Resource Allocation in Multicarrier NOMA Systems Based on Optimal Channel Gain Ratios

Estela Carmona Cejudo, *Graduate Student Member, IEEE*, Huiling Zhu, *Member, IEEE*, and Jiangzhou Wang, *Fellow, IEEE*

***Abstract*—The application of non-orthogonal multiple access (NOMA) to multicarrier systems can improve the spectrum efficiency and enable massive connectivity in future mobile systems. Resource allocation in multicarrier NOMA systems is a non-deterministic polynomial time-hard problem requiring exhaustive search, which has prohibitive computational complexity. Instead, efficient algorithms that provide a good trade-off between system performance and implementation practicality are needed. In this paper, exact values of the optimal channel gain ratios between a pair of NOMA users are presented for the first time for quadrature amplitude modulation (QAM) schemes. Further, numerical limits are derived for the values of channel gain ratios that fulfill the system constraints. These findings are used to propose a user pairing algorithm with quasi-linear complexity. Further, a novel scheme for data rate and continuous power allocation is proposed. Through numerical simulations, it is proved that the proposed scheme yields an achievable sum-rate close to the performance of exhaustive search, and it outperforms other suboptimal resource allocation schemes.**

***Index Terms*—Non-orthogonal multiple access (NOMA), channel gain ratios, resource allocation, user pairing, multicarrier, bit error rate (BER).**

I. Introduction

OVER the last few years, the evolution of wireless networks has been driven by the need for the fifth generation (5G) of cellular systems to provide a 1000-fold increase in capacity with respect to fourth generation (4G) long-term evolution (LTE). It is envisioned that sixth generation (6G) systems will have to provide ten times the connectivity density of 5G, reaching up to 10^7 devices/km^2, with an area traffic capacity of up to 1 Gb/s/m^2, along with up to a 100-fold increase in network energy efficiency and a 10-fold increase in spectrum efficiency [1].

Non-orthogonal multiple access (NOMA) can boost the spectrum efficiency and support massive connectivity in future mobile systems. NOMA can be realized in different domains, such as power or code. Power domain NOMA has been maturely studied and has become the most promising NOMA technique [2]. In power domain NOMA, referred to as NOMA hereafter, multiple access is achieved in a non-orthogonal manner by using multi-layer modulation at the transmitter and successive interference cancellation (SIC) at the receiver [3]–[5]. Radio resources can be allocated to multiple users in the same frequency and time slot through power-domain multiplexing. Hence, NOMA can provide service to a greater number of users than orthogonal multiple access (OMA) technologies.

High data rate wireless communication systems have extensively adopted multicarrier multiple access techniques. Specifically, orthogonal frequency division multiple access (OFDMA) [6] provides multiple users' transmissions through the dynamic allocation of broadband radio resources. Based on orthogonal frequency division multiplexing (OFDM), OFDMA inherits the benefit of converting a frequency selective fading channel into multiple flat fading subchannels.

By applying NOMA to subcarrier-based schemes, its capabilities can be further extended. The key to achieve the full potential of NOMA is resource allocation, which optimizes the assignment of radio resources to users. The joint optimization of subcarrier assignment and power allocation in NOMA systems leads to a mixed-integer problem, which was proved to be nondeterministic polynomial time (NP)-hard in [7]. Hence, the optimal solution can only be found through exhaustive search, which is a combinatorial optimization problem. One approach to select the best user set is to search over all possible combinations of users and select the one that maximizes the sum-rate [3], which yields prohibitive computational complexity. A more practical approach is to separate the problems of subcarrier and power allocation, fix one of them and optimize the other [7]–[9]. This leads to suboptimal but practical and efficient solutions.

In terms of the subcarrier allocation problem, existing literature focused on user grouping, i.e. roughly classifying users as strong or weak, depending on their channel gains, and multiplexing them according to this classification. In [10] it was proved that, in order to maximize individual data rates in NOMA systems with fixed power allocation, it is preferable to pair two users whose channel gains are significantly distinctive. These findings have since been applied to the development of user-pairing and user-grouping schemes, which offer a more efficient solution to the subcarrier allocation problem. For example, in [11], a low-complexity algorithm for user pairing was proposed, in which a priority coefficient was derived in terms of data rate constraints, and the user with the largest priority coefficient was paired with the user with the largest channel gain. In [12], a dynamic user clustering and power allocation scheme was proposed. Users were clustered and then paired within each cluster depending on their channel gains. The procedure in [13] was to randomly allocate users in groups, then pair the users with largest and smallest channel gains within each group. Further,

E. Carmona Cejudo is with i2CAT Foundation, Barcelona, 08034, Spain (email: estela.carmona@i2cat.net).

H. Zhu and J. Wang are with the School of Engineering and Digital Arts, University of Kent, Canterbury, CT2 7NT, United Kingdom (e-mail: {h.zhu, j.z.wang}@kent.ac.uk).

a *fair* NOMA power allocation scheme was proposed in [14], where users were opportunistically paired irrespective of their channel gains. However, these studies did not reveal how a certain ratio of users' channel gains affects parameters such as the system sum-rate, bit error rate (BER) and outage probability. In addition, although the aim of all these works was to maximize the system throughput or single-user data rates, only one carrier was considered in their proposed schemes, which did not take full advantage of the capabilities of multicarrier NOMA.

Resource allocation in multicarrier NOMA systems has been studied for different performance metrics, with the problem of sum-rate maximization the most commonly addressed in the literature. Due to the complexity of this problem, most researches divided it into two subproblems, namely subcarrier and power allocation, and only proposed methods based on heuristic user grouping for pairing users sharing the same subcarrier. [15] provided a comprehensive investigation of resource allocation in downlink NOMA systems, under different performance criteria. In [16], the authors formulated a resource allocation problem for downlink OFDM-based NOMA, with the objective of maximizing the system sum-rate. For subcarrier allocation, users with similar channel gains were allocated into a group. NOMA was applied to users from different groups in a greedy manner, by imposing a constraint to ensure a large enough difference in mean channel gain ratio between any two groups. [17] investigated resource allocation for the maximization of the weighted sum throughput of full-duplex multicarrier NOMA, and a low-complexity suboptimal algorithm based on successive convex approximation was proposed, with a performance close to optimal. A near-optimal solution to the sum-rate maximization problem was proposed in [7], through the discretization of the user power budget. However, in practical systems with a large number of power levels, the achievable computational complexity remains too high. In [18], optimal and approximate algorithms were proposed for joint subcarrier and power allocation, which relied on using a precomputation procedure in order to reduce the complexity of the power control algorithm. Therefore, this solution is not viable for a practical system where the channel gains of all users are time variant. In addition, these works assumed continuous data rate allocation, based on theoretical data rate expressions, and did not consider practical discrete modulation levels.

In summary, existing researches did not reveal the exact effect that a given ratio of user channel gains in NOMA has on system parameters such as the sum-rate and BER. Most existing user pairing schemes in the literature considered broadly classifying users according to their channel gains and randomly selecting users from a group according to the system optimization metric. Moreover, novel resource allocation schemes are needed in NOMA which are capable of providing a performance close to optimal in a practical and simplified manner. In this work, a resource allocation problem in multicarrier NOMA is considered, with the objective of sum-rate maximization, under maximum BER and transmit power constraints. Exact values of the optimal channel gain ratios between a pair of NOMA users are presented for the first time for quadrature amplitude modulation (QAM) schemes. Further, numerical limits are derived for the values of channel gain ratios that fulfill the system constraints. These findings are used to develop a user pairing scheme with quasi-linear complexity that finds the pair of NOMA users that maximize the sum-rate. Further, a strategy is proposed for power and data rate allocation. Unlike existing works, practical modulation schemes are considered and continuous power allocation is assumed in order to achieve a performance close to optimal with reduced computational complexity. The contributions in this paper are summarized as follows:

1) Theoretical BER expressions are presented for the BER in NOMA, assuming multi-layer, multi-level quadrature amplitude modulation (QAM). The theoretical BER expressions are used for calculating the optimal channel gain ratios for a pair of NOMA users, i.e. the ratios of channel gains that maximize the achievable rate for a given BER constraint. Further, boundaries are derived for the values of channel gain ratios that fulfill BER constraints.
2) The findings about optimal channel gain ratios are applied to propose a user pairing algorithm that achieves quasi-linear complexity.
3) Accurate BER approximations are presented in the form of exponential functions. These are used for proposing a scheme for selecting the power allocation and transmission rate per subcarrier, in terms of the BER, transmit power constraints, and channel gain ratio. Unlike existing works, continuous power levels and discrete modulation schemes are considered.
4) Through numerical simulations, it is proved that the proposed scheme yields a performance close to optimal, and it outperforms suboptimal schemes such as FTPC [3]. In addition, it is shown that the benefit of pairing NOMA users with very distinct channel gains is lost under practical QAM schemes, due to the inability of users with poor channel conditions to fulfill individual BER constraints.

The remainder of this paper is organized as follows. In Section II, the system model is described. In Section III, theoretical BER expressions are derived for multi-layer, multi-level QAM in NOMA, and these are used to derive the optimal channel gain ratios between two NOMA users. In Section IV, the problem of resource allocation in multicarrier NOMA systems is formulated. Accurate BER approximations are presented in the form of exponential functions. Based on them, algorithms are proposed for user pairing and power and rate allocation in terms of channel gain ratios. Numerical results are given in Section V. Section VI concludes the paper.

Notation: $\mathbb{C}$, $\mathbb{R}$ and $\mathbb{I}^+$ denote the sets of complex, real and positive integer values, respectively. $|\cdot|$ denotes the absolute value of a complex scalar. $\mathbb{E}\{\cdot\}$ denotes statistical expectation. The circularly symmetric complex Gaussian distribution with mean a and variance σ^2 is denoted by $\mathcal{CN}(a, \sigma^2)$. Boldface symbols denote a finite set of elements.

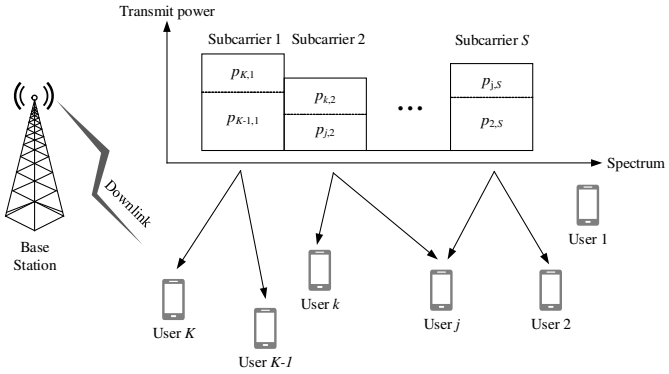


Fig. 1: Multicarrier NOMA system model.

II. System Model

Consider a single-cell, multicarrier downlink NOMA system as depicted in Fig. 1, where it is assumed that the downlink channel gains of all subcarriers are estimated by each user through pilot channels and channel state information. The scenario under consideration is assumed to have a set of active users $\mathcal{K} = \{1, \cdots, K\}$. The entire bandwidth of W Hertz is partitioned into S orthogonal subcarriers contained in the set $\mathcal{S} = \{1, \cdots, S\}$. It is assumed that there is no interference among subcarriers due to the orthogonal frequency partitioning. For all users, all subcarriers are assumed to be Rayleigh fading channels with additive white Gaussian noise (AWGN), with double-sided spectral density $N_0/2$. The channel response of user j on subcarrier s is given by $h_{j,s}$, where $\mathbb{E}\{|h_{j,s}|^2\} = 1$, and $|h_{j,s}|$ is Rayleigh distributed, $\forall j, s$. Further, $h_{j,s}$ and $h_{k,s}$ are independent for $j \neq k$. Is is assumed that one user may be simultaneously assigned to several subcarriers, and that a particular subcarrier might be unused during a certain time resource. Without loss of generality, it is assumed that the users' channels on subcarrier s have been ordered as $|h_{1,s}|^2 \leq \cdots \leq |h_{K,s}|^2$. Assume that users $\{j, k\} \in \mathcal{K}$ are jointly selected to perform NOMA on subcarrier $s \in \mathcal{S}$. The base station transmits a signal of the form

$$x_{\{j,k\},s} = \sqrt{p_{j,s}} x_{j,s} + \sqrt{p_{k,s}} x_{k,s}, \tag{1}$$

where $x_{j,s} \in \mathbb{C}$ denotes user j's transmit symbol, and $p_{j,s}$ is the transmit power to user j on subcarrier s. Further, $\mathbb{E}\{|x_{j,s}|^2\} = \mathbb{E}\{|x_{k,s}|^2\} = 1$ and $p_{j,s} = \alpha_{j,s} p_s$, where p_s is the total transmit power on subcarrier s and $\alpha_{j,s}$ is the power allocation factor of user j on that subcarrier. For any two multiplexed users j and k, the conditions $p_{j,s} + p_{k,s} = p_s$ and $\alpha_{j,s} + \alpha_{k,s} = 1$ hold for subcarrier s. Hence, the received signals at users j and k on subcarrier s can be expressed as

$$y_{j,s} = \sqrt{p_{j,s}} h_{j,s} x_{j,s} + \sqrt{p_{k,s}} h_{j,s} x_{k,s} + z_{j,s}, \tag{2}$$

$$y_{k,s} = \sqrt{p_{k,s}} h_{k,s} x_{k,s} + \sqrt{p_{j,s}} h_{k,s} x_{j,s} + z_{k,s}, \tag{3}$$

respectively, where $z_{j,s} \sim \mathcal{CN}(0, \sigma^2_{z_{j,s}})$, $z_{k,s} \sim \mathcal{CN}(0, \sigma^2_{z_{k,s}})$ denote the AWGN.

User j comes first in the SIC decoding order since $|h_{j,s}|^2 \leq |h_{k,s}|^2$, and it can decode its own signal directly, disregarding user k's signal as noise [10]. Hence, on subcarrier s, it can theoretically achieve a maximum data rate of

$$R_{j,s} = \log_2 \left(1 + \frac{\alpha_{j,s} |h_{j,s}|^2 p_s}{\alpha_{k,s} |h_{j,s}|^2 p_s + N_0}\right). \tag{4}$$

User k decodes user j's signal first, with an achievable data rate of

$$R_{j \to k,s} = \log_2 \left(1 + \frac{\alpha_{j,s} |h_{k,s}|^2 p_s}{\alpha_{k,s} |h_{k,s}|^2 p_s + N_0}\right). \tag{5}$$

After decoding user j's signal, user k can subtract it from the receive signal and then decode its own symbols. The theoretically maximum achievable data rate when decoding its own signal, under perfect SIC, is given by

$$R_{k,s} = \log_2 \left(1 + \alpha_{k,s} |h_{k,s}|^2 p_s\right). \tag{6}$$

Resource allocation is performed at the BS under the constraints of total downlink transmit power, p_{total}, and identical maximum BER per user, β_0. During each resource allocation instant, the BS multiplexes at most two users[1] into each subcarrier according to their channel gain ratios, such that the NOMA principle can be successfully applied. In the event that the application of NOMA is not feasible, a subcarrier may be unused. The BS feeds forward information about subcarrier and modulation level assignment to each active user through downlink control signaling. In addition, where two users are multiplexed according to the NOMA principle, information about user j's assigned power and modulation level is also fed forward to user k, thus enabling SIC decoding at the receiver.

III. BER Analysis in NOMA

Expressions (4) and (6) are the theoretical limits to the achievable data rates in NOMA under continuous modulation levels. Below, a BER analysis is carried out in order to theoretically determine the achievable data rates in NOMA under practical discrete QAM levels. This analysis also yields the optimal channel gain ratio in NOMA, i.e. the ratio of channel gains of the pair of NOMA users that maximize the sum-rate for a given BER constraint. These findings are used in Section IV to solve the problem of user pairing in NOMA, and to propose a novel resource allocation strategy for multicarrier NOMA systems.

Consider the situation where two NOMA users j and k are multiplexed on subcarrier s. For simplicity of notation, the subscript s corresponding to subcarrier s is omitted throughout this section. The BS transmits a superposed signal $x = x_{\{j,k\}}$ as given in (1). Let x_j be the symbol intended for user j, from a M_j-QAM signal constellation $\mathcal{X}_j$, and let x_k be user k's symbol, from the M_k-QAM constellation $\mathcal{X}_k$. The *supersymbol* $x = x_{\{j,k\}}$ is the result of superposing symbols x_j and x_k, where x_j belongs to the lower layer of the superposed constellation, and x_k to the upper layer. Further, supersymbol x belongs to a *superconstellation* (i.e. a

[1]SIC is applied at the user equipment in order to cancel inter-user interference. However, the high implementation complexity of SIC remains an open research challenge [19], [20]. Therefore, the superposition of a maximum of two NOMA users is considered in order to reduce SIC implementation complexity.

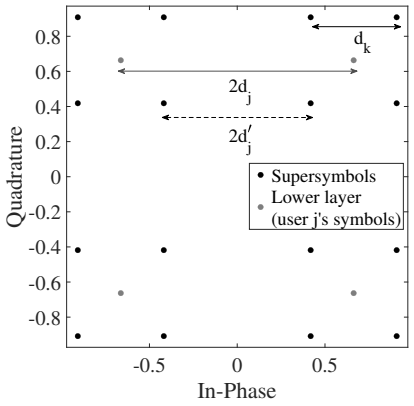


Fig. 2: 4-QAM + 4-QAM superconstellation.

superposed constellation) $\mathcal{X}$ of size M_jM_k. The concept of superconstellations in NOMA is similar to that of hierarchical constellations in digital broadcasting systems [21], [22], where information is grouped into data streams with different relative importance. The location of the superconstellation points in NOMA is determined by the power allocation factors α_j and α_k. The average transmit symbol energy of the resulting superconstellation is normalized to $E_{sym} = 1$. See Fig. 2, which represents two superposed 4-QAM constellations.

To determine the symbol error rate (SER) at each user, the layout and number of superconstellation points must be considered, as well as the minimum Euclidean distance among them. The SER at user k is conditioned on the incoming interference from user j's signal, which can either reduce or increase user k's minimum Euclidean distance. User k uses an SIC receiver, so an error in decoding user j's symbol means that the decoding of user k's own signal is also unsuccessful, due to SIC error propagation. Hence, the SER at user k, $P_k(e)$, is affected by the SER when decoding user j's transmit symbols, i.e.

$$P_k(e) = P_k(e|correct_{x_j})P_k(correct_{x_j}) + P_k(e|error_{x_j})P_k(error_{x_j}), \quad (7)$$

where $P_k(correct_{x_j})$ is user j's symbols correct detection rate at user k, $P_k(error_{x_j})$ is the error rate whilst detecting user j's symbols at user k, $P_k(e|correct_{x_j})$ is the error rate at user k under the condition that no error occurred while detecting user j's symbols, and $P_k(e|error_{x_j})$ is the error rate under the condition that an error occurred whilst detecting user j's symbols. Therefore, the term $P_k(e|error_{x_j})P_k(error_{x_j})$ in (7) models the SIC error propagation at user k.

Below, closed-form expressions are derived for $P_j(e)$ and $P_k(e)$. It is assumed that the lower level constellation is assigned to user j, and that the average transmit power is always normalized to one for the NOMA superconstellation, i.e. $p = p_j + p_k = 1$. $E_{sym,j} = 1$, $E_{sym,k} = 1$ denote users j's and k's normalized symbol energy before NOMA power allocation and superposition. $E_{bit,j}$ and $E_{bit,k}$ refer to the average bit energy of users j and k, respectively. Further, $E_{bit,j} = E_{sym,j}/\log_2 M_j = 1/\log_2 M_j$ and $E_{bit,k} = E_{sym,k}/\log_2 M_k = 1/\log_2 M_k$. It is assumed that symbol errors in decoding are caused by incoming interference from the nearest neighbor, and that Gray coding is independently applied to each user's layer.

A. BER Analysis

Assume that user j is assigned with square M_j-QAM, and user k with square M_k-QAM. The resulting superconstellation is made up of M_jM_k symbols, and it is symmetrical in both dimensions. Hence, the in-phase and quadrature components yield the same error rate. Let d_j and d_k represent the Euclidean distances of the layers belonging to user j's and user k's symbols, respectively, after scaling by α_j and α_k. Further, let d'_j represent the smallest Euclidean distance between symbols in the lower layer of the superconstellation, as represented in Fig. 2.

User j decodes symbols from the lower layer directly, disregarding interference from superposed upper-layer symbols as noise. During SIC, user k decodes symbols in the lower layer first, removes them from the data stream, and then decodes its own symbols in the upper layer.

Lemma 1. *The SERs at user j and user k are respectively given by*

$$P_j(e) = 1 - \left[1 - \frac{2\sqrt{M_j}-2}{\sqrt{M_jM_k}} \sum_{i=0}^{\sqrt{M_k}-1} Q\left(2\frac{d'_j + 2id_k}{\sqrt{2N_0}}\right)\right]^2, \quad (8)$$

$$P_k(e) \lessapprox 1 - \left[1 - 2\left(1 - \frac{1}{\sqrt{M_k}}\right) Q\left(\sqrt{\frac{2}{N_0}}d_k\right) \cdot \left(1 - \frac{2\sqrt{M_j}-2}{\sqrt{M_jM_k}} \sum_{i=0}^{\sqrt{M_k}-1} Q\left(2\frac{d'_j + 2id_k}{\sqrt{2N_0}}\right)\right) - \frac{2\sqrt{M_j}-2}{\sqrt{M_jM_k}} \sum_{i=0}^{\sqrt{M_k}-1} Q\left(2\frac{d'_j + 2id_k}{\sqrt{2N_0}}\right)\right]^2, \quad (9)$$

where $Q(x)$ is known as the Q-function, and it is defined as $Q(x) = 1/\sqrt{2\pi}\int_x^\infty \exp(-u^2/2)\,\mathrm{d}u$.

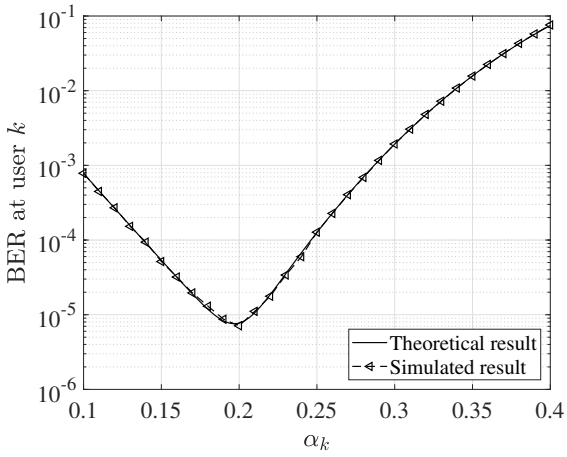
Proof. See Appendix A. □

For small error rates, the BER at user j and user k can be calculated from the SER as $\beta_j \approx P_j(e)/\log_2 M_j$ and $\beta_k \approx P_k(e)/\log_2 M_k$, respectively [23, Chapter 6.1].

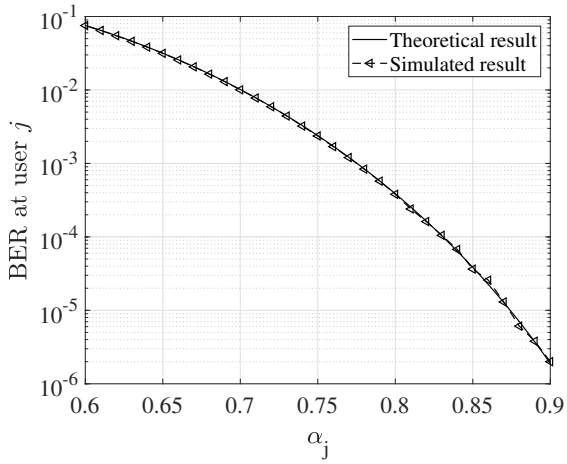
In order to ensure a manageable level of inter-user interference at user j and successful SIC decoding at user k, the distance among symbols in the lower layer must be equal or greater than the symbols in the upper layer, i.e. $d'_j \geq d_k$. From (A1)–(A3), this yields

$$\alpha_k \leq \frac{M_k - 1}{M_jM_k - 1}. \quad (10)$$

Equivalently, the value of α_k for which the BER constraint can be met with minimum transmit power always fulfills (10), as otherwise the interference in the lower layer increases, resulting in a larger SIC error rate at user k due to larger inter-user interference. Under the assumption that the lower modulation level is always assigned to user j in the lower



(a) BER at user k, SNR=20dB.



(b) BER at user j, SNR=17dB.

Fig. 3: BER versus power allocation factor for 4-QAM + 4-QAM.

superconstellation layer, (10) guarantees that the effect of inter-user interference can be overcome.

Fig. 3 represents the BER at users j and k versus α_j and α_k for fixed received signal-to-noise ratio (SNR) at each user. It can be seen that a large α_j, or a small α_k, is always preferable at user j. However, at user k, the lowest BER is achieved for the boundary value of α_k in (10). For a larger value of α_k, the BER worsens due to increased interference from user j's symbols, which negatively impacts the SIC error rate.

B. Optimal Channel Gain Ratio

Having prior knowledge about the optimal ratio between the channel gains of the two NOMA users is useful for simplifying the problem of user pairing. Therefore, a channel gain ratio $g_{j,k}$ for user j and user k is defined as follows,

$$g_{\{j,k\}} = \frac{|h_k|^2}{|h_j|^2}. \tag{11}$$

Let $\mathcal{G}(M_j, M_k, \beta_0)$ denote the optimal channel gain ratio for two NOMA users, assuming modulation levels M_j and M_k, and a BER constraint β_0. A pair of users $\{j, k\}$ is an optimal pair if $g_{j,k} = \mathcal{G}(M_j, M_k, \beta_0)$. Numerical results in

Optimal channel gain ratio $\mathcal{G}(M_j, M_k, \beta_0)$ (dB)				
Modulation \ β_0	10^{-3}	10^{-4}	10^{-5}	10^{-6}
4-QAM + 4-QAM	1.548	1.342	1.264	1.206
16-QAM + 4-QAM	1.760	1.456	1.324	1.253
64-QAM + 4-QAM	2.225	1.686	1.467	1.343
256-QAM + 4-QAM	3.167	2.060	1.685	1.500
16-QAM + 16-QAM	2.113	1.640	1.443	1.342
64-QAM + 16-QAM	2.732	1.886	1.604	1.342

TABLE I: Optimal channel gain ratio $\mathcal{G}(M_j, M_k, \beta_0)$ (dB) for M_k-QAM + M_j-QAM.

Table I show optimal channel gain ratio $\mathcal{G}(M_j, M_k, \beta_0)$ values calculated from the numerical evaluation of the analytical expressions (8) and (9), assuming that the BER constraint and expression (10) are marginally met, i.e. $\beta_j = \beta_k = \beta_0$, and $\alpha_k = (M_k - 1)/(M_j M_k - 1)$. These results are used for developing a user pairing strategy in Section IV.

The results from Table I can be interpreted as follows. Take, for example, a BER constraint $\beta_0 = 10^{-3}$ and a 4-QAM + 4-QAM constellation, where $\mathcal{G}(4, 4, 10^{-3}) = 1.548$ dB. Any pair of NOMA users j and k that fulfill $g_{j,k} = 1.548$ dB can simultaneously meet β_0, provided that the transmit power is large enough. Moreover, if user k fulfills β_0 for a given transmit SNR and modulation level, any user j with a channel gain $|h_j^s|^2 \geq |h_k^s|^2 / g_{\{j,k\}}$ is guaranteed to fulfill β_0 too. In general, for any given BER constraint β_0, a larger difference between M_j and M_k requires a larger channel gain ratio between users j and k.

In [10], it was proved that the achievable data rate in NOMA is enlarged by pairing users with more distinctive channel gains, which is consistent with the results in Table I. However, Table I also shows that there exists a limit on how distinctive the channel gains of two NOMA users can be in order for users with a poor channel condition to fulfill their individual BER constraint for a given modulation level.

C. Extension to N-User NOMA Scenarios

This work may be extended to N-user NOMA scenarios by defining N individual power allocation factors, such that $\alpha_1 + \ldots + \alpha_{N-1} + \alpha_N = 1$. A larger number of users N translates into the superposition of N data streams and, therefore, an increased level of inter-user interference. The BER performance at each user will be dependent on the power allocation factor and modulation level assigned to itself and each one of the other users.

In N-user NOMA, the lowest-layer user decodes its signal directly, without applying SIC, and its power allocation factor must be as large as possible in order to reduce the inter-user interference and achieve a good BER performance. A larger number of users N introduce additional superconstellation points, translating into an increased level of inter-user interference among users. Therefore, as N increases, a larger channel gain is needed at the lowest-layer user in order to meet its BER constraint.

Since the highest-layer user decodes and removes $N - 1$ interfering data streams before decoding its own data, it is critical to achieve a good BER performance at each decoding

stage in order to minimize SIC error propagation. The optimal BER performance at the highest-layer user is achieved when expression (10) is marginally met, which occurs when all superconstellation points are equidistant. In N-user NOMA, expression (10) will become a set of $N-1$ power allocation factor conditions for the $N-1$ upper constellation layers. In general, as N increases, lower transmit power is available at every user, and therefore larger channel gains are required at all users to meet their BER constraint.

Further, in N-user NOMA, there will exist $N-1$ optimal channel gain ratios: between users N and $N-1$, between users $N-1$ and $N-2$, and so on. These can be calculated numerically, as in the 2-user case, by assuming that the BER constraints and the set of $N-1$ power allocation factor conditions are marginally met. According to the numerical results in Table I, the optimal channel gain ratio between two users increases when their modulation levels are more distinctive, and when the Euclidean distances among superconstellation symbols are smaller (i.e. for larger superposed modulation levels). Thus, as N increases, the channel gain ratios required between every pair of users will increase with respect to those given in Table I for the 2-user case.

IV. Resource Allocation Based on Channel Gain Ratios

The problem formulation in this paper is the maximization of the system sum-rate under individual BER and total transmit power constraints, as given by

$$\max_{R_{j,s},R_{k,s},\delta_{j,s},\delta_{k,s}} \quad \sum_{s=1}^{S}\sum_{\substack{j=1\\ j\neq k}}^{K}\sum_{k=1}^{K}\delta_{j,s}R_{j,s}+\delta_{k,s}R_{k,s} \tag{12a}$$

$$\text{s.t.} \quad \text{C1: } 0\leq p_{j,s},\ \forall j,s, \tag{12b}$$

$$\text{C2: } \delta_{j,s}\in\{0,1\},\ \forall j,s, \tag{12c}$$

$$\text{C3: } \sum_{j=1}^{K}\delta_{j,s}\leq 2,\ \forall s, \tag{12d}$$

$$\text{C4: } \sum_{s=1}^{S}(p_{j,s}+p_{k,s})\leq p_{total}, \tag{12e}$$

$$\text{C5: } \beta_{j,s}\leq\beta_0,\forall j,s, \tag{12f}$$

where $\delta_{j,s}$ is a binary variable defined as $\delta_{j,s}=1$ if user j is assigned to subcarrier s, or $\delta_{j,s}=0$ otherwise. Constraint C1 ensures non-negative transmit power. Constraints C2 and C3 guarantee that a maximum of two users may be allocated to one subcarrier. Constraint C4 is the maximum transmit power allowance at the BS, p_{total}. Further, constraint C5 guarantees that the achievable BER at all users on subcarrier s, $\beta_{j,s}$, is lower than the BER threshold, β_0.

The problem in (12) is NP-hard, thus its optimal solution can only be found through exhaustive search, which is not practical due to its large computational complexity [3], [7]. In order to make the maximization problem (12) more tractable, the proposed solution is to divide it into the subproblems of subcarrier and power allocation.

A channel gain gap expression is derived first in terms of the channel gain ratio and the achievable sum-rate for a pair of NOMA users. This definition is useful for solving the problem of user pairing in a simplified manner. Further, by using the channel gain gap expression, a procedure for power and rate allocation is proposed.

A. Channel Gain Gap

As the levels of M_j and M_k increase, the error probability expressions (8) and (9) become more complicated, with additional Q-function terms. These expressions are neither easily invertible nor easily differentiable in their arguments, and therefore cannot be used for adaptive rate and modulation design. Therefore, BER approximations are introduced with only one Q-function term. These approximations are used for finding the analytical relationship between the channel gain ratio of a pair of NOMA users and their achievable data rates. This knowledge is later applied to the subcarrier, power and rate allocation scheme.

Assume that the power allocation factor condition given by (10) is marginally met for the user pair $\{j,k\}$ on subcarrier s, i.e.

$$\alpha_{k,s}=\frac{M_{k,s}-1}{M_{j,s}M_{k,s}-1}. \tag{13}$$

In this case, according to (8), the term $Q(\sqrt{2/N_0}d'_{j,s})$ determines the value of the error probability $P_{j,s}(e)$, and therefore $P_{j,s}(e)$ is approximately proportional to $Q(\sqrt{2/N_0}d'_{j,s})$, i.e.

$$P_{j,s}(e)\underset{\sim}{\propto}Q(\sqrt{2/N_0}d'_{j,s}). \tag{14}$$

Further, the argument in all Q-function terms in (9) is $\sqrt{2N_0}d_{k,s}$. Therefore,

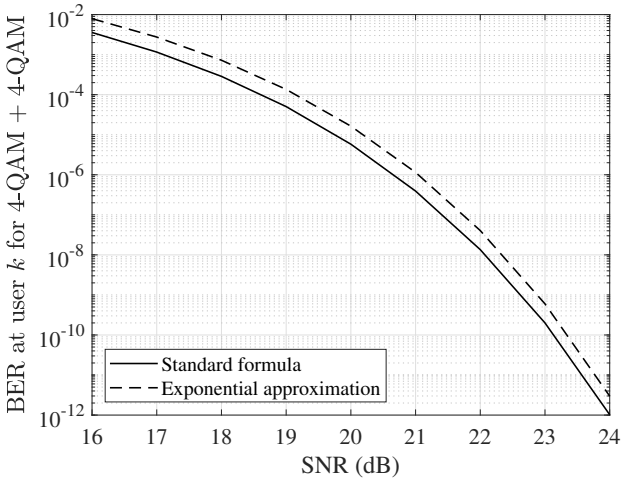
$$P_{k,s}(e)\underset{\sim}{\propto}Q(\sqrt{2/N_0}d_{k,s}). \tag{15}$$

Following eq. (9) in [24], the BERs at user j and user k on subcarrier s under Rayleigh fading can be approximated from (8) and (9) as

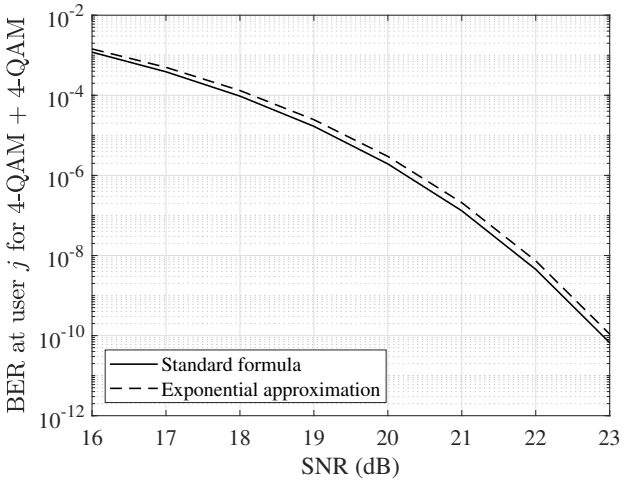
$$\beta_{j,s}\approx 0.14\frac{1.6+\sqrt{0.18\log_2 M_{j,s}\log_2 M_{k,s}}}{\sqrt{M_{j,s}M_{k,s}}}\cdot\exp\left(\frac{-1.54|h_{j,s}|^2T_{sym}p_s}{N_0(M_{j,s}M_{k,s}-1)}\right), \tag{16}$$

$$\beta_{k,s}\approx 0.77\frac{1.6+\sqrt{0.18\log_2 M_{j,s}\log_2 M_{k,s}}}{\sqrt{M_{j,s}M_{k,s}}}\cdot\exp\left(\frac{-1.54|h_{k,s}|^2T_{sym}p_s}{N_0(M_{j,s}M_{k,s}-1)}\right), \tag{17}$$

respectively, where T_{sym} is the symbol duration. The accuracy of these expressions is shown in Figs. 4–7 for some modulation levels. At user j, in the worst case, for 256-QAM + 4-QAM, (16) is accurate to within 0.51 dB with respect to (8). The approximation error introduced by (16) decreases for increasing $M_{j,s}$ and decreasing $M_{j,s}M_{k,s}$. At user k, the exponential approximation (17) is more accurate with respect to (9) for increasing values of $M_{j,s}$ and $M_{k,s}$. In the worst

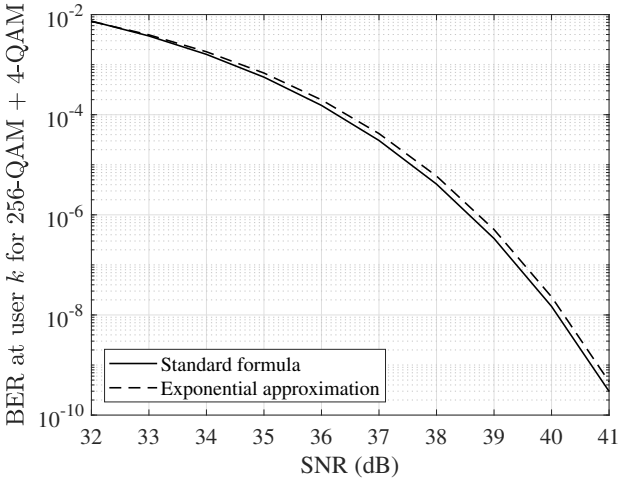


(a) BER at user k

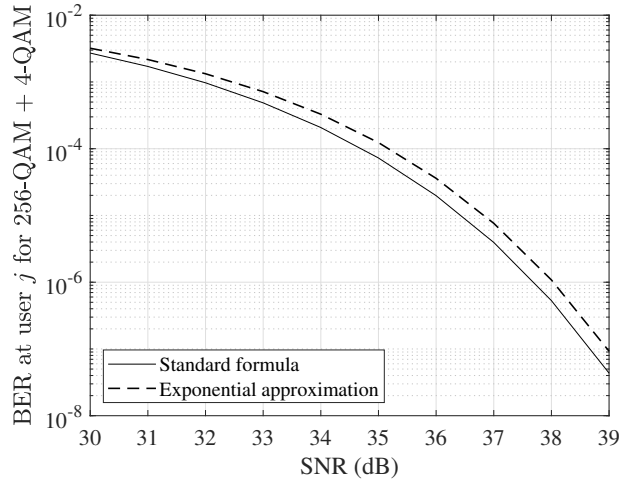


(b) BER at user j

Fig. 4: BER approximations for 4-QAM + 4-QAM



(a) BER at user k



(b) BER at user j

Fig. 5: BER approximations for 256-QAM + 4-QAM

case, for 4-QAM + 4-QAM and 16-QAM + 4QAM, (17) is accurate to within 0.67dB with respect to (9).

Expressions (16) and (17) can be used for selecting the modulation level for a pair of NOMA users at each subcarrier. Assuming that the BER constraint β_0 is marginally met at both user j and user k on subcarrier s, and that condition (13) is also met, (16) and (17) can be written in terms of $R_{j,s}$ and $R_{k,s}$ as

$$\exp\left(\frac{c|h_{j,s}|^2 p_s}{2^{R_{j,s}+R_{k,s}}-1}\right) = 7.14\beta_0 \cdot \frac{2^{0.5(R_{j,s}+R_{k,s})}}{1.6+\sqrt{0.18R_{j,s}R_{k,s}}}, \tag{18}$$

$$\exp\left(\frac{c|h_{k,s}|^2 p_s}{2^{R_{j,s}+R_{k,s}}-1}\right) = 1.30\beta_0 \cdot \frac{2^{0.5(R_{j,s}+R_{k,s})}}{1.6+\sqrt{0.18R_{j,s}R_{k,s}}}, \tag{19}$$

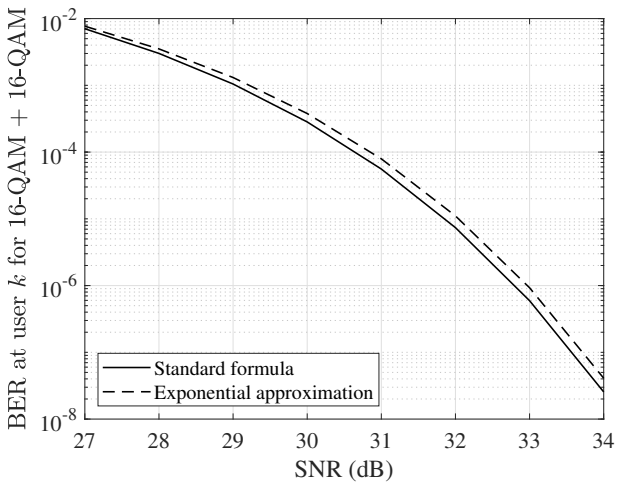
respectively, where $c = -1.54T_{sym}/N_0$. After some algebraic manipulations, the channel gains at users j and k on subcarrier s can be expressed in terms of $R_{j,s}$, $R_{k,s}$ and β_0 as

$$|h^s_{u_{j,s}}|^2 = \ln\left(2.86\beta_0 \frac{2^{0.5(R_{j,s}+R_{k,s})}}{1.6+\sqrt{0.18R_{j,s}R_{k,s}}}\right) \cdot \frac{2^{R_{j,s}+R_{k,s}}-1}{cp_s}, \tag{20}$$

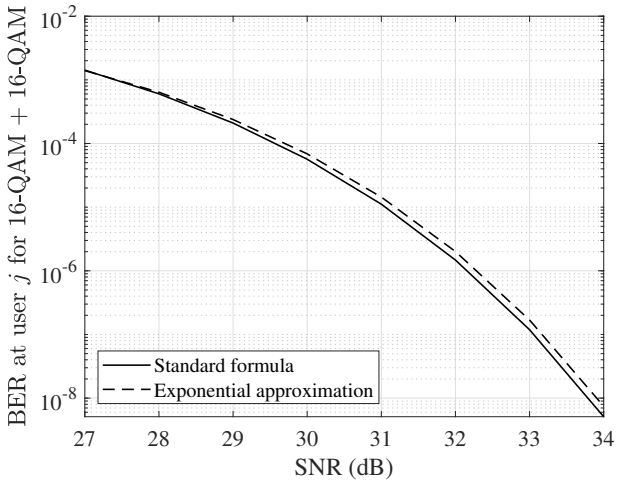
$$|h^s_{u_{k,s}}|^2 = \ln\left(1.11\beta_0 \frac{2^{0.5(R_{j,s}+R_{k,s})}}{1.6+\sqrt{0.18R_{j,s}R_{k,s}}}\right) \cdot \frac{2^{R_{j,s}+R_{k,s}}-1}{cp_s}, \tag{21}$$

respectively. It is clear that, for a fixed p_s and β_0, a larger data rate requires larger a channel gain at both user j and user k. Further, dividing (19) over (18) yields

$$\exp\left(\frac{cp_s\left(|h_{k,s}|^2-|h_{j,s}|^2\right)}{2^{R_{j,s}+R_{k,s}}-1}\right) = 0.18. \tag{22}$$

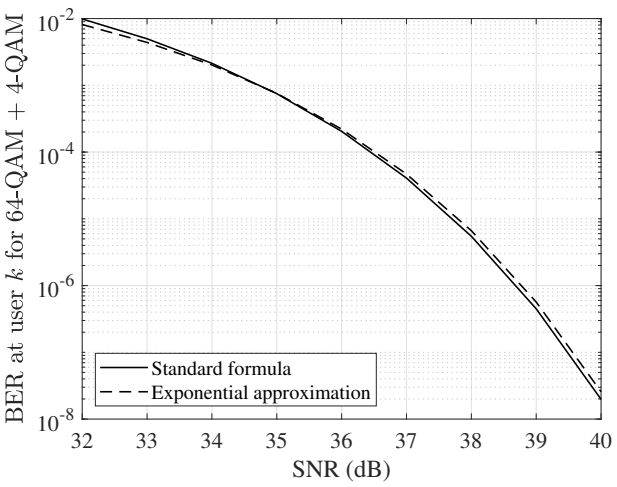


(a) BER at user k

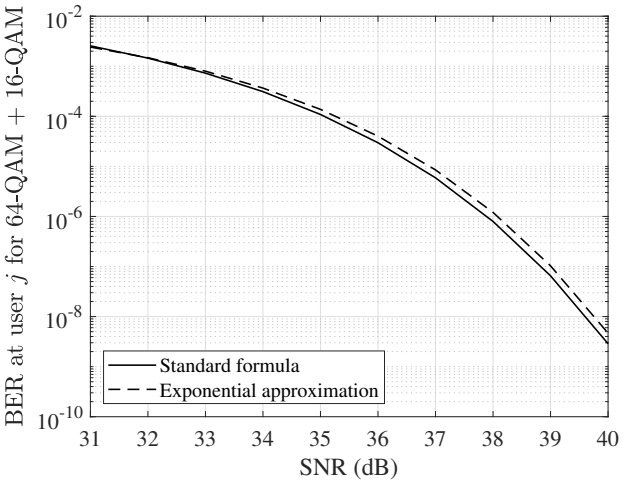


(b) BER at user j

Fig. 6: BER approximations for 16-QAM + 16-QAM



(a) BER at user k



(b) BER at user j

Fig. 7: BER approximations for 64-QAM + 16-QAM

Therefore,

$$H_{\{j,k\},s} = \ln 0.18 \cdot \frac{2^{R_{j,s}+R_{k,s}} - 1}{cp_s}. \tag{23}$$

From (11), by writing $|h_{j,s}|^2$ as $|h_{k,s}|^2/g_{\{j,k\},s}$,

$$H_{\{j,k\},s} = |h_{k,s}|^2 - |h_{j,s}|^2 = |h_{k,s}|^2\left(1 - \frac{1}{g_{\{j,k\},s}}\right), \tag{24}$$

where $H_{\{j,k\},s}$ is the channel gain gap of user j and user k on subcarrier s. From (23), it is clear that, for a fixed p_s, the sum-rate on subcarrier s increases with an increasing value of $H_{\{j,k\},s}$. However, according to (24), $H_{\{j,k\},s}$ is upper-bounded by the channel gain ratio $g_{\{j,k\},s}$, which must be small enough to guarantee that the BER constraint β_0 is simultaneously met at user j and user k for $M_{j,s}$ and $M_{k,s}$, i.e. $g_{\{j,k\},s} \leq \mathcal{G}(M_{j,s}, M_{k,s}, \beta_0)$.

In addition, $g_{\{j,k\},s}$ must be large enough to maximize $H_{\{j,k\},s}$, and therefore the achievable sum-rate $R_{j,s} + R_{k,s}$. Consider the scenario where the BER constraint is marginally met at user k on subcarrier s, and user j is selected as a partner such that the BER constraint is met by a larger margin. In this case, $P_{j,s}(e) \leq P_{k,s}(e)$. From (14) and (15), this condition translates into $d'_{j,s}|h_{j,s}| \geq d_{k,s}|h_{k,s}|$. By substituting (A2) and (A3), this yields

$$g_{\{j,k\},s} \geq 1, \forall \beta_0, M_{j,s}, M_{k,s}. \tag{25}$$

Therefore, from (24), $H_{\{j,k\},s} \geq 0$. Moreover, any two users j and k can be paired on subcarrier s if and only if

$$1 \leq g_{\{j,k\},s} \leq \mathcal{G}(M_{j,s}, M_{k,s}, \beta_0), \tag{26}$$

for a BER constraint β_0 and modulation levels $M_{j,s}$ and $M_{k,s}$.

A relevant conclusion that can be extracted from (24)–(26) is that, in a scenario with a finite number of users, there might not be a suitable NOMA partner for the user with the largest channel gain. Therefore, in order to optimize $H_{\{j,k\},s}$, it might be feasible to select a user with a lower channel gain, but for whom a suitable partner that yields an adequate value of $g_{\{j,k\},s}$ can be found. The power allocation factor can be re-adjusted accordingly, as explained in IV-D.

B. Subcarrier Allocation

The main idea of subcarrier allocation is to allocate the best pair of NOMA users to a subcarrier, with no prior information on the transmit power, such that the sum-rate is maximized. It is assumed that one user can be assigned to more than one subcarrier at the same time. Based on the numerical results in Table I, an efficient user pairing algorithm is proposed to carry out optimal user pairing at each subcarrier. In the user pairing algorithm, the channel gains of users are ordered from smallest to largest on subcarrier s, i.e. $|h_{1,s}|^2 \leq \cdots \leq |h_{K-1,s}|^2 \leq |h_{K,s}|^2$. Starting with the user with the largest channel gain, the set $\mathbf{J}$ of possible partners that fulfill condition (26) is found, then the user j that provides the largest $H_{\{j,k\},s}$ value, $\forall j$, is selected as the partner to user K, i.e. $j = \min(\mathbf{J})$. If no users fulfill (26) when paired with user K, the procedure is repeated for user $K-1$, and so on, until a suitable pair of users is found. The user pairing algorithm is summarized in Algorithm 1, and it is applied in Section IV-F to select the initial pair of NOMA users on every subcarrier for any given transmit power and BER constraints. The computational complexity of the user pairing algorithm is provided in Section IV-H.

Algorithm 1: User Pairing Algorithm

initialization;
order users from lowest to largest channel gain, as $|h_{1,s}|^2 \leq \cdots \leq |h_{K-1,s}|^2 \leq |h_{K,s}|^2$;
set $i = K$ and $j = \emptyset$;
while $j = \emptyset$ *and* $i > 1$ **do**
  set $k = i$;
  find set $\mathbf{J}$ of possible partners for user k that fulfill (26);
  if $\mathbf{J} = \emptyset$ **then**
    i = i-1;
  else
    set j as element from $\mathbf{J}$ that yields the maximum $H_{\{j,k\},s}$, i.e. $j = \min(\mathbf{J})$;
Result: pair of users $\{j, k\}$.

A smaller channel gain ratio between user j and user k on subcarrier s yields a smaller $H_{\{j,k\},s}$ and therefore, according to (23) and Table I, a smaller achievable sum-rate at subcarrier s. In the power allocation procedure in Section IV-C, this is penalized by allocating less power to subcarriers with a smaller $H_{\{j,k\},s}$. However, since (10) and (26) are fulfilled, it is guaranteed that the effect of inter-user interference is manageable at user j, even for small channel gain ratios.

C. Power Allocation

After all subcarriers are allocated, the original objective in (12) leads to the development of power allocation in all subcarriers under the given subcarrier allocation result. Assume that the pairs of NOMA users allocated to each subcarrier are collected in the subcarrier allocation vector $\mathbf{U}$. The optimization problem in (12) can be solved using the Lagrangian method [25].

Lemma 2. *Consider the Lagrangian function*

$$\mathcal{L} = \sum_{s=1}^{S} \left(\hat{R}_{j,s} + \hat{R}_{k,s} \right) - \lambda \left(\sum_{s=1}^{S} p_s - p_{total} \right), \tag{27}$$

where λ is the Lagrange multiplier for constraint (12e) *and p_s is the total power allocated to subcarrier s. In order to simplify the analysis, $\{R_{j,s}, R_{k,s}\} \in \mathbb{I}^+$ in* (12)*, which are positive integers, are transformed into their equivalent real versions $\{\hat{R}_{j,s}, \hat{R}_{k,s}\} \in \mathbb{R}$, respectively. The solution for the optimal power allocation, under the subcarrier allocation result $\mathbf{U}$, is given by*

$$p_s = \overline{p} + \frac{\ln 0.18}{cS} \sum_{t=1}^{S} \frac{1}{H_{\{j,k\},t}} - \frac{\ln 0.18}{cH_{\{j,k\},s}} \geq 0, \tag{28}$$

where $\overline{p} = p_{total}/S$ is the average transmit power per subcarrier, and $(1/S) \cdot \sum_{t=1}^{S} 1/H_{\{j,k\},t}$ is the average of $1/H_{\{j,k\},t}$ over all subcarriers $t = 1, \cdots S$.

Proof. See Appendix B. □

According to (26) and (28), the power allocated to subcarrier s increases when the channel gain gap $H_{\{j,k\},s}$ of users j and k on subcarrier s increases. This fact follows the water-filling principle in multiuser environments [23]. Note that, due to constraint (12b), if (28) is negative then p_s is set to zero. In the event that one or more subcarriers are allocated with zero power, power allocation is carried out again among subcarriers allocated with non-zero power.

D. Power correction factor

The approximation error introduced by (16) and (17) with respect to (8) and (9) results in excess power allocated to subcarriers. Let $\varepsilon_{j,s}$ denote the excess power introduced on subcarrier s by (16) with respect to (8), and let $\varepsilon_{k,s}$ denote the excess power introduced on subcarrier s by (17) with respect to (9). By reducing the powers allocated to user j and user k by factors of $\varepsilon_{j,s}$ and $\varepsilon_{k,s}$ respectively, the same sum rate can be achieved while fulfilling the BER constraint β_0. The excess powers allocated to user j and user k on subcarrier s become equal by setting the power allocation factor at user j to $\check{\alpha}_{j,s} = \alpha_{j,s}\sqrt{\varepsilon_{j,s}/\varepsilon_{k,s}}$. At user k, the power allocation factor is set to $\check{\alpha}_{k,s} = \alpha_{k,s}\sqrt{\varepsilon_{k,s}/\varepsilon_{j,s}}$. Therefore, a power correction factor F_s is defined as

$$\begin{aligned} \mathcal{F}_s &= \sqrt{\varepsilon_{k,s}/\varepsilon_{j,s}}, \text{ if } \varepsilon_{k,s}/\varepsilon_{j,s} \geq 1 \\ \mathcal{F}_s &= \sqrt{\varepsilon_{j,s}/\varepsilon_{k,s}}, \text{ otherwise.} \end{aligned} \tag{29}$$

$\mathcal{F}_s$ is calculated from (29) by applying the numerical values for $\varepsilon_{k,s}/\varepsilon_{j,s}$, as given on Table II. The transmit power per subcarrier is readjusted by applying $\mathcal{F}_s$ to (28), i.e.

$$\check{p}_s = \left(\overline{p} + \frac{\ln 0.18}{cS} \sum_{t=1}^{S} \frac{1}{H_{\{j,k\},t}} - \frac{\ln 0.18}{cH_{\{j,k\},s}} \right) \cdot \frac{1}{\mathcal{F}_s} \geq 0. \tag{30}$$

$\mathcal{F}_s$ is applied after data rate allocation, given that prior knowledge about $M_{j,s}$ and $M_{k,s}$ is needed.

Ratio $\varepsilon_{k,s}/\varepsilon_{j,s}s$				
Modulation \ β_0	10^{-3}	10^{-4}	10^{-5}	10^{-6}
4-QAM + 4-QAM	1.101	1.088	1.074	1.066
16-QAM + 4-QAM	1.124	1.226	1.084	1.073
64-QAM + 4-QAM	1.110	1.096	1.083	1.068
256-QAM + 4-QAM	0.926	0.928	0.936	0.943
16-QAM + 16-QAM	1.023	1.040	1.038	1.036
64-QAM + 16-QAM	0.989	0.970	0.965	0.968

TABLE II: Numerical evaluation of the ratio $\varepsilon_{k,s}/\varepsilon_{j,s}$ for M_k-QAM + M_j-QAM.

E. Modulation Level Selection

The selection of user k's modulation level is carried out after subcarrier allocation. From (23), the data rate at user k on subcarrier s can be expressed in terms of the variable $H_{\{j,k\},s}$ as

$$2^{\hat{R}_{j,s}+\hat{R}_{k,s}} = \frac{cp_s}{\ln 0.18} H_{\{j,k\},s} + 1. \tag{31}$$

From (28), it is found that (31) is equivalent to

$$2^{\hat{R}_{j,s}+\hat{R}_{k,s}} = \frac{cH_{\{j,k\},s}}{\ln 0.18} \cdot \overline{p} + \frac{1}{S}\sum_{t=1}^{S} \frac{H_{\{j,k\},s}}{H_{\{j,k\},t}}. \tag{32}$$

Therefore,

$$\hat{R}_{j,s}+\hat{R}_{k,s} = \log_2\left(\frac{cH_{\{j,k\},s}}{\ln 0.18} \cdot \overline{p} + \frac{1}{S}\sum_{t=1}^{S} \frac{H_{\{j,k\},s}}{H_{\{j,k\},t}}\right). \tag{33}$$

From (33), it is clear that the largest sum data rate corresponds to the pair of NOMA users with the largest $H_{\{j,k\},s}$ value, which is consistent with the results obtained in (22) and Table I.

F. Iterative Resource Allocation Algorithm

According to (23) and the numerical results from Table I, the maximum achievable sum-rate depends on the optimal choice of a pair of NOMA users on each subcarrier. However, in order to select the optimal pair of users, prior knowledge about the achievable data rate on each subcarrier is needed. Therefore, an iterative resource allocation (IRA) algorithm is proposed in order to perform user pairing, power allocation and modulation selection in an iterative manner.

The subset $\mathbf{V} \subset \mathbf{U}$ is defined as the collection of users j allocated to each subcarrier, whereas $\mathbf{W} \subset \mathbf{U}$ is defined as the subset of users k allocated to each subcarrier, and $\mathbf{V} \cup \mathbf{W} = \mathbf{U}$. The first user in $\mathbf{V}$ is paired with the first user in $\mathbf{W}$ on the first subcarrier, the second user in $\mathbf{V}$ is paired with the second user in $\mathbf{W}$ on the second subcarrier, and so on. In IRA, an initial pair of NOMA users is selected first for each subcarrier, based on the average power per subcarrier $\overline{p}$. Expression (24) can be expressed in terms of $\overline{p}$ as follows,

$$\exp\left(\frac{c\overline{p}H_{\{j,k\},s}}{2^{\hat{R}^s_{u_{j,s}}+\hat{R}^s_{u_{k,s}}}-1}\right) = 0.18. \tag{34}$$

According to (24), the expectation of the channel gain gap $H_{\{j,k\},s}$ across all subcarriers can be expressed as

$$\mathbb{E}\{H_{\{j,k\},s}\} = \left(1 - \frac{1}{g_{\{j,k\},s}}\right)\mathbb{E}\{|h_{k,s}|^2\}, \tag{35}$$

where $\mathbb{E}\{|h_{k,s}|^2\} = 1$. Therefore, from (34), the average data rate on subcarrier s is given by

$$\begin{aligned}\mathbb{E}\{\hat{R}_{j,s}+\hat{R}_{k,s}\} &= \log_2\left(\frac{c\overline{p}\mathbb{E}\{H_{\{j,k\},s}\}}{\ln 0.18}+1\right)\\ &= \log_2\left(\frac{c\overline{p}\left(1-1/g_{\{j,k\},s}\right)}{\ln 0.18}+1\right).\end{aligned} \tag{36}$$

The average data rates of the sets of users $\mathbf{V}$ and $\mathbf{W}$ across all subcarriers are given by $\overline{R}_{\mathbf{V}} = \mathbb{E}\{\hat{R}_{k,s}\}$ and $\overline{R}_{\mathbf{W}} = \mathbb{E}\{\hat{R}_{j,s}\}$, respectively. The average modulation levels are given by $\overline{M}_{\mathbf{V}} = 2^{\overline{R}_{\mathbf{V}}}$ and $\overline{M}_{\mathbf{W}} = 2^{\overline{R}_{\mathbf{W}}}$, respectively. Note that there is an implicit dependency between $\overline{R}_{\mathbf{V}}$, $\overline{R}_{\mathbf{W}}$ and the BER constraint β_0, which is given by the value of the channel gain ratio $g_{\{j,k\},s}$.

In IRA, the initial pair of NOMA users allocated to subcarrier s corresponds to the pair $\{j,k\}$ such that $H_{\{j,k\},s} = \max\{\mathbf{H}_s\}$, where $\mathbf{H}_s$ is the set of all possible channel gain gaps on subcarrier s that fulfill the channel gain ratio requirement for $\mathcal{G}(\lfloor\overline{M}_{\mathbf{W}}\rfloor, \lfloor\overline{M}_{\mathbf{V}}\rfloor, \beta_0)$. Note that the use of the function $\lfloor\cdot\rfloor$ is necessary given that $\{\overline{M}_{\mathbf{W}}, \overline{M}_{\mathbf{V}}\} \in \mathbb{R}$ and the fact that the channel gain ratio values on Table I are given for positive integer modulation values.

Let $\mathbf{V}_{(0)}$ and $\mathbf{W}_{(0)}$ denote the initial sets of users on each subcarrier. The initial power allocation set is denoted as $\hat{\mathbf{P}}_{(0)} = \{p_{total}/S, \cdots, p_{total}/S\}$. The set of achievable data rates for the set of users $\mathbf{V}_{(0)}$ is expressed as $\hat{\mathbf{R}}_{\mathbf{V}_{(0)}} = \{\overline{R}_{\mathbf{V}}, \cdots, \overline{R}_{\mathbf{V}}\}$, and that for the set of users $\mathbf{W}_{(0)}$ is given by $\hat{\mathbf{R}}_{\mathbf{W}_{(0)}} = \{\overline{R}_{\mathbf{W}}, \cdots, \overline{R}_{\mathbf{W}}\}$. Further, the initial modulation level of users $\mathbf{V}_{(0)}$ is given by the set $\{\overline{M}_{\mathbf{V}}, \cdots, \overline{M}_{\mathbf{V}}\}$, and that of the set of users $\mathbf{W}_{(0)}$ is given by $\{\overline{M}_{\mathbf{W}}, \cdots, \overline{M}_{\mathbf{W}}\}$. In IRA, user pairing is applied initially to derive $\mathbf{V}_{(0)}$ and $\mathbf{W}_{(0)}$ in terms of $\overline{M}_{\mathbf{W}}$, $\overline{M}_{\mathbf{V}}$ and β_0.

After the initial set of users is paired to each subcarrier, the achievable modulation level on each subcarrier is calculated in an iterative manner according to (33), and a new pair of users $\{l,m\}$ is allocated on subcarrier s during each iteration. The algorithm finishes when an equilibrium state is found, i.e. when the sets of users $\mathbf{V}_{(i)}$ and $\mathbf{W}_{(i)}$ remain unchanged after the i-th iteration. IRA is summarized in Algorithm 2.

The performance achieved with IRA depends on the distribution of the channel gains of the users within each subcarrier; therefore, IRA yields the same result independently of the optimization order. During IRA, expressions (28), (30) and (36) are computed. However, these expressions are derived from (23), which is ultimately obtained from the BER exponential approximations (16) and (17). Equations (16) and (17) introduce an approximation error with respect to the theoretical BER expressions (8) and (9), and therefore IRA yields an approximation to the optimal solution.

G. Data Rate Conversion

The elements in the sets of resulting data rates from IRA are real, i.e. $\{\hat{\mathbf{R}}_{\mathbf{V}}, \hat{\mathbf{R}}_{\mathbf{W}}\} \subset \mathbb{R}$. In order to obtain discrete modulation levels, it is necessary to carry out a conversion process to transform the sets $\hat{\mathbf{R}}_{\mathbf{V}}$ and $\hat{\mathbf{R}}_{\mathbf{W}}$ into equivalent sets of positive integers, i.e. $\tilde{\mathbf{R}}_{\mathbf{V}} = \{\tilde{R}_{k,1}, \cdots, \tilde{R}_{k,s}\} \subset \mathbb{I}^+$ and

Algorithm 2: Iterative Resource Allocation (IRA) Algorithm

derive $\hat{\mathbf{R}}_{\mathbf{V}_{(0)}}$, $\hat{\mathbf{R}}_{\mathbf{W}_{(0)}}$, $\hat{\mathbf{M}}_{\mathbf{V}_{(0)}}$ and $\hat{\mathbf{M}}_{\mathbf{W}_{(0)}}$ from $\hat{\mathbf{P}}_{(0)}$ and β_0, according to (36);
execute Algorithm 1 to derive $\mathbf{V}_{(0)}$ and $\mathbf{W}_{(0)}$ from $\hat{\mathbf{M}}_{\mathbf{V}_{(0)}}$, $\hat{\mathbf{M}}_{\mathbf{W}_{(0)}}$ and β_0;
derive $\hat{\mathbf{P}}_{(1)}$ from $\mathbf{V}_{(0)}$ and $\mathbf{W}_{(0)}$ according to (28);
derive $\hat{\mathbf{R}}_{\mathbf{V}_{(1)}}$, $\hat{\mathbf{R}}_{\mathbf{W}_{(1)}}$, $\hat{\mathbf{M}}_{\mathbf{V}_{(1)}}$ and $\hat{\mathbf{M}}_{\mathbf{W}_{(1)}}$ from $\mathbf{P}_{(1)}$, then update $\mathbf{P}_{(1)}$ according to (30);
execute Algorithm 1 to derive $\mathbf{V}_{(1)}$ and $\mathbf{W}_{(1)}$ from $\hat{\mathbf{M}}_{\mathbf{V}_{(1)}}$ and $\hat{\mathbf{M}}_{\mathbf{W}_{(1)}}$, then set $i = 1$;
while $\mathbf{V}_{(i)} \neq \mathbf{V}_{(i-1)}$ *or* $\mathbf{W}_{(i)} \neq \mathbf{W}_{(i-1)}$ **do**
  $i = i + 1$;
  derive $\hat{\mathbf{P}}_{(i)}$ from $\mathbf{V}_{(i-1)}$ and $\mathbf{W}_{(i-1)}$;
  derive $\hat{\mathbf{R}}_{\mathbf{V}_{(i)}}$, $\hat{\mathbf{R}}_{\mathbf{W}_{(i)}}$, $\hat{\mathbf{M}}_{\mathbf{V}_{(i)}}$ and $\hat{\mathbf{M}}_{\mathbf{W}_{(i)}}$ from $\hat{\mathbf{P}}_{(i)}$, then update $\hat{\mathbf{P}}_{(i)}$ according to (30);
  execute Algorithm 1 to derive $\mathbf{V}_{(i)}$ and $\mathbf{W}_{(i)}$ from $\hat{\mathbf{M}}_{\mathbf{V}_{(i)}}$ and $\hat{\mathbf{M}}_{\mathbf{W}_{(i)}}$;
make $\hat{\mathbf{P}} = \hat{\mathbf{P}}_{(i)}$, $\hat{\mathbf{R}}_{\mathbf{V}} = \hat{\mathbf{R}}_{\mathbf{V}_{(i)}}$, $\hat{\mathbf{R}}_{\mathbf{W}} = \hat{\mathbf{R}}_{\mathbf{W}_{(i)}}$, $\hat{\mathbf{M}}_{\mathbf{V}} = \hat{\mathbf{M}}_{\mathbf{V}_{(i)}}$, $\hat{\mathbf{M}}_{\mathbf{W}} = \hat{\mathbf{M}}_{\mathbf{W}_{(i)}}$, $\mathbf{V} = \mathbf{V}_{(i)}$, $\mathbf{W} = \mathbf{W}_{(i)}$;
Result: $\hat{\mathbf{P}}$, $\hat{\mathbf{R}}_{\mathbf{V}}$, $\hat{\mathbf{R}}_{\mathbf{W}}$, $\hat{\mathbf{M}}_{\mathbf{V}}$, $\hat{\mathbf{M}}_{\mathbf{W}}$, $\mathbf{V}$ and $\mathbf{W}$.

Algorithm 3: Data Rate Selection (DRS) Algorithm

Data: $\tilde{\mathbf{P}}$, $\hat{\mathbf{R}}_{\mathbf{V}}$, $\hat{\mathbf{M}}_{\mathbf{V}}$, $\mathbf{V}$ and $\mathbf{W}$.
convert $\hat{\mathbf{R}}_{\mathbf{V}} \subset \mathbb{R}$ to $\tilde{\mathbf{R}}_{\mathbf{V}} \subset \mathbb{I}^+$ and derive p_{unused} according to (37);
derive $\Upsilon = \{\Delta p_1, \cdots, \Delta p_s, \cdots, \Delta p_S\}$ according to (38);
while $\exists r : \Delta p^r < p_{unused}$ **do**
  select s such that $\Delta p_s = \min \Upsilon$;
  if $g_{\{j,k\},s} \notin \left[1, \mathcal{G}(\tilde{M}_{j,s}, M^*_{k,s}, \beta_0)\right]$ **then**
    execute Algorithm 1 to find pair of users $\{l, m\}$ such that $g_{\{l,m\},s} \in [1, \mathcal{G}(\tilde{M}_{j,s}, M^*_{k,s}, \beta_0)]$;
    if $H_{\{l,m\},s} \geq H_{\{j,k\},s}$ **then**
      allocate users $\{l, m\}$ to subcarrier s;
      make $\tilde{R}_{k,s} \in \tilde{\mathbf{R}}_{\mathbf{V}}$ equal to $R^*_{m,s}$;
      re-calculate Δp_s for $H_{\{l,m\},s}$;
      make $p_{unused} = p_{unused} - \Delta p_s$ and $\tilde{p}^s = \tilde{p}^s + \Delta p_s$;
    else
      remove Δp_s from Υ;
  else
    make $\tilde{R}_{k,s} \in \tilde{\mathbf{R}}_{\mathbf{V}}$ equal to $R^*_{m,s}$;
    update Δp_s and make $p_{unused} = p_{unused} - \Delta p_s$, $\tilde{p}^s = \tilde{p}^s + \Delta p_s$;
Result: $\tilde{\mathbf{R}}_{\mathbf{V}}$, $\tilde{\mathbf{P}}$, $\mathbf{V}$, $\mathbf{W}$ and p_{unused}.

$\tilde{\mathbf{R}}_{\mathbf{W}} = \{\tilde{R}_{j,1}, \cdots, \tilde{R}_{j,s}\} \subset \mathbb{I}^+$. $\tilde{R}_{j,s}$ and $\tilde{R}_{k,s}$ are, initially, the largest integer values less than or equal to $\hat{R}_{j,s}$ and $\hat{R}_{k,s}$, respectively.

In order to carry out data rate conversion after IRA, the modulation levels $\tilde{M}_{j,s}$ and $\tilde{M}_{k,s}$ are allocated to users j and k, respectively, on subcarrier s. Let $\tilde{p}_s$ express the necessary power at subcarrier s, and denote $\tilde{\mathbf{P}} = \{\tilde{p}_1, \cdots, \tilde{p}_s, \cdots, \tilde{p}_S\}$. After initial modulation selection, given that $\tilde{R}_{j,s} \leq \hat{R}_{j,s}$ and $\tilde{R}_{k,s} \leq \hat{R}_{k,s}$, the required transmit power is given by $\sum_{s=1}^{S} \tilde{p}_s \leq \sum_{s=1}^{S} \hat{p}_s = p_{total}$, i.e. not all available power is used. The unused power after data rate conversion can be evaluated from (30) as

$$p_{unused} = \sum_{s=1}^{S} (\hat{p}_s - \tilde{p}_s) = \frac{\ln 0.18}{c} \sum_{s=1}^{S} \left(\frac{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}}{\hat{\mathcal{F}}_s H_{\{j,k\},s}} - \frac{2^{\tilde{R}_{j,s}+\tilde{R}_{k,s}}}{\tilde{\mathcal{F}}_s H_{\{j,k\},s}} \right). \quad (37)$$

The unused power can be assigned to certain subcarriers in order to further increase the data rate at user k, without affecting user j's allocated data rate on that subcarrier.

Let $R^*_{k,s} \in \mathbb{I}^+$ denote $\lceil \hat{R}_{k,s} \rceil$, i.e. $R^*_{k,s} = \tilde{R}_{k,s} + 1$. In order to achieve $M^*_{k,s} = 2^{R^*_{k,s}}$, if $g_{\{j,k\},s} \notin [1, \mathcal{G}(M^*_{k,s}, \beta_0)]$ it is necessary to apply user pairing in order to select a different pair of NOMA users $\{l, m\}$ on subcarrier s such that $g_{\{l,m\},s} \in [1, \mathcal{G}(M^*_{k,s}, \beta_0)]$. The pair $\{l, m\}$ is allocated to subcarrier s if and only if $H_{\{l,m\},s} \geq H_{\{j,k\},s}$. Otherwise, the pair $\{j, k\}$ remains allocated to subcarrier s and the data rate is set to $\tilde{R}_{k,s}$.

In the case where the pair $\{l, m\}$ is allocated to subcarrier s, the required power increase for a data rate of $R^*_{k,s}$ with respect to $\hat{R}_{k,s}$ is given by

$$\Delta p_s = \frac{\ln 0.18 \cdot 2^{\tilde{R}_{j,s}}}{c} \left(\frac{2^{R^*_{m,s}}}{\mathcal{F}^*_s H_{\{l,m\},s}} - \frac{2^{\hat{R}_{k,s}}}{\hat{\mathcal{F}}_{k,s} H_{\{j,k\},s}} \right). \quad (38)$$

A lower factor of Δp_s means that lower additional extra power is required to increase the data rate at user k on subcarrier s. Therefore, in order to maximize the system sum-rate, extra power is assigned to subcarriers in a strictly increasing order of Δp_s while the total power constraint (12e) is met or, equivalently, while $p_{unused} > 0$. A data rate selection (DRS) algorithm is proposed in Algorithm 3 to assign the optimal integer data rate to each subcarrier, and to allocate p_{unused} to selected subcarriers in order to maximize the sum-rate.

H. Computational Complexity and Convergence

The computational complexity of the user pairing algorithm has been computed over $6.4 \cdot 10^4$ repetitions, for a varying number of users, and it is given in Fig. 8. The complexity of the user pairing algorithm is quasi-linear, of the order of $\mathcal{O}(1.54K)$ for $K = 55$ users. In contrast, the complexity of the user pairing prodecure in FTPC [3] is of the order of $\mathcal{O}(K^2)$ per subcarrier. The user pairing algorithm uses prior knowledge about the optimal channel gain ratio from (26) and Table I and therefore, the achieved complexity is greatly reduced compared to FTPC.

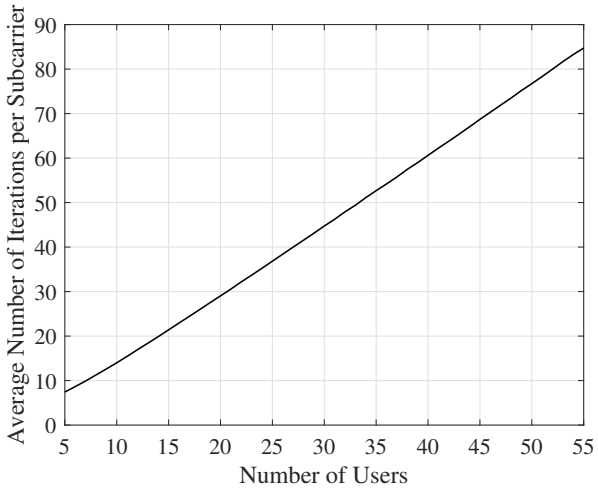


Fig. 8: Complexity of user pairing algorithm.

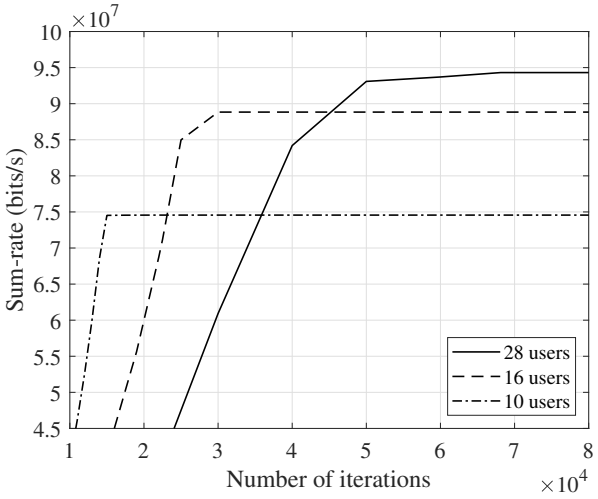


Fig. 9: IRA-DRS convergence performance.

Fig. 9 shows the convergence performance of IRA-DRS, for different numbers of users and $S = 64$. The number of iterations required to achieve the maximum sum-rate performance is approximately linear with the number of users, of the order of $\mathcal{O}(1.30K)$.

V. Numerical Results

The performance of IRA-DRS is evaluated in this section. The system model is formed by a single cell with one BS, $S = 64$ subcarriers, and number of users K between 4 and 32. The BS and all users are equipped with a single transmit antenna. The downlink carrier frequency is 2GHz, the transmission bandwidth is $W = 10$MHz, and the channel bandwidth is given by $W_s = W/S$. The broadband channel is assumed to be frequency selective and independent for all users in all subcarriers. The normalized channel fading factors of all users on each subcarrier follow a Rayleigh distribution with mean square of one. The total power budget is 12.8 W, and the average transmit SNR per subcarrier is 23dB. Continuous power levels are assigned to all subcarriers and users. The BER constraint is $\beta_0 = 10^{-3}$ for all users. In order to ensure statistical significance, all results have been averaged out over the transmission of 5×10^5 time slots.

The performance of IRA-DRS is compared to those of exhaustive search, UB-LDDP [7] and FTPC [3]. UB-LDDP serves as a very tight upper-bound for the optimum achievable value of exhaustive search, and therefore it is used as a theoretical framework for performance evaluation. In UB-LDDP, the transmit power constraints are first relaxed, and then the sum-rate utility is maximized by multiplexing the users with the best channel gains per subcarrier and applying water-filling power allocation per subcarrier; dynamic programming is then performed to allocate power across all subcarriers. In FTPC, the user with the largest channel gain is opportunistically paired with a user that can meet the BER constraint. FTPC employs similar transmission power control to that used in LTE [3], where power is allocated to users according to a decay factor that is assumed fixed for every subcarrier, such that more power is assigned to the user with a lower channel gain. All subcarriers are allocated with equal transmit power, as in [3]. FTPC is a commonly used framework in the literature for sub-optimal resource allocation schemes [7], since it offers a low-complexity solution that can be easily integrated into multi-carrier systems, and which yields a better sum-rate performance than OMA schemes. There exist other schemes in the literature that provide better gains than FTPC, but this is at a cost of very large computational complexity. For example, the work in [18] proposed a heuristic based on gradient descent for the optimization of the sum-rate in multicarrier NOMA; however, this scheme relies on an initial stage of precomputation where exhaustive search is applied, and subsequent allocations are evaluated based on their cost with respect to the optimal solution. This yields a performance within 0.1% of the optimal, but at a computational cost of $\mathcal{O}(K^2 + 2CK)$, where C is the number of discrete power allocation factor levels, of the order of $10^2 - 10^3$.

Fig. 10 represents the sum-rate achieved by IRA-DRS, for different numbers of users, for $S = 8$ and $S = 64$. In IRA-DRS, discrete-level modulation is employed. Therefore, there will always exist a performance gap with respect to exhaustive search and UB-LDDP, since the optimal sum-rate performance is based on the theoretical capacity, which cannot be achieved under discrete modulation schemes. For a system with $K = 32$, IRA-DRS has a performance loss within 2% of the maximum rate achieved with exhaustive search, and a performance gain of over 28% with respect to FTPC. However, for a system with $K < 8$, IRA-DRS is outperformed by FTPC. The explanation is as follows. For a small number of users, there is a small probability of finding a pair of users with a suitable channel condition ratio in IRA-DRS, and it might not be possible to apply the NOMA principle to some of the subcarriers. Further, in IRA-DRS, a minimum modulation level of 4-QAM + 4-QAM is assumed. In FTPC, a continuous modulation level is assumed, so it might be feasible to apply NOMA to a subcarrier yielding a total modulation level smaller than $4 \cdot 4$ in a situation where NOMA could not be applied through IRA-DRS. Nevertheless, IRA-DRS greatly benefits from multi-user diversity, as the probability of finding two users with a channel condition ratio close to

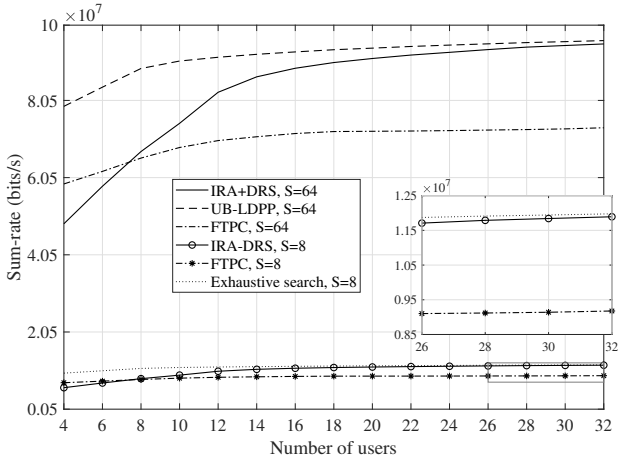


Fig. 10: Sum-rate versus number of users.

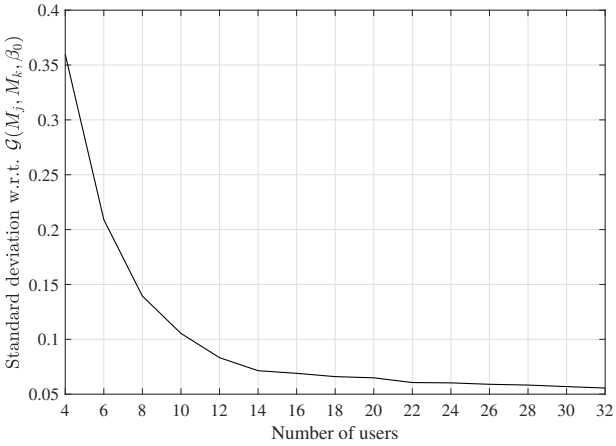


Fig. 12: Standard deviation w.r.t. $\mathcal{G}(M_j, M_k, \beta_0)$.

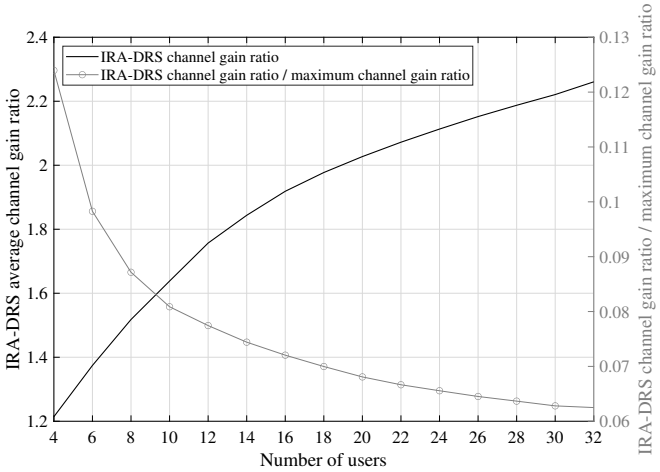


Fig. 11: IRA-DRS average channel gain ratio.

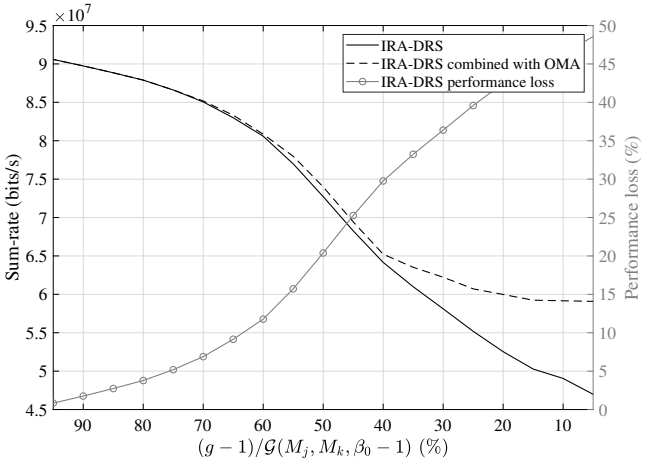


Fig. 13: Sum-rate versus channel gain ratio.

the optimal value increases with K. The performance gap with respect to exhaustive search decreases for increasing K, until multi-user diversity gain saturation is reached for a system with $K = 28$. When $K > 28$, the performance gap remains approximately constant. Overall, the IRA-DRS scheme provides an excellent tradeoff between achievable performance and system complexity for $K \geq 8$.

Fig. 11 shows the average channel gain ratio for IRA-DRS. The quotient of this curve over the maximum channel gain ratio is also presented, where the maximum channel gain ratio is calculated as follows. A set of randomized channel gains are obtained for each subcarrier over 10^7 realizations. Only the channel gain values between the 5th and 95th percentiles of the resulting distribution are considered. Then, for each subcarrier, the ratio between the largest and smallest channel gains is calculated, and this value is averaged over all the realizations. When the number of users increases, the IRA-DRS average channel gain ratio increases, but the coefficient between the IRA-DRS average channel gain ratio and the maximum channel gain ratio decreases. Therefore, the maximum channel gain ratio increases faster than the IRA-DRS channel gain ratio. This is due to the fact that, for a fixed transmit SNR, there exists a limit on the sum-rate that IRA-DRS can achieve. Therefore, there also exists a limit to the IRA-DRS maximum average channel gain ratio. Moreover, Fig. 11 illustrates the fact that, when practical QAM schemes are applied to NOMA, the benefit of pairing users with the most distinct channel condition [10] is lost, due to the inability of users with poor channel conditions to meet BER constraints. In the IRA-DRS setting simulated in Fig. 11, the average maximum channel gain ratio between NOMA pairs is of the order of 2.1 as the optimal sum-rate is approached. This result is meaningful since it implies that NOMA user pairing schemes can be greatly simplified, by only searching among a handful of users that fulfill certain channel gain ratio conditions.

Fig. 12 presents the standard deviation of the IRA-DRS average channel gain ratio with respect to the optimal value, $\mathcal{G}(M_j, M_k, \beta_0)$. The curve decreases with the number of users due to multi-user diversity. By comparing Figs. 10 and 12, it is observed that IRA-DRS outperforms FTPC when the standard deviation with respect to $\mathcal{G}(M_j, M_k, \beta_0)$ is smaller than 0.15.

Fig. 13 shows the system performance versus the ratio $(g-1)/(\mathcal{G}(M_j, M_k, \beta_0) - 1)$, for $K = 20$ users, where a value of 100% is equivalent to $g = \mathcal{G}(M_j, M_k, \beta_0)$, and a value of 0%

is equivalent to $g = 1$, which are the boundaries for the value of the channel gap ratio according to (26). The performance loss of IRA-DRS becomes more critical for values of $(g-1)/(\mathcal{G}(M_j, M_k, \beta_0)-1)$ under 65%, when the negative slope of the sum-rate curve becomes steeper, and the performance loss reaches 10%. For values of $(g-1)/(\mathcal{G}(M_j, M_k, \beta_0)-1)$ under 40%, a large benefit can be obtained by applying OMA to unused subcarriers. Further, when $(g-1)/(\mathcal{G}(M_j, M_k, \beta_0)-1)$ falls under 50%, the performance achieved by IRA-DRS and IRA-DRS combined with NOMA falls under the achieved by FTPC, as given in Fig. 10. Therefore, by increasing the lower boundary for the channel gain ratio in (26), a better trade-off between performance and simplicity can be achieved, for example by applying OMA or FTPC to subcarriers where the value of the NOMA channel gain ratio falls under a given threshold.

VI. Conclusions

In this work, the problem of resource allocation in multi-carrier NOMA with BER and transmit power constraints was studied. Exact values of the optimal channel gain ratios between a pair of NOMA users were derived for practical QAM levels, along with numerical limit values of channel gain ratios that fulfill BER constraints. Based on these findings, a user pairing algorithm with quasi-linear complexity was presented, and an IRA-DRS algorithm was proposed for data rate and continuous power allocation. Numerical results showed that the benefit of pairing users with very distinct channel gains is lost under practical QAM schemes, due to the inability of users with poor channel conditions to fulfill BER constraints. IRA-DRS benefits from multi-user diversity. It yields an achievable sum-rate close to that of exhaustive search, and it outperforms other resource allocation schemes such as FTPC in terms of system sum-rate.

Appendix A: SER Derivation

The Euclidean distances between superconstellation points are given by

$$d_j = \sqrt{\frac{1.5\log_2 M_j}{M_j-1}\alpha_j E_{bit,j}} = \sqrt{\frac{1.5}{M_j-1}\alpha_j E_{sym,j}}, \quad \text{(A1)}$$

$$d_k = \sqrt{\frac{1.5\log_2 M_k}{M_k-1}\alpha_k E_{bit,k}} = \sqrt{\frac{1.5}{M_k-1}\alpha_k E_{sym,k}}, \quad \text{(A2)}$$

$$d'_j = d_j - \left(\sqrt{M_k}-1\right)d_k. \quad \text{(A3)}$$

A. User j's SER

For clarity, let us consider the in-phase dimension first. In the lower superconstellation layer, there are $\sqrt{M_j}$ decision regions from which user j decodes its symbols. There are two outer decision regions which are limited by only one decision boundary. In addition, there are $\sqrt{M_j}-2$ decision regions enclosed within two decision boundaries. Enclosed within each decision region is a M_k-QAM constellation, which represents user k's symbols. In terms of decoding at user j, each upper constellation point introduces a certain amount of interference, due to the varying distance to user j's decision boundary. This results in a certain error rate during decoding. Averaging out the corresponding error rate per branch yields

$$P_j^{(I)}(e) = P_j^{(Q)}(e) = \frac{2\sqrt{M_j}-2}{\sqrt{M_j M_k}}\sum_{i=0}^{\sqrt{M_k}-1} Q\left(2\frac{d'_j+2id_k}{\sqrt{2N_0}}\right), \quad \text{(A4)}$$

due to in-phase and quadrature symmetry. The joint SER is given by

$$P_j(e) = 1-\left[1-P_j^{(I)}(e)\right]^2 = 1-\left[1-P_j^{(Q)}(e)\right]^2. \quad \text{(A5)}$$

The level of error due to interference from user k's symbols is given by

$$P_{k\to j}(e) = P_j(e) - 2Q\left(\sqrt{\frac{2}{N_0}}d_j\right), \quad \text{(A6)}$$

where $2Q(\sqrt{(2/N_0)}d_j)$ is the SER of an M_j-QAM constellation with transmit power $\alpha_j p$.

B. User k's SER

During the first stage of SIC decoding, user k removes user j's symbols first. Errors in SIC are caused by symbols received on the wrong side of the lower layer decision boundaries. Therefore, the SER due to SIC errors in the in-phase dimension is given by

$$P_k^{(I)}(e|error_{x_j})P_k^{(I)}(error_{x_j}) = \frac{2\sqrt{M_j}-2}{\sqrt{M_j M_k}}\sum_{i=0}^{\sqrt{M_k}-1} Q\left(2\frac{d'_j+2id_k}{\sqrt{2N_0}}\right), \quad \text{(A7)}$$

where $P_k^{(I)}(e|error_{x_j}) = 1$, i.e., an error in decoding user j's symbols at user k necessarily causes an error when user k decodes its own symbols. If SIC is successful, user k decodes its own symbols from an M_k-QAM constellation, yielding [24]

$$P_k^{(I)}(e|correct_{x_j}) = 2\left(1-\frac{1}{\sqrt{M_k}}\right)Q\left(\sqrt{\frac{2}{N_0}}d_k\right). \quad \text{(A8)}$$

Thus, from (7),

$$P_k^{(I)}(e) \lessapprox 2\left(1-\frac{1}{\sqrt{M_k}}\right)Q\left(\sqrt{\frac{2}{N_0}}d_k\right) \cdot\left(1-\frac{2\sqrt{M_j}-2}{\sqrt{M_j M_k}}\sum_{i=0}^{\sqrt{M_k}-1} Q\left(2\frac{d'_j+2id_k}{\sqrt{2N_0}}\right)\right) + \frac{2\sqrt{M_j}-2}{\sqrt{M_j M_k}}\sum_{i=0}^{\sqrt{M_k}-1} Q\left(2\frac{d'_j+2id_k}{\sqrt{2N_0}}\right). \quad \text{(A9)}$$

As the SER on each dimension is identical, $P_k^{(I)}(e) = P_k(e)^{(Q)}$, and the joint SER can be calculated by substituting (A9) into

$$P_k(e) = 1-\left[1-P_k^{(I)}(e)\right]^2 = 1-\left[1-P_k^{(Q)}(e)\right]^2. \quad \text{(A10)}$$

APPENDIX B: TRANSMIT POWER ALLOCATION

The solution to the optimal power allocation can be found by differentiating $\mathcal{L}$ in (27) with respect to p_s and equaling each derivation to zero, i.e.

$$\frac{\partial \mathcal{L}}{\partial p_s} = \frac{\mathrm{d}\hat{R}_{j,s}}{\mathrm{d}p_s} + \frac{\mathrm{d}\hat{R}_{k,s}}{\mathrm{d}p_s} - \lambda = 0. \tag{B1}$$

In order to obtain $\mathrm{d}\hat{R}_{j,s}/\mathrm{d}p_s$ and $\mathrm{d}\hat{R}_{k,s}/\mathrm{d}p_s$, both sides of (22) are differentiated with respect to p_s, yielding

$$\begin{aligned}&\exp\left(\frac{cH_{\{j,k\},s}p_s}{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1}\right)\cdot cH_{\{j,k\},s}\\ &\cdot\frac{\left(2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1\right)-\ln 2\cdot 2^{\hat{R}_{j,s}+\hat{R}_{k,s}}p_s\left(\frac{\mathrm{d}\hat{R}_{j,s}}{\mathrm{d}p_s}+\frac{\mathrm{d}\hat{R}^s_{u_{k,s}}}{\mathrm{d}p_s}\right)}{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1}\\ &=0.\end{aligned} \tag{B2}$$

In order for (B2) to be zero, the following condition must be fulfilled:

$$\begin{aligned}&\left(2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1\right)\\ &\qquad-\ln 2\cdot 2^{\hat{R}_{j,s}+\hat{R}_{k,s}}p_s\left(\frac{\mathrm{d}\hat{R}_{j,s}}{\mathrm{d}p_s}+\frac{\mathrm{d}\hat{R}_{k,s}}{\mathrm{d}p_s}\right)=0.\end{aligned} \tag{B3}$$

Hence, the derivative of the data rate at user $u_{k,s}$ with respect to the transmit power can be expressed as

$$\frac{\mathrm{d}\hat{R}_{j,s}}{\mathrm{d}p_s}+\frac{\mathrm{d}\hat{R}_{k,s}}{\mathrm{d}p_s}=\frac{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1}{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}p_s\ln 2}. \tag{B4}$$

Substituting (B4) in (B1) yields

$$\lambda=\frac{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1}{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}p_s\ln 2},\ s=1,\cdots,S. \tag{B5}$$

Thus, the transmit power of subcarrier s is derived as

$$p_s=\frac{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1}{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}}\cdot\frac{1}{\lambda\cdot\ln 2} \tag{B6}$$

Consider now that the constraint given in (12e) is marginally met, i.e.

$$\sum_{s=1}^{S}p_s=p_{total}. \tag{B7}$$

After substituting (B6) into (B7), λ is derived as

$$\lambda=\frac{1}{\ln 2\cdot p_{total}}\sum_{s=1}^{S}\frac{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1}{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}}. \tag{B8}$$

By equaling the value of λ in (B6) and (B8), the transmit power allocated to subcarrier s is derived as

$$\begin{aligned}p_s&=\frac{p_{total}\left(2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1\right)}{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}\sum_{t=1}^{S}\left(1-\frac{1}{2^{\hat{R}_{j,t}+\hat{R}_{k,t}}}\right)}\\ &=\frac{p_{total}\left(2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1\right)}{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}S-\sum_{t=1}^{S}\left(\frac{2^{\hat{R}_{j,s}+\hat{R}_{k,s}}}{2^{\hat{R}_{j,t}+\hat{R}_{k,t}}}\right)}.\end{aligned} \tag{B9}$$

From (23) and (B6), the variable $H_{\{j,k\},s}$ can be written in terms of λ. Then, by replacing $H_{\{j,k\},s}$ into (B9),

$$p_s=\frac{p_{total}\left(2^{\hat{R}_{j,s}+\hat{R}_{k,s}}-1\right)}{2^{\hat{R}_{j,s}+\hat{R}^s_{u_{k,s}}}S-\sum_{t=1}^{S}\frac{H_{\{j,k\},s}}{H_{\{j,k\},t}}}. \tag{B10}$$

Further, by expressing $2^{\hat{R}_{j,s}+\hat{R}_{k,s}}$ in terms of $H_{\{j,k\},s}$ and p_s according to (23), the transmit power allocation at subcarrier s is finally given by

$$p_s=\overline{p}+\frac{\ln 0.18}{cS}\sum_{t=1}^{S}\frac{1}{H_{\{j,k\},t}}-\frac{\ln 0.18}{cH_{\{j,k\},s}}\geq 0, \tag{B11}$$

where $\overline{p}=p_{total}/S$ is the average transmit power per subcarrier, and $(1/S)\cdot\sum_{t=1}^{S}1/H_{\{j,k\},t}$ is the average of $1/H_{\{j,k\},t}$ over all subcarriers $t=1,\cdots S$.

REFERENCES

[1] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, 2019.

[2] J. Zeng, T. Lv, R. P. Liu, X. Su, M. Peng, C. Wang, and J. Mei, "Investigation on evolving single-carrier NOMA into multi-carrier NOMA in 5G," *IEEE Access*, vol. 6, pp. 48 268–48 288, 2018.

[3] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Int. Symp. Personal Indoor Mobile Radio Commun.*, Sep. 2013, pp. 611–615.

[4] R. Zhang and L. Hanzo, "A unified treatment of superposition coding aided communications: Theory and practice," *IEEE Commun. Surveys Tutorials*, vol. 13, no. 3, pp. 503–520, Third Quarter 2011.

[5] H. Haci, H. Zhu, and J. Wang, "Performance of non-orthogonal multiple access with a novel asynchronous interference cancellation technique," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1319–1335, Mar. 2017.

[6] H. Zhu and J. Wang, "Chunk-based resource allocation in OFDMA systems - part II: joint chunk, power and bit allocation," *IEEE Trans. Commun.*, vol. 60, no. 2, pp. 499–509, Feb. 2012.

[7] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, 2016.

[8] J. Tseng, Y. Chen, and C. Wang, "User selection and resource allocation algorithms for multicarrier NOMA systems on downlink beamforming," *IEEE Access*, vol. 8, pp. 59 211–59 224, 2020.

[9] J. Shi, W. Yu, Q. Ni, W. Liang, Z. Li, and P. Xiao, "Energy efficient resource allocation in hybrid non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3496–3511, 2019.

[10] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. PP, no. 99, pp. 1–1, Sep. 2015.

[11] J. He and Z. Tang, "Low-complexity user pairing and power allocation algorithm for 5G cellular network non-orthogonal multiple access," *Electron. Lett.*, vol. 53, no. 9, pp. 626–627, May 2017.

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

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

[14] J. A. Oviedo and H. R. Sadjadpour, "A fair power allocation approach to NOMA in multi-user SISO systems," *IEEE Trans. Veh. Technol*, vol. PP, no. 99, pp. 1–1, Mar. 2017.

[15] J. Zhu, J. Wang, Y. Huang, S. He, X. You, and L. Yang, "On optimal power allocation for downlink non-orthogonal multiple access systems," *IEEE J. Sel. Areas Commun.*, vol. PP, no. 99, pp. 1–1, 2017.

[16] W. Cai, C. Chen, L. Bai, Y. Jin, and J. Choi, "Subcarrier and power allocation scheme for downlink OFDM-NOMA systems," *IET Signal Process.*, vol. 11, no. 1, pp. 51–58, 2017.

[17] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.
[18] L. Salaün, M. Coupechoux, and C. S. Chen, "Joint subcarrier and power allocation in NOMA: Optimal and approximate algorithms," *IEEE Trans. Signal Process.*, vol. 68, pp. 2215–2230, 2020.
[19] W. Liang, Z. Ding, Y. Li, and L. Song, "User pairing for downlink non-orthogonal multiple access networks using matching algorithm," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5319–5332, 2017.
[20] B. Makki, K. Chitti, A. Behravan, and M. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Society*, vol. 1, pp. 179–189, 2020.
[21] P. K. Vitthaladevuni and M. Alouini, "BER computation of 4/M-QAM hierarchical constellations," *IEEE Trans. Broadcast.*, vol. 47, no. 3, pp. 228–239, 2001.
[22] E. Carmona Cejudo, H. Zhu, and O. Alluhaibi, "On the power allocation and constellation selection in downlink NOMA," in *Proc. IEEE Veh. Technol. Conf.*, Sept. 2017, pp. 1–5.
[23] A. Goldsmith, *Wireless communications*. Cambridge University Press, 2005.
[24] S. T. Chung and A. J. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view," *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1561–1571, Sep. 2001.
[25] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

Estela Carmona Cejudo (S'16-GS'21) received the B.S. degree in Communications Engineering from University of Malaga, Malaga, Spain, in 2011, and the M.S. and Ph.D. degrees in Electronic Engineering from University of Kent, Canterbury, United Kingdom, in 2015 and 2021, respectively. She is currently a Researcher with the Software Networks Department, i2CAT Foundation, Barcelona, Spain. Her research interests include wireless/mobile networks, software defined networks, network function virtualization, machine learning and artificial intelligence driven network automation, and signal processing techniques. She was a recipient of the Rohde & Schwarz's 2015 Best Wireless Communications Project, and the BAE Systems' 2015 Best Master's Degree Final Project.

Huiling Zhu (SM'17) received the B.S degree from Xidian University, China, and the Ph.D. degree from Tsinghua University, China. She is currently a Reader (Associate Professor) in the School of Engineering and Digital Arts, University of Kent, United Kingdom. Her research interests are in the area of wireless communications. She was holding European Commission Marie Curie Fellowship from 2014 to 2016. She received the best paper award from IEEE Globecom 2011. She was Symposium Co-Chair for IEEE Globecom 2015 and IEEE ICC 2018, and Track Co-Chair of IEEE VTC2016-Spring and VTC2018-Spring. Currently, she serves as an Editor for IEEE Transactions on Vehicular Technology.

Jiangzhou Wang (Fellow, IEEE) has been a Professor since 2005 at the University of Kent, U.K. He has published over 300 papers and 4 books in the areas of wireless communications.

Professor Wang is a Fellow of the Royal Academy of Engineering, U.K., Fellow of the IEEE, and Fellow of the IET. He was a recipient of the Best Paper Award from the IEEE GLOBECOM2012. He was an IEEE Distinguished Lecturer from 2013 to 2014. He was the Technical Program Chair of the 2019 IEEE International Conference on Communications (ICC2019), Shanghai, the Executive Chair of the IEEE ICC2015, London, and the Technical Program Chair of the IEEE WCNC2013. He has served as an Editor for a number of international journals, including IEEE Transactions on Communications from 1998 to 2013.