}} = \mathbf{1}_{\frac{2H_1 \times q_o(H_2)}{\sqrt{[q_o(H_1)+q_o(H_2)]^2+4q_o(H_1)q_o^2(H_2)P}+[q_o(H_1)+q_o(H_2)]}<\frac{\beta}{P}}$$

$$\leq \mathbf{1}_{\frac{2(q_o(H_1)-\Delta)\times q_o(H_2)}{\sqrt{[q_o(H_1)+q_o(H_2)]^2+4q_o(H_1)q_o^2(H_2)P}+[q_o(H_1)+q_o(H_2)]}<\frac{\beta}{P}} = \mathbf{1}_{g_{\geq}(q_o(H_1),q_o(H_2))<\frac{\beta}{P}\times\frac{1}{1-\frac{\Delta}{q_o(H_1)}}} \tag{21}$$

$$\leq \mathbf{1}_{g_{\geq}(q_o(H_1),q_o(H_2))<\frac{\beta}{P}\times\left(1+\frac{2\Delta}{q_o(H_1)}\right)} \leq \mathbf{1}_{g_{\geq}(q_o(H_1),q_o(H_2))<\frac{\beta}{P}\times\left(1+\frac{2\Delta}{q_o(H_2)}\right)} \tag{22}$$

$$\leq \mathbf{1}_{g_{\geq}(q_o(H_1),q_o(H_2))<\frac{\beta}{P}\times\left(1+\frac{2\Delta}{H_2}\right)} \leq \mathbf{1}_{g_{\geq}(H_1,H_2)<\frac{\beta}{P}\times\left(1+\frac{2\Delta}{H_2}\right)}, \tag{23}$$

where (21) follows from $q_o(H_1) \leq H_1 + \Delta$, (22) follows from $\left(1-\frac{\Delta}{q_o(H_1)}\right) \times \left(1+\frac{2\Delta}{q_o(H_1)}\right) \geq 1$ because $q_o(H_1) \geq 2\Delta > q_o(H_2) = \Delta$, and (23) follows from $q_o(H_2) \geq H_2$ and $g_{\geq}(q_o(H_1), q_o(H_2)) \geq g_{\geq}(H_1,H_2)$. Then, we obtain $\mathscr{F}_{2,1} \leq \int_{I_2^{'}=I_2\cap\left\{(H_1,H_2):g_{\geq}(H_1,H_2)<\frac{\beta}{P}\times\left(1+\frac{2\Delta}{H_2}\right)\right\}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i$.

We change the integration variables from $(H_1,H_2)$ to $(\phi,H_2)$ where $\phi = g_{\geq}(H_1,H_2)$. Then, $H_1 = \frac{\phi^2 P+\phi}{H_2-\phi} \times H_2$, and the Jacobian matrix is $\left|\frac{\mathrm{d}H_1}{\mathrm{d}\phi}\right| = \frac{2\phi P H_2+H_2-\phi^2 P}{(H_2-\phi)^2} \times H_2 \leq \frac{2\phi P H_2+H_2}{(H_2-\phi)^2} \times H_2 \leq \frac{2\phi P H_2+2H_2}{(H_2-\phi)^2} \times H_2 = \frac{2(\phi P+1)}{(H_2-\phi)^2} \times H_2^2$. For any $(H_1,H_2) \in I_2^{'}$, we have: (i) $\frac{\beta}{P} \leq \phi = g_{\geq}(H_1,H_2) \leq H_2$ and $\phi < \frac{\beta}{P} \times \left(1+\frac{2\Delta}{H_2}\right)$; (ii) since $H_1 \geq H_2$, $H_1 = \frac{\phi^2 P+\phi}{H_2-\phi} \times H_2 \geq H_2$, then, $H_2 \leq \phi^2 P + 2\phi$. Therefore, $\mathscr{F}_{2,1}$ is derived as $\mathscr{F}_{2,1} \leq \int_{I_2^{''}=\left\{(H_1,H_2):\frac{\beta}{P}\leq H_2\leq\phi^2 P+2\phi,\frac{\beta}{P}\leq\phi\leq\min\left\{H_2,\frac{\beta}{P}\left(1+\frac{2\Delta}{H_2}\right)\right\}\right\}} \prod_{i=1}^{2} f_{H_i}(H_i)\mathrm{d}H_i$. The integration region $I_2^{''}$ is demonstrated in Fig. 12 as the shaded area surrounded by the points $A, E, D$ and $C$. It can be strictly proven that $I_2^{''}$ is within the region surrounded the points $A, B, D$

and $C$. Recall that $H_1 = \frac{\phi^2 P+\phi}{H_2-\phi} \times H_2$ and $\left|\frac{\mathrm{d}H_1}{\mathrm{d}\phi}\right| \leq \frac{2(\phi P+1)}{(H_2-\phi)^2} \times H_2^2$. Then, we have

$$\begin{aligned}
\mathscr{F}_{2,1} &\leq \int_{\frac{\beta}{P}}^{\frac{\beta+\sqrt{\beta^2+8\Delta\beta}}{2P}} \int_{\phi}^{\phi^2 P+2\phi} \frac{e^{-\frac{H_2}{\lambda_2}}}{\lambda_2} \times \frac{e^{-\frac{1}{\lambda_1}\times\frac{\phi^2P+\phi}{H_2-\phi}\times H_2}}{\lambda_1} \times \frac{2\left(\phi P+1\right)}{\left(H_2-\phi\right)^2} \times H_2^2 \mathrm{d}\phi \mathrm{d}H_2 \\
&\overset{z=H_2-\phi}{=} D_{21} \int_{\frac{\beta}{P}}^{\frac{\beta+\sqrt{\beta^2+8\Delta\beta}}{2P}} \int_0^{\phi^2 P+\phi} \underbrace{e^{-\frac{z}{\lambda_2}-\frac{\phi}{\lambda_2}}}_{\leq e^{-\frac{z}{\lambda_2}} \times e^{-\frac{\frac{\beta}{P}}{\lambda_2}}} \times \underbrace{e^{-\frac{1}{\lambda_1}\times\frac{\phi^2P+\phi}{z}\times(z+\phi)}}_{\leq e^{-\frac{\phi^2(\phi P+1)}{\lambda_1 z}}} \times \frac{\phi P+1}{z^2} \times (z+\phi)^2 \mathrm{d}\phi \mathrm{d}z \\
&\leq D_{21} \times e^{-\frac{\beta}{\lambda_2 P}} \int_{\frac{\beta}{P}}^{\frac{\beta+\sqrt{\beta^2+8\Delta\beta}}{2P}} \int_0^{\phi^2 P+\phi} e^{-\frac{z}{\lambda_2}} e^{-\frac{\phi^2(\phi P+1)}{\lambda_1 z}} \times (\phi P+1) \times \left[1+\frac{2\phi}{z}+\frac{\phi^2}{z^2}\right] \mathrm{d}\phi \mathrm{d}z. \qquad (24)
\end{aligned}$$

Using the inequalities: (i) $\int_0^\infty x^{\nu-1} e^{-\frac{\beta}{x}-\gamma x} \mathrm{d}x = 2\left(\frac{\beta}{\gamma}\right)^{\frac{\nu}{2}} \mathscr{K}_\nu\left(2\sqrt{\beta\gamma}\right)$ [28, Eq. (3.471.9)] with $\mathscr{K}_\nu(z)$ being the modified bessel function of the second kind, (ii) $\mathscr{K}_0(x) \leq \frac{2}{x}$ and $\mathscr{K}_{-1}(x) = \mathscr{K}_1(x) \leq \frac{1}{x}$ for $x > 0$ [29, Eq. (27)], after lengthy but basic calculations, we obtain $\mathscr{F}_{2,1} \leq D_{22} \times e^{-\frac{D_{23}}{P}} \times \frac{\Delta+\sqrt{\Delta}}{P}$. Detailed calculations for (24) can be found in Appendix B of [30].

For $\mathscr{F}_{2,2}$, because $H_1 > H_2$ and $q_o(H_1) > q_o(H_2) = \Delta \geq H_2$, we have

$$\begin{aligned}
\alpha_{q_o} &\leq \frac{2q_o(H_2)}{\sqrt{\left[q_o(H_2)+q_o(H_2)\right]^2 + 4q_o(H_2)q_o^2(H_2)P} + q_o(H_2) + q_o(H_2)} \\
&= \frac{1}{\sqrt{q_o(H_2)P+1}+1} = \frac{1}{\sqrt{P\Delta+1}+1}. \qquad (25)
\end{aligned}$$

Since $r_2\left(\alpha_{q_o}\right)$ is decreasing on $\alpha_{q_o}$, we obtain $r_2\left(\alpha_{q_o}\right) \geq r_2\left(\frac{1}{\sqrt{P\Delta+1}+1}\right)$ and $\mathbf{1}_{r_2\left(\alpha_{q_o}\right)<r_{\text{th}}} \leq \mathbf{1}_{r_2\left(\frac{1}{\sqrt{P\Delta+1}+1}\right)<r_{\text{th}}} = \mathbf{1}_{\frac{H_2\left(1-\frac{1}{\sqrt{P\Delta+1}+1}\right)}{PH_2\frac{1}{\sqrt{P\Delta+1}+1}+1}<\frac{\beta}{P}} = \mathbf{1}_{\frac{H_2\sqrt{P\Delta+1}}{PH_2+1+\sqrt{P\Delta+1}}<\frac{\beta}{P}} \leq \mathbf{1}_{\frac{H_2\sqrt{P\Delta+1}}{P\Delta+1+\sqrt{P\Delta+1}}<\frac{\beta}{P}} = \mathbf{1}_{H_2\leq\frac{\beta\left(\sqrt{P\Delta+1}+1\right)}{P}}$. Similar to (20), we will have $\mathscr{F}_{2,2} \leq \int_{I_2} \mathbf{1}_{H_2<\beta\frac{\sqrt{P\Delta+1}+1}{P}} \prod_{i=1}^2 f_{H_i}(H_i)\mathrm{d}H_i \leq D_{24} \times e^{-\frac{D_{25}}{P}} \times \frac{\sqrt{P\Delta}+1}{P} \times \Delta$. Together with the upper bound on $\mathscr{F}_{2,1}$, we obtain $\mathscr{F}_2 \leq \mathscr{F}_{2,1} + \mathscr{F}_{2,2} \leq D_{22} \times e^{-\frac{D_{23}}{P}} \times \frac{\Delta+\sqrt{\Delta}}{P} + D_{24} \times e^{-\frac{D_{25}}{P}} \times \frac{\sqrt{P\Delta}+1}{P} \times \Delta \leq D_9 \times e^{-\frac{D_{10}}{P}} \times \frac{1+\sqrt{P}}{P} \times \max\left\{e^{-\frac{T\Delta}{\lambda_1}}, \Delta^{\frac{1}{2}}, \Delta^{\frac{3}{2}}\right\}$.

For $\mathscr{F}_3$, since $q_o(H_1) \geq q_o(H_2)$ and $q_o(H_i) - \Delta \leq H_i \leq q_o(H_i)$ for $i = 1, 2$, we obtain

$$\begin{aligned}
r_1\left(\alpha_{q_o}\right) &= \log_2\left(1+PH_1\times\alpha_{q_o}\right) \geq \log_2\left(1+P\times\left(q_o(H_1)-\Delta\right)\times\alpha_{q_o}\right) \\
&= \log_2\left(1+P\times q_o(H_1)\times\alpha_{q_o} - P\times\Delta\times\alpha_{q_o}\right) \\
&= \log_2\left(1+P\times g_{\geq}(q_o(H_1),q_o(H_2)) - P\times g_{\geq}(q_o(H_1),q_o(H_2))\times\frac{\Delta}{q_o(H_1)}\right) \\
&= \log_2\left(1+P\times g_{\geq}(q_o(H_1),q_o(H_2))\times\left(1-\frac{\Delta}{q_o(H_1)}\right)\right)
\end{aligned}$$

$$\begin{aligned}
&\geq \log_2\left(1+P\times g_{\geq}(q_o(H_1),q_o(H_2))\times\left(1-\frac{\Delta}{q_o(H_2)}\right)\right)\\
&\geq \log_2\left(1+P\times g_{\geq}(H_1,H_2)\times\left(1-\frac{\Delta}{q_o(H_2)}\right)\right),\\
r_2\left(\alpha_{q_o}\right) &= \log_2\left(1+\frac{H_2\left(1-\alpha_{q_o}\right)}{H_2\alpha_{q_o}+\frac{1}{P}}\right) = \log_2\left(1+\frac{(q_o(H_2)-\Delta)\times(1-\alpha_{q_o})}{(q_o(H_2)-\Delta)\times\alpha_{q_o}+\frac{1}{P}}\right) \qquad (26)\\
&\geq \log_2\left(1+\frac{(q_o(H_2)-\Delta)\times(1-\alpha_{q_o})}{q_o(H_2)\times\alpha_{q_o}+\frac{1}{P}}\right) = \log_2\left(1+\frac{q_o(H_2)\times(1-\alpha_{q_o})}{q_o(H_2)\times\alpha_{q_o}+\frac{1}{P}}-\frac{\Delta\times(1-\alpha_{q_o})}{q_o(H_2)\times\alpha_{q_o}+\frac{1}{P}}\right)\\
&= \log_2\left(1+P\times g_{\geq}(q_o(H_1),q_o(H_2))\times\left(1-\frac{\Delta}{q_o(H_2)}\right)\right)\\
&\geq \log_2\left(1+P\times g_{\geq}(H_1,H_2)\times\left(1-\frac{\Delta}{q_o(H_2)}\right)\right),
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathbf{1}_{\min\left\{r_1\left(\alpha_{q_o}\right),r_2\left(\alpha_{q_o}\right)\right\}<r_{\text{th}}} &\leq \mathbf{1}_{\log_2\left(1+P\times g_{\geq}(H_1,H_2)\times\left(1-\frac{\Delta}{q_o(H_2)}\right)\right)<r_{\text{th}}} = \mathbf{1}_{g_{\geq}(H_1,H_2)<\frac{\beta}{P\left(1-\frac{\Delta}{q_o(H_2)}\right)}}\\
&\leq \mathbf{1}_{g_{\geq}(H_1,H_2)<\frac{\beta}{P}\left(1+\frac{2\Delta}{q_o(H_2)}\right)} \leq \mathbf{1}_{g_{\geq}(H_1,H_2)<\frac{\beta}{P}\left(1+\frac{2\Delta}{H_2}\right)}, \qquad (27)
\end{aligned}$$

where (27) is because $\left(1-\frac{\Delta}{q_o(H_2)}\right)\times\left(1+\frac{2\Delta}{q_o(H_2)}\right)=1+\frac{\Delta}{q_o(H_2)}-2\left(\frac{\Delta}{q_o(H_2)}\right)^2\geq 1$ since $q_o(H_2)\geq 2\Delta$ for $(H_1,H_2)\in I_3$, and $q_o(H_2)\geq H_2$. Similar to (23) and (24), we can obtain an upper bound on $\mathscr{F}_3$ (the detailed derivation is omitted due to similarity). For $\mathscr{F}_4$, its upper bound can be developed in the same way as the upper bound on $\mathscr{F}_3$.

For $\mathscr{F}_5$, when $H_1\geq H_2\geq\Delta$, since $g_{\geq}(H_1,H_2)\geq\frac{2H_1H_2}{\sqrt{(H_1+H_1)^2+4H_1^2H_2P}+H_1+H_1}=\frac{H_2}{\sqrt{PH_2+1}+1}$, we obtain from (27) that

$$\mathbf{1}_{\min\left\{r_1\left(\alpha_{q_o}\right),r_2\left(\alpha_{q_o}\right)\right\}<r_{\text{th}}} \leq \mathbf{1}_{g_{\geq}(H_1,H_2)<\frac{\beta}{P}\left(1+\frac{2\Delta}{H_2}\right)} \leq \mathbf{1}_{g_{\geq}(H_1,H_2)<\frac{\beta}{P}\left(1+\frac{2\Delta}{\Delta}\right)=\frac{3\beta}{P}} \leq \mathbf{1}_{\frac{H_2}{\sqrt{1+H_2P}+1}<\frac{3\beta}{P}} = \mathbf{1}_{H_2<\frac{D_{26}}{P}}, \qquad (28)$$

where $D_{26}=(3\beta+1)^2-1$. Similarly, when $H_1<H_2$, we have $\mathbf{1}_{\min\left\{r_1\left(\alpha_{q_o}\right),r_2\left(\alpha_{q_o}\right)\right\}<r_{\text{th}}}\leq\mathbf{1}_{H_1<\frac{D_{26}}{P}}$. Therefore, an upper bound on $\mathscr{F}_5$ is

$$\begin{aligned}
\mathscr{F}_5 &\leq \int_{I_4\cap\{(H_1,H_2):H_1\geq H_2\}}\mathbf{1}_{H_2<\frac{D_{26}}{P}}\times\prod_{i=1}^{2}f_{H_i}(H_i)\mathrm{d}H_i+\int_{I_4\cap\{(H_1,H_2):H_1<H_2\}}\mathbf{1}_{H_1<\frac{D_{26}}{P}}\times\prod_{i=1}^{2}f_{H_i}(H_i)\mathrm{d}H_i\\
&\leq \underbrace{\int_{T\Delta}^{\infty}\frac{1}{\lambda_1}e^{-\frac{H_1}{\lambda_1}}\mathrm{d}H_1}_{=e^{-\frac{T\Delta}{\lambda_1}}}\int_{\frac{\beta}{P}}^{\frac{D_{26}}{P}}\frac{1}{\lambda_2}\underbrace{e^{-\frac{H_2}{\lambda_2}}}_{\leq e^{-\frac{\beta}{P\lambda_2}}\leq e^{-\frac{\beta}{P\lambda_1}}}\mathrm{d}H_2+\underbrace{\int_{T\Delta}^{\infty}\frac{1}{\lambda_2}e^{-\frac{H_2}{\lambda_2}}\mathrm{d}H_2}_{=e^{-\frac{T\Delta}{\lambda_2}}\leq e^{-\frac{T\Delta}{\lambda_1}}}\int_{\frac{\beta}{P}}^{\frac{D_{26}}{P}}\frac{1}{\lambda_1}\underbrace{e^{-\frac{H_1}{\lambda_1}}}_{\leq e^{-\frac{\beta}{P\lambda_1}}}\mathrm{d}H_1 \qquad (29)
\end{aligned}$$

$$\leq e^{-\frac{T\Delta}{\lambda_1}} \times \frac{1}{\lambda_2} \times e^{-\frac{\beta}{P\lambda_1}} \times \frac{D_{26}-\beta}{P} + e^{-\frac{T\Delta}{\lambda_1}} \times \frac{1}{\lambda_1} \times e^{-\frac{\beta}{P\lambda_1}} \times \frac{D_{26}-\beta}{P} \leq D_{27} \times e^{-\frac{D_{28}}{P}} \times \frac{1}{P} \times e^{-\frac{T\Delta}{\lambda_1}}$$

$$\leq D_{15} \times e^{-\frac{D_{16}}{P}} \times \frac{1+\sqrt{P}}{P} \times \max\left\{ e^{-\frac{T\Delta}{\lambda_1}}, \Delta^{\frac{1}{2}}, \Delta^{\frac{3}{2}} \right\},$$

where (29) is based on the assumption that $\lambda_1 \geq \lambda_2$. This completes the proof of the upper bound on $\texttt{out}_{\text{loss},q_o}$ in (12). ■

## Appendix C: Proof of Lemma 4

It is trivial to obtain the maximum diversity order for both receivers is 1 in the full-CSI case.[13] When $q_o(\cdot)$ is employed, the outage probability of Receiver $i$ is $\texttt{out}_{q_o,i} = \int \mathbf{1}_{r_i(\alpha_{q_o})<r_{\text{th}}} \prod_{i=1}^{2} f_{H_i}(H_i) \mathrm{d}H_i$ for $i = 1,2$. Following the derivations of $\mathscr{F}_i$ for $i = 1,\ldots,5$ in Appendix B, we will obtain $\texttt{out}_{q_o,1} \leq \texttt{out}_{\min} + D_{29} \times e^{-\frac{D_{30}}{P}} \times \left[ \frac{\sqrt{\Delta}+e^{-\frac{T\Delta}{\lambda_1}}}{P} + \frac{\Delta^{\frac{3}{2}}}{\sqrt{P}} \right]$ and $\texttt{out}_{q_o,2} \leq \texttt{out}_{\min} + D_{31} \times e^{-\frac{D_{32}}{P}} \times \frac{D_{33}+\Delta+e^{-\frac{T\Delta}{\lambda_1}}}{P}$.[14] Therefore, for fixed $\Delta$, the diversity orders of $\frac{1}{2}$ and 1 are achievable for Receivers 1 and 2, respectively.

For Receiver 1, $\frac{\Delta^{\frac{3}{2}}}{\sqrt{P}}$ in the upper bound on $\texttt{out}_{q_o,1}$ is the bottleneck for diversity gains. If we scale $\Delta$ as $\Delta^{\frac{3}{2}} \sim_P \frac{1}{\sqrt{P}}$, i.e., $\Delta \sim_P P^{-\frac{1}{3}}$, the diversity order of 1 is also achievable for Receiver 1. ■

## Appendix D: Proof of Lemma 5

Given $K$ and $\beta > 0$, define the following two optimization problems:

**(P1)** $r^{\star}_{\max}(K,\beta) = \max\limits_{\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_K]} \min\limits_{k=1,\ldots,K} r_k(\boldsymbol{\alpha})$, subject to $0 \leq \alpha_k \leq \beta$ and $\sum_{k=1}^{K} \alpha_k = \beta$.

**(P2)** $r^{\dagger}_{\max}(K,\beta) = \max\limits_{\boldsymbol{\alpha}=[\alpha_1,\ldots,\alpha_K]} \min\limits_{k=1,\ldots,K} r_k(\boldsymbol{\alpha})$, subject to $r_1(\boldsymbol{\alpha}) = \cdots = r_K(\boldsymbol{\alpha})$, $0 \leq \alpha_k \leq \beta$, and $\sum_{k=1}^{K} \alpha_k = \beta$,

where **(P1)** is the original optimization problem in (15) when $\beta = 1$. We will show that the maximum minimum rates of **(P1)** and **(P2)** are the same, i.e., $r^{\star}_{\max}(K,\beta) = r^{\dagger}_{\max}(K,\beta)$, which proves the lemma.

[13] Detailed derivations for the maximum diversity order can be found in Appendix C of [30].

[14] Note that when we derive the diversity order for $\mathscr{F}_{2,2}$, we will not use its upper bound here. From (25), we obtain $\alpha_{q_o} \leq \frac{1}{\sqrt{P\Delta+1}+1} \leq \frac{1}{\sqrt{PH_2+1}+1}$, and $\mathbf{1}_{r_2(\alpha_{q_o})<r_{\text{th}}} \leq \mathbf{1}_{r_2\left(\frac{1}{\sqrt{PH_2+1}+1}\right)<r_{\text{th}}} = \mathbf{1}_{H_2<\frac{\beta^2+\beta}{P}}$, then, it is trivial to obtain that $\mathscr{F}_{2,2} \leq D_{34} \times \frac{e^{-\frac{D_{35}}{P}}}{P}$.

Denote the optimal power allocations for **(P1)** and **(P2)** by $\boldsymbol{\alpha}_K^\star(\beta)=\left[\alpha_{1,K}^\star(\beta),\ldots,\alpha_{K,K}^\star(\beta)\right]$ and $\boldsymbol{\alpha}_K^\dagger(\beta)=\left[\alpha_{1,K}^\dagger(\beta),\ldots,\alpha_{K,K}^\dagger(\beta)\right]$, respectively. Since $r_{\max}^\star(K,\beta)\geq r_{\max}^\dagger(K,\beta)$, it is sufficient to prove that $r_{\max}^\star(K,\beta)\leq r_{\max}^\dagger(K,\beta)$.

The proof for $K=2$ is provided in the proof of Theorem 1. By induction, assume $r_{\max}^\star(K,\beta)=r_{\max}^\dagger(K,\beta)$ holds for $K=K_1$. When $K=K_1+1$, there are two possibilities:

(i) If $r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)\geq r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\dagger(\beta)\right)$, since $r_{K_1+1}(\boldsymbol{\alpha})=\log_2\left(1+\frac{\alpha_{K_1+1}}{\sum_{i=1}^{K_1}\alpha_i+\frac{1}{PH_{K_1+1}}}\right)=\log_2\left(1+\frac{\alpha_{K_1+1}}{\beta-\alpha_{K_1+1}+\frac{1}{PH_{K_1+1}}}\right)$ for any $\boldsymbol{\alpha}$ satisfying $\sum_{i=1}^{K_1+1}\alpha_i=\beta$, it must have $\alpha_{K_1+1,K_1+1}^\star(\beta)\geq\alpha_{K_1+1,K_1+1}^\dagger(\beta)$, then, $\beta_1=\sum_{k=1}^{K_1}\alpha_{k,K_1+1}^\star(\beta)=\beta-\alpha_{K_1+1,K_1+1}^\star(\beta)\leq\beta-\alpha_{K_1+1,K_1+1}^\dagger(\beta)=\sum_{k=1}^{K_1}\alpha_{k,K_1+1}^\dagger(\beta)=\beta_2$. Next, we obtain

$$\begin{aligned}
r_{\max}^\star(K_1+1,\beta)&=\min\left\{\left\{\min_{k=1,\ldots,K_1}r_k\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)\right\},r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)\right\}\\
&\leq\min\left\{r_{\max}^\star\left(K_1,\beta_1\right),r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)\right\} &(30)\\
&=\min\left\{r_{\max}^\dagger\left(K_1,\beta_1\right),r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)\right\} &(31)\\
&\leq\min\left\{r_{\max}^\dagger\left(K_1,\beta_2\right),r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)\right\} &(32)\\
&=\min\left\{r_{\max}^\dagger\left(K_1+1,\beta\right),r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)\right\} &(33)\\
&=\min\left\{r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\dagger(\beta)\right),r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)\right\}\\
&=r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\dagger(\beta)\right)=r_{\max}^\dagger\left(K_1+1,\beta\right).
\end{aligned}$$

Thus, $r_{\max}^\star(K_1+1,\beta)\leq r_{\max}^\dagger(K_1+1,\beta)$. The inequality (30) is due to the optimality of $r_{\max}^\star\left(K_1,\beta_1\right)$; (31) arises from the assumption that $r_{\max}^\star(K,\beta_1)=r_{\max}^\dagger(K,\beta_1)$ when $K=K_1$; (32) is because $r_{\max}^\dagger(K,\beta)$ is non-decreasing on $\beta$; (33) holds since $r_{\max}^\dagger\left(K_1,\beta_2\right)=r_{\max}^\dagger\left(K_1+1,\beta\right)$.

(ii) If $r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)<r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\dagger(\beta)\right)$, we have $r_{\max}^\star(K_1+1,\beta)\leq r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\star(\beta)\right)<r_{K_1+1}\left(\boldsymbol{\alpha}_{K_1+1}^\dagger(\beta)\right)=r_{\max}^\dagger(K_1+1,\beta)$, which completes the proof of Lemma 5. ■

## REFERENCES

[1] X. Liu and H. Jafarkhani, “Two-user downlink non-orthogonal multiple access with limited feedback,” accepted by *IEEE International Symposium on Information Theory (ISIT)*, 2017.

[2] 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.

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

[4] K. Hoiguchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.*, vol. E98-B, no. 3, pp. 403–414, 2015.

[5] 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.

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

[7] 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.

[8] 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.

[9] J. Choi, "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.

[10] 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.

[11] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE Access*, vol. 4, pp. 6325–6343, 2016.

[12] 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.

[13] S. Liu and C. Zhang, "Downlink non-orthogonal multiple access system with limited feedback channel," in *International Conference on Wireless Communications Signal Processing (WCSP)*, Oct. 2015, pp. 1–5.

[14] 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.

[15] 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.

[16] 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.

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

[18] Y. Yu, H. Chen, Y. Li, Z. Ding, and B. Vucetic, "Antenna selection for mimo non-orthogonal multiple access systems,," https://arxiv.org/pdf/1609.07978v1.pdf, 2016.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] H. Jafarkhani, *Space-Time Coding: Theory and Practice*, 1st ed. New York, NY, USA: Cambridge University Press, 2005.

[27] Z. Chen, Z. Ding, X. Dai, and R. Zhang, “A mathematical proof of the superiority of NOMA compared to conventional OMA,” https://arxiv.org/pdf/1612.01069.pdf, 2016.

[28] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed., A. Jeffrey and D. Zwillinger, Eds. Academic Press, Mar. 2007.

[29] E. Koyuncu, Y. Jing, and H. Jafarkhani, “Distributed beamforming in wireless relay networks with quantized feedback,” *IEEE J. Select. Areas Commun.*, vol. 26, no. 8, pp. 1429–1439, Oct. 2008.

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