Spatial Modulation-Aided Cooperative NOMA: Performance Analysis and Comparative Study

Qiang Li, Miaowen Wen, *Senior Member, IEEE,* Ertugrul Basar, *Senior Member, IEEE,* H. Vincent Poor, *Fellow, IEEE,* and Fangjiong Chen, *Member, IEEE*

***Abstract*— The combination of non-orthogonal multiple access (NOMA) and multiple-input multiple-output (MIMO) signaling is capable of providing significantly improved spectral efficiency and effective multiple access. However, conventional MIMO-NOMA entails significantly increased power consumption and implementation complexity. Inspired by the essence of spatial modulation (SM), in this paper, a novel three-node cooperative relaying system (CRS) using SM-aided NOMA, termed CRS-SM-NOMA, is proposed, in which the messages for two users are conveyed by two different information-bearing units of SM. As the reference scheme, conventional CRS-NOMA is further re-investigated with multiple receive antennas. Besides the achievable rate that is adopted by most of the existing NOMA works as the performance metric, the bit error rate (BER) of CRS-SM-NOMA and CRS-NOMA is analyzed assuming the maximum-likelihood detection. Approximate bit error probability expressions are derived in closed form for both users. Simulation results verify the analysis and show that the proposed CRS-SM-NOMA outperforms conventional CRS-NOMA and SM-OMA in terms of BER as well as ergodic sum rate.**

***Index Terms*— Bit error rate, cooperative relaying, non-orthogonal multiple access, spatial modulation, sum rate.**

I. Introduction

With the increasing growth of mobile data traffic and required connectivity among very large number of devices, wireless resources are becoming increasingly scarce. Moreover, in conventional orthogonal multiple access (OMA) systems, each user should occupy a distinct time/frequency channel, which limits the spectral efficiency (SE) of the system. To improve the SE, the concept of non-orthogonal multiple access (NOMA) has been proposed, in which multiple users having different channel conditions are served by one base station (BS) through the same resource block [1]. The users in NOMA are typically multiplexed in the power domain by using superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver [2]. It has been shown that compared with OMA techniques, NOMA achieves larger system throughput, lower latency, and greater fairness [3]–[5]. Due to these advantages, NOMA has been widely recognized as one of the key enabling technologies to meet the stringent requirements in fifth generation (5G) wireless networks [6], [7], and various aspects of NOMA have been extensively studied in the past few years.

Since near users with better channel conditions need to decode not only their own signals but also the signals intended for far users with poorer channel conditions in NOMA systems, they can act as relays to help far users, thus forming the paradigm of cooperative NOMA [8]–[15]. Specifically, in [8], a cooperative NOMA system with one BS and an arbitrary number of users is proposed, in which all users can achieve the maximum diversity gain. The study of [9] applies the technique of simultaneous wireless information and power transfer to NOMA networks, in which a near user serves as an energy harvesting node as well as an information relay. In [10] and [11], NOMA is introduced into a three-node cooperative relaying system (CRS) that comprises a source, a relay, and a destination to enhance the SE. The outage probability and ergodic sum rate of a NOMA-based relaying network over Nakagami-m fading channels are investigated in [12] and [13], where the BS communicates with multiple cell-edge mobile users simultaneously with the help of a cell-center relay. Unlike the studies in [8]–[13] where relays operate in the half-duplex mode, [14] and [15] analyze downlink NOMA systems with cooperative full-duplex relaying. Besides one-way relaying, NOMA-based two-way relay networks are discussed in [16] and [17].

On the other hand, multiple-input multiple-output (MIMO) in the form of spatial multiplexing is another important SE enhancing technology. It has been shown that at high signal-to-noise ratios (SNRs), the capacity of MIMO systems increases linearly with the minimum number of transmit and receive antennas. Hence, NOMA has been studied in multi-antenna scenarios. In [18], multi-antenna NOMA is categorized into beamformer-based and cluster-based structures, depending on whether one beamformer serves one user or multiple users. For the beamformer-based structure, [19] proposes two novel coordinated beamforming techniques for multi-cell MIMO-NOMA, in which beamforming vectors of two base stations are jointly optimized to enhance the performance of cell-edge users. For the cluster-based structure, [20] presents a framework, in which users are randomly grouped into several clusters, and precoding and detection matrices are carefully designed to cancel the inter-cluster interference. However, due to zero-forcing detection, the number of antennas at each user is required to be greater than that at the BS, which limits the application of this framework to massive MIMO-NOMA [21]. To remove this constraint, a novel cluster-based MIMO-NOMA design is proposed in [22] by allowing the existence

Q. Li, M. Wen, and F. Chen are with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: eeqiangli@mail.scut.edu.cn; eemwwen@scut.edu.cn; eefjchen@scut.edu.cn).

E. Basar is with Communications Research and Innovation Laboratory (CoreLab), Department of Electrical and Electronics Engineering, Koç University, Sariyer 34450, Istanbul, Turkey (e-mail: ebasar@ku.edu.tr).

H. V. Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA (e-mail: poor@princeton.edu).

of inter-cluster interference. In the above-mentioned MIMO-NOMA systems, however, the power and spatial domains should be jointly exploited to eliminate the user interference via beamforming, which complicates the beamforming design and SIC [23]. Besides, the requirement of multiple radio-frequency (RF) chains at the transmitter further lowers the energy efficiency and increases the implementation complexity. Recently, spatial modulation (SM), which activates only a single transmit antenna for each transmission, has been regarded as a competitive multi-antenna technique for 5G and beyond networks [24]–[26]. SM completely avoids inter-antenna interference and the requirement of multiple RF chains. In SM systems, the information is conveyed by the index of an active antenna as well as an M-ary modulated symbol. As a special realization of SM, space shift keying (SSK) solely utilizes the index of an active antenna to carry information [27]. Applying SM to NOMA systems is capable of improving the SE without increasing the power consumption and implementation complexity. Moreover, in current multi-user SM systems, the frequently used space division multiple access (SDMA) suffers from severe inter-user interference. Recalling that NOMA can effectively mitigate inter-user interference, it appears to be a promising multiple access method for SM-based multiuser communications. Hence, the combination of SM and NOMA has been promoted by the attempt to exploit the potential benefits of both SM and NOMA [28]–[31]. In [28] and [29], the bit stream for each user is split into two parts: the first part is used to activate a specific antenna and the second part is mapped to an M-ary modulated symbol. This, however, necessitates the use of multiple RF chains, which contradicts the inherent principle of SM. In order to keep a single transmit antenna active, in [30], the active antenna is selected according to the data for one specific user, and then the superposition coded symbol of all users is emitted by that antenna. It is worth noting that, in [28]–[30], superposition coding and SIC are still needed, thus imposing challenges on practical implementations of SM-NOMA. The study of [31] investigates a different prototype of downlink SM-NOMA, which avoids the utilization of SIC at the receiver. In this setup, the antennas at the BS are partitioned into several groups, each of which serves a specific user pair through mapping two independent bit streams to different information-bearing units. However, [31] provides relatively limited computer simulation results and performance analysis, and it fails to investigate the potential of cooperative relaying in SM-NOMA systems.

Against the background, not only to overcome the disadvantages of the existing techniques but also to enhance the system performance further through joint SM and cooperative relaying, we propose a novel CRS for Rayleigh flat-fading channels using SM-NOMA in this paper, termed CRS-SM-NOMA, which consists of one BS and two users. Two information-bearing units of SM are utilized to convey the messages for two users, respectively.[1] The user closer to the BS also acts as a relay to improve the performance of the other user. All nodes are equipped with multiple transmit/receive antennas. The main contributions of this paper are summarized as follows:

- The performance of CRS-SM-NOMA is theoretically analyzed in terms of bit error rate (BER) and achievable rate. Approximate closed-form bit error probability (BEP) expressions are derived for both users employing M-ary pulse amplitude modulation (PAM)/quadrature amplitude modulation (QAM). The asymptotic achievable rate of CRS-SM-NOMA is analyzed and compared with that of CRS-NOMA.
- Conventional CRS-NOMA [11] is re-investigated as the reference scheme. Different from the case in [11], multiple receive antennas are assumed for each user. Moreover, besides the sum rate, general BEP expressions are derived in closed form for both users, where the effects of superposition coding and SIC are properly taken into account. To the best of our knowledge, only [32] and [33] try to analytically evaluate the BER performance of NOMA systems in the literature. However, all derived results are restricted to some specific constellation combinations and additive white Gaussian noise (AWGN) channels. From this perspective, this work also provides a comprehensive background on the theoretical BEP analysis of NOMA-based systems.
- Extensive computer simulations are performed to validate the theoretical analysis and compare the performance of CRS-SM-NOMA, CRS-NOMA, and SM-OMA. Simulation results accurately agree with the analysis and show that CRS-SM-NOMA outperforms CRS-NOMA and SM-OMA in terms of BER and ergodic sum rate.

The remainder of this paper is organized in the following. The system model of CRS-SM-NOMA is presented in Section II, followed by the performance analysis in Section III. Section IV describes the reference scheme, namely CRS-NOMA, and evaluates its BER and achievable rate performance. Computer simulation results are given in Section V, and finally Section VI concludes the paper.

Notation: Scalar variables are in italic letters. Column vectors and matrices are denoted by lowercase and uppercase boldface letters, respectively. Superscripts ^{T} and ^{H} stand for transpose and Hermitian transpose, respectively. h_{ij} and $\mathbf{h}_i$ represent the (i,j)-th element and the i-th column of $\mathbf{H}$, respectively. The $n \times n$ identity matrix is symbolized by $\mathbf{I}_n$. The probability of an event and the probability density function (PDF) are denoted by $\Pr(\cdot)$ and $p(\cdot)$, respectively. $\mathcal{CN}(\mu, \sigma^2)$ denotes the complex Gaussian distribution with mean μ and variance σ^2. $Q(\cdot)$, $\Gamma(\cdot)$, and $E[\cdot]$ denote the Gaussian Q-function, gamma function, and expectation operation, respectively. $\binom{\cdot}{\cdot}$ and $\|\cdot\|$ stand for the binomial coefficient and Frobenius norm, respectively. !! is the double factorial notation denoting the product of only odd integers. $\lfloor \cdot \rfloor$ denotes the floor function. $\mathrm{mod}(x, y)$ returns the remainder after division of x by y.

[1]CRS-SM-NOMA still belongs to NOMA systems in the sense that two users simultaneously occupy the same time-and-frequency resources. Further, due to the advantages of SM, the messages for one user are implicitly embedded in the signal for the other user, getting rid of the requirement of power allocation and SIC.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2898099, IEEE Journal of Selected Topics in Signal Processing

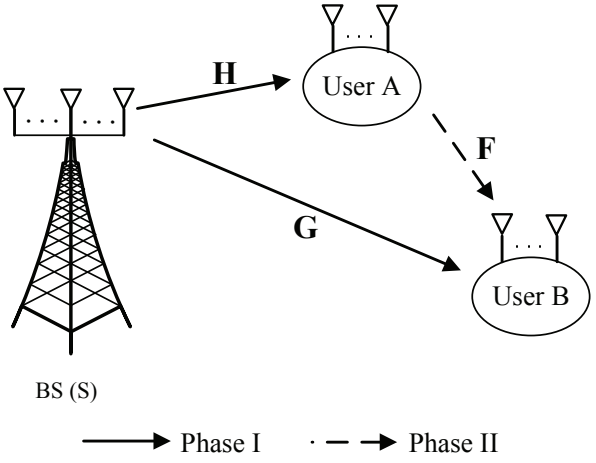

Fig. 1. System model of CRS-SM-NOMA.

II. System Model

The system model of CRS-SM-NOMA is given in Fig. 1, in which a BS (S) serves two users (A and B) simultaneously. All nodes operate in the half-duplex mode. Both the BS and user A have N_t transmit antennas, while users A and B are equipped with N_r^A and N_r^B receive antennas, respectively. There is no constraint on N_t and N_r^A. The channel matrices of S-to-A, S-to-B, and A-to-B links are denoted by $\mathbf{H} \in \mathbb{C}^{N_r^A \times N_t}$, $\mathbf{G} \in \mathbb{C}^{N_r^B \times N_t}$, and $\mathbf{F} \in \mathbb{C}^{N_r^B \times N_t}$, respectively. The entries of $\mathbf{H}$, $\mathbf{G}$, and $\mathbf{F}$ are modeled as independent and identically distributed (i.i.d) complex Gaussian random variables following the distributions $\mathcal{CN}(0, \lambda^{SA})$, $\mathcal{CN}(0, \lambda^{SB})$, and $\mathcal{CN}(0, \lambda^{AB})$, respectively. Provided that A is the near user and B is the far user, since the variance of channel coefficients is typically inversely proportional to the propagation distance [4], we have $\lambda^{SA} > \lambda^{SB}$.

In CRS-SM-NOMA, there are two phases involved in each transmission of $b_1 + b_2$ bits to users A and B, where b_1 (b_2) bits are dedicated for user A (B). During the first time slot, the BS generates an SM symbol

$$\mathbf{x} = \big[\underbrace{0, \ldots, 0}_{l-1}, s, \underbrace{0, \ldots, 0}_{N_t - l}\big]^T, \tag{1}$$

and broadcasts it to users A and B, where s is an M-ary PAM/QAM symbol determined by the b_1 bits with $E[|s|^2] = 1$, and $l \in \{1, \ldots, N_t\}$ denotes the active antenna index determined by the b_2 bits. Obviously, we have $b_1 = \log_2(M)$ and $b_2 = \log_2(N_t)$.[2] By normalizing the transmit power to unity, the received signals at users A and B can be expressed as

$$\mathbf{y}^{SA} = \mathbf{H}\mathbf{x} + \mathbf{n}^{SA} \tag{2}$$

and

$$\mathbf{y}^{SB} = \mathbf{G}\mathbf{x} + \mathbf{n}^{SB}, \tag{3}$$

[2] N_t in this paper is assumed to be an arbitrary integer power of two. Hence, l can be easily obtained by converting the b_2 bits into the decimal representation.

respectively, where $\mathbf{n}^{SA} \in \mathbb{C}^{N_r^A \times 1}$ and $\mathbf{n}^{SB} \in \mathbb{C}^{N_r^B \times 1}$ are the AWGN vectors at users A and B, which follow the distributions $\mathcal{CN}(\mathbf{0}, N_0 \mathbf{I}_{N_r^A})$ and $\mathcal{CN}(\mathbf{0}, N_0 \mathbf{I}_{N_r^B})$, respectively. We define the transmit SNR as $\rho = 1/N_0$. User B preserves the received signal $\mathbf{y}^{SB}$ for decoding purposes at the second phase, while user A adopts the maximum-likelihood (ML) detection to jointly decode s and l, namely

$$\begin{aligned}(\hat{s}, \hat{l}) &= \arg\min_{s,l} \left\|\mathbf{y}^{SA} - \mathbf{H}\mathbf{x}\right\|^2 \\ &= \arg\min_{s,l} \left\|\mathbf{y}^{SA} - \mathbf{h}_l s\right\|^2,\end{aligned} \tag{4}$$

where $\hat{s}$ and $\hat{l}$ are the estimates of s and l at user A, respectively. User A then recovers its own information bits from $\hat{s}$ and preserves $\hat{l}$ for cooperation.

At the second phase, user A forwards $\hat{l}$ to B via an SSK signal as

$$\mathbf{y}^{AB} = \mathbf{f}_{\hat{l}} + \mathbf{n}^{AB}, \tag{5}$$

where $\mathbf{n}^{AB} \in \mathbb{C}^{N_r^B \times 1}$ is the AWGN vector at user B distributed according to $\mathcal{CN}(\mathbf{0}, N_0 \mathbf{I}_{N_r^B})$. By combining $\mathbf{y}^{AB}$ in (5) and $\mathbf{y}^{SB}$ in (3), user B obtains its own message from

$$\tilde{l} = \arg\min_{l} \left\{ \left\|\mathbf{y}^{AB} - \mathbf{f}_l\right\|^2 + \min_{s} \left\|\mathbf{y}^{SB} - \mathbf{g}_l s\right\|^2 \right\}, \tag{6}$$

where $\tilde{l}$ is the estimate of l at user B.

III. Performance Analysis

In this section, we theoretically analyze the performance of the proposed CRS-SM-NOMA scheme in terms of BER and sum rate.

A. BEP Analysis

In this subsection, approximate BEP expressions are derived in closed form for users A and B in CRS-SM-NOMA systems.

1) BEP of User A: User A extracts s from the received signal $\mathbf{y}^{SA}$ according to the ML detection in (4). Obviously, the error events for user A can be classified into two complementary types, depending on whether the index of the active antenna is detected correctly or not. Thus, the overall BEP of user A can be evaluated as

$$P_A \approx \frac{1}{2} P_{IA} + (1 - P_{IA})\, P_{Ab}, \tag{7}$$

where P_{IA} is the error probability of the active antenna index detection at user A, P_{Ab} is the BEP of M-ary PAM/QAM demodulation over N_r^A i.i.d. Rayleigh fading channels, and the factor $1/2$ results from the fact that there is no symbol transmitted from inactive antennas.

We first study the conditional pairwise error probability (PEP) from (4), which is the probability of detecting $\mathbf{x}$ as $\hat{\mathbf{x}}$ conditioned on $\mathbf{H}$, s, and $\hat{s}$, namely

$$\begin{aligned}\Pr\left(\mathbf{x} \to \hat{\mathbf{x}} \,|\mathbf{H}, s, \hat{s}\right) &= \Pr\left(\left\|\mathbf{y}^{SA} - \mathbf{h}_l s\right\|^2 > \left\|\mathbf{y}^{SA} - \mathbf{h}_{\hat{l}} \hat{s}\right\|^2\right) \\ &= Q\left(\sqrt{\frac{\kappa}{2N_0}}\right),\end{aligned} \tag{8}$$

where $\kappa = \left\|\mathbf{h}_l s - \mathbf{h}_{\hat{l}}\hat{s}\right\|^2 = \sum_{i=1}^{2N_r^A} \alpha_i^2$ with $\alpha_i \sim \mathcal{CN}\big(0, \lambda^{SA}(|s|^2+|\hat{s}|^2)/2\big)$. Hence, κ is chi-squared distributed with $2N_r^A$ degrees of freedom, whose PDF is given by

$$p_\kappa(x) = \frac{1}{2^{N_r^A}\Gamma\left(N_r^A\right)\sigma^{2N_r^A}} x^{N_r^A-1}\exp\left(-\frac{x}{2\sigma^2}\right), \quad (9)$$

where $\sigma^2 = \lambda^{SA}(|s|^2 + |\hat{s}|^2)/2$. By averaging $\Pr(\mathbf{x}\to\hat{\mathbf{x}}\,|\mathbf{H}, s, \hat{s}\,)$ over κ, we arrive at [34, Eq. (64)]

$$\begin{aligned}\Pr(\mathbf{x}\to\hat{\mathbf{x}}\,|s,\hat{s}\,) &= \int_0^{+\infty} Q\left(\sqrt{\frac{x}{2N_0}}\right) p_\kappa(x)\,dx \\ &= [P(c)]^{N_r^A}\sum_{k=0}^{N_r^A-1}\binom{N_r^A-1+k}{k}[1-P(c)]^k, \quad (10)\end{aligned}$$

where

$$P(c) = \frac{1}{2}\left(1-\sqrt{\frac{c}{1+c}}\right) \quad (11)$$

with $c = \sigma^2/(2N_0)$. According to the union bounding technique, an upper bound on the error probability of active antenna index detection at user A is given by

$$\begin{aligned}P_{IA} &\le \frac{1}{N_t M}\sum_{\substack{l,\hat{l}\\ l\neq\hat{l}}}\sum_{s,\hat{s}}\Pr(\mathbf{x}\to\hat{\mathbf{x}}\,|s,\hat{s}\,) \\ &= \frac{N_t-1}{M}\sum_{s,\hat{s}}\Pr(\mathbf{x}\to\hat{\mathbf{x}}\,|s,\hat{s}\,). \quad (12)\end{aligned}$$

For the calculation of P_{Ab}, let us first consider the M-PAM constellation. The conditional error probability for the m-th bit $(m = 1,\ldots,\log_2(M))$ can be formulated as [35]

$$\begin{aligned}P_{Ab}\left(m\,\middle|\,\{\gamma_k^{SA}\}_{k=1}^{N_r^A}\right) = \frac{2}{M}\sum_{i=0}^{(1-2^{-m})M-1}\Bigg\{&(-1)^{\left\lfloor\frac{i\cdot 2^{m-1}}{M}\right\rfloor} \\ &\times\left(2^{m-1}-\left\lfloor\frac{i\cdot 2^{m-1}}{M}+\frac{1}{2}\right\rfloor\right) \\ &\times Q\left((2i+1)\,d\sqrt{2\gamma_{tot}^{SA}}\right)\Bigg\}, \quad (13)\end{aligned}$$

where $d = \sqrt{3/(M^2-1)}$ is the half of the minimum distance between two signal points of the normalized M-PAM constellation, and $\gamma_{tot}^{SA} = \sum_{k=1}^{N_r^A}\gamma_k^{SA}$ is the total instantaneous receive SNR with $\gamma_k^{SA} = |h_{kl}|^2/N_0$. To facilitate the calculation of the average error probability, i.e., $P_{Ab}(m)$, the Gaussian Q-function in (13) is replaced with the alternative representation [34]

$$Q(x) = \frac{1}{\pi}\int_0^{\pi/2}\exp\left(-\frac{x^2}{2\sin^2\phi}\right)d\phi, \quad (14)$$

which leads to (15), shown at the top of the next page. By statistically averaging (15) over the joint PDF of $\{\gamma_k^{SA}\}_{k=1}^{N_r^A}$ which is given by

$$\begin{aligned}p\left(\gamma_1^{SA},\ldots,\gamma_{N_r^A}^{SA}\right) &= \prod_{k=1}^{N_r^A} p_{\gamma_k^{SA}}\left(\gamma_k^{SA}\right) \\ &= \prod_{k=1}^{N_r^A}\frac{1}{\bar{\gamma}^{SA}}\exp\left(-\frac{\gamma_k^{SA}}{\bar{\gamma}^{SA}}\right) \quad (16)\end{aligned}$$

with $\bar{\gamma}^{SA} = \lambda^{SA}/N_0$, we have [36, Eq. (5A.4b)]

$$\begin{aligned}P_{Ab}(m) = \frac{2}{M}\sum_{i=0}^{(1-2^{-m})M-1}\Bigg\{&(-1)^{\left\lfloor\frac{i\cdot 2^{m-1}}{M}\right\rfloor} \\ &\times\left(2^{m-1}-\left\lfloor\frac{i\cdot 2^{m-1}}{M}+\frac{1}{2}\right\rfloor\right)\cdot T(i)\Bigg\}, \quad (17)\end{aligned}$$

where $T(i)$ is shown at the top of the next page as (18) with

$$\mu_c(i) = \sqrt{\frac{\lambda^{SA}((2i+1)\,d)^2}{N_0+\lambda^{SA}((2i+1)\,d)^2}}. \quad (19)$$

Finally, a closed-form BEP expression for M-PAM over N_r^A i.i.d. Rayleigh fading channels can be obtained by averaging (17) over all constellation points, yielding

$$P_{Ab} = \frac{1}{\log_2(M)}\sum_{m=1}^{\log_2(M)} P_{Ab}(m). \quad (20)$$

For the M-QAM constellation with $M = M_I M_Q$, since M-ary QAM consists of two independent PAM signals, namely M_I-ary PAM of the in-phase (I-) signal and M_Q-ary PAM of the quadrature (Q-) signal, the derivation of P_{Ab} for M-ary QAM is similar to that for M-ary PAM. Particularly, the probabilities that the m_I-th bit $(m_I = 1,\ldots,\log_2(M_I))$ in the I- component and the m_Q-th bit $(m_Q = 1,\ldots,\log_2(M_Q))$ in the Q- component are in error, denoted by $P_{Ab-I}(m_I)$ and $P_{Ab-Q}(m_Q)$, can be formulated as (17) except that M and m should be replaced with M_I (M_Q) and m_I (m_Q), respectively, for the I (Q)- dimension. Note that the parameter d in $\mu_c(i)$ for M-QAM should also be revised as $d = \sqrt{3/(M_I^2+M_Q^2-2)}$. Finally, we are led to P_{Ab} for M-QAM, namely

$$\begin{aligned}P_{Ab} = \frac{1}{\log_2(M)}\Bigg(&\sum_{m_I=1}^{\log_2(M_I)} P_{Ab-I}(m_I) \\ &+\sum_{m_Q=1}^{\log_2(M_Q)} P_{Ab-Q}(m_Q)\Bigg). \quad (21)\end{aligned}$$

After obtaining P_{IA} and P_{Ab}, the overall BEP expression for user A can be derived by substituting (12) and (20) or (21) into (7). We observe from (10) and (18) that both P_{IA} and P_{Ab} achieve a diversity order of N_r^A, such that the overall diversity order of user A is also given by N_r^A.

2) BEP of User B: In the active antenna index detection, an error happens once the index of the active antenna is detected incorrectly as one of the remaining N_t-1 indices. A different incorrect index may result in a different number of erroneous bits, and we observe that there are in total of $\binom{b_2}{t}$ error events that correspond to t erroneous index bits, where $t\in\{1,\ldots,b_2\}$. Thus, on average, the erroneous detection of the index of the active antenna produces

$$n_e = \frac{1}{N_t-1}\sum_{t=1}^{b_2} t\cdot\binom{b_2}{t} \quad (22)$$

erroneous bits out of b_2 bits. Recalling that the bit errors of user B result from two complementary conditions: (a) user A

$$P_{Ab}\left(m\left|\left\{\gamma_k^{SA}\right\}_{k=1}^{N_r^A}\right.\right)=\frac{2}{M}\sum_{i=0}^{(1-2^{-m})M-1}\left\{(-1)^{\left\lfloor\frac{i\cdot 2^{m-1}}{M}\right\rfloor}\left(2^{m-1}-\left\lfloor\frac{i\cdot 2^{m-1}}{M}+\frac{1}{2}\right\rfloor\right)\right.$$
$$\left.\times\frac{1}{\pi}\int_0^{\pi/2}\prod_{k=1}^{N_r^A}\exp\left(-\frac{\left((2i+1)\,d\right)^2\gamma_k^{SA}}{\sin^2\phi}\right)d\phi\right\}. \tag{15}$$

$$\begin{aligned}T(i)&=\frac{1}{\pi}\int_0^{\pi/2}\left[\int_0^{\infty}\exp\left(-\frac{\left((2i+1)\,d\right)^2\gamma_k^{SA}}{\sin^2\phi}\right)p_{\gamma_k^{SA}}\left(\gamma_k^{SA}\right)d\gamma_k^{SA}\right]^{N_r^A}d\phi\\&=\frac{1}{\pi}\int_0^{\pi/2}\left[\frac{\sin^2\phi}{\left((2i+1)\,d\right)^2\bar{\gamma}^{SA}+\sin^2\phi}\right]^{N_r^A}d\phi\\&=\left(\frac{1-\mu_c(i)}{2}\right)^{N_r^A}\sum_{k=0}^{N_r^A-1}\binom{N_r^A-1+k}{k}\left(\frac{1+\mu_c(i)}{2}\right)^k\end{aligned} \tag{18}$$

detects l correctly; (b) user A detects l incorrectly, we can obtain the approximate BEP expression for user B as

$$P_B\approx\left(\left(1-P_{IA}\right)P_{IB}+P_{IA}\right)\cdot\frac{n_e}{b_2}, \tag{23}$$

where P_{IB} is the error probability of active antenna index detection at user B on condition that user A correctly detects that index. Note that in (23), we assume that an error in active antenna index detection at user A always causes an error at user B, which is a common behavior for decode-and-forward relaying-based systems.

To calculate P_B in (23), we turn to calculate P_{IB} since P_{IA} has been given in (12). Let us consider (6) to derive the conditional PEP, which becomes

$$\begin{aligned}\Pr\left(\mathbf{x}\to\hat{\mathbf{x}}\,|\mathbf{F},\mathbf{G},s,\hat{s}\right)&=\Pr\left(\left\|\mathbf{y}^{AB}-\mathbf{f}_l\right\|^2+\left\|\mathbf{y}^{SB}-\mathbf{g}_l s\right\|^2\right.\\&\left.>\left\|\mathbf{y}^{AB}-\mathbf{f}_{\hat{l}}\right\|^2+\left\|\mathbf{y}^{SB}-\mathbf{g}_{\hat{l}}\hat{s}\right\|^2\right)\\&=Q\left(\sqrt{\frac{\eta}{2N_0}}\right),\end{aligned} \tag{24}$$

where $\eta=\left\|\mathbf{f}_l-\mathbf{f}_{\hat{l}}\right\|^2+\left\|\mathbf{g}_l s-\mathbf{g}_{\hat{l}}\hat{s}\right\|^2$. We observe that η is the sum of two independent central chi-square random variables both of which are with $2N_r^B$ degrees of freedom. Hence, the moment generating function (MGF) of η is given by [37]

$$\mathcal{M}_\eta(\omega)=\left[\frac{1}{\left(1-2\omega\sigma_1^2\right)\left(1-2\omega\sigma_2^2\right)}\right]^{N_r^B}, \tag{25}$$

where $\sigma_1^2=\lambda^{AB}$ and $\sigma_2^2=\lambda^{SB}(|s|^2+|\hat{s}|^2)/2$. Resorting to the approximation of Q-function, namely $Q(x)\approx 1/12\cdot e^{-x^2/2}+1/4\cdot e^{-2x^2/3}$, and the MGF approach, we have

$$\begin{aligned}\Pr\left(\mathbf{x}\to\hat{\mathbf{x}}\,|s,\hat{s}\right)\approx\frac{1}{12}&\left[\frac{1}{\left(1+\frac{\sigma_1^2}{2N_0}\right)\left(1+\frac{\sigma_2^2}{2N_0}\right)}\right]^{N_r^B}\\&+\frac{1}{4}\left[\frac{1}{\left(1+\frac{2\sigma_1^2}{3N_0}\right)\left(1+\frac{2\sigma_2^2}{3N_0}\right)}\right]^{N_r^B}.\end{aligned} \tag{26}$$

Therefore, the erroneous active antenna index detection probability at user B in case (a) can be expressed as

$$\begin{aligned}P_{IB}&\approx\frac{1}{N_t M}\sum_{\substack{l,\hat{l}\\l\neq\hat{l}}}\sum_{s,\hat{s}}\Pr\left(\mathbf{x}\to\hat{\mathbf{x}}\,|s,\hat{s}\right)\\&=\frac{N_t-1}{M}\sum_{s,\hat{s}}\Pr\left(\mathbf{x}\to\hat{\mathbf{x}}\,|s,\hat{s}\right).\end{aligned} \tag{27}$$

Finally, we obtain the BEP of user B by substituting (12), (22), and (27) into (23). Since the diversity order of P_{IB} in (26) is $2N_r^B$, user B achieves a diversity order of $\min\{2N_r^B,N_r^A\}$ from (23).

B. Sum Rate Analysis Under Gaussian Input

In this subsection, we analyze the achievable sum rate of CRS-SM-NOMA systems with Gaussian input.

1) Achievable Rate of User A: Since the information of user A is only conveyed by classical amplitude phase modulation (APM) symbols, the instantaneous achievable rate of user A can be easily given by [38]

$$C_A=\frac{1}{2}I\left(X;Y^{SA}\,|X_{ch}\right)=\frac{1}{2N_t}\sum_{l=1}^{N_t}\log_2\left(1+\rho\|\mathbf{h}_l\|^2\right), \tag{28}$$

where X and X_{ch} denote APM and antenna symbol spaces, respectively, Y^{SA} is the output symbol space, and $1/2$ is due to two-phase transmission.

2) Achievable Rate of User B: Since the index of the active antenna should be detected at user A as well as user B, the achievable rate associated with user B is obtained as

$$C_B = \frac{1}{2}\min\left\{I\left(X_{ch};Y^{SA}\right), I\left(X_{ch};Y^{SB},Y^{AB}\right)\right\}. \quad (29)$$

Next, we deal with the calculations of $I\left(X_{ch};Y^{SA}\right)$ and $I\left(X_{ch};Y^{SB},Y^{AB}\right)$. Specifically,

$$\begin{aligned} I\left(X_{ch};Y^{SA}\right) &= H\left(X_{ch}\right) - H\left(X_{ch}\left|Y^{SA}\right.\right) \\ &= \log_2\left(N_t\right) + \frac{1}{N_t}\sum_{l=1}^{N_t}\int p\left(\mathbf{y}^{SA}\left|X_{ch=l}\right.\right) \\ &\quad \times \log_2\left(p\left(X_{ch=l}\left|\mathbf{y}^{SA}\right.\right)\right) d\mathbf{y}^{SA} \\ &= \log_2\left(N_t\right) + \frac{1}{N_t}\sum_{l=1}^{N_t}\int p\left(\mathbf{y}^{SA}\left|X_{ch=l}\right.\right) \\ &\quad \times \log_2\left(\frac{p\left(\mathbf{y}^{SA}\left|X_{ch=l}\right.\right)}{N_t\cdot p\left(\mathbf{y}^{SA}\right)}\right) d\mathbf{y}^{SA}, \end{aligned} \quad (30)$$

where

$$p\left(\mathbf{y}^{SA}\left|X_{ch=l}\right.\right) = \frac{1}{\pi^{N_r^A}\det\left(\mathbf{\Sigma}_l\right)}\exp\left\{-\left(\mathbf{y}^{SA}\right)^H\mathbf{\Sigma}_l^{-1}\mathbf{y}^{SA}\right\} \quad (31)$$

and

$$p\left(\mathbf{y}^{SA}\right) = \frac{1}{N_t}\sum_{l=1}^{N_t}p\left(\mathbf{y}^{SA}\left|X_{ch=l}\right.\right) \quad (32)$$

with $\mathbf{\Sigma}_l = N_0\mathbf{I}_{N_r^A} + \mathbf{h}_l\mathbf{h}_l^H$. Similarly, we have

$$\begin{aligned} &I\left(X_{ch};Y^{SB},Y^{AB}\right) = H\left(X_{ch}\right) - H\left(X_{ch}\left|Y^{SB},Y^{AB}\right.\right) \\ &= \log_2\left(N_t\right) + \frac{1}{N_t}\sum_{i=l}^{N_t}\int p\left(\mathbf{y}^{SB},\mathbf{y}^{AB}\left|X_{ch=l}\right.\right) \\ &\quad \times \log_2\left(p\left(X_{ch=l}\left|\mathbf{y}^{SB},\mathbf{y}^{AB}\right.\right)\right) d\mathbf{y}^{SB}d\mathbf{y}^{AB} \\ &= \log_2\left(N_t\right) + \frac{1}{N_t}\sum_{l=1}^{N_t}\int p\left(\mathbf{y}^{SB},\mathbf{y}^{AB}\left|X_{ch=l}\right.\right) \\ &\quad \times \log_2\left(\frac{p\left(\mathbf{y}^{SB},\mathbf{y}^{AB}\left|X_{ch=l}\right.\right)}{N_t\cdot p\left(\mathbf{y}^{SB},\mathbf{y}^{AB}\right)}\right) d\mathbf{y}^{SB}d\mathbf{y}^{AB}, \end{aligned} \quad (33)$$

where

$$\begin{aligned} p\left(\mathbf{y}^{SB},\mathbf{y}^{AB}\left|X_{ch=l}\right.\right) &= \frac{1}{\pi^{2N_r^B}\det\left(\mathbf{\Xi}_l\right)\det\left(\mathbf{\Omega}_l\right)} \\ &\times \exp\left\{-\left(\mathbf{y}^{SB}\right)^H\mathbf{\Xi}_l^{-1}\mathbf{y}^{SB} - \left(\mathbf{y}^{AB}\right)^H\mathbf{\Omega}_l^{-1}\mathbf{y}^{AB}\right\} \end{aligned} \quad (34)$$

and

$$p\left(\mathbf{y}^{SB},\mathbf{y}^{RB}\right) = \frac{1}{N_t}\sum_{l=1}^{N_t}p\left(\mathbf{y}^{SB},\mathbf{y}^{AB}\left|X_{ch=l}\right.\right) \quad (35)$$

with $\mathbf{\Xi}_l = N_0\mathbf{I}_{N_r^B} + \mathbf{g}_l\mathbf{g}_l^H$ and $\mathbf{\Omega}_l = N_0\mathbf{I}_{N_r^B} + \mathbf{f}_l\mathbf{f}_l^H$. Unfortunately, both (30) and (33) cannot be exactly calculated in closed form. For this reason, we employ numerical evaluation to estimate them. Finally, the instantaneous achievable sum rate of CRS-SM-NOMA systems can be readily obtained as

$$C_{\text{CRS-SM-NOMA}} = C_A + C_B. \quad (36)$$

At extremely high SNRs, since the index of the active antenna is likely to be detected correctly, we have

$$C_B \approx \frac{1}{2}\log_2\left(N_t\right), \quad (37)$$

and

$$C_{\text{CRS-SM-NOMA}} \approx \frac{1}{2N_t}\sum_{l=1}^{N_t}\log_2\left(\rho\|\mathbf{h}_l\|^2\right) + \frac{1}{2}\log_2\left(N_t\right). \quad (38)$$

IV. On Conventional CRS-NOMA

In this section, we take the conventional CRS-NOMA [11] as the reference scheme. The considered system model is the same as that of [11] except that users A and B are equipped with multiple receive antennas. Besides the sum rate performance, we investigate the BER performance of CRS-NOMA and derive closed-form BEP expressions for both users. Without incurring confusions, we use the same notations for received signals, channel matrices and noise vectors as those in Section II. Note that since each node has one transmit antenna and multiple receive antennas, channel matrices are degenerated into column vectors. For fair comparisons, channel coefficients and noise vectors follow the same distributions as those in CRS-SM-NOMA.

At the first phase in CRS-NOMA, the BS transmits the superposition coded symbol $s_C = \sqrt{a}s_A + \sqrt{1-a}s_B$ to users A and B instantaneously, where $0 < a \leq 0.5$ is the power allocation factor for user A, and $s_A \in \mathcal{S}_A$ ($s_B \in \mathcal{S}_B$) is intended for user A (B). Here, $\mathcal{S}_A$ ($\mathcal{S}_B$) is an M_A-ary (M_B-ary) normalized PAM/QAM constellation, and the signal constellation of s_C thus can be expressed as $\mathcal{S}_C = \{\sqrt{a}s_A + \sqrt{1-a}s_B | s_A \in \mathcal{S}_A, s_B \in \mathcal{S}_B\}$, which is an irregular M_AM_B-ary constellation. The total transmit power is normalized to unity and thus the received signals at users A and B are respectively expressed as

$$\mathbf{y}^{SA} = \mathbf{h}\left(\sqrt{a}s_A + \sqrt{1-a}s_B\right) + \mathbf{n}^{SA} \quad (39)$$

and

$$\mathbf{y}^{SB} = \mathbf{g}\left(\sqrt{a}s_A + \sqrt{1-a}s_B\right) + \mathbf{n}^{SB}. \quad (40)$$

User A first detects symbol s_B by treating symbol s_A as user interference from (39) and then cancels s_B using SIC to obtain the desired symbol s_A from

$$\hat{\mathbf{y}}^{SA} = \mathbf{h}\left(\sqrt{a}s_A + \sqrt{1-a}\left(s_B - \hat{s}_B\right)\right) + \mathbf{n}^{SA}, \quad (41)$$

where $\hat{s}_B$ is the estimate of s_B at user A. User B conserves $\mathbf{y}^{SB}$ and does not detect any symbols during the first phase.

At the second phase, user A relays $\hat{s}_B$ to user B with full transmit power such that the received signal at user B is given by

$$\mathbf{y}^{AB} = \mathbf{f}\hat{s}_B + \mathbf{n}^{AB}. \quad (42)$$

According to the maximal ratio combining (MRC), user B combines $\mathbf{y}^{SB}$ in (40) and $\mathbf{y}^{AB}$ in (42) with weights $\mathbf{g}^H\sqrt{1-a}/(\|\mathbf{g}\|^2 a + N_0)$ and $\mathbf{f}^H/N_0$, respectively, resulting in the combined signal

$$\begin{aligned} y^B &= \frac{\mathbf{g}^H\sqrt{1-a}}{\|\mathbf{g}\|^2 a + N_0}\mathbf{y}^{SB} + \frac{\mathbf{f}^H}{N_0}\mathbf{y}^{AB} \\ &= \frac{\|\mathbf{g}\|^2(1-a)}{\|\mathbf{g}\|^2 a + N_0}s_B + \frac{\|\mathbf{f}\|^2}{N_0}\hat{s}_B + \frac{\|\mathbf{g}\|^2\sqrt{a(1-a)}}{\|\mathbf{g}\|^2 a + N_0}s_A \\ &\quad + \frac{\sqrt{1-a}\mathbf{g}^H\mathbf{n}^{SB}}{\|\mathbf{g}\|^2 a + N_0} + \frac{\mathbf{f}^H\mathbf{n}^{AB}}{N_0}. \end{aligned} \tag{43}$$

From (43), user B estimates its own symbol s_B assuming correct detection of s_B at user A via

$$\tilde{s}_B = \arg\min_{s_B}\left| y^B - \left(\frac{\|\mathbf{g}\|^2(1-a)}{\|\mathbf{g}\|^2 a + N_0} + \frac{\|\mathbf{f}\|^2}{N_0}\right)s_B\right|^2. \tag{44}$$

A. BEP Analysis for CRS-NOMA

In this subsection, the analytical BEP expressions are derived for users A and B in closed form. Due to SIC, the BEP of user A is highly affected by the error probability of detecting s_B at user A. According to the law of total probability, the BEP of user A can be approximated as

$$P_A \approx \left(1 - P_{s_B}^A\right)P_{Ab} + \frac{P_{s_B}^A}{\log_2(M_A)}, \tag{45}$$

where $P_{s_B}^A$ is the symbol error probability (SEP) of s_B at user A, and P_{Ab} denotes the BEP of M-ary PAM/QAM demodulation over N_r^A i.i.d. Rayleigh fading channels with transmit power a. Therefore, P_{Ab} is the same as (20) for PAM and (21) for QAM except that $\mu_c(i)$ in (19) should be revised as

$$\mu_c(i) = \sqrt{\frac{a\lambda^{SA}((2i+1)d)^2}{N_0 + a\lambda^{SA}((2i+1)d)^2}}. \tag{46}$$

Similar to (45), the BEP of user B can be given by

$$P_B \approx \frac{\left(1 - P_{s_B}^A\right)\cdot P_{s_B}}{\log_2(M_B)} + \frac{P_{s_B}^A}{\log_2(M_B)}, \tag{47}$$

where P_{s_B} denotes the SEP of detecting s_B from (44) under the condition that user A correctly detects s_B. Since it is difficult to exactly evaluate the error propagation when s_B is detected incorrectly, we assume that an incorrectly detected s_B yields one erroneous bit for s_A at user A in (45) and for s_B at user B in (47). Further, in (47), when s_B is detected correctly at user A, the BEP of user B is approximated as P_{s_B} divided by $\log_2(M_B)$. These approximations lead to the slight deviations between analytical results and their simulation counterparts, which will be shown in Section V.

1) PAM Constellation: We first consider the scenario in which both s_A and s_B are PAM symbols transmitted over the I- channel, i.e., $M_A = M_{AI}$ and $M_B = M_{BI}$. They are independently selected from the constellations $\mathcal{S}_A = \{\pm d_A, \pm 3d_A, \pm 5d_A, \ldots, \pm(M_{AI}-1)d_A\}$ and $\mathcal{S}_B = \{\pm d_B, \pm 3d_B, \pm 5d_B, \ldots, \pm(M_{BI}-1)d_B\}$, respectively, where $d_A = \sqrt{3/(M_{AI}^2-1)}$ and $d_B = \sqrt{3/(M_{BI}^2-1)}$.

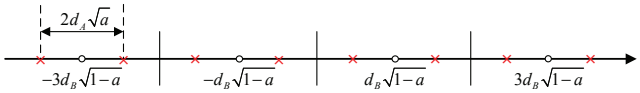

Fig. 2. Signal constellation of superposition coded symbols, where s_A and s_B are 2-PAM and 4-PAM symbols, respectively.

We observe that $\mathcal{S}_C$ is an irregular $M_{AI}M_{BI}$-PAM constellation, and the power allocation factor, i.e., a, should satisfy $(M_{AI}-1)d_A\sqrt{a} < d_B\sqrt{1-a}$ to ensure that all M_{AI} superposition coded symbols associated with each s_B are located in the decision region for that s_B. An example of $\mathcal{S}_C$ is given in Fig. 2, where $M_{AI}=2$, $M_{BI}=4$, and vertical lines denote decision boundaries for s_B.

Due to the superposition coding, the SEP of s_B is equal to that of s_C with the decision boundaries for s_B. Furthermore, because the constellation diagram of $\mathcal{S}_C$ is symmetrical, we only need to consider the error probability of those constellation points in the left half of the plane. To begin with, let us define two vectors that respectively contain the decision boundaries and positions of s_C from left to right in the left half of the plane, namely

$$\begin{aligned} \mathbf{u}_{BI} &= [u_{BI}(0), \ldots, u_{BI}(M_{BI}/2-1)]^T \\ &= d_B\sqrt{1-a}[-(M_{BI}-2), -(M_{BI}-4), \ldots, 0]^T, \end{aligned} \tag{48}$$

$$\mathbf{v}_{CI} = [v_{CI}(0), \ldots, v_{CI}(M_{AI}M_{BI}/2-1)]^T, \tag{49}$$

where $v_{CI}(\alpha) = \sqrt{a}\cdot v_{AI}(\mathrm{mod}(\alpha, M_{AI})) + \sqrt{1-a}\cdot v_{BI}(\lfloor\alpha/M_{AI}\rfloor)$, $\alpha = 0, \ldots, M_{AI}M_{BI}/2-1$. Here, $[v_{AI}(0), \ldots, v_{AI}(M_{AI}-1)]^T \triangleq \mathbf{v}_{AI}$ and $[v_{BI}(0), \ldots, v_{BI}(M_{BI}-1)]^T \triangleq \mathbf{v}_{BI}$ are two vectors containing the positions of signal points from left to right in M_{AI}-PAM and M_{BI}-PAM constellations, respectively.

Note that there is only a single decision boundary, i.e., $u_{BI}(0)$, for $v_{CI}(\alpha), \alpha \in \{0, \ldots, M_{AI}-1\}$. By contrast, there are two decision boundaries which lie at $u_{BI}(\lfloor\alpha/M_{AI}\rfloor - 1)$ and $u_{BI}(\lfloor\alpha/M_{AI}\rfloor)$, respectively, for $v_{CI}(\alpha), \alpha \in \{M_{AI}, \ldots, M_{AI}M_{BI}/2-1\}$. The distances between each signal point and the corresponding decision boundaries can be easily obtained from (48) and (49), and the error probability of each s_C can also be derived by considering that an error occurs if the noise exceeds either distance. Therefore, it can be shown that the conditional SEP of s_B at user A for PAM is formulated as

$$\begin{aligned} P_{s_B-I|\gamma_{tot}^{SA}}^A = \frac{2}{M_{AI}M_{BI}}\Bigg\{ &\sum_{\alpha=0}^{M_{AI}-1} Q\left(\sqrt{2\gamma_{tot}^{SA}}d_{\alpha 0-I}\right) \\ &+ \sum_{\alpha=M_{AI}}^{M_{AI}M_{BI}/2-1}\left\{Q\left(\sqrt{2\gamma_{tot}^{SA}}d_{\alpha 1-I}\right)\right. \\ &\left.\left. + Q\left(\sqrt{2\gamma_{tot}^{SA}}d_{\alpha 2-I}\right)\right\}\right\}, \end{aligned} \tag{50}$$

where $d_{\alpha 0-I} = |v_{CI}(\alpha) - u_{BI}(0)|$, $d_{\alpha 1-I} = |v_{CI}(\alpha) - u_{BI}(\lfloor\frac{\alpha}{M_{AI}}\rfloor - 1)|$, and $d_{\alpha 2-I} = |v_{CI}(\alpha) - u_{BI}(\lfloor\frac{\alpha}{M_{AI}}\rfloor)|$; $\gamma_{tot}^{SA} = \sum_{k=1}^{N_r^A}\gamma_k^{SA}$ is the total instantaneous

TABLE I
OBTAINED $\hat{a}_{opt}$ VALUES FROM COMPUTER SEARCH FOR DIFFERENT CONSTELLATION COMBINATIONS OF s_A AND s_B.

$(\mathcal{S}_A, \mathcal{S}_B)$	(2-PAM, 2-PAM) (4-QAM, 4-QAM)	(4-QAM, 8-QAM)	(4-QAM, 16-QAM)	(8-QAM, 8-QAM)	(8-QAM, 16-QAM)
$\hat{a}_{opt}$	0.22	0.09	0.051	0.061	0.038

receive SNR with $\gamma_k^{SA} = |h_k|^2/N_0$. The average SEP, i.e., $P_{s_B-I}^A$, then can be derived by averaging (50) over the joint PDF of $\{\gamma_k^{SA}\}_{k=1}^{N_r^A}$. After several mathematical manipulations similar to (14)-(19), we obtain

$$P_{s_B-I}^A = \frac{2}{M_{AI}M_{BI}}\left\{\sum_{\alpha=0}^{M_{AI}-1} T_{0-I}(\alpha) + \sum_{\alpha=M_{AI}}^{M_{AI}M_{BI}/2-1}\left[T_{1-I}(\alpha)+T_{2-I}(\alpha)\right]\right\}, \quad (51)$$

where

$$T_{i-I}(\alpha) = \left(\frac{1-\mu_{i-I}(\alpha)}{2}\right)^{N_r^A}\sum_{k=0}^{N_r^A-1}\left[\binom{N_r^A-1+k}{k} \times \left(\frac{1+\mu_{i-I}(\alpha)}{2}\right)^k\right], \quad i=0,1,2 \quad (52)$$

with

$$\mu_{i-I}(\alpha) = \sqrt{\frac{\lambda^{SA}d_{\alpha i-I}^2}{N_0+\lambda^{SA}d_{\alpha i-I}^2}}. \quad (53)$$

Finally, substituting (51) into (45) yields a closed-form expression for P_A.

For the calculation of P_B, we first address the derivation of P_{s_B} as follows. Since the MRC at user B is the optimal diversity combiner, it can be shown that the SEP conditioned on γ_{tot}^{SB} and γ_{tot}^{AB} is given at the top of the next page as (54), where $z_{\alpha 0-I} = |v_{BI}(0) - u_{BI}(0)|$, $z_{\alpha 1-I} = |v_{BI}(\lfloor\frac{\alpha}{M_{AI}}\rfloor) - u_{BI}(\lfloor\frac{\alpha}{M_{AI}}\rfloor - 1)|$, and $z_{\alpha 2-I} = |v_{BI}(\lfloor\frac{\alpha}{M_{AI}}\rfloor) - u_{BI}(\lfloor\frac{\alpha}{M_{AI}}\rfloor)|$; $d_{\alpha 0-I}$, $d_{\alpha 1-I}$, and $d_{\alpha 2-I}$ are defined in (50); $\gamma_{tot}^{SB} = \sum_{k=1}^{N_r^A}\gamma_k^{SB} = \sum_{k=1}^{N_r^A}|g_k|^2/N_0$ and $\gamma_{tot}^{AB} = \sum_{\tau=1}^{N_r^A}\gamma_\tau^{AB} = \sum_{\tau=1}^{N_r^A}|f_\tau|^2/N_0$. Since random variables $\{\gamma_k^{SB}\}_{k=1}^{N_r^B}$ and $\{\gamma_\tau^{AB}\}_{\tau=1}^{N_r^B}$ are assumed to be independent, the average SEP, i.e., P_{s_B-I}, becomes

$$P_{s_B-I} = \underbrace{\int_0^\infty\cdots\int_0^\infty}_{2N_r^B-fold} P_{s_B-I|\gamma_{tot}^{SB},\gamma_{tot}^{AB}} \times \prod_{k=1}^{N_r^B} p_{\gamma_k^{SB}}\left(\gamma_k^{SB}\right)\cdot\prod_{\tau=1}^{N_r^B} p_{\gamma_\tau^{AB}}\left(\gamma_\tau^{AB}\right) \times d\gamma_1^{SB}\cdots d\gamma_{N_r^B}^{SB}d\gamma_1^{AB}\cdots d\gamma_{N_r^B}^{AB}, \quad (55)$$

where $p_{\gamma_k^{SB}}\left(\gamma_k^{SB}\right) = \exp(-\gamma_k^{SB}/\bar{\gamma}^{SB})/\bar{\gamma}^{SB}$ and $p_{\gamma_\tau^{AB}}\left(\gamma_\tau^{AB}\right) = \exp(-\gamma_\tau^{AB}/\bar{\gamma}^{AB})/\bar{\gamma}^{AB}$ are the PDFs of γ_k^{SB} and γ_τ^{AB}, respectively, with $\bar{\gamma}^{SB} = \lambda^{SB}/N_0$ and $\bar{\gamma}^{AB} = \lambda^{AB}/N_0$. By using the alternative representation of Gaussian Q-function in (14), (55) can be expressed as [36, Eq. (5A.58)]

$$P_{s_B-I} = \frac{2}{M_{AI}M_{BI}}\left\{\sum_{\alpha=0}^{M_{AI}-1} W_{0-I}(\alpha) + \sum_{\alpha=M_{AI}}^{M_{AI}M_{BI}/2-1}\left[W_{1-I}(\alpha)+W_{2-I}(\alpha)\right]\right\}, \quad (56)$$

where $W_{i-I}(\alpha)$ is given at the top of the next page as (57) with

$$c_1 = \lambda^{SB}d_{\alpha i-I}^2/N_0, \quad c_2 = \lambda^{AB}z_{\alpha i-I}^2/N_0, \quad (58)$$

$$B_k = \frac{A_k}{\binom{2N_r^B-1}{k}}, \quad C_k = \sum_{n=0}^{N_r^B-1}\frac{\binom{k}{n}}{\binom{2N_r^B-1}{n}}A_n, \quad (59)$$

$$A_k = (-1)^{N_r^B-1+k}\frac{\binom{N_r^B-1}{k}}{(N_r^B-1)!}\prod_{\substack{n=1\\ n\neq k+1}}^{N_r^B}(2N_r^B-n), \quad (60)$$

and

$$I_k(c) = 1-\sqrt{\frac{c}{1+c}}\left[1+\sum_{n=1}^{k}\frac{(2n-1)!!}{n!2^n(1+c)^n}\right]. \quad (61)$$

Afterwards, we can formulate P_B based on (51) and (56).

2) QAM Constellation: In this case, we consider that s_A and s_B are drawn from $M_{AI}\times M_{AQ}(=M_A)$ and $M_{BI}\times M_{BQ}(=M_B)$ rectangular QAM constellations, respectively. Since an M_A-QAM (M_B-QAM) constellation is composed of the quadrature combination of M_{AI}-PAM (M_{BI}-PAM) in the I- channel and M_{AQ}-PAM (M_{BQ}-PAM) in the Q- channel, the BEP analysis for QAM constellations is similar to that for PAM constellations addressed above. We redefine some parameters for the Q- dimension including the decision boundaries, positions of s_A, s_B and s_C, in a similar fashion. Based on $P_{s_B-I}^A$ in (51), the error probability for the Q- dimension, i.e., $P_{s_B-Q}^A$, is straightforward. Note that for QAM constellations, d_A and d_B become $d_A = \sqrt{3/(M_{AI}^2+M_{AQ}^2-2)}$ and $d_B = \sqrt{3/(M_{BI}^2+M_{BQ}^2-2)}$, respectively. Since a symbol error of QAM demodulation occurs when an incorrect symbol decision is made on the I- or Q- channel, the SEP of s_B at user A for QAM is given by

$$P_{s_B}^A = 1-\left(1-P_{s_B-I}^A\right)\left(1-P_{s_B-Q}^A\right). \quad (62)$$

Likewise, P_{s_B} for QAM is obtained as

$$P_{s_B} = 1-\left(1-P_{s_B-I}\right)\left(1-P_{s_B-Q}\right). \quad (63)$$

Finally, the BEP expressions for users A and B with QAM can be derived by substituting (62) into (45) and (63) into (47),

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2898099, IEEE Journal of Selected Topics in Signal Processing

$$P_{s_B-I|\gamma_{tot}^{SB},\gamma_{tot}^{AB}} = \frac{2}{M_{AI}M_{BI}}\Bigg\{\sum_{\alpha=0}^{M_{AI}-1} Q\left(\sqrt{2\left(\gamma_{tot}^{SB}d_{\alpha 0-I}^2+\gamma_{tot}^{AB}z_{\alpha 0-I}^2\right)}\right) + \sum_{\alpha=M_{AI}}^{M_{AI}M_{BI}/2-1}\Big[Q\left(\sqrt{2\left(\gamma_{tot}^{SB}d_{\alpha 1-I}^2+\gamma_{tot}^{AB}z_{\alpha 1-I}^2\right)}\right) + Q\left(\sqrt{2\left(\gamma_{tot}^{SB}d_{\alpha 2-I}^2+\gamma_{tot}^{AB}z_{\alpha 2-I}^2\right)}\right)\Big]\Bigg\} \tag{54}$$

$$\begin{aligned} W_{i-I}(\alpha) &= \frac{1}{\pi}\int_0^{\pi/2}\left[\int_0^{\infty}\exp\left(-\frac{\gamma_k^{SB}d_{\alpha i}^2}{\sin^2\phi}\right)p_{\gamma_k^{SB}}\left(\gamma_k^{SB}\right)d\gamma_k^{SB}\right]^{N_r^B}\left[\int_0^{\infty}\exp\left(-\frac{\gamma_\tau^{AB}z_{\alpha i}^2}{\sin^2\phi}\right)p_{\gamma_\tau^{AB}}\left(\gamma_\tau^{AB}\right)d\gamma_\tau^{AB}\right]^{N_r^B}d\phi \\ &= \frac{1}{\pi}\int_0^{\pi/2}\left(\frac{\sin^2\phi}{\bar{\gamma}^{SB}d_{\alpha i-I}^2+\sin^2\phi}\right)^{N_r^B}\left(\frac{\sin^2\phi}{\bar{\gamma}^{AB}z_{\alpha i-I}^2+\sin^2\phi}\right)^{N_r^B}d\phi \\ &= \frac{\left(\frac{c_1}{c_2}\right)^{N_r^B-1}}{2\left(1-\frac{c_1}{c_2}\right)^{2N_r^B-1}}\left\{\sum_{k=0}^{N_r^B-1}\left(\frac{c_2}{c_1}-1\right)^k B_k I_k(c_2)-\frac{c_1}{c_2}\sum_{k=0}^{N_r^B-1}\left(1-\frac{c_1}{c_2}\right)^k C_k I_k(c_1)\right\} \end{aligned} \tag{57}$$

respectively. Similar to the observation in CRS-SM-NOMA, the diversity orders of users A and B for CRS-NOMA are given by N_r^A and $\min\{2N_r^B, N_r^A\}$, respectively.

Since the value of power allocation factor, i.e., a, highly impacts the BER performance of both users, it should be optimized to minimize their BER values. Unfortunately, it seems impossible to obtain a value of a that optimizes the BER performance of both users simultaneously. As a compromise, we select

$$\hat{a}_{opt} = \arg\min_a \max\{P_A, P_B\} \tag{64}$$

as the optimal a in the sense of minimizing the system BER values. Due to the complexity of P_A and P_B, we perform exhaustive computer search to derive $\hat{a}_{opt}$ for different constellation combinations, whose results are summarized in Table I.

B. Sum Rate Analysis for CRS-NOMA Under Gaussian Input

In this subsection, the achievable sum rate of CRS-NOMA systems is studied under Gaussian input.

During the first time slot, symbol s_B is first decoded with the receive SNR

$$\gamma_B^A = \frac{(1-a)\rho\|\mathbf{h}\|^2}{a\rho\|\mathbf{h}\|^2+1}, \tag{65}$$

and then symbol s_A is detected after the SIC at user A with the receive SNR

$$\gamma_A = a\rho\|\mathbf{h}\|^2. \tag{66}$$

On the condition of correct detection of s_B by user A at the first phase, it can be figured out from (43) that the receive SNR for s_B at user B during the second time slot is given by

$$\gamma_B^B = \frac{(1-a)\rho\|\mathbf{g}\|^2}{a\rho\|\mathbf{g}\|^2+1}+\rho\|\mathbf{f}\|^2, \tag{67}$$

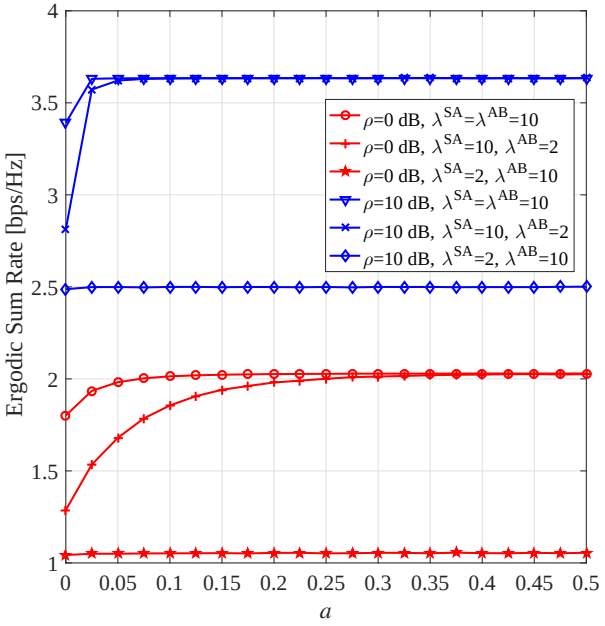

Fig. 3. Ergodic sum rate versus a for CRS-NOMA ($N_r^A = N_r^B = 2$ and $\lambda^{SB} = 1$).

which is the sum of SNRs provided by $\mathbf{y}^{SB}$ and $\mathbf{y}^{AB}$. Therefore, the instantaneous sum rate of CRS-NOMA is explicitly expressed as

$$C_{\text{CRS-NOMA}} = C_A + C_B, \tag{68}$$

where

$$C_A = \frac{1}{2}\log_2(1+\gamma_A), \tag{69}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2898099, IEEE Journal of Selected Topics in Signal Processing

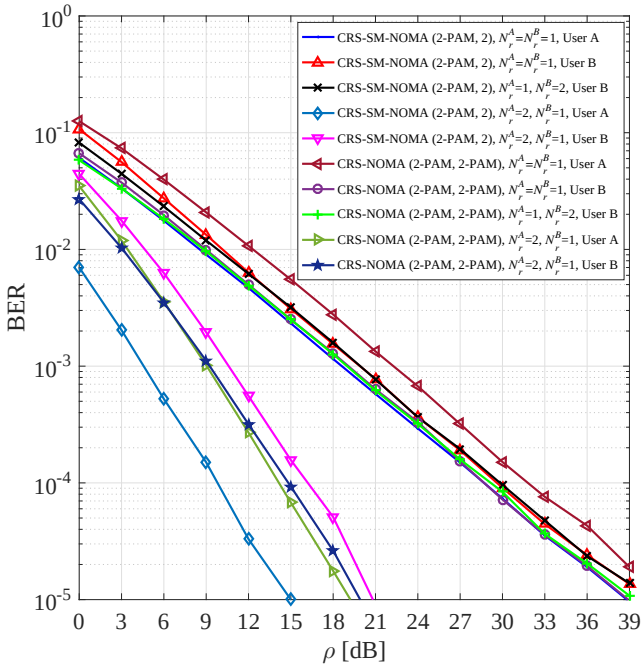

Fig. 4. BER performance comparison between "CRS-SM-NOMA (2-PAM, 2)" and "CRS-NOMA (2-PAM, 2-PAM)", where $N_r^A = 1, 2$, $N_r^B = 1, 2$, and $a = 0.22$.

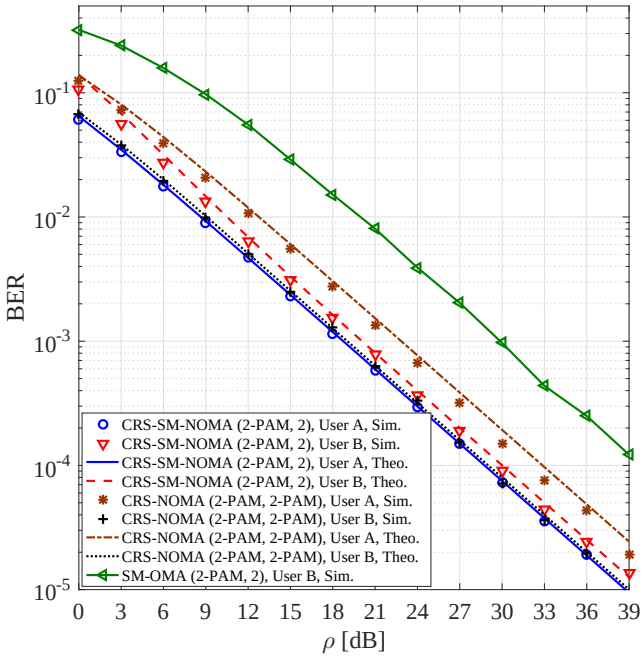

Fig. 5. BER performance comparison among "CRS-SM-NOMA (2-PAM, 2)", "CRS-NOMA (2-PAM, 2-PAM)", and "SM-OMA (2-PAM, 2)", where $N_r^A = N_r^B = 1$ and $a = 0.22$.

and

$$C_B = \frac{1}{2}\log_2\left(1 + \min\left(\gamma_B^A, \gamma_B^B\right)\right). \tag{70}$$

To see how the power allocation factor affects the ergodic sum rate in CRS-NOMA, we present the curves of the ergodic sum rate versus a with different system setups in Fig. 3, where $N_r^A = N_r^B = 2$ and $\lambda^{SB} = 1$. As seen from Fig. 3, a higher transmit SNR yields a larger ergodic sum rate as expected. Furthermore, the ergodic sum rate increases with a, especially when $\lambda^{SA} \geq \lambda^{AB}$. Therefore, we select $\tilde{a}_{opt} = 0.5$ as the optimal power allocation coefficient in the sense of maximizing the ergodic sum rate.

When $a = 0.5$, at high SNR values, we have

$$C_B \approx \frac{1}{2}, \tag{71}$$

such that (68) reduces to

$$\begin{aligned} C_{\text{CRS-NOMA}} &\approx \frac{1}{2}\log_2\left(0.5\rho\left\|\mathbf{h}\right\|^2\right) + \frac{1}{2} \\ &= \frac{1}{2}\log_2\left(\rho\|\mathbf{h}\|^2\right). \end{aligned} \tag{72}$$

By comparing (38) with (72), we conclude that our proposed CRS-SM-NOMA almost achieves an increase of $\log_2(N_t)/2$ bps/Hz in sum rate over CRS-NOMA due to the use of spatial domain in data transmission.

V. Simulation Results and Discussion

In this section, we perform Monte Carlo simulations to examine the uncoded BER and the achievable rate performance of CRS-SM-NOMA, and compare them with those of CRS-NOMA and SM-OMA. In SM-OMA, the BS transmits two SM signals to users A and B, respectively, in two consecutive time slots, and there is no cooperation between two users. From the received signals, users A and B extract the M-ary symbol and the index bits as the desired information, respectively. In all simulations, the channels are assumed to be Rayleigh flat-faded and perfectly known to users A and B. The average powers of S-to-A, S-to-B, and A-to-B links are set to be $\lambda^{SA} = 10$, $\lambda^{SB} = 1$, and $\lambda^{AB} = 10$, respectively. For brevity, we will refer to "CRS-SM-NOMA/SM-OMA (M-PAM/QAM, N_t)" as the CRS-SM-NOMA/SM-OMA scheme in which the BS employs M-PAM/QAM with N_t transmit antennas, and "CRS-NOMA (M_A-PAM/QAM, M_B-PAM/QAM)" as the CRS-NOMA scheme in which the BS maps the information bits for users A and B into M_A-PAM/QAM and M_B-PAM/QAM symbols, respectively.

A. Uncoded BER Performance

In this subsection, the uncoded BER performance of CRS-SM-NOMA and CRS-NOMA is compared, assuming that users A and B in all considered schemes employ ML or joint ML detection. Since user A of SM-OMA performs the same process as that of CRS-SM-NOMA, the BER curves of user A for SM-OMA are not presented.

Fig. 4 shows the BER versus ρ curves of users A and B for "CRS-SM-NOMA (2-PAM, 2)" and "CRS-NOMA (2-PAM, 2-PAM)", where $N_r^A = 1, 2$, $N_r^B = 1, 2$, and $a = 0.22$. To see the effects of the number of receive antennas clearly, the theoretical BEP curves are not included. As seen from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2898099, IEEE Journal of Selected Topics in Signal Processing

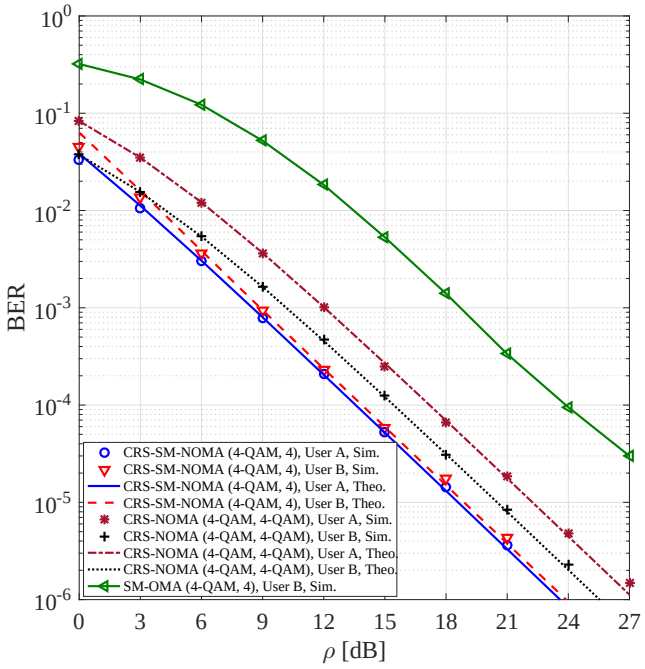

Fig. 6. BER performance comparison among "CRS-SM-NOMA (4-QAM, 4)", "CRS-NOMA (4-QAM, 4-QAM)", and "SM-OMA (4-QAM, 4)", where $N_r^A = N_r^B = 2$ and $a = 0.22$.

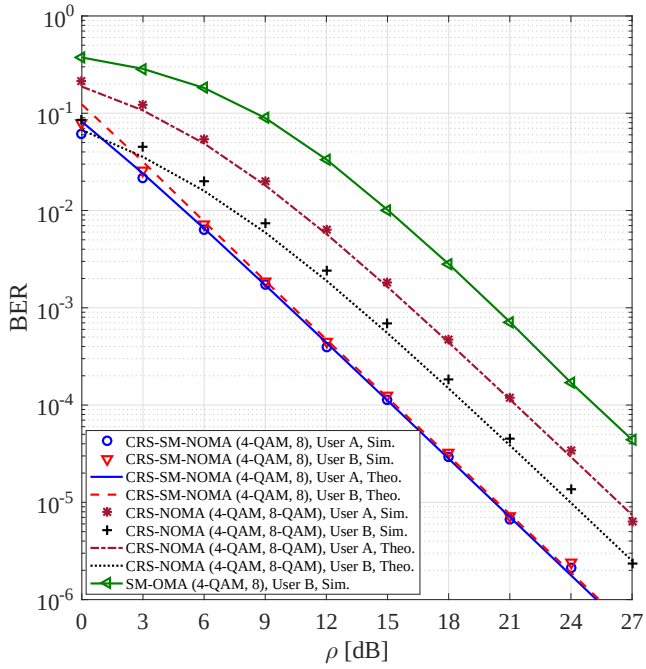

Fig. 7. BER performance comparison among "CRS-SM-NOMA (4-QAM, 8)", "CRS-NOMA (4-QAM, 8-QAM)", and "SM-OMA (4-QAM, 8)", where $N_r^A = N_r^B = 2$ and $a = 0.09$.

Fig. 4, since for both CRS-SM-NOMA and CRS-NOMA, the diversity orders of users A and B are given by N_r^A and $\min\{2N_r^B, N_r^A\}$, respectively, when $N_r^A \leq 2N_r^B$, increasing N_r^A improves the diversity orders of both users, while a larger N_r^B merely enhances the BER performance of user B in the low SNR region slightly. Although the curve with $N_r^A = N_r^B = 2$ is not provided in the figure, it is expected to be the same as that with $N_r^A = 2$ and $N_r^B = 1$ for both CRS-SM-NOMA and CRS-NOMA, except that a minor improvement can be observed in the low SNR region for user B. Because user A is influenced by the error propagation of SIC in CRS-NOMA, we observe that user A of CRS-SM-NOMA outperforms that of CRS-NOMA, achieving SNR gains of about 3.5 dB and 4.5 dB when $N_r^A = 1$ and $N_r^A = 2$, respectively, at a BER value of 10^{-4}. On the contrary, user B of CRS-SM-NOMA performs worse than that of CRS-NOMA. Fortunately, the performance loss is less than 1.5 dB.

In Fig. 5, we compare the BER performance of "CRS-SM-NOMA (2-PAM, 2)", "CRS-NOMA (2-PAM, 2-PAM)", and "SM-OMA (2-PAM, 2)", where $N_r^A = N_r^B = 1$ and theoretical BEP results are also presented. It can be seen from Fig. 5 that the theoretical BEP curves accurately match with their simulation counterparts over the considered SNR region. In CRS-SM-NOMA, the BER performance of user A is superior to that of user B, while the situation is reversed in CRS-NOMA. Although not both users of CRS-SM-NOMA perform better than those of CRS-NOMA, we observe that the performance gap between two users in CRS-SM-NOMA, which is about 1 dB, is much smaller than that in CRS-NOMA, which is about 3 dB. This means that CRS-SM-NOMA provides a higher user fairness in error performance than CRS-NOMA. Moreover, user B of CRS-SM-NOMA achieves about 10 dB SNR gain over that of SM-OMA.

Fig. 6 depicts the performance comparisons among "CRS-SM-NOMA (4-QAM, 4)", "CRS-NOMA (4-QAM, 4-QAM)", and "SM-OMA (4-QAM, 4)", where $N_r^A = N_r^B = 2$ and $a = 0.22$. As shown in Fig. 6, all theoretical BEP results agree with their computer simulation counterparts quite well in the entire SNR region. User B of SM-OMA performs the worst among all schemes without cooperating with user A. Different from the observations in Figs. 4 and 5, both users of CRS-SM-NOMA outperform those of CRS-NOMA. More specifically, users A and B in CRS-SM-NOMA obtain SNR gains of about 5 dB and 2.5 dB over those in CRS-NOMA, respectively, at a BER value of 10^{-5}. Moreover, a better BER fairness between two users is provided by CRS-SM-NOMA.

Fig. 7 illustrates the comparison results among "CRS-SM-NOMA (4-QAM, 8)", "CRS-NOMA (4-QAM, 8-QAM)", and "SM-OMA (4-QAM, 8)", where $N_r^A = N_r^B = 2$ and $a = 0.09$. Similar to Fig. 6, the approximate analytical BEP curves predict the simulation results accurately for all SNR values, and the BER performance of CRS-SM-NOMA is superior to that of CRS-NOMA for both users. User B of CRS-SM-NOMA preforms significantly better than that of SM-OMA, achieving an SNR gain of about 10 dB. Particularly, by comparing with Fig. 6, we observe that the performance improvements achieved by CRS-SM-NOMA over CRS-NOMA become more prominent in Fig. 7 due to higher constellations. Approximately 6 dB SNR gain is obtained by user A and 4 dB SNR gain by user B, at a BER value of 10^{-5}. Meanwhile,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2019.2898099, IEEE Journal of Selected Topics in Signal Processing

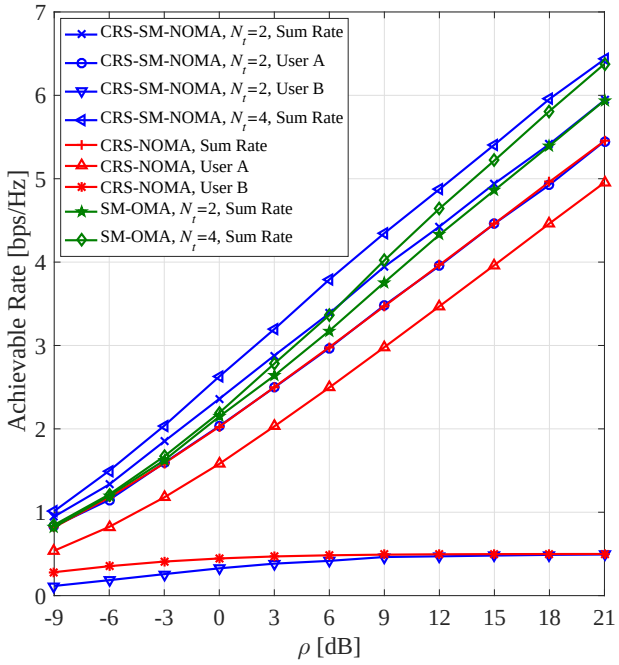

Fig. 8. Achievable rate comparison among CRS-SM-NOMA, CRS-NOMA, and SM-OMA, where $N_r^A = N_r^B = 2$, $N_t = 2, 4$, and $a = 0.5$.

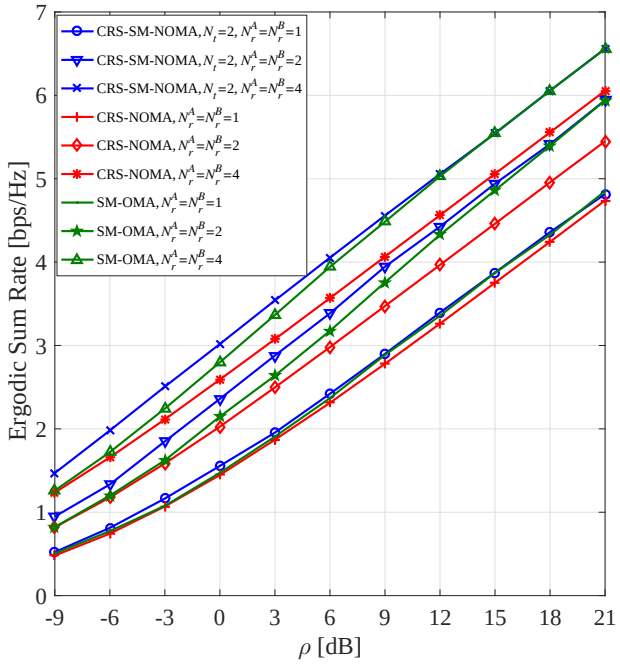

Fig. 9. Ergodic sum rate comparison among CRS-SM-NOMA, CRS-NOMA, and SM-OMA, where $N_t = 2$, $N_r^A = N_r^B = 1, 2, 4$, and $a = 0.5$.

CRS-SM-NOMA provides a higher fairness between two users, since their BER curves almost overlap.

B. Achievable Rate Performance

In Fig. 8, we make comparisons among CRS-SM-NOMA, CRS-NOMA, and SM-OMA in terms of achievable rate, where $N_r^A = N_r^B = 2$, $N_t = 2, 4$, and $a = 0.5$. As shown in Fig. 8, for both CRS-SM-NOMA with $N_t = 2$ and CRS-NOMA, the achievable rate of user B is saturated to 0.5 bps/Hz, which corroborates the analysis in (37) and (71). For all considered schemes, the sum rate as well as the achievable rate of user A grows linearly with ρ. The ergodic sum rate of CRS-NOMA nearly equals the achievable rate of user A in CRS-SM-NOMA with $N_t = 2$, and CRS-SM-NOMA achieves a sum rate improvement of about $\log_2(N_t)/2$ bps/Hz over CRS-NOMA for a fixed ρ, which can be explained by (38) and (72). Further, as indicated by (38), the sum rate of CRS-SM-NOMA improves with N_t, and an increase of about 0.5 bps/Hz can be obtained when N_t is doubled. The ergodic sum rate of CRS-SM-NOMA is larger than that of SM-OMA in the low SNR region, while at high SNR, they are nearly equal.

Fig. 9 gives the comparison results among the ergodic sum rates of CRS-SM-NOMA, CRS-NOMA, and SM-OMA, where $N_t = 2$, $N_r^A = N_r^B = 1, 2, 4$, and $a = 0.5$. It can be observed from Fig. 9 that increasing the number of receive antennas promotes significantly the ergodic sum rates for all schemes. However, different from the effect of doubling N_t in Fig. 8, the increments in ergodic sum rate gradually reduce as N_r^A or N_r^B doubles. Moreover, with $N_r^A = N_r^B = 1$, CRS-SM-NOMA performs slightly better than CRS-NOMA. By contrast, CRS-SM-NOMA achieves an increase of about $\log_2(N_t)/2 = 0.5$ bps/Hz in ergodic sum rate over CRS-NOMA when $N_r^A = N_r^B = 2, 4$. These can be explained as follows. With $N_r^A = N_r^B = 1$, the index of the active antenna is very likely to be detected erroneously at user B of CRS-SM-NOMA, which is very common for IM-based systems with single receive antenna, and thus its achievable rate is far less than $\log_2(N_t)/2$. Increasing N_r^A and N_r^B to 2 or 4 greatly improves the reliability of the active antenna index detection such that the approximation in (37) makes sense. Similar to the observation in Fig. 8, CRS-SM-NOMA obtains a larger ergodic sum rate than SM-OMA at low SNR, and the gap narrows with increasing SNR.

VI. Conclusion

In this paper, we have proposed a new three-node CRS using SM-NOMA, in which the information of two users per transmission is carried by the index of the active antenna and the ordinary APM symbol, respectively. The near user also functions as a relay to assist the far user. As the reference scheme, conventional CRS-NOMA has been re-investigated with multiple receive antennas. For both CRS-SM-NOMA and CRS-NOMA, approximate BEP expressions have been derived in closed form for both users employing PAM/QAM over Rayleigh flat fading channels, and the system sum rates under Gaussian input have also been developed. Computer simulations have verified the performance analysis, and shown that the proposed CRS-SM-NOMA achieves better BER performance and a larger sum rate than CRS-NOMA. We conclude that the proposed CRS-SM-NOMA can be considered as a candidate for a fair, reliable and high throughput multiple access technique towards next-generation wireless networks.

This work focuses only on the two-user scenario. Actually, CRS-SM-NOMA can be extended to more than two users through power multiplexing or SDMA, which involves resource allocation, user pairing, beamforming design, etc. We leave this extension for future study.

REFERENCES

[1] L. Dai, B. Wang, Y. Yuan, S. Han, C.-L. I, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[2] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[3] 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.

[4] 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.

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

[6] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Non-orthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

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

[8] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[9] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[10] J.-B. Kim and I.-H. Lee, "Capacity analysis of cooperative relaying systems using non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1949–1952, Nov. 2015.

[11] M. Xu, F. Ji, M. Wen, and W. Duan, "Novel receiver design for the cooperative relaying system with non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1679–1682, Aug. 2016.

[12] J. Men, J. Ge, and C. Zhang, "Performance analysis of nonorthogonal multiple access for relaying networks over Nakagami-m fading channels," *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1200–1208, Feb. 2017.

[13] D. Wan, M. Wen, F. Ji, Y. Liu, and Y. Huang, "Cooperative NOMA systems with partial channel state information over Nakagami-m fading channels," *IEEE Trans. Commun.* vol. 66, no. 3, pp. 947–958, Mar. 2018.

[14] C. Zhong and Z. Zhang, "Non-orthogonal multiple access with cooperative full-duplex relaying," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2478–2481, Dec. 2016.

[15] L. Zhang, J. Liu, M. Xiao, G. Wu, Y.-C. Liang, and S. Li, "Performance analysis and optimization in downlink NOMA systems with cooperative full-duplex relaying," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2398–2412, Oct. 2017.

[16] B. Zheng, X. Wang, M. Wen, and F. Chen, "NOMA-based multi-pair two-way relay networks with rate splitting and group decoding," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2328–2341, Oct. 2017.

[17] X. Yue, Y. Liu, S. Kao, A. Nallanathan, and Y. Chen, "Modeling and analysis of two-way relay non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3784–3796, Sep. 2018.

[18] Y. Liu, H. Xing, C. Pan, A. Nallanathan, M. Elkashlan, and L. Hanzo, "Multiple-antenna-assisted non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 17–23, Apr. 2018.

[19] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee, and H. V. Poor, "Coordinated beamforming for multi-cell MIMO-NOMA," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 84–87, Jan. 2017,

[20] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[21] D. Zhang, Z. Zhou, C. Xu, Y. Zhang, J. Rodriguez, and T. Sato, "Capacity analysis of NOMA with mmWave massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol .35, no. 7, pp. 1606–1618, Jul. 2017

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

[23] Y. Huang, C. Zhang, J. Wang, Y. Jing, L. Yang, and X. You, "Signal processing for MIMO-NOMA: Present and future challenges," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 32–38, Apr. 2018.

[24] R. Mesleh, H. Haas, S. Sinanovic, C. Ahn, and S. Yun, "Spatial modulation," *IEEE Trans. Veh. Technol.*, vol. 57, no. 4, pp. 2228–2241, Jul. 2008.

[25] M. Wen, X. Cheng, and L. Yang, *Index Modulation for 5G Wireless Communications*. Berlin, Germany: Springer, 2017.

[26] E. Basar, M. Wen, R. Mesleh, M. Di Renzo, Y. Xiao, and H. Haas, "Index modulation techniques for next-generation wireless networks," *IEEE Access*, vol. 5, pp. 16693–16746, Sep. 2017.

[27] J. Jeganathan, A. Ghrayeb, L. Szczecinski, and A. Ceron, "Space shift keying modulation for MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 7, pp. 3692–3703, Jul. 2009.

[28] X. Wang, J. Wang, L. He, Z. Tang, and J. Song, "On the achievable spectral efficiency of spatial modulation aided downlink non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 1937–1940, Sep. 2017.

[29] X. Zhu, Z. Wang, and J. Cao, "NOMA-based spatial modulation," *IEEE Access*, vol. 5, pp. 3790–3800, Mar. 2017.

[30] Y. Chen, L. Wang, Y. Ai, B. Jiao, and L. Hanzo, "Performance analysis of NOMA-SM in vehicle-to-vehicle massive MIMO channels," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2653–2666, Dec. 2017.

[31] C. Zhong, X. Hu, X. Chen, D. W. K. Ng, and Z. Zhang, "Spatial modulation assisted multi-antenna non-orthogonal multiple access," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 61–67, Apr. 2018.

[32] X. Wang, F. Labeau, and L. Mei, "Closed-form BER expressions of QPSK constellation for uplink non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2242–2245, Oct. 2017.

[33] E. C. Cejudo, H. Zhu, and O. Alluhaibi, "On the power allocation and constellation selection in downlink NOMA," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Toronto, Canada, Sep. 2017, pp. 1–5.

[34] M. -S. Alouini and A. J. Goldsmith, "A unified approach for calculating error rates of linearly modulated signals over generalized fading channels," *IEEE Trans. Commun.*, vol. 47, no. 9, pp. 1324–1334, Sep. 1999.

[35] K. Cho and D. Yoon, "On the general BER expression of one- and two-dimensional amplitude modulations," *IEEE Trans. Commun.*, vol. 50, no. 7, pp. 1074–1080, Jul. 2002.

[36] M. K. Simon and M.-S. Alouini, *Digital Communication Over Fading Channels*, 2nd ed. New York, NY, USA: Wiley, 2005.

[37] M. K. Simon, *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers and Scientists*. Berlin, Germany: Springer, 2006.

[38] Y. Yang and B. Jiao, "Information-guided channel-hopping for high data rate wireless communication," *IEEE Commun. Lett.*, vol. 12, no. 4, pp. 225–227, Apr. 2008.