NOMA Decoding: Successive Interference Cancellation or Maximum Likelihood Detection?

Yue Qi[†] and Mojtaba Vaezi[‡]
[†]Samsung Research America, 6625 Excellence Way, Plano, TX 75023, USA
[‡]Wireless Networking Lab, Villanova University, Villanova, PA 19085, USA
Email: {yqi, mvaezi}@villanova.edu

2024 58th Annual Conference on Information Sciences and Systems (CISS) | 979-8-3503-6929-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/CISS59072.2024.10480218

***Abstract*—Is successive interference cancellation (SIC) decoding always the optimal choice in non-orthogonal multiple access (NOMA) systems? While the answer is positive based on Shannon theory, which is applicable to infinite-length codewords drawn from a Gaussian distribution, this may not universally hold for systems with finite-alphabet inputs. Specifically, in this paper, we demonstrate that for quadrature amplitude modulation (QAM)-based NOMA, SIC decoding fails for certain values of power allocation coefficient α, used to divide power among NOMA users. With this observation, we propose employing maximum likelihood (ML) detection to decode QAM-NOMA. While SIC decoding for QAM-NOMA requires allocating higher power to the user with a weaker channel to prevent symbol crossing in super-constellations, ML detection can successfully handle a broader range of power allocation coefficients. We then derive closed-form symbol error rates for quadrature phase shift keying-based NOMA systems across any α and validate them through simulations. The results demonstrate the effectiveness of ML detection, particularly in scenarios where SIC decoding fails.**

***Index Terms*—NOMA, successive interference cancellation (SIC), finite-alphabet input, maximum likelihood detection.**

I. Introduction

Non-orthogonal multiple access (NOMA) holds the promise of expanding the number of users and improving spectral efficiency of wireness networks [1]–[4]. Key techniques for achieving the capacity region of the single-input and single-output (SISO) NOMA, also referred to as power-domain NOMA [2]–[5], are *superposition coding (SC)* and *successive interference cancellation (SIC)*. This SISO NOMA model is essentially analogous to the widely recognized broadcast channel (BC) [6]–[8]. Similar to other Shannon-theoretic capacity limits, in the context of the Gaussian BC achievability is established by employing Gaussian codewords with lengths approaching infinity.

Inspired by the above, SC-SIC is applied to finite-length inputs built on finite alphabet constellations such as quadrature amplitude modulation (QAM) constellations [2], [3], [9]–[12]. Nonetheless, the practical application of such principles may not align with the same theoretical assertions [13], [14]. This discrepancy have contributed to misconceptions, such as the notion that users with smaller channel gains should be allocated higher power [3], [9], [10]. While these misconceptions are refuted based on theoretically optimal inputs [15], this paper delves into the discussion of such confusions specifically for finite-alphabet inputs, with a focus on QAM constellations.

This work was supported by the U. S. National Science Foundation under Grant ECCS-2301778.

The two premises, namely, NOMA with theoretically optimal inputs and finite-alphabet inputs, differ in several aspects. Consider a two-user SISO NOMA and let αP, $(0 \leq \alpha \leq 1)$, be a fraction of the transmitter power P allocated to the user with the weaker channel gain. Here are key observations:

- In theory, for any arbitrary value of α, decoding can be completed successfully in both users [15]. However, with inputs drawn from QAM constellations, careful selection of α is essential to avoid constellation overlap and prevent erroneous decisions.
- In theory, utilizing SIC at the user with the stronger channel is always optimal for achieving the capacity region with an arbitrarily small error probability. However, with finite-alphabet inputs, this approach is effective only for a subset of power allocation values.
- While theory strictly mandates the order of performing SIC, with the weak user consistently treating interference as noise and the strong user executing SIC, with QAM-based inputs, depending on the value of α, the order of SIC can be exchanged, allowing the weak user to perform SIC in certain cases.

The main message of this paper is that SIC is not always the best decoding method when finite-alphabet inputs are used in NOMA. Some other work has avoided SIC by applying lattice [16] and index modulations [17], and other methods [18]. Our claim is different in that the alternative decoding method based on *maximum-likelihood* (ML) can support more values of α in QAM modulations. The contributions and conclusions of this paper can be summarized as follows:

- We demonstrate that ML decoding has the potential to outperform SIC-based decoding for finite-alphabet inputs. Notably, ML works in some scenarios where SIC fails.
- We reaffirm that NOMA itself, even with finite inputs, does not mandate the allocation of higher power to the weaker user. In fact, when finite-alphabet constellations are employed for NOMA users, less power can be allocated to the weak user [13], [17]. There is no restriction on power allocation parameter α.
- We derive a closed-form symbol error rate for a two-user quadrature phase shift keying (QPSK)-based NOMA

system and validate it through simulations. An intriguing observation is that SIC can be employed for the weaker user when the power allocation coefficient is below a certain threshold.

The remainder of this paper is organized as follows: In the following section, we present the system model. Section III is dedicated to a detailed mathematical and graphical exploration of ML detection, including the analysis of symbol error rate and discussions on the SIC order. Numerical results are provided in Section IV, and we conclude the paper in Section V.

II. System Model

Consider a downlink NOMA system with one transmitter (Tx) and two users, where all nodes are equipped with a single antenna. The channels between the Tx and users are assumed to experience quasi-static flat Rayleigh fading, i.e., channels are constant within each transmission. In each transmission, we always name the user equipment (UE) with the weaker channel gain UE1 and the one with stronger channel gain as UE2. The data for UE1 and UE2 are modulated using QAM modulations with orders M_1 and M_2, and the corresponding alphabets of these two constellations are denoted as $\mathcal{A}$ and $\mathcal{B}$.

The symbols for the users are denoted by x_1 and x_2, and are assumed to have unit (average) powers. Then, to transmit the two signals simultaneously, a fraction $\alpha \in [0,1]$ of the total power P is assigned to UE1, and a fraction $\bar{\alpha} \triangleq 1-\alpha$ of the power is allocated to UE2. The superposition of the scaled symbols is given by

$$x = \sqrt{\alpha P}x_1 + \sqrt{\bar{\alpha} P}x_2. \tag{1}$$

Let the complex channel gains for UE1 and UE2 be h_1 and h_2, and assume $|h_1|^2 \leq |h_2|^2$. The received signal at user k, $k = 1, 2$, is given as

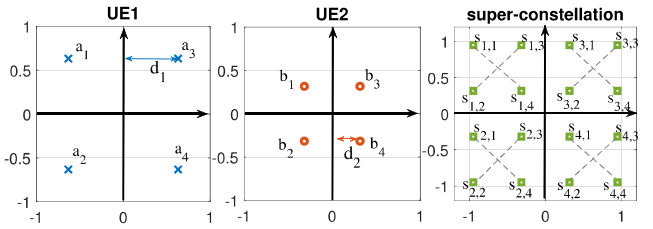
$$y_k = h_k x + n_k = h_k(\sqrt{\alpha P}x_1 + \sqrt{\bar{\alpha} P}x_2) + n_k. \tag{2}$$

Where n_k represents a complex noise. The real and imaginary parts of the noise are independent and identically distributed (i.i.d.) Gaussian random variables for user k, thus $\mathcal{R}\{n_k\}$ and $\mathcal{I}\{n_k\} \sim \mathcal{N}(0, \sigma_k^2)$. The received signal after equalizing the estimated channel $(\hat{h}_k)$ can be written as

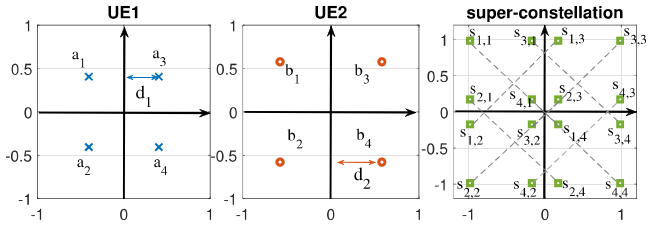
$$r_k = \frac{y_k}{\hat{h}_k} = \sqrt{\alpha P}x_1 + \sqrt{\bar{\alpha} P}x_2 + \hat{n}_k, \tag{3}$$

where $\hat{n}_k \triangleq \frac{n_k}{\hat{h}_k}$ is a scaled complex white Gaussian random noises with zero mean and variance $\hat{\sigma}_k^2 = \frac{2\sigma_k^2}{||\hat{h}_k||^2}$.

The two NOMA users decode the received symbols differently. UE1 directly decodes x_1 by treating interference as noise (TIN) whereas UE2 employs SIC to retrieve its symbols. With QAM constellations, a widely accepted assumption for SIC is to allocate more power to the weaker user in order to avoid constellation overlapping [3], [9], [10], [19]. In the following section, we show an alternative decoding method in which more values of α in QAM modulations can be supported.



(a) $\alpha = 0.84$



(b) $\alpha = 0.33$

Fig. 1: The construction of the superimposed symbols with different power allocation values.

III. Maximum-Likelihood Detection

Using QAM modulations of orders M_1 and M_2, respectively for UE1 and UE2, we denote

$$x_1 \in \mathcal{A} \triangleq \{a_1, \ldots, a_i, \ldots, a_{M_1}\}, \tag{4a}$$

$$x_2 \in \mathcal{B} \triangleq \{b_1, \ldots, b_j, \ldots, b_{M_2}\}, \tag{4b}$$

in which a_i and b_j are the ith and jth symbols in $\mathcal{A}$ and $\mathcal{B}$, respectively. All symbols are assumed to have an equal probability. The superimposed constellation consists of at most $M_1 \cdot M_2$ symbols, generated by (1). Thus, the superimposed symbol can be written as

$$x \in \mathcal{A} \otimes \mathcal{B} \triangleq \{s_{1,1}, \ldots, s_{i,j}, \ldots, s_{M_1,M_2}\} \tag{5}$$

in which $\otimes$ is tensor product and symbol $s_{i,j} = \sqrt{\alpha P}a_i + \sqrt{\bar{\alpha} P}b_j$. Once the power allocation coefficient α is fixed, the superimposed alphabet set is generated. The two users can decode their messages by applying the ML detection, i.e., measuring the minimum squared Euclidean distance between the superimposed alphabet set and the received signals, which for user k is

$$(i^*, j^*) = \arg\min_{i,j} ||r_k - s_{i,j}||^2. \tag{6}$$

Then, the decisions are $\hat{x}_1 = a_{i^*}$ and $\hat{x}_2 = b_{j^*}$, for UE1 and UE2, respectively. That is, we find the indices of the symbols.

A. Symbol Error Rate Analysis

We analyze the symbol error rates (SERs) for both users for the special case of $M_1 = M_2 = 4$, i.e., QPSK constellations or 4-QAM. Errors happen when the superimposed constellation points are overlapped. Let us define d_1 and d_2 as the distances of the symbols to the y-axis, as shown in Fig. 1. The distances are affected by the transmit power and α. The literature has

assumed $d_1 > d_2$ to avoid the constellation overlapping. We define the cut-off threshold of overlapping as α_0. In the two-QPSK NOMA system, there is one cut-off threshold $\alpha_0 = 0.5$, which is obtained from $d_1 = \sqrt{\frac{3\alpha P}{2(M_1-1)}} = d_2 = \sqrt{\frac{3\bar{\alpha} P}{2(M_2-1)}}$. Two constellation examples for $\alpha > \alpha_0$ and $\alpha \leq \alpha_0$ are illustrated in Fig. 1(a) and Fig. 1(b), respectively.

In [9], SERs for UE1 and UE2 are provided only for $\alpha > \alpha_0$, since SIC fails for $\alpha \leq \alpha_0$. As a result, it is concluded that NOMA can work only for $\alpha > \alpha_0 = 0.5$. This is a common myth that can be found in many other NOMA papers.

To overcome this deficiency, in the following, we use ML detection in both users and analyze the two error probabilities. To this end, we use ϵ_k to represent the SER for UEk. $\mathbb{P}(\hat{x}_k = x_k|\phi)$ denotes the probability of successfully decoding UEk under the condition that event ϕ happens. Also, the $Q(x)$ is the Gaussian Q-function defined as $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$.

Proposition 1. *For the two-user NOMA with QPSK is used for both users, the SER at UE1 and UE2 for $\alpha \leq \alpha_0$ can be expressed as*

$$\epsilon_1 = 1 - \mathbb{P}^2(\hat{x}_1 = x_1|\alpha \leq 0.5), \tag{7a}$$

$$\epsilon_2 = 1 - \mathbb{P}^2(\hat{x}_2 = x_2|\alpha \leq 0.5), \tag{7b}$$

in which $\mathbb{P}(\hat{x}_1 = x_1|\alpha \leq 0.5)$ and $\mathbb{P}(\hat{x}_2 = x_2|\alpha \leq 0.5)$ are given in (8) and (9) on top of next page.

Proof. The QAM can be considered as two independent components on real and imaginary axes and have identical probability [9]. Thus, a QAM-based NOMA symbol can be decomposed into two pulse amplitude modulation (PAM) symbols. The derivation is based on the superimposition of two independent PAMs. A simple way is to apply the symmetric property of the two 4-QAMs. Then, we can obtain the SERs by swapping M_1 and M_2, $\hat{\sigma}_1^2$ in [9, Proposition 3, 4], which is derived only for $\alpha > \alpha_0$. □

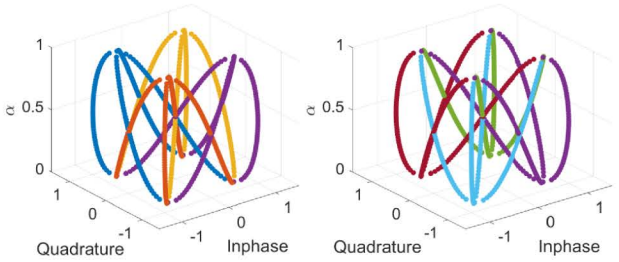
Remark 1. When $\alpha \leq 0.5$, SERs for UE1 and UE2 in two 4-QAM-based NOMA systems are ϵ_1 and ϵ_2 in (7). When $\alpha > 0.5$, the SERs are given in [9, Proposition 3, 4].

To demonstrate that UE1, whose channel gain smaller than UE2's channel gain, can operate effectively without necessitating a higher power allocation in the above scenario, we compute the derivative of (8) as illustrated in (10). This facilitates an analysis of the characteristics of (10) within the range $\alpha \leq 0.5$. Particularly, we can show that

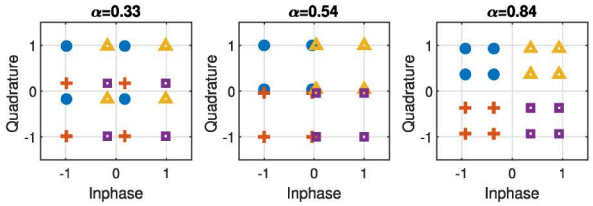
$$\left(\frac{d\mathbb{P}_1}{d\alpha}|_{\alpha\to 0}\right) \cdot \left(\frac{d\mathbb{P}_1}{d\alpha}|_{\alpha=0.5}\right) < 0,$$

indicating that (8) is non-monotonous and has an extremum.

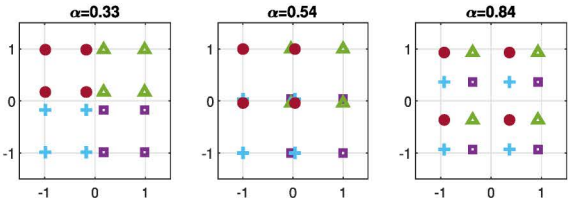
Remark 2. Proposition 1 and Remark 1 are for 4-QAM. For higher-order modulations, there may exist more than one cut-off threshold. The proposition still holds for some thresholds.



(a) The super-constellations versus α of UE1 (left) and UE2 (right).



(b) UE1



(c) UE2

Fig. 2: Superimposed constellation for different values of α observed at a) UE1, and b) UE2.

B. Illustration of the ML Detection

The ML method works by finding the nearest symbol of the superimposed constellation to the received signal. It can be applied for more values of α than the SIC decoding. We elaborate on this by considering a case for a QPSK-based NOMA system without noise. The superimposed constellations for UE1 and UE2 with continuous different αs are denoted in Fig 2(a), and three different values of α (the slices of the Fig 2(a) at $\alpha = 0.33, 0.54, 0.84$) are depicted in Fig 2. The cut-off thresholds are the cross-points in Fig. 2(a). The colors show the decision. For example, the blue circles in Fig. 2(b) for UE1 are $\{s_{1,1}, s_{1,2}, s_{1,3}, s_{1,4}\}$. If we receive any symbol that is close to this blue circle set, the symbols result in $\hat{x}_1 = a_1$. Similarly, the yellow triangles are $\{s_{3,1}, s_{3,2}, s_{3,3}, s_{3,4}\}$ and thus give $\hat{x}_1 = a_3$. On the other hand, the red circles in Fig. 2(c) are $\{s_{1,1}, s_{2,1}, s_{3,1}, s_{4,1}\}$ which all result in $\hat{x}_2 = b_1$. When α approaches 0.5, the 16-QAM-like super-constellation reduces to 9-QAM, because of symbols overlapping. In this case, there is no method to distinguish all symbols successfully.

C. SIC Order

Based on the theory of the BC, UE1 (which has a weaker channel) should employ TIN and UE2 should perform SIC. However, if this is followed for $\alpha \leq \alpha_0$, both users fail to

$$\mathbb{P}_1 = \mathbb{P}(\hat{x}_1 = x_1 | \alpha \le 0.5) = 1 - \frac{2(M_2-1)}{M_1 M_2} \times \sum_{n=1}^{M_1/2} \left[Q\Big(\frac{d_1 - (2n-1)d_2}{\sqrt{\hat{\sigma}_2^2/2}} \Big) + Q\Big(\frac{d_1 + (2n-1)d_2}{\sqrt{\hat{\sigma}_2^2/2}} \Big) \right].$$

$$= 1 - \frac{2(M_2-1)}{M_1 M_2} \times \sum_{n=1}^{M_1/2} \left[Q\Big(\frac{\sqrt{\frac{3\alpha P}{2(M_1-1)}} - (2n-1)\sqrt{\frac{3\bar{\alpha} P}{2(M_2-1)}}}{\sqrt{\hat{\sigma}_2^2/2}} \Big) + Q\Big(\frac{\sqrt{\frac{3\alpha P}{2(M_1-1)}} + (2n-1)\sqrt{\frac{3\bar{\alpha} P}{2(M_2-1)}}}{\sqrt{\hat{\sigma}_2^2/2}} \Big) \right]. \quad (8)$$

$$\mathbb{P}_2 = \mathbb{P}(\hat{x}_2 = x_2 | \alpha \le 0.5) = 1 - \frac{2(M_1-1)}{M_1} Q\Big(\frac{d_2}{\sqrt{\hat{\sigma}_1^2/2}} \Big) + \frac{2}{M_1 M_2} \sum_{n=1}^{M_2-1} (M_2 - n) \Big\{ (M_1-1) \Big[Q\Big(\frac{2nd_1 - d_2}{\sqrt{\hat{\sigma}_1^2/2}} \Big)$$
$$- Q\Big(\frac{2nd_1 + d_2}{\sqrt{\hat{\sigma}_1^2/2}} \Big) \Big] + \Big[Q\Big(\frac{(2n-1)d_1 + (M_1-1)d_2}{\sqrt{\hat{\sigma}_1^2/2}} \Big) + Q\Big(\frac{(2n-1)d_1 - (M_1-1)d_2}{\sqrt{\hat{\sigma}_1^2/2}} \Big) \Big] \Big\}$$
$$= 1 - \frac{2(M_1-1)}{M_1} Q\Big(\frac{\sqrt{\frac{3\bar{\alpha} P}{2(M_2-1)}}}{\sqrt{\hat{\sigma}_1^2/2}} \Big) \frac{2}{M_1 M_2} \sum_{n=1}^{M_2-1} (M_2 - n) \Big\{ (M_1-1) \Big[Q\Big(\frac{2n\sqrt{\frac{3\alpha P}{2(M_1-1)}} - \sqrt{\frac{3\bar{\alpha} P}{2(M_2-1)}}}{\sqrt{\hat{\sigma}_1^2/2}} \Big)$$
$$- Q\Big(\frac{2n\sqrt{\frac{3\alpha P}{2(M_1-1)}} + \sqrt{\frac{3\bar{\alpha} P}{2(M_2-1)}}}{\sqrt{\hat{\sigma}_1^2/2}} \Big) \Big] + \Big[Q\Big(\frac{(2n-1)\sqrt{\frac{3\alpha P}{2(M_1-1)}} + (M_1-1)\sqrt{\frac{3\bar{\alpha} P}{2(M_2-1)}}}{\sqrt{\hat{\sigma}_1^2/2}} \Big)$$
$$+ Q\Big(\frac{(2n-1)\sqrt{\frac{3\alpha P}{2(M_1-1)}} - (M_1-1)\sqrt{\frac{3\bar{\alpha} P}{2(M_2-1)}}}{\sqrt{\hat{\sigma}_1^2/2}} \Big) \Big] \Big\}. \quad (9)$$

$$\frac{d\mathbb{P}_1}{d\alpha} = -\frac{2(M_2-1)}{M_1 M_2} \times \sum_{n=1}^{M_1/2} \Big[-\frac{1}{\sqrt{2\pi}} e^{-\frac{(d_1 - (2n-1)d_2)^2}{\hat{\sigma}_2^2}} \Big(\frac{1}{\sqrt{\hat{\sigma}_2^2/2}} \sqrt{\frac{3P}{8(M_1-1)\alpha}} + \frac{(2n-1)}{\sqrt{\hat{\sigma}_2^2/2}} \sqrt{\frac{3P}{8(M_2-1)\bar{\alpha}}} \Big)$$
$$- \frac{1}{\sqrt{2\pi}} e^{-\frac{(d_1 + (2n-1)d_2)^2}{\hat{\sigma}_2^2}} \Big(\frac{1}{\sqrt{\hat{\sigma}_2^2/2}} \sqrt{\frac{3P}{8(M_1-1)\alpha}} - \frac{(2n-1)}{\sqrt{\hat{\sigma}_2^2/2}} \sqrt{\frac{3P}{8(M_2-1)\bar{\alpha}}} \Big) \Big]. \quad (10)$$

decode their symbols. This is because the theory is based on Gaussian codewords whose lengths go to infinity. This assumption is not valid when finite constellations like QAM are used. Interestingly, with QAM inputs, for $\alpha \le \alpha_0$ if the order of SIC is exchanged, i.e., SIC is applied in the user with a weaker channel, both users' signals can be decoded.

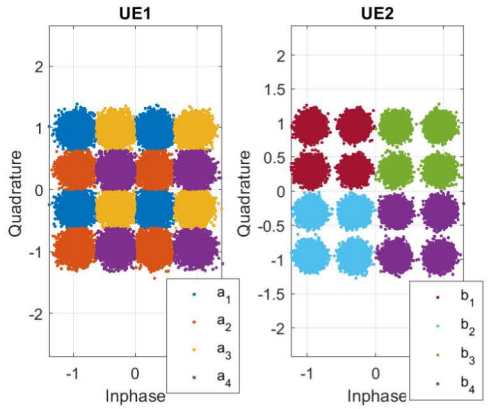
IV. Simulations

In the following simulations, we assume that the noise power is identical at the two receivers, $\sigma_1^2 = \sigma_2^2 = \frac{N_0}{2}$. The input signal has 10^5 samples.

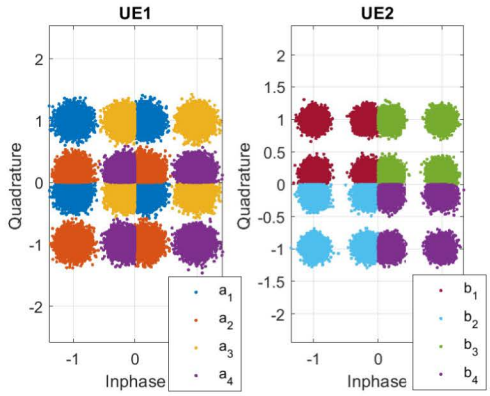
Example 1: We first consider the case with $M_1 = M_2 = 4$, where $h_1 = 2{+}3j$, $h_2 = 3{+}4j$, and $\frac{P}{N_0} = 5$dB. The estimated symbols of the ML method for both users can be found in Fig. 3 for $\alpha = 0.21, 0.33, 0.84$. Similar to the analysis in Section III-B, the received signal with noise is decided by the ML detection, and the colors represent the decoded symbols. Given different values of α, the QAM symbols of each user in the superimposed constellations are distinguishable using ML method. The SERs of UE1 and UE2 are plotted in Fig. 4 where theoretical derivation, ML, and SIC (UE1 employs TIN and UE2 performs SIC) are compared. It is seen that the analytical SERs of both users match the simulated SERs of the ML method. This validates the SER expressions of Propositions 1 and Remark 1. For $\alpha > 0.5$, ML and SIC curves are matched. For $\alpha \le 0.5$, TIN at UE1 and SIC at UE2 fail to decode. Specifically, the SERs of ML method are $[\epsilon_1, \epsilon_2] = [0.627, 0.005]\%$ for $\alpha = 0.21$, $[\epsilon_1, \epsilon_2] = [5.766, 1.456]\%$ for $\alpha = 0.33$, and $[\epsilon_1, \epsilon_2] = [0.052, 0.037]\%$ for $\alpha = 0.84$. The SERs of SIC for $\alpha = 0.21$ and 0.33 are $[75.012, 50.961]\%$ and $[72.063, 74.196]\%$, respectively.

The effect of the order of SIC is shown in Fig. 5. Assume the required maximum SER is 10^{-2}. First, UE1 performs TIN and UE2 employs SIC (this is the normal order of SIC). This works for $0.75 \le \alpha \le 0.92$ (marked by vertical black dash lines in Fig. 5). However, if we change the order of SIC, i.e., perform SIC at UE1 and TIN at UE2. Decoding can still work for $0.17 \le \alpha \le 0.24$, which implies that the weaker user can perform SIC. Finally, ML works for a higher range of α.

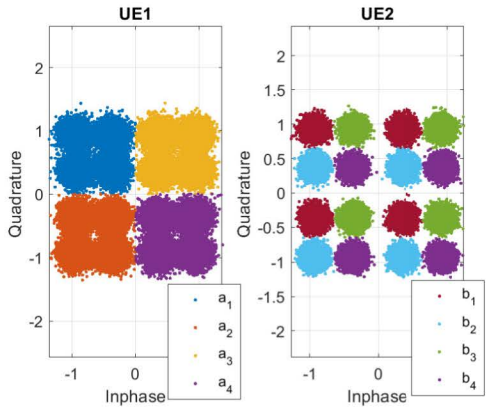
Example 2: We use the same settings as [9] in which $M_1 = 16$, $M_2 = 4$, $h_1 = 0.5{+}0.6j$, $h_2 = 0.7{+}0.8j$, and $\frac{P}{N_0} = 23$dB.



(a) $\alpha = 0.21$



(b) $\alpha = 0.33$



(c) $\alpha = 0.84$

Fig. 3: The super-constellation at UE1 and UE2 for three different values of α and $h_1 = 2 + 3j$, $h_2 = 3 + 4j$, and $\frac{P}{N_0} = 5$dB.

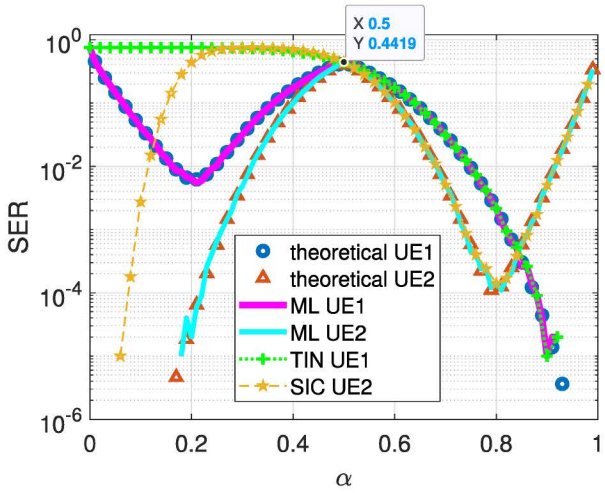


Fig. 4: SER comparison of the theoretical, ML, and SIC for $M_1 = M_2 = 4$.

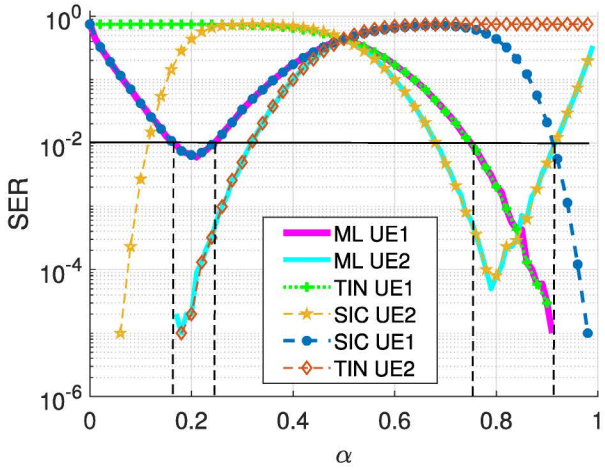


Fig. 5: SER comparison of the ML and SIC for $M_1 = M_2 = 4$.

The SERs of ML and SIC are plotted in Fig. 6. There are three cut-off thresholds which are $\alpha_0 = 0.83, 0.56, 0.36$. First, we obtain the same results of TIN at UE1 and SIC at UE2 as [9, Fig. 8] for $\alpha > 0.83$ (i.e., $d_1 > d_2$). The ML and SC-SIC can achieve the same performance in this range. On the other hand, exchanging the order of SIC, i.e., using SIC at UE1 and TIN at UE2, can achieve the same results as the ML method for $\alpha < 0.36$ (i.e., $d_1 < \frac{1}{3}d_2$). SIC fails for $0.36 \leq \alpha \leq 0.83$ (i.e., $\frac{1}{3}d_2 \leq d_1 \leq d_2$) while ML can work better to achieve local optimums. The SER curves of both users are quasi-convex and the minimum values of the curves are at $\alpha_{\min} = 0.95$ for UE2 when $\alpha \geq 0.5$, and $\alpha_{\min} = 0.21$ for UE1 when $\alpha < 0.5$ by an exhaustive search. The local-minimums are 0.44 and 0.69 for $\frac{1}{2}d_2 \leq d_1 \leq d_2$ and $\frac{1}{3}d_2 < d_1 \leq \frac{1}{2}d_2$, respectively.

Next, we demonstrate the influence of $\frac{P}{N_0}$ on SERs for two power allocation coefficients, $\alpha = 0.95$ and $\alpha = 0.21$, The results are shown in Fig. 7. The analytical SERs of both users match the simulation results, which validates Proposition 1 for $\alpha = 0.21$ and [9, Proposition 3, 4] for $\alpha = 0.95$. The results also verify that ML supports more values of α, only $\alpha = 0.95$ was considered as an option [9] and small value of power allocation such as $\alpha = 0.21$ was never qualified as a solution.

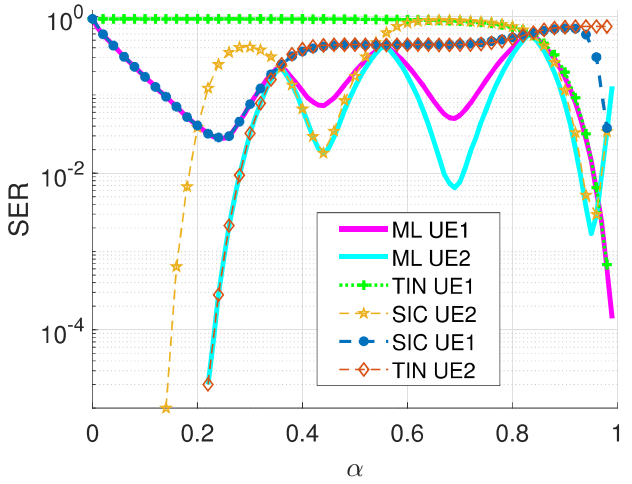


Fig. 6: SER comparison between ML and SIC for $M_1 = 16$ and $M_2 = 4$.

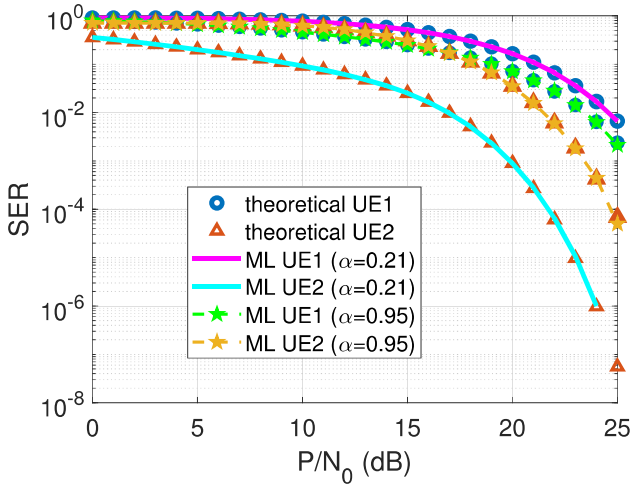


Fig. 7: Comparison of the theoretical and simulated SERs for $\alpha = 0.95$ and $\alpha = 0.21$.

V. Conclusions

In this study, we have revealed new insights into NOMA decoding with finite-alphabet inputs. We have demonstrates that employing maximum likelihood detection for SISO NOMA with finite-alphabet constellations can address certain limitations of successive interference cancellation decoding. We have also shown that decoding error is highly sensitive to the power allocation coefficient when successive interference cancellation is used for decoding; it also works well for a limited range of coefficients whereas the maximum likelihood decoding decoding is effective for more extensive of coefficients. Furthermore, contrary to what Shannon theory suggests where successive interference cancellation is only applicable to stronger users, in finite-alphabet NOMA, successive interference cancellation can be also applied to weaker user when the power allocation coefficient α is smaller than a certain threshold. However, such a rate is typically far from Shannon capacity.

References

[1] M. Vaezi, Z. Ding, and H. V. Poor, *Multiple access techniques for 5G wireless networks and beyond*. Cham, Switzerland: Springer, 2019.

[2] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access," in *Proc. International Symposium on Personal, Indoor, and Mobile Radio Communications*, 2013, pp. 611–615.

[3] S. 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 Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2016.

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

[5] G. Mazzini, "Power division multiple access," in *Proc. of IEEE International Conference on Universal Personal Communications*, vol. 1. IEEE, 1998, pp. 543–546.

[6] A. El Gamal and Y. H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[7] M. Vaezi and H. V. Poor, "NOMA: An information-theoretic perspective," in *Multiple Access Techniques for 5G Wireless Networks and Beyond*. Cham, Switzerland: Springer, 2019, pp. 167–193.

[8] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge, U.K.: Cambridge niversity press, 2005.

[9] Q. He, Y. Hu, and A. Schmeink, "Closed-form symbol error rate expressions for non-orthogonal multiple access systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 6775–6789, 2019.

[10] T. Assaf, A. J. Al-Dweik, M. S. El Moursi, H. Zeineldin, and M. Al-Jarrah, "Exact bit error-rate analysis of two-user NOMA using QAM with arbitrary modulation orders," *IEEE Communications Letters*, vol. 24, no. 12, pp. 2705–2709, 2020.

[11] Y. Qi, X. Zhang, and M. Vaezi, "Over-the-air implementation of NOMA: New experiments and future directions," *IEEE Access*, vol. 9, pp. 135 828–135 844, 2021.

[12] A. Salari, M. Shirvanimoghaddam, M. B. Shahab, Y. Li, and S. Johnson, "NOMA joint channel estimation and signal detection using rotational invariant codes and GMM-based clustering," *IEEE Communications Letters*, vol. 26, no. 10, pp. 2485–2489, 2022.

[13] C. Huppert and M. Bossert, "On achievable rates in the two user AWGN broadcast channel with finite input alphabets," in *Proc. IEEE International Symposium on Information Theory*, 2007, pp. 2581–2585.

[14] N. Deshpande and B. S. Rajan, "Constellation constrained capacity of two-user broadcast channels," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, 2009, pp. 1–6.

[15] M. Vaezi, R. Schober, Z. Ding, and H. V. Poor, "Non-orthogonal multiple access: Common myths and critical questions," *IEEE Wireless Communications*, vol. 26, no. 5, pp. 174–180, 2019.

[16] M. Qiu, Y.-C. Huang, and J. Yuan, "Downlink non-orthogonal multiple access without SIC for block fading channels: An algebraic rotation approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3903–3918, 2019.

[17] A. Almohamad, M. O. Hasna, S. Althunibat, and K. Qaraqe, "A novel downlink IM-NOMA scheme," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 235–244, 2021.

[18] K. Chung, "Correlated superposition coding: Lossless two-user NOMA implementation without SIC under user-fairness," *IEEE Wireless Communications Letters*, vol. 10, no. 9, pp. 1999–2003, 2021.

[19] F. Kara and H. Kaya, "BER performances of downlink and uplink NOMA in the presence of SIC errors over fading channels," *IET Communications*, vol. 12, no. 15, pp. 1834–1844, 2018.

[20] H. Jafarkhani, H. Maleki, and M. Vaezi, "Next generation multiple access: Modulation and coding," *Proceedings of the IEEE*, 2024.