NOMA Meets Finite Resolution Analog Beamforming in Massive MIMO and Millimeter-Wave Networks

Zhiguo Ding, *Senior Member, IEEE*, Linglong Dai, *Senior Member, IEEE*, Robert Schober, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*— Finite resolution analog beamforming (FRAB) has been recognized as an effective approach to reduce hardware costs in massive multiple-input multiple-output and millimeter-wave networks. However, the use of FRAB means that the beamformers are not perfectly aligned with the users' channels and multiple users may be assigned similar or even identifical beamformers. This letter shows how non-orthogonal multiple access (NOMA) can be used to exploit this feature of FRAB, in which a single FRAB-based beamformer is shared by multiple users. Both analytical and simulation results are provided to demonstrate the excellent performance achieved by this new NOMA transmission scheme.**

***Index Terms*— Non-orthogonal multiple access (NOMA), massive MIMO, millimeter-wave (mmWave) networks, finite resolution analog beamforming (FRAB).**

I. Introduction

NON-ORTHOGONAL multiple access (NOMA) is a promising multiple access technique for next generation wireless networks [1]–[3], and has been shown to be compatible with many important 5G technologies, including massive multiple-input multiple-output (MIMO) and millimeter wave (mmWave) transmission [4]–[6].

A recent development in massive MIMO and mmWave networks is the use of finite resolution analog beamforming (FRAB), which reduces hardware costs significantly compared to purely digital precoding [7], [8]. Analog beamforming does not alter the amplitude of a signal, but modifies its phase only, which is different from digital beamforming. Analog beamforming can be either used alone by a transmitter or combined with conventional digital beamforming, which results in so-called hybrid precoding. The finite resolution constraint on analog beamforming is due to the fact that the number of phase shifts supported by a practical circuit is finite [7], [9]. An example for one-bit resolution analog beamforming is provided in Table 1, where $\mathcal{S}_1$ and $\mathcal{S}_2$ denote two distinct subsets of users to be described in more detail below. Depending on the values of a user's complex-valued channel coefficients, 1 or -1 will be chosen as the beamformer coefficients, as shown in Table I. The reduced hardware costs of FRAB are at the expense of performance losses since the obtained beamformers are not perfectly aligned with the target users' channels.

TABLE I

An Example for Finite Resolution Analog Beamforming

	user 1 in $\mathcal{S}_1$	user 2 in $\mathcal{S}_1$	user 1 in $\mathcal{S}_2$	user 2 in $\mathcal{S}_2$
channel vectors	-0.19 + 0.66j	-0.49 + 0.16j	-0.27 - 0.11j	-0.33 + 0.25j
	-0.06 - 0.53j	-0.35 + 0.22j	-0.06 + 0.58j	-0.45 + 0.10j
	0.34 - 0.03j	-0.10 - 0.62j	0.31 - 0.05j	-0.20 + 0.59j
	0.31 - 0.18j	-0.06 + 0.41j	0.34 - 0.60j	-0.45 - 0.15j
FRAB beam-formers	-1	-1	-1	-1
	-1	-1	-1	-1
	1	- 1	1	-1
	1	- 1	1	-1

The purpose of this letter is to demonstrate that the characteristics of FRAB favor the use of NOMA. Consider again the example shown in Table 1. To clearly illustrate the benefits of the combination of FRAB and NOMA, the users' channel vectors are chosen to be *orthogonal* to each other. Assume that the users in $\mathcal{S}_1$ need to be served first because of their high priority traffic. If digital beamforming with perfect resolution were used, the beamforming vector for user 1 in $\mathcal{S}_1$ would be simply this user's channel vector, and therefore this beamformer could not be used by the two users in $\mathcal{S}_2$, since this beamformer would be orthogonal to these two users' channel vectors. On the other hand, if FRAB is used, the beamformer preferred by user i in $\mathcal{S}_1$ is exactly the same as that of user i in $\mathcal{S}_2$, even though the two users have orthogonal channel vectors. As a result, the use of NOMA ensures that all four users can communicate concurrently.

In this letter, a new NOMA transmission scheme that exploits the features of FRAB is proposed, and analytical results for the corresponding outage probabilities and diversity gains of the users are presented. The provided simulation results demonstrate not only the excellent performance of the proposed NOMA scheme, but also the accuracy of the developed analytical results. We note that the developed analytical results concerning the diversity gains are also applicable to conventional MIMO scenarios without NOMA, and hence shed light on the performance loss caused by FRAB in a general MIMO network.

II. System Model

Consider a NOMA downlink scenario, in which the base station is equipped with M antennas. Assume that there are two groups of single-antenna users in the network. Denote by $\mathcal{S}_1$ a set of users with strict quality of service (QoS)

Manuscript received February 28, 2017; revised April 12, 2017; accepted April 25, 2017. Date of publication May 2, 2017; date of current version August 10, 2017. The work of Z. Ding was supported by the UK EPSRC under grant number EP/L025272/1 and by H2020-MSCA-RISE-2015 under grant number 690750. The work of L. Dai was supported by the National Natural Science Foundation of China (Grant No. 61571270). The work of R. Schober was supported by the Alexander von Humboldt Professorship Program. The work of H. V. Poor was supported by the U.S. National Science Foundation under Grants CCF-1420575 and ECCS-1343210. The associate editor coordinating the review of this letter and approving it for publication was Y. Liu. *(Corresponding author: Zhiguo Ding.)*

Z. Ding is with the School of Computing and Communications, Lancaster University, Lancashire LA1 4WA, U.K.

L. Dai is with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

R. Schober is with the Institute for Digital Communications, University of Erlangen-Nurnberg, D-91058 Erlangen, Germany.

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

Digital Object Identifier 10.1109/LCOMM.2017.2700846

requirements, whose distances to the base station are denoted by d_{yk} and are assumed to be fixed. Denote by $\mathcal{S}_2$ a set of users to be served opportunistically, and these users are uniformly distributed in a disk-shaped area with radius r_1, where the base station is at the center. Denote the distances of the users in $\mathcal{S}_2$ to the base station by d_{xi}. The $M \times 1$ channel vector of a user in $\mathcal{S}_1$ ($\mathcal{S}_2$) is denoted by $\mathbf{h}_k$ ($\mathbf{g}_i$). Two types of channel models are considered in this letter, namely Rayleigh fading which is commonly used to model massive MIMO channels, and the mmWave model [8]. The mmWave channel vector is modelled as follows:

$$\mathbf{h}_k = \frac{a_k}{1+d_{yk}^{\alpha}}\left[1 \; e^{-j\pi\theta_k} \; \cdots e^{-j\pi(M-1)\theta_k}\right]^T. \tag{1}$$

Here, α denotes the path loss exponent, θ_k is the normalized direction, and a_k denotes the fading attenuation coefficient. Note that for the purpose of illustration, only the line-of-sight path is considered for the mmWave model.

A. Implementation of Finite Resolution Analog Beamforming

Suppose that the users in $\mathcal{S}_1$ are served via FRAB. Denote by $\mathbf{f}_k$ the $M \times 1$ beamforming vector for user k, where each element of $\mathbf{f}_k$ is drawn from the following vector:

$$\bar{\mathbf{f}} = \left[1 \; e^{j\frac{2\pi}{N_q}} \; \cdots \; e^{j\frac{(N_q-1)2\pi}{N_q}}\right], \tag{2}$$

where N_q denotes the number of supported phase shifts.

The i-th element of $\bar{\mathbf{f}}$ is chosen as the m-th element of $\mathbf{f}_k$ based on the following criterion:

$$i_{k,m}^* = \arg\min_{i\in\{1,\cdots,N_q\}} \left|\bar{f}_i - \frac{h_{k,m}}{|h_{k,m}|}\right|^2, \tag{3}$$

where $\bar{f}_i$ denotes the i-th element of $\bar{\mathbf{f}}$, and $h_{k,m}$ denotes the m-th element of user k's channel vector.

B. Implementation of NOMA

To reduce the system complexity, suppose that only one user from $\mathcal{S}_2$ will be paired with user k from $\mathcal{S}_1$ and denote this user by user i_k^*. The base station sends a superposition of the messages of the two users on each beam. User k in $\mathcal{S}_1$ treats its partner's message as noise and decodes its own message with the following signal-to-interference-plus-noise ratio (SINR):

$$\mathrm{SINR}_k = \frac{|\mathbf{h}_k^H\mathbf{f}_k|^2\alpha_{0,k}^2}{|\mathbf{h}_k^H\mathbf{f}_k|^2\alpha_{1,k}^2 + \sum\limits_{l\in\mathcal{S}_1\backslash k}|\mathbf{h}_k^H\mathbf{f}_l|^2 + \frac{M}{\rho}}, \tag{4}$$

where the factor M in the denominator is due to the transmit power normalization, and the power allocation coefficients are denoted by $\alpha_{n,k}$. Note that $\sum_{n=0}^{1}\alpha_{n,k}^2 = 1$ and $\alpha_{0,k} \geq \alpha_{1,k}$.

By applying successive interference cancellation (SIC), user i_k^* can decode its partner's message with the following SINR: $\mathrm{SINR}_k^{k\to i_k^*} = \frac{|\mathbf{g}_{i_k^*}^H\mathbf{f}_k|^2\alpha_{0,k}^2}{|\mathbf{g}_{i_k^*}^H\mathbf{f}_k|^2\alpha_{1,k}^2 + \sum_{l\in\mathcal{S}_1\backslash k}|\mathbf{g}_{i_k^*}^H\mathbf{f}_l|^2 + \frac{M}{\rho}}$. Let $\epsilon_i = 2^{R_i}-1$, $i \in \{0,1\}$, where R_0 and R_1 denote the targeted data rates for user k and user i_k^*, respectively. If $\mathrm{SINR}_k^{k\to i_k^*} \geq \epsilon_0$, SIC can be carried out successfully at user i_k^* and the SINR for decoding its own message is given by

$$\mathrm{SINR}_k^{i_k^*} = \frac{|\mathbf{g}_{i_k^*}^H\mathbf{f}_k|^2\alpha_{1,k}^2}{\sum\limits_{l\in\mathcal{S}_1\backslash k}|\mathbf{g}_{i_k^*}^H\mathbf{f}_l|^2 + \frac{M}{\rho}}. \tag{5}$$

We use the following user selection criterion:

$$i_k^* = \arg\max\{\mathrm{SINR}_k^{k\to 1}, \cdots, \mathrm{SINR}_k^{k\to|\mathcal{S}_2|}\}. \tag{6}$$

Note that this criterion selects that user which maximizes the probability of successful intra-NOMA interference cancellation, a key stage for SIC. Since the users in $\mathcal{S}_2$ are served opportunistically, we allow one user from $\mathcal{S}_2$ to be included in more than one pair. More sophisticated user scheduling algorithms can be designed to realize fairness for the users in $\mathcal{S}_2$; these are not presented here due to space limitations.

III. Performance Analysis

To the best of the authors' knowledge, the impact of FRAB on the diversity gain has not been analyzed yet, not even for scenarios without NOMA. In order to obtain insight into the performance of the proposed NOMA scheme, in this section, we focus on the special case of $N_q = 2$, $|\mathcal{S}_1| = 1$, and Rayleigh fading channels. Note that $N_q = 2$ represents the case of one-bit resolution analog beamforming [9].

A. Performance of the User in $\mathcal{S}_1$

When there is a single beam, i.e., $|\mathcal{S}_1| = 1$, the outage probability achieved by the user in $\mathcal{S}_1$ is given by[1]

$$\mathrm{P}_k^o = \mathrm{P}\left(|\mathbf{h}_k^H\mathbf{f}_k|^2 < \phi_0\right),$$

where $\phi_i = \frac{\frac{M\epsilon_i}{\rho}}{\alpha_{i,k}^2 - \epsilon_i\sum_{n=i+1}^{1}\alpha_{n,k}^2}$, $i \in \{0,1\}$. Note that $\alpha_{i,k}^2 > \epsilon_i\sum_{n=i+1}^{1}\alpha_{n,k}^2$ is assumed in this letter, since otherwise the outage probability is always one. In order to find the cumulative distribution function (CDF) of $|\mathbf{h}_k^H\mathbf{f}_k|^2$, the following proposition is provided first.

Proposition 1: Consider M independent and identically distributed (i.i.d.) random variables, denoted by z_m, each of which follows the folded normal distribution, i.e., z_m is the absolute value of a Gaussian variable with mean 0 and variance $\frac{1}{2}$. The CDF of $z_\Sigma \triangleq \left|\sum_{m=1}^{M} z_m\right|^2$ can be approximated as follows:

$$F_{z_\Sigma}(z) \approx \frac{2^M}{\pi^{\frac{M}{2}}}\frac{z^{\frac{M}{2}}}{M!}, \tag{7}$$

when $z \to 0$.

The following lemma provides an asymptotic approximation for the CDF of the effective channel gains of the user in $\mathcal{S}_1$.

Lemma 1: For user k in $\mathcal{S}_1$, the CDF of its effective channel gain on beam $\mathbf{f}_k$ can be approximated as follows:

$$F_{|\mathbf{h}_k^H\mathbf{f}_k|^2}(y) \approx \frac{2^M\left[y(1+d_{yk}^{\alpha})\right]^{\frac{M+1}{2}} B\left(\frac{3}{2},\frac{M}{2}\right)}{\pi^{\frac{M}{2}}(M-1)!M^{\frac{1}{2}}\Gamma\left(\frac{1}{2}\right)}, \tag{8}$$

for $y \to 0$, where $B(\cdot)$ denotes the beta function and $\Gamma(\cdot)$ denotes the gamma function.

By using Lemma 1 and with some algebraic manipulations, the following corollary can be obtained.

Corollary 1: For the proposed NOMA system with Rayleigh fading, user k from $\mathcal{S}_1$ achieves a diversity gain of $\frac{M+1}{2}$.

[1] Since $|\mathcal{S}_1| = 1$, there is only one user in $\mathcal{S}_1$. To keep the notation consistent with that in the previous section, this user is denoted by user k in the remainder of this section.

Remark 1: With M antennas at the base station, the full diversity gain for the considered scenario is M, but only a diversity gain of $\frac{M+1}{2}$ is achieved by the proposed scheme. This performance loss is mainly due to the use of FRAB.

Remark 2: It is important to point out that Corollary 1 is general and applicable to conventional MIMO networks without NOMA as well, since the users in $\mathcal{S}_1$ do not perform SIC.

B. Performance of the Paired User in $\mathcal{S}_2$

When there is a single beam, the outage probability achieved by the user in $\mathcal{S}_2$ paired with user k is $\mathcal{S}_1$ becomes $\mathrm{P}^o_{i_k^*} = 1 - \mathrm{P}\left(|\mathbf{g}^H_{i_k^*}\mathbf{f}_k|^2 > \max(\phi_0, \phi_1)\right)$. For the case of $|\mathcal{S}_1| = 1$, the proposed user selection criterion shown in (6) simplifies to

$$i_k^* = \arg\max\{|\mathbf{g}_1^H\mathbf{f}_k|^2, \cdots, |\mathbf{g}^H_{|\mathcal{S}_2|}\mathbf{f}_k|^2\}, \tag{9}$$

since $f(x) \triangleq \frac{x\alpha^2_{0,k}}{x\alpha^2_{1,k}+\frac{M}{\rho}}$ is a monotonically increasing function of x. Therefore, the CDF of $|\mathbf{g}^H_{i_k^*}\mathbf{f}_k|^2$ can be obtained as follows.

Denote a user randomly chosen from $\mathcal{S}_2$ by user $\pi(i)$. The fading and path loss components of its composite channel gain, $\mathbf{g}_{\pi(i)}$, can be decomposed as $\mathbf{g}_{\pi(i)} = \frac{\bar{\mathbf{g}}_{\pi(i)}}{1+d^{\alpha}_{x\pi(i)}}$. Therefore, the effective fading gain of this user, $|\bar{\mathbf{g}}^H_{\pi(i)}\mathbf{f}_k|^2$, is exponentially distributed, since $\bar{\mathbf{g}}_{\pi(i)}$ and $\mathbf{f}_k$ are independent and a unitary transformation of a Gaussian vector is still Gaussian distributed. It is important to point out that $|\bar{\mathbf{g}}^H_{\pi(i)}\mathbf{f}_k|^2$ is exponentially distributed with parameter $\frac{1}{M}$, instead of 1 as in [10]. Based on this observation and following steps similar to those in [10], the composite channel gain has the following approximate CDF:

$$F_{\pi(i)}(y) = \sum_{n=1}^{N} w_n(1 - e^{-c_n\frac{y}{M}}), \tag{10}$$

where N is a parameter for the Gauss-Chebyshev approximation, $w_n = \frac{\pi}{2N}\sqrt{1-\eta_n^2}(\eta_n+1)$, $\eta_n = \cos\left(\frac{2n-1}{2N}\pi\right)$, and $c_n = 1 + \left(\frac{r_1}{2}\eta_n + \frac{r_1}{2}\right)^{\alpha}$.

After applying the simplified criterion in (9) and assuming that the users' channels are i.i.d., the outage probability of user i_k^* for decoding its message delivered on beam k is $\left(F_{\pi(i)}(\max(\phi_0, \phi_1))\right)^{|\mathcal{S}_2|}$. After some algebraic manipulations, we obtain the following corollary.

Corollary 2: For the proposed NOMA system with Rayleigh fading, the full diversity gain of $|\mathcal{S}_2|$ is achievable by user i_k^* in $\mathcal{S}_2$.

IV. Numerical Results

In this section, the performance of the proposed NOMA-MIMO scheme is evaluated by using computer simulations. The NOMA power coefficients are set as $\alpha^2_{0,k} = 3/4$ and $\alpha^2_{1,k} = 1/4$. The noise power is -30dBm. Without loss of generality, assume that the users in $\mathcal{S}_1$ lie on a circle with radius r_y, and the base station is located at its center.

Note that for the NOMA scheme, two users are served on each beam, and they have their own target data rates, R_0 and R_1, respectively. However, for the baseline scheme without NOMA, a single user is served on each beam. For a fair comparison, the user's targeted data rate is $R_0 + R_1$ for the case without NOMA. In Fig. 1, two types of channel models, namely Rayleigh fading and the mmWave model,

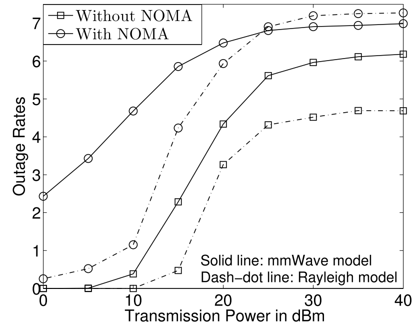

Fig. 1. Outage rates achieved by the considered schemes for different channel models. $M = 30$, $|\mathcal{S}_1| = 3$, $|\mathcal{S}_2| = 300$, $N_q = 2$, $r_1 = 40$m, $r_y = r_1$, $\alpha = 3$, $R_0 = 1$ bit per channel use (BPCU), and $R_1 = 1.5$ BPCU.

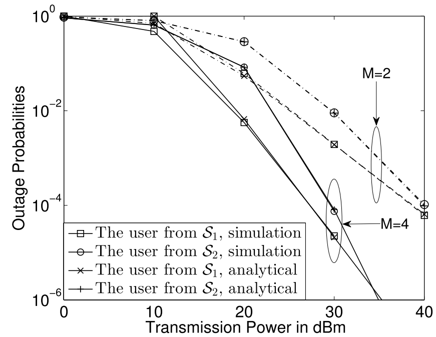

Fig. 2. Outage probabilities achieved by the considered schemes for the Rayleigh fading channel model. $|\mathcal{S}_1| = 1$, $|\mathcal{S}_2| = M$, $N_q = 2$, $r_1 = 40$m, $r_y = \frac{r_1}{2}$, $\alpha = 3$, $R_0 = 1$ BPCU, and $R_1 = 1$ BPCU.

are considered, where there are three users in $\mathcal{S}_1$. As can be observed from the figure, the use of NOMA can result in a significant performance gain compared to the scheme without NOMA. For example, when the transmission power of the base station is 15dBm, the use of NOMA can offer a rate improvement of 3 bits per channel use (BPCU) over the conventional MIMO scheme, for both considered channel models. In Fig. 2, we focus on the special case $|\mathcal{S}_1| = 1$, where the diversity gain achieved by NOMA is studied. To facilitate this diversity analysis, we set $|\mathcal{S}_2| = M$, which means that the diversity gains for the users in $\mathcal{S}_1$ and $\mathcal{S}_2$ are $\frac{M+1}{2}$ and M, respectively. From the figure, one can clearly observe the loss of diversity gain for the user from $\mathcal{S}_1$ due to FRAB. Note that the curves for the analytical results match those of the simulations, which verifies the accuracy of our analysis.

V. Conclusions

In this letter, NOMA has been proposed as a means of mitigating the reduced degrees of freedom induced by FRAB in massive MIMO and mmWave networks. The developed analytical and simulation results have demonstrated the superior performance of the proposed NOMA scheme.

Appendix A
Proof of Proposition 1

Recall that the probability density function (pdf) of a random variable generated by folding a normally distributed variable with zero mean and variance one is $f_{z_m}(x) = \frac{2}{\sqrt{\pi}}e^{-x^2}$. Therefore, the CDF of the square of the sum of z_m is given by

$$F_{z_{\sum},M}(z) = \frac{2^M}{\pi^{\frac{M}{2}}} \int_{\sum_{m=1}^{M} x_m < \sqrt{z}} \cdots \int \prod_{m=1}^{M} e^{-x_m^2} dx_m, \tag{11}$$

where the subscript M is added to show that the CDF is a function of M, and to facilitate the following analysis.

The proposition can be proved by using the inductive method. For the case $M = 1$, the approximate expression in the proposition can be simplified as follows:

$$F_{z_{\sum,1}}(z) \approx \frac{2}{\pi^{\frac{1}{2}}} z^{\frac{1}{2}}. \tag{12}$$

By calculating the integral of the pdf of z_m and applying $e^{-x} \approx 1-x$ for $x \to 0$, one can verify that (12) is a valid approximate expression for the CDF.

Assuming that the approximation is correct for the case of $M = n$, $F_{z_{\sum,n+1}}(z)$ can be expressed as follows:

$$F_{z_{\sum,n+1}}(z) = \frac{2^{n+1}}{\pi^{\frac{n+1}{2}}} \int_0^{\sqrt{z}} \int \cdots \int_{\sum_{m=1}^{n} x_m < \sqrt{z} - x_{n+1}} \prod_{m=1}^{n} e^{-x_m^2} dx_m \times e^{-x_{n+1}^2} dx_{n+1}. \tag{13}$$

By using $F_{z_{\sum,n}}(z) \approx \frac{2^n}{\pi^{\frac{n}{2}}} \frac{(\sqrt{z})^n}{n!}$, $F_{z_{\sum,n+1}}(z)$ can be approximated as follows:

$$\begin{aligned} F_{z_{\sum,n+1}}(z) &\approx \frac{2^{n+1}}{\pi^{\frac{n+1}{2}}} \int_0^{\sqrt{z}} \frac{(\sqrt{z} - x_{n+1})^n}{n!} e^{-x_{n+1}^2} dx_{n+1} \\ &\approx \frac{2^{n+1}}{\pi^{\frac{n+1}{2}}} \frac{(\sqrt{z})^{n+1}}{(n+1)!}. \end{aligned} \tag{14}$$

Therefore, the approximate expression is correct for $M = n + 1$, and the proof is complete via induction.

Appendix B
Proof for Lemma 1

With FRAB, the user's effective channel gain can be expressed as follows:

$$|\mathbf{h}_k^H \mathbf{f}_k|^2 = \left| \sum_{m=1}^{M} h_{k,m} e^{-j \frac{(i_{k,m}^* - 1) 2\pi}{N_q}} \right|^2. \tag{15}$$

When $N_q = 2$, we separate the real and imaginary parts of the user's channel as follows:

$$\begin{aligned} |\mathbf{h}_k^H \mathbf{f}_k|^2 &= \left| \sum_{m=1}^{M} (h_{k,m,real} + j h_{k,m,imag}) \bar{f}_{i_{k,m}^*} \right|^2 \\ &= \left| \sum_{m=1}^{M} (|h_{k,m,real}| + j \operatorname{sign}(h_{k,m,real}) h_{k,m,imag}) \right|^2, \end{aligned} \tag{16}$$

where sign$(\cdot)$ is the sign operation. This separation leads to the following expression:

$$|\mathbf{h}_k^H \mathbf{f}_k|^2 = \left| \sum_{m=1}^{M} \bar{h}_m \right|^2 + \left| \bar{h}_0 \right|^2, \tag{17}$$

where $\bar{h}_0 = \sum_{m=1}^{M} \operatorname{sign}(h_{k,m,real}) h_{k,m,imag}$ and $\bar{h}_m = |h_{k,m,real}|$. Note that $\operatorname{sign}(h_{k,m,real})$ is independent of $|h_{k,m,real}|$, because the phase and the amplitude of a complex Gaussian random variable are independent. Therefore, $\left|\sum_{m=1}^{M} \bar{h}_m\right|^2$ and $\left|\bar{h}_0\right|^2$ are independent.

Define $z_0 = |\bar{h}_0|^2$ whose pdf can be easily obtained as follows: Recall that the sum of i.i.d. Gaussian random variables is still a Gaussian variable, which means $\bar{h}_0$ is Gaussian with zero mean and variance $\frac{M}{2(1+d_{yk}^{\alpha})}$. Therefore, the CDF of z_0 is $F_{z_0}(z) = \frac{\gamma\left(\frac{1}{2}, \frac{(1+d_{yk}^{\alpha})z}{M}\right)}{\Gamma\left(\frac{1}{2}\right)}$, and the pdf is $f_{z_0}(z) = \frac{dF_{z_0}(z)}{dz}$, where $\gamma(x, y)$ denotes the incomplete gamma function. Note that when $x \to 0$, this CDF can be approximated as follows:

$$F_{z_0}(z) \approx \frac{2(1 + d_{yk}^{\alpha})^{\frac{1}{2}} z^{\frac{1}{2}}}{M^{\frac{1}{2}} \Gamma\left(\frac{1}{2}\right)}. \tag{18}$$

Therefore, the CDF of $|\mathbf{h}_k^H \mathbf{f}_k|^2$ can be expressed as follows:

$$F_{|\mathbf{h}_k^H \mathbf{f}_k|^2}(y) = \iint_{x+z<y} f_{z_0}(z) f_{\bar{z}_{\sum}}(x) dz dx. \tag{19}$$

According to Proposition 1, when $z \to 0$, the pdf of $\bar{z}_{\sum} \triangleq \left|\sum_{m=1}^{M} \bar{h}_m\right|^2$ can be approximated as follows:

$$f_{\bar{z}_{\sum}}(x) \approx \frac{2^{M-1}(1 + d_{yk}^{\alpha})^{\frac{M}{2}}}{\pi^{\frac{M}{2}}} \frac{x^{\frac{M-2}{2}}}{(M-1)!}. \tag{20}$$

Therefore, when $y \to 0$, we have the following approximation:

$$F_{|\mathbf{h}_k^H \mathbf{f}_k|^2}(y) \approx \int_0^y \frac{2^{M-1}(1 + d_{yk}^{\alpha})^{\frac{M}{2}}}{\pi^{\frac{M}{2}}} \frac{x^{\frac{M-2}{2}}}{(M-1)!} F_{z_0}(y - x) dx,$$

which yields the following approximation:

$$F_{|\mathbf{h}_k^H \mathbf{f}_k|^2}(y) \approx \frac{2^M (1 + d_{yk}^{\alpha})^{\frac{M+1}{2}} \int_0^y x^{\frac{M-2}{2}} (y - x)^{\frac{1}{2}} dx}{\pi^{\frac{M}{2}} (M-1)! M^{\frac{1}{2}} \Gamma\left(\frac{1}{2}\right)}.$$

Using the definition of the beta function in the above expression, the lemma is proved.

References

[1] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[2] L. Song, Y. Li, Z. Ding, and H. V. Poor, "Resource management in non-orthogonal multiple access networks for 5G and beyond," *IEEE Netw.*, to be published.

[3] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, to be published.

[4] 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.

[5] X. Liu and X. Wang, "Efficient antenna selection and user scheduling in 5G massive MIMO-NOMA system," in *Proc. IEEE Veh. Tech. Conf.*, Nangjing, China, May 2016, pp. 1–5.

[6] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, to be published.

[7] A. Alkhateeb, Y.-H. Nam, J. Zhang, and R. W. Heath, "Massive MIMO combining with switches," *IEEE Wireless Commun. Lett.*, vol. 5, no. 3, pp. 232–235, Jun. 2016.

[8] X. Gao, L. Dai, Y. Sun, S. Han, and C.-L. I, "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. Commun.*, Paris, France, May 2017.

[9] H. Yang *et al.*, "A 1-bit 10 × 10 reconfigurable reflectarray antenna: Design, optimization, and experiment," *IEEE Trans. Antennas Propag.*, vol. 64, no. 6, pp. 2246–2254, Jun. 2016.

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