Efficient Bit Loading Algorithm for OFDM-NOMA Systems with BER Constraints

T. Assaf, Member, IEEE, A. Al-Dweik, Senior Member, IEEE, M. S. El Moursi, Senior Member, IEEE, and H. Zeineldin, Senior Member, IEEE

Abstract—This paper proposes a novel bit loading algorithm to multicarrier non-orthogonal multiple access (NOMA) systems to maximize the system total throughput while satisfying the users’ individual quality of service (QoS) constraints. Although bit loading is generally a non-deterministic polynomial-time (NP)-hard problem, even for orthogonal multiple access (OMA), the mutual interference between the users, and the dependence of power coefficients and modulation orders are additional challenges that add substantial complexity to the optimization problem. Therefore, in this paper, we propose an efficient bit loading algorithm for multicarrier NOMA systems and compare the complexity and throughput with a corresponding OMA system. The obtained results show that for the case of NOMA, bit loading provides an additional degree of freedom that allows non-uniform spectrum sharing among the users. In particular scenarios, the system may experience hybrid modes of operation, where the system switches between NOMA and OMA. The obtained numerical results show that NOMA can outperform OMA by 100% in terms of spectral efficiency, for a two-user scenario. The complexity of the loading process for NOMA is noticeably higher than OMA, which is due to the high complexity associated with the bit error rate (BER) computation in the case of NOMA.

Index Terms—Bit loading, non-orthogonal multiple access (NOMA), orthogonal multiple access (OMA), bit error rate (BER), spectrum efficiency, computational complexity, orthogonal frequency-division multiplexing (OFDM), adaptive modulation.

I. INTRODUCTION

The expeditious development of the mobile Internet and Internet of things (IoT) has introduced various capacity, data rate, and power consumption challenges for the fifth generation (5G) wireless communications systems [1]. Consequently, extensive research has been focused in the last few years to develop technologies that can overcome such challenges, such as energy-aware communication and energy harvesting [2], [3], cloud-based radio access networks [4], wireless network virtualization [5], advanced antenna systems [6], [7], and efficient error correction coding [8]. Furthermore, significant advancements have been achieved in terms of multiple access technologies such as the non-orthogonal multiple access (NOMA), which may enhance the spectral efficiency with respect to orthogonal multiple access (OMA), by allowing multiple users to share the transmission resources simultaneously. Generally speaking, NOMA schemes can be classified into two main categories, code-domain NOMA [9], [10] and power-domain NOMA [11], which is the focus of this paper.

Towards further increasing the spectrum efficiency of NOMA, efficient modulation schemes such as orthogonal frequency-division multiplexing (OFDM) have been proposed in the literature [12], [13]. Despite the several advantages of OFDM, its bit error rate (BER) performance in fading channels is similar to single-carrier systems. Therefore, adopting adaptive modulation, commonly known as bit loading, can enhance the performance of OFDM-NOMA systems in terms of BER and consequently spectrum efficiency. Both BER and throughput in OFDM-NOMA depend on the modulation order and operating environments. In the low signal-to-noise ratio (SNR) regime, increasing the modulation order does not contribute to the total throughput due to high data detection errors. In contrast, in the high-SNR regime, increasing the modulation order may significantly improve the total throughput. However, applying bit loading, or adaptive modulation, for OMA [14]–[17] is drastically different from NOMA, because the loading processes for all OMA users are generally independent, i.e., changing the modulation order for any user does not affect the BER of other users. For NOMA, the mutual inter-user interference (IUI) between all users creates BER dependency, i.e., the BER of any user depends on the modulation order of all other users. Moreover, changing the modulation order for a particular subcarrier makes it necessary in certain scenarios to change the power assignment for that subcarrier, which may affect the power assignment for other users as well. Consequently, the bit loading process for NOMA is computationally more demanding than OMA, nevertheless, the throughput gain can be significant.

A. Related Work

Although NOMA has been considered widely in the literature, based on extensive literature search, and to the best of the authors’ knowledge, very little work has considered the throughput maximization problem for NOMA systems using bit loading. For example, user pairing and adaptive turbo trellis coded modulation scheme for cognitive radio (CR) in downlink NOMA are proposed in [18]. The adaptive modulation is implemented by switching between multiple modulation orders based on a predefined set of SNR thresholds, and thus, the throughput is not maximized. In [19], a hybrid OMA/NOMA
downlink system is considered, where the authors proposed a multi-stage algorithm for user grouping, subcarrier allocation, and bit allocation schemes while considering data-rate and error-rate constraints. However, the number of bits allocated to each OFDM subcarrier is predefined and the modulation order assigned for each user depends on the associated SNR. Therefore, no throughput maximization process is performed. Moreover, the study focuses on power reduction with no effort to maximize the throughput. Luo and Teh [20] considered maximizing the throughput of a cooperative NOMA system by adaptively switching between OMA and NOMA at the relay node. However, no error rate is guaranteed and the switching is performed by selecting the mode that maximizes the throughput based on channel gains.

B. Motivation and Contribution

Although adaptive modulation, or bit loading, has been widely considered for OMA as reported in [14]–[17], [21] and the references listed therein, it is not the case for NOMA where only very limited research work considered the problem, and with several simplifying assumptions. Therefore, this paper considers the bit loading problem in OFDM-NOMA systems and proposes an efficient bit loading algorithm for a two-user scenario, where the objective is to maximize the total system throughput while satisfying the users’ individual BER constraints. The considered bit loading problem is a non-convex integer non-linear optimization problem, and thus it is NP-hard. Moreover, the mutual interference between the users and dependence on the power coefficients and modulation orders are critical challenges that introduce additional complexity to the optimization process. For example, given that the number of possible modulation orders is $M$, and the number of subcarriers is $K$ per OFDM symbol, then the total number of combinations to find the optimal modulation orders for $N$ OMA users is $N \times M^K$ because the subcarriers of each user are loaded independently from other users. In the case of NOMA, the subcarriers of all users have to be loaded simultaneously, and thus, the total number of combinations is $M^{K \times N}$, which is substantially larger than the OMA case.

To resolve the complexity problem, we propose an iterative greedy algorithm that loads the bits of each user separately. The proposed solution is based on the incremental allocation algorithm (IAA) for OFDM-OMA with BER constraints [22]. However, the IAA is suitable only for single-user applications, and hence, it should be modified to handle multi-user scenarios. Specifically, the bit loading is performed based on the sorted channel coefficients rather than the BER at every subcarrier where different starting points and loading directions, incremental or decremental, are considered. Therefore, the algorithm is greedy within each iteration, but may change directions in different iterations. The impact of the initial loading is also considered to reduce the number of iterations required to provide reliable results. The performance is presented in terms of the throughput per user, system total throughput, and computational complexity, and compared to OMA. The obtained results show that NOMA significantly outperforms OMA in terms of throughput due to the cooperation between the two users. This behavior can be considered as virtual cognitive radio in the sense that the utilized spectrum by one user is utilized by the other user. In such scenarios, the resources of the user with poor channel conditions can be utilized by the other NOMA user, while they are wasted for OMA. Moreover, the bit loading naturally imposes a hybrid OMA/NOMA mode of operation because only one user might be fully allocated to the subcarrier. The complexity of the loading process for NOMA is noticeably higher than OMA, which is due to the high computational complexity required to compute the BER for the NOMA systems. Moreover, some additional complexity is introduced because of the iterative process. To reduce the complexity, the bit loading process is evaluated using two different initial starting points, namely the fully loaded (FL) and partially loaded (PL). The obtained complexity results show that the PL approach can reduce the complexity by about 50% for several cases of interest.

C. Notations

The notations used throughout the paper are as follows. Boldface uppercase and lowercase symbols, such as $\mathbf{X}$ and $\mathbf{x}$ will denote matrices and row/column vectors, respectively. The transpose is denoted by $(\cdot)^T$, $\mathbb{E}[\cdot]$ is the statistical expectation, and the complex Gaussian random variable with a zero mean and $\sigma^2$ variance is denoted as $\mathcal{C}\mathcal{N}(0, \sigma^2)$.

D. Paper Organization

The rest of the paper is organized as follows. In Sec. II, the system and channel models are represented. The proposed bit loading algorithm is detailed in Sec. III. The computational complexity of the bit loading process for OMA and NOMA systems is derived in Sec. IV. Numerical and simulation results are given in Sec. V, and the work is concluded in Sec. VI. The lists of acronyms and symbols are given in Appendices I and II, respectively.

II. SYSTEM AND CHANNEL MODELS

This work considers a downlink power-domain NOMA system serving two users, $U_1$ and $U_2$, where the user equipment (UE) and the base station (BS) are equipped with a single antenna. The users share $K$ subcarriers of an OFDM symbol through superposition modulation (SM). The multiplexed baseband NOMA symbol can be described as

$$ y = \sqrt{P_T} \beta_1 x^{(1)} + \sqrt{P_T} \beta_2 x^{(2)} $$

where $y = [y_0, y_1, \ldots, y_{K-1}]^T$, $x^{(n)} = [x^{(n)}_0, x^{(n)}_1, \ldots, x^{(n)}_{K-1}]^T$ denotes the modulated data sequence of the $n$th user, $P_T$ is the total transmit power at the BS, and $\beta_n \in [0,1]$ is the power allocation coefficient for the $n$th user. For the rest of the paper, the transmit power $P_T$ is normalized to unity and $\sum_{n=1}^{N} \beta_n = 1$. The data symbols $x^{(n)}_k \forall \{k,n\}$ have unit energy, i.e., $\mathbb{E}\left[|x^{(n)}_k|^2\right] = 1$ where $\mathbb{E}[\cdot]$ denotes the statistical
Each data symbol $x_k^{(n)}$ can be loaded with $u_k^{(n)}$ bits using a particular modulation with an order $M_k^{(n)} = 2^{u_k^{(n)}}$. For a proper detection of NOMA symbols, the power allocation coefficients must satisfy the following constraint [23],

$$\frac{\beta_1}{\beta_2} > \frac{M_k^{(1)} - 1}{M_k^{(2)} - 1} \left( \sqrt{M_k^{(2)} - 1} \right)^2 \forall k. \tag{2}$$

The NOMA stream $y$ is converted to an OFDM symbol using the inverse discrete Fourier transform (IDFT), and a cyclic prefix (CP) is added to prevent inter-symbol interference (ISI). At the receiver, the CP is removed and discrete Fourier transform (DFT) is employed for demodulation. For the $n$th user, assuming that the channel remains fixed during one OFDM symbol period and the CP is larger than the maximum delay spread of the channel, the DFT output can be represented as,

$$r^{(n)} = G^{(n)} y + w^{(n)} \tag{3}$$

where $\{r^{(n)}, w^{(n)}\} \in \mathbb{C}^{K \times 1}$, $w^{(n)} = [w_0^{(n)}, w_1^{(n)}, \ldots, w_{K-1}^{(n)}]$ is the additive white Gaussian noise (AWGN) vector whose elements $w_k^{(n)} \sim \mathcal{CN}(0, \sigma_w^2)$, and $G^{(n)} \in \mathbb{C}^{K \times K}$ is the $n$th user channel frequency response matrix, which is given by

$$G^{(n)} = \text{diag}\left\{\left[G_0^{(n)}, G_1^{(n)}, \ldots, G_{K-1}^{(n)}\right]\right\} \tag{4}$$

and

$$G_k^{(n)} = \sqrt{d_n \xi} \sum_{i=0}^{Q_k(n)} g_i^{(n)} \exp \left( -j \frac{2 \pi i k}{K} \right) \forall k \tag{5}$$

where the distance between the $n$th user and the BS is denoted as $d_n$, $\xi$ is the path loss exponent, $g_i^{(n)}$ is the $i$th multipath component gain and $Q_k^{(n)} + 1$ represents the number of multipath components. In the NOMA system considered in this work we assume that the first user, has lower average channel gain, i.e., $\mathbb{E} \left[ \frac{1}{K} \sum_k \alpha_k^{(1)} \right] < \mathbb{E} \left[ \frac{1}{K} \sum_k \alpha_k^{(2)} \right]$, where $\alpha_k^{(n)} \triangleq |G_k^{(n)}|^2$. Thus, the power coefficients should be allocated in the opposite order of the channel amplitudes, i.e., $\beta_1 > \beta_2$. It should be noted that the coherence time is assumed to be larger than the OFDM symbol duration. After the DFT, the information detection can be performed using the joint multi-user detection (JMuD) [24] or successive interference cancellation (SIC) [25]. It is worth noting that for the considered system model, both the SIC and JMuD detection schemes have equal BER performance while their computational and delay properties are different [24].

### III. BIT LOADING FOR NOMA

Generally speaking, bit loading for OFDM-NOMA is substantially more complicated than OFDM-OMA, due to the mutual IUI, complexity of computing the BER, and the relation between the modulation order and power coefficients allocated for each user. Therefore, we propose in this work a low complexity spectrally-efficient algorithm to enable maximizing the total OFDM-NOMA throughput while satisfying the BER constraints for each user. The proposed algorithm is based on the IAA [14], [15], [22], which is modified to suit the special structure of OFDM-NOMA. In particular, the proposed algorithm maximizes the system total throughput by assigning each user a certain number of bits per subcarrier based on the BER constraints of both users, since the number of bits allocated for one of the users affects the BER of the other due to the mutual IUI. The algorithm allows allocating one or more subcarriers exclusively for one user if the fading is severe for the other user and its BER constraint cannot be maintained. On the other hand, if the fading is not severe, then both users can share that subcarrier, but not necessarily with the same number of bits per user. In this section, the throughput maximization problem is formulated and the proposed solution is presented. However, an overview of the conventional IAA is shown for the sake of completeness.

#### A. Overview of IAA for OMA

The bit loading problem for OFDM with OMA and squared quadrature amplitude modulation (QAM) that aims at maximizing the system throughput can be formulated as

$$\max_{b_k} T_T = \frac{1}{K} \sum_{k=0}^{K-1} b_k \tag{6}$$

subject to:

$$C_1 : b_k \in \{0, 2, 4, 6, 8, \ldots, b_{\text{max}}\} \forall k \tag{7a}$$

$$C_2 : \bar{P} = \frac{\sum_{k=0}^{K-1} b_k P_k}{\sum_{k=0}^{K-1} b_k} \leq \bar{P}_{\text{Th}} \tag{7b}$$

where $T_T$ is the total throughput normalized by the total number of subcarriers, constraint $C_1$ in (7a) is used to force the modulation to be a squared QAM, while $C_2$ in (7b) is used to ensure that the average BER per OFDM symbol $\bar{P}$ is less than a predefined threshold $\bar{P}_{\text{Th}}$. Moreover, $b_k$ represents the number of bits for the $k$th subcarrier, $b = [b_0, b_1, \ldots, b_{K-1}]$, $b_{\text{max}}$ is the maximum possible number of bits that can be allocated for any subcarrier. The conditional BER per subcarrier for squared QAM can be closely approximated by [22],

$$P_k = \frac{4}{\log_2(M_k)} \varphi_k [1 - \varphi_k] \tag{8}$$

where $M_k = 2^{b_k}$, $\varphi_k = \left(1 - \frac{1}{\sqrt{M_k}}\right) Q \left(\frac{3\alpha_k^2 \gamma}{\gamma \sqrt{M_k} - 1}\right)$, and $Q(\cdot)$ is the complementary cumulative distribution function of the standard normal Gaussian distribution. The average SNR $\bar{\gamma} = 1/\sigma_w^2$ and $\alpha_k^2$ is the channel gain of the $k$th subcarrier. In the context of OMA with fixed power allocation, the power factor is normalized to unity, i.e., $\beta = 1$.

The IAA can be used to efficiently solve the optimization problem in (6), where the process starts by allocating all subcarriers $b_{\text{max}}$ bits. Given that channel state information (CSI), i.e., $\bar{\gamma}$ and $\alpha_k \forall k$ are available at the transmitter, then they can be used to compute $\bar{P}$. If $\bar{P} \leq \bar{P}_{\text{Th}}$, the algorithm stops, otherwise, we search for the subcarrier with the maximum $P_k$ and reduce its modulation order. Then $\bar{P}$ is computed again and compared to $\bar{P}_{\text{Th}}$, if the condition...
Algorithm 1: IAA

| Step | Description |
|------|-------------|
| 1    | Input: $\alpha, \gamma, \beta, P_{Th}$, $b = [b_{1}, \ldots, b_{K}]$ |
| 2    | Compute $P_{k}$ for all $k$ using (8), $P_{k}$ |
| 3    | if $P \leq P_{Th}$ then |
| 4    | Stop |
| 5    | else |
| 6    | $l = \arg\max_{l \in \{1, \ldots, K\}} P_{l}$ |
| 7    | if $b_{l} > 0$ then |
| 8    | $b_{l} = b_{l} - 2$ |
| 9    | Compute $P_{l}$ using (8), $P_{l}$ |
| 10   | Go to: 3 |
| 11   | end |
| 12   | end |
| 13   | Output: $b$ |

is satisfied the algorithm stops, otherwise the algorithm goes into a new iteration by reducing the modulation order of the subcarrier with the maximum $P_{k}$. The algorithm stops when the $P \leq P_{Th}$, or when $b = [0, 0, \ldots, 0]$. The IAA can be explained for OMA using Algorithm 1, where the channel envelope vector $\alpha = [\alpha_{0}, \alpha_{1}, \ldots, \alpha_{K-1}]$. It is worth noting that the operation $l = \arg\max_{l \in \{1, \ldots, K\}} P_{l}$ in Algorithm 1 corresponds to finding the subcarrier with the maximum conditional BER. Moreover, it can be noted that the number of bits per subcarrier is decremented by two bits in each step to limit the modulation scheme to squared QAM.

The complexity of the IAA algorithm is mostly dominated by two main operations. The first is that the algorithm initially calculates the BER for all subcarriers, and then the BER will be computed for each subcarrier that its allocation changes. The second is the search process for the subcarrier with maximum probability of error. The overall complexity depends on the number of times the two operations are repeated, which is a random variable that depends on the channel conditions.

### B. Bit Loading for NOMA: Problem Formulation

Similar to the single user (SU) case, the adaptive two-user bit loading problem that aims at maximizing the total throughputs while satisfying the users' BER constraints can be formulated as

$$b_{\text{max}}^{(1)} = b_{\text{max}}^{(2)} = b_{\text{max}}.$$  

subject to:

$$C_{1} : b^{(n)}_{k} \in \{0, 2, 4, 6, \ldots, b_{\text{max}}^{(n)}\} \forall (n, k)$$  

$$C_{2} : \tilde{P}^{(n)} = \sum_{k=0}^{K-1} \frac{b^{(n)}_{k} P^{(n)}_{k}}{\sum_{k=0}^{K-1} b^{(n)}_{k}} \leq \tilde{P}_{Th}^{(n)} \forall n.$$  

The objective function in (9) aims at maximizing the total throughput over $K$ subcarriers for the two users. Constraint $C_{1}$ in (10a) specifies the possible number of bits per subcarrier which are restricted to squared QAM with $M_{k}^{(n)}$ modulation order. Moreover, for national simplicity we consider that $b_{\text{max}}^{(1)} = b_{\text{max}}^{(2)} = b_{\text{max}}$. Constraint $C_{2}$ in (10b) is used to enforce BER fairness by setting $P^{(n)}_{k}$ based on the quality of service (QoS) requirements for each user. The bit loading problem in (9) allows for three modes of transmission, NOMA, OMA or OFF. In the NOMA mode both users are active, in the OMA mode only one of the two users is active, while in the OFF mode both users must switch off, i.e., $b_{k}^{(1)} = b_{k}^{(2)} = 0$. The value of $P^{(n)}_{k}$ in (10b) can be computed as

$$P^{(n)}_{k} = \left(1 - \Gamma_{k}\right) P^{(n)}_{k} |\Gamma_{k} = 0 + \Gamma_{k} P^{(n)}_{k} |\Gamma_{k} = 1$$

where $\Gamma_{k} \in \{0, 1\}$ is a binary variable that allows computing $P^{(n)}_{k}$ regardless of the number of bits per user in each subcarrier. More specifically,

$$\Gamma_{k} = \begin{cases} 0, & \text{if min} \{b_{k}^{(1)}, b_{k}^{(2)}\} = 0 \\ 1, & \text{otherwise} \end{cases} (11)$$

$P^{(n)}_{k} |\Gamma_{k}$ can be expressed as [26], [27].

$$P^{(n)}_{k} |\Gamma_{k} = 0 = \frac{2}{v^{(n)}_{k} \bar{v}_{k}^{(n)}} \sum_{i=1}^{v^{(n)}_{k}} \sum_{t=0}^{L^{(n)}_{i,i,k}} D^{(n)}_{t,k}$$

$$\times Q \left(g_{n,k}^{+} \left(2t + 1, 2l - \bar{v}_{k}^{(2)} + 1\right)\right), \quad n = 1, 2$$

(12)

and

$$P^{(1)}_{k} |\Gamma_{k} = 1 = \frac{2}{v^{(1)}_{k} \sqrt{M^{(1)}_{k} T^{(1)}_{k}}} \sum_{i=1}^{v^{(1)}_{k}} \sum_{t=0}^{L^{(1)}_{i,i,k}} A^{(1)}_{t,k}$$

$$\times Q \left(g_{1,k}^{+} \left(2t + 1, 2l - \bar{v}_{k}^{(2)} + 1\right)\right)$$

(13)

and

$$P^{(2)}_{k} |\Gamma_{k} = 1 = \frac{2}{v^{(2)}_{k} \sqrt{M^{(2)}_{k} T^{(2)}_{k}}} \sum_{i=1}^{v^{(2)}_{k}} \sum_{t=0}^{L^{(2)}_{i,i,k}} 2A^{(1)}_{t,k}$$

$$\times Q \left(g_{2,k}^{+} \left(c_{l}, 2t + 1\right)\right)$$

$$- \sum_{t=0}^{L^{(2)}_{i,i,k}} 2A^{(1)}_{t,k} \sum_{l=1}^{S_{k} D^{(2)}_{2,k} D^{(1)}_{3,k}} Q \left(g_{2,k}^{+} \left(c_{l}, 2t + 1\right)\right)$$

(14)

where all the symbols in (12), (13) and (14) are defined in Appendix II. It is worth noting that $P^{(n)}_{k} |\Gamma_{k} = 0$ corresponds to the OMA case and (12) is its exact conditional BER expression, which is used in order to be consistent with the NOMA case where the exact BER is used as well. As can be noted from (9), the bit loading problem is a combinatorial non-linear non-deterministic polynomial-time (NP)-hard optimization problem. Therefore, in this paper, an efficient bit loading algorithm is proposed to solve the problem.

### C. Proposed Bit Loading Algorithm for NOMA

The proposed bit loading algorithm is based on the principle of the IAA for OFDM-OMA with BER constraints [14].
However, the IAA is suitable only for single-user applications, and hence, it should be modified to handle multi-user scenarios. The main points that should be considered when applying the IAA to the two-user OFDM-NOMA are:

1) The allocation cannot be performed jointly for the two users because the channels for the two users are different.
2) The order of allocation has a major impact on the user fairness. For example, if the allocation starts with $U_1$, the interference from $U_2$ will be maximum because all subcarriers are loaded with $b_{\text{max}}$. The same scenario is also applicable if the loading process starts with $U_2$. In such scenarios, the user that is loaded first might have less throughput when compared with the user who is loaded second.
3) After completing the loading for both users, most likely there will be a possibility for adding extra bits for the user that was initially loaded due to the reduced interference that results after reducing the total bits for the user that was loaded second.

To take these points into consideration, we propose a modified IAA (MIAA). Unlike the IAA, the MIAA can be applied using an arbitrary initial loading vector for each user, denoted as $b^{(1)}$ and $b^{(2)}$. The MIAA is bidirectional in the sense that it can increase or decrease the number of bits per subcarrier. Moreover, the MIAA does not require the search for the subcarrier with the maximum error probability after each subcarrier allocation operation, which can reduce the time and computational complexity significantly. It is worth noting that the IAA outperforms the MIAA in terms of throughput when both algorithms are used with a single iteration. However, they produce roughly the same throughput when the number of iterations is more than one. The justification for such behavior is that the IAA generally produces better throughput because it is directly based on the BER, while the MIAA will reduce the total number of bits per OFDM symbol because it is based on SNR. In the consequent iterations, the MIAA will be able to add some of the missing bits, while the IAA will have less improvement because its outcome in the first iteration is close to the final loading solution. Consequently, both techniques will converge roughly to the same solution. Moreover, the loading process for a particular user should also consider the impact of the loading process on the other user, because increasing the bits for a user might increase the probability of error for the other user. In the MIAA, the number of bits is decreased by 2 in each iteration because (13) or (14) are valid only for squared QAM. Extending the algorithm for other modulation schemes or constellation shapes is straightforward. To simplify the presentation of the MIAA, it can be divided into two small algorithms MIAA-A and MIAA-D.

1) **Algorithm 2 (MIAA-A):** The MIAA-A is generally similar to the conventional IAA where it starts from a certain initial loading vector $b^{(n)}$, and then starts decreasing the number of bits per subcarrier iteratively. To avoid the repetitive search for the subcarrier with maximum BER, the subcarriers are sorted based on their fading coefficients vector $\alpha^{(n)}$ in an ascending order, which implies that only one sorting process is needed. Therefore, the first subcarrier will be the one with worst fading conditions. Then, the allocation process is performed sequentially by reducing the number of bits per subcarrier until $P^{(n)} \leq P_{Th}$, or when $b^{(n)} = \{0, 0, \ldots, 0\}$. In Algorithm 2, the indicator $c$ is used to avoid an unnecessary sorting iteration of $\alpha^{(n)}$ in good channel conditions where $P^{(n)} \leq P_{Th}$ for $b^{(n)} = \{b_{\text{max}}, \ldots, b_{\text{max}}\}$. The power for the two users is defined as $\beta = [\beta_1, \beta_2]$.

2) **Algorithm 3 (MIAA-D):** The MIAA-D starts from an initial bit loading vector, and then increases the allocation per subcarrier iteratively while satisfying $P^{(n)} \leq P_{Th}$ for $b^{(n)} = \{b_{\text{max}}, \ldots, b_{\text{max}}\}$. Therefore, increasing the number of bits for one user should not cause BER constraint violation for the other user. The bit loading process starts by sorting the subcarriers based on their channel gains vector $\alpha^{(n)}$ in a descending order. Thus, the first subcarrier is the one with the best fading conditions.

3) **Algorithm 4 (I-MIAA):** Based on Algorithm 2 and 3, we propose the iterative MIAA (I-MIAA) Algorithm 4. The I-MIAA has four steps that start by applying the MIAA-A for both users to obtain an initial loading vectors $b^{(1)}$ and $b^{(2)}$ that are generated such that $P^{(1)} \leq P_{Th}^{(1)}$ and $P^{(2)} \leq P_{Th}^{(2)}$. However, the total bits in $b^{(1)}$ and $b^{(2)}$ can be increased because the MIAA-A is strongly affected by the mutual interference between the users, which initially provides pessimistic results. In particular, the user that was loaded first would suffer from a strong interference from the other user. Therefore, the MIAA-D is applied to the user that was loaded first to overcome the severe bit reduction that was obtained in the initial MIAA-A. Because the MIAA-A is not as efficient as the IAA, the MIAA-D should be applied to the user was loaded second as well.

D. Using Partially Loaded Initial Bit Loading Vectors

As can be noted from the MIAA-A in **Algorithm 2**, the initial solution for the bit loading process is $b^{(n)} = \{b_{\text{max}}, b_{\text{max}}, \ldots, b_{\text{max}}\}$, and thus, this vector is called a FL
vector. In such a scenario, converging to the final solution for \( b^{(n)} \) might take several iterations and hence increase the system complexity. To expedite the convergence process to the final solution, the initial loading vectors for both users should be closer to the final solution as compared to the FL case where \( b^{(n)} = b_{\text{max}} \forall k \). Towards this goal, we propose using an initial solution for the MIAA-A that may have a small difference as compared to the final bit allocation. By noting that computing \( P_1 \) incur significant computational complexity as a result of using (13) and (14), the proposed solution for the \( n \)th user can be obtained by successively applying the MIAA-A and MIAA-D algorithms while considering that \( \beta_{3-n} = 0 \), i.e., there is no mutual interference between the users and the low complexity (12) will be used to compute \( P_k \). The output bit loading vectors of this preliminary operation are considered as the initial loading vectors, instead of \( b^{(n)} = b_{\text{max}} \forall k \), for the actual MIAA-A and MIAA-D. Because the loading vectors in this case will mostly have subcarriers with \( b_k < b_{\text{max}} \), the vectors are denoted as PL vectors.

### IV. Computational Complexity

In this section, the computational complexity of the proposed bit loading algorithm is provided. The computational complexity of bit loading algorithms with BER constraints is mostly dominated by the required extensive BER calculations performed during the iterative optimization process [22]. In particular, the complexity is computed in terms of the number of real multiplication (\( \times \)) and addition/subtraction (\( \pm \)) operations, and number of times the \( Q() \) and \( |\cdot| \) functions are evaluated in the BER expressions (12), (13) and (14). For more informative comparison, the overall equivalent complexity is also presented [28]. Moreover, the computational complexity for the exponent term, \( 2^v \) where \( v \) is a real integer, in the BER expressions is neglected as they can be implemented using simple shift registers. Table I presents the computational complexity for the scenario where a single user is transmitting at subcarrier \( k \). Tables II and III show the computational complexity of the BER calculations for \( U_1 \) and \( U_2 \), respectively, where both are transmitting per subcarrier. It can be noticed that the computational complexity is a function of the modulation order \( M_k^{(n)} \) and user’s index. Clearly, the computational complexities of \( P_k^{(1)} \mid r_k = 1 \) and \( P_k^{(2)} \mid r_k = 1 \) are substantially larger than \( P_k^{(n)} \mid r_k = 0 \) which is due to the UII introduced in NOMA systems. In order to quantify the number of operations required, Tables IV, V and VI represent numerical computational complexity results for various \( M_k^{(n)} \) values. The difference between the number of operations between \( P_k^{(n)} \mid r_k = 0 \) and \( P_k^{(n)} \mid r_k = 1 \) are, respectively, 2 and 6 times larger than \( P_k^{(n)} \mid r_k = 0 \), for \( M_k^{(n)} = 4 \). It can be noticed that the computational complexity increases exponentially as \( M_k^{(n)} \) increases when \( U_1 \) and \( U_2 \) are both transmitting while linearly when one of them is transmitting. Consequently, there is a critical trade-off between the computational complexity and throughput requirements for OFDM-NOMA systems which should be carefully tackled in terms of \( M_k^{(n)} \), BER constraints and user’s index.

In order to fairly compare the computational complexity of the IAA and the proposed MIAA-A/D, Algorithm 4 is implemented using two different approaches, the first is based on the MIAA-A/D, and the other is using the original IAA. In both cases FL initial vectors are considered for the loading process. The implementation of Algorithm 4 using IAA initially requires computing \( P_k^{(1)} \forall k \) and \( P^{(n)} \). Then if \( C_2 \) is not satisfied, we search for the subcarrier with the maximum \( P_k^{(n)} \) and reduce its allocated bits by 2. After that the BER for the \( k \)th subcarrier and \( P^{(n)} \) are recomputed to validate \( C_2 \), and the process continues until it is satisfied. The same procedure is applied for the other user. Due to the mutual interference in NOMA systems, the \( n \)th user final output of the aforementioned steps can be further increased by first searching for the subcarrier with the minimum BER then increasing its bits by 2, and computing \( P_k^{(2)} \) and \( P^{(n)} \) to check \( C_2 \). This process continues until \( C_2 \) is violated for both users. As can be noted, the implementation of Algorithm 4 using IAA is mostly determined by two main operations. The first is that the algorithm initially calculates the BER for all subcarriers, and then the BER will be computed for each subcarrier that its allocation changes. The second is the search process for the subcarrier with minimum/maximum probability of error. The overall complexity depends on the number of times the two operations are repeated, which is random. In the worst case scenario, the BER and maximum search processes might be repeated \( 4 \times K^M \) times, which is expected to occur at low SNRs. It should be noted that the complexity of the ascending or descending sorting processes can be represented by the number of maximum/minimum searches. Therefore, as compared to the proposed Algorithm 4 using MIAA-A/D, the BER computations and search processes may be performed...
Algorithm 4: I-MIAA
1. Input: $n, \gamma, \bar{P}_{Th}^{(n)}, \bar{P}_{Th}^{(3-n)}, \beta, \alpha^{(n)}, \alpha^{(3-n)}, b^{(n)}, b^{(3-n)}$
2. $b_1^{(n)} = \text{MIAA-A} \left(n, \gamma, \bar{P}_{Th}^{(n)}, \beta, \alpha^{(n)}, b^{(n)}, b^{(3-n)} \right)$
3. $b_1^{(3-n)} = \text{MIAA-A} \left(3 - n, \gamma, \bar{P}_{Th}^{(3-n)}, \beta, \alpha^{(3-n)}, b^{(3-n)}, b_1^{(n)} \right)$
4. $b_2^{(n)} = \text{MIAA-D} \left(n, \gamma, \bar{P}_{Th}^{(1)}, \bar{P}_{Th}^{(2)}, \beta, \alpha^{(1)}, \alpha^{(2)}, b_1^{(n)}, b_1^{(3-n)} \right)$
5. $b_2^{(3-n)} = \text{MIAA-D} \left(3 - n, \gamma, \bar{P}_{Th}^{(1)}, \bar{P}_{Th}^{(2)}, \beta, \alpha^{(1)}, \alpha^{(2)}, b_1^{(3-n)}, b_2^{(n)} \right)$
6. Output: $b_1^{(1)} = b_2^{(1)}, b_2^{(2)} = b_2^{(2)}$

TABLE I

| $M_k^{(n)}$ | Total Complexity | $Q(\cdot)$ | $\lfloor \cdot \rfloor$ |
|---------|-----------------|-----------|----------------|
| $4$     | $62$            | $1$       | $2$            |

TABLE II

| $M_k^{(1)}$ | $M_k^{(2)}$ | Total Complexity | $Q(\cdot)$ | $\lfloor \cdot \rfloor$ |
|------------|-------------|-----------------|-----------|----------------|
| $4, 4$     | $63$        | $2$             | $4$       |
| $16, 16$   | $602$       | $20$            | $40$      |

V. Numerical Results

This section presents the performance of the proposed algorithm concerning throughput and complexity. The simulation results are obtained using a computing machine that runs Intel Xeon CPU E5-2640 processor, clock frequency of 2.5 GHz, 16 GB RAM, and 64 bit operating system. The considered OFDM-NOMA system has a total of $K = 64$ subcarriers for $U_1$ and $U_2$. In order to obtain a reliable BER performance, the possible modulation orders for each user are set to either QPSK or 16-QAM. Moreover, a subcarrier can be turned off for both users. Therefore, if both users are transmitting at the $k$th subcarrier, the superposed signal modulation order is going to be either 16-QAM or 256-QAM. The BER thresholds $\bar{P}_{Th}^{(1)}, \bar{P}_{Th}^{(2)} \in \{10^{-2}, 10^{-3}\}$. For the case where both users are using a certain subcarrier, $\beta_1$ is chosen based on (2), $\beta_1 \in \{0.92, 0.95, 0.99\}$. These values are selected because they are valid for all the modulation orders that will be used in this work, where the power should be selected such that $\beta_1 \geq 0.9$ because $b_{\text{max}} = 4$ for both subcarriers as shown in (2) or [29] for more general scenarios. If the subcarrier is allocated to only one user, then $\beta_n = 1$ for that user. The numerical results are presented for the two possible initial solutions, i.e., FL and PL.

number of required iterations to obtain the final bit loading.
The $n$th channel is modeled as a quasi-static Rayleigh frequency-selective fading channel, where the channel remains fixed during a given OFDM symbol, but changes randomly between adjacent symbols. More specifically, the channel is modelled as a 15-tap multipath channel with normalized delays of $[0, 1, 4, 5, 12, 14]$ and average gains of $[0.4584, 0, 0, 0, 0.147, 0.0928, 0, 0, 0, 0, 0, 0.1851, 0, 0.1167]$ [22]. The distance ratio between the two users and the BS is $d_1/d_2 = 1.3$, and the path loss exponent is assumed $\xi = 2$. The CSI is assumed to be known perfectly at the transmitter side, unless mentioned otherwise. The results are obtained by running and averaging Monte Carlo simulation for $10^4$ channel realizations. For each channel realization, Algorithm 4 is run at 26 different average SNR values, $\bar{\gamma}$, ranging from 0 to 50 dB.

Figures 1 and 2 represent the throughput of $U_1$ and $U_2$, respectively, for various power coefficients where $\tilde{P}_{Th}^{(1)} = \tilde{P}_{Th}^{(2)} = 10^{-2}$. In the presented results, the starting user and initial solution type are varied and their impacts are investigated. The initial solution type and starting user are denoted as $[\text{FL}, U_n]$ and $[\text{PL}, U_n]$. The SU bit loading solution for the $n$th user is obtained by running Algorithm 4 using $\beta_n = 1$ and $\beta_{3-n} = 0$. This setup can provide a near-optimal solution because it is equivalent to OMA [22], and thus can be considered as an upper bound for Algorithm 4 due to the absence of the IUI. For fair comparison with the proposed algorithm, another SU solution is provided for $U_1$ where $\beta_1 \in \{0.92, 0.95, 0.99\}$ and $\beta_2 = 0$ as shown in Fig. 1. Similarly, Fig. 2 shows the SU solution for $U_2$ using $\beta_2 \in \{0.08, 0.05, 0.01\}$ while $\beta_1 = 0$.

As can be noted from Figures 1 and 2, the throughput per user is not monotonically increasing versus SNR as in the case of OMA [22]. The reason for such behavior is that both users are competing for the same frequency bands, and these bands will be allocated to the user that can contribute more to the total throughput. Therefore, increasing SNR can be beneficial for one user more than the other, which allows that user to be allocated more bits and hence increase the total throughput. For example, for the case of $[\text{FL}, 1]$ in both figures in the range of $\bar{\gamma} \in [14, 18]$ dB, $U_1$ loses about 0.378 bits while $U_2$ gains 0.671, and hence, the total throughput increases by 0.29 bits. The obtained trend also depends on $\beta_1$.

For the SU performance, the two SU solutions do not upper bound the proposed algorithm solutions due to the reduced power coefficients, specifically, over the low SNR region. For SU in Fig. 1 the main reason is that the near user $U_2$, will have generally good channel conditions, which implies that most of its subcarriers will not be switched off. Therefore, the power allocated to the SU is similar to the one used for the proposed algorithm. In some rare cases at low SNRs, a few subcarriers for $U_2$ will be switched off, and thus all the power will be allocated to $U_1$. Therefore, the proposed algorithm will slightly outperform the SU case. For the SU in Fig. 2, the same justification is also applicable. However, the subcarriers of $U_1$ will be switched off more frequently, which gives the chance for $U_2$ to utilize the full power, i.e., $\beta_2 = 1$. Therefore, the throughput of $U_2$ using the proposed algorithm in this case will outperform SU. For the high-SNR region, the impact of the reduced power allocation becomes less significant on the BER, therefore, both SU solutions converge and become upper bounds for Algorithm 4. Moreover, it can be noted that increasing $\beta_1$ for $U_1$ increases the throughput for all solutions and reduces the gap with the corresponding SU solution. This is due to the fact that increasing $\beta_1$ reduces the impact of the IUI [26]. Therefore, by excessively increasing $\beta_1$, the throughput of $U_1$ becomes almost independent of the other user as shown in Fig. 1, particularly for $\beta_1 = 0.99$. On contrary, increasing $\beta_2$ does not necessarily guarantee enhanced throughput performance as can be noted for $\beta_2 = 0.05$ and 0.08 in Fig. 2. This is due to the error performance of the two stages of SIC and the associated error propagation. Therefore, it is important to properly allocate power coefficients to provide optimal BER performance at both stages to enhance the corresponding throughput.

Because the BER in NOMA depends on the IUI and $\beta_n$, four different solutions are provided for both users where the initial solution type and starting user are varied as, $[\text{FL}, 1]$, $[\text{FL}, 2]$, $[\text{PL}, 1]$, and $[\text{PL}, 2]$. In Fig. 1 for $\beta_1 = 0.92$, both $[\text{FL}, 2]$ and $[\text{PL}, 2]$ outperform the solutions that started $U_1$ in the range of $\bar{\gamma} = 0$ to 34 dB. In fact, starting the bit loading with $U_2$ reduces the IUI for $U_1$, which results in a throughput performance similar to the near-optimal SU solution with unity power. On the other hand, the, $[\text{FL}, 1]$ and $[\text{PL}, 1]$ are obtained where high IUI from $U_2$ exists, which reduces the chance to satisfy $\tilde{P}_{Th}^{(1)}$, and hence, decreases the throughput. In the high-SNR range, $\bar{\gamma} = 35$ to 50 dB, the throughput for all cases becomes roughly identical because the BER is reduced substantially and the BER threshold can be satisfied regardless of the initial solution or starting user. Similar observations can be made when increasing $\beta_1$. An opposite behavior can be noticed in Fig. 2 where higher throughput performance is achieved when the algorithm starts with $U_1$, which reduces the IUI for $U_2$. Moreover, both FL and PL solutions result in similar throughput.

Figures 3 and 4 present the total throughput of the proposed system for different power coefficients, where $\tilde{P}_{Th}^{(1)} = \tilde{P}_{Th}^{(2)} = 10^{-2}$ in Fig. 3, while $\tilde{P}_{Th}^{(1)} = 10^{-2}$ and $\tilde{P}_{Th}^{(2)} = 10^{-3}$ in Fig. 4. It can be noted that both the FL and PL initial solutions perform almost identically over the full SNR range. Due to the conflicting BER requirements of the individual users, increasing $\beta_1$ does not necessarily increase the total throughput over the full SNR range. It can be noted that using $\beta_1 = 0.95$ provides the best throughput. Therefore, the need for proper power allocation is crucial for OFDM-NOMA systems to maximize the system total throughput. It is also interesting that OFDM-NOMA significantly outperforms the OFDM-OMA in terms of the throughput for the entire range of the considered SNR values and all the considered scenarios. The difference between the two solutions increases as the SNR increases because both NOMA users will be able to have fully loaded subcarrier. As can be noted from both figures, NOMA with bit loading can offer a throughput improvement that may reach 100% for a wide range of SNRs and operating scenarios.
Fig. 1. The throughput for $U_1$, where $P_{T_h}^{(1)} = P_{T_h}^{(2)} = 10^{-2}$, and $\beta_1 \in \{0.92, 0.95, 0.99\}$.

Figures 5 and 6 depict the modulation distribution for the proposed system for $\tilde{\gamma} = 10$ and 30 dB, $\beta_1 = 0.95$, and $P_{T_h}^{(1)} = P_{T_h}^{(2)} = 10^{-2}$. Also, Figures 7 and 8 present the modulation distribution at $\tilde{\gamma} = 10$ and 30 dB, $\beta_1 = 0.95$, and $P_{T_h}^{(1)} = 10^{-2}$ and $P_{T_h}^{(2)} = 10^{-3}$. The modulation order when single user transmission is decided is denoted as $M_k^{(n)+}$. As stated earlier, a subcarrier $k$ can be turned off for both users especially at low SNR region which is referred to as "OFF". For low SNR values, single user transmission is more likely to occur as it can provide less BER as can be seen in Fig. 5. Moreover, because $U_2$ has much lower power than $U_1$, it is expected to have lower throughput and lower modulation orders. For the case where $\tilde{\gamma} = 30$ dB, the OFF mode percentage is almost zero because $C_2$ can be satisfied for both users. Also, at such high SNRs, the simultaneous transmission from both users is highly probable with superposed modulation orders, 16-QAM or 256-QAM. Due to the low power assigned to $U_2$, lower modulation orders are frequently assigned. The results show that as $P_{T_h}^{(n)/n}$ increases, the modulation orders and simultaneous transmission increases as well.

The computational efficiency is a critical requirement in several delay-sensitive applications, hence, we provide the average computational time required for several cases of interest as shown in Table VII. The time required to obtain the throughput results varies depending on the initial solution type and starting user. It can be noted that starting with a PL solution reduces the computational time as compared to starting with a FL solution. To quantify the reduction, the ratio percentage is calculated as $R_1 = \frac{T_{FL, 1}}{T_{PL, 1}}$, where $T_{PL}$ and $T_{FL}$ are the time needed to compute Algorithm 4 when starting with PL and FL initial solutions, respectively. The computational time decreases as $\tilde{\gamma}$ increases as it becomes easier to satisfy the BER constraints for both users. Moreover, as $P_{T_h}^{(n)/n}$ decreases, the computational time increases as $U_n$ requires more time to satisfy $C_2$. As can be concluded from the table, it is more desirable to start with $U_2$ and PL initial solution for all considered scenarios. The table also shows that the simulation time for the bit loading process is generally large as compared to single-user IAA [22]. Such performance is obtained because the time requirements to compute BER for the NOMA is time demanding. To reduce the overall time-requirements for the bit allocation problem, the exact NOMA BER formulas can be replaced by more efficient approximation or lookup tables. Moreover further complexity reduction techniques such as the subcarrier grouping [22] can be adopted.

Fig. 9 shows the individual and total throughput using Algorithm 4 based on the IAA and MIAA-A/D for $\beta_1 = 0.95$ and $P_{T_h}^{(1)} = P_{T_h}^{(2)} = 10^{-2}$. In other words, Algorithm 4 is implemented using two different approaches, the first is based on the proposed MIAA, and the other is using the original IAA. In both cases we use FL initial vectors for the loading process. As can be noted from the figure, using the IAA provides
higher total throughput when compared to the proposed MIAA-D as detailed in Sec. III-C. However, applying the second iteration to the IAA, which aims at increasing the bit allocation, does not necessarily enhance the throughput as in the case of the proposed MIAA-D. For \( U_1 \), the throughput using the proposed algorithm outperforms the other algorithm over low and moderate SNR ranges. However, for \( U_2 \), the result obtained using Algorithm 4 based IAA performs better than the other Algorithm over the range \( \gamma = [4, 22] \) dB. As for the total throughput results, the proposed Algorithm outperforms the IAA slightly over the SNR range except \( \gamma = [8, 18] \). It can be noted that the proposed algorithm using MIAA throughput is generally equivalent to that using the IAA, but with lower complexity.

Fig. 10 shows the impact of the imperfect CSI on the proposed bit loading scheme. The CSI imperfection corresponds to the SNR estimates where the noise power is assumed to be biased by an error factor of \( \sigma_E^2 \). As can be noted from the figure, the transmitter underestimates the noise effect, which causes throughout inflation in certain regions of the SNR axis. Therefore, the communications link will actually experience higher BER than anticipated which can cause QoS constraint violation. This problem is less severe in SNR regions where the throughput plateaus. On the contrary, if \( \sigma_E^2 \) is negative, the system will be under loaded, which may cause throughout degradation, but the QoS constraint will still be satisfied.

Fig. 2. The throughput for \( U_2 \), where \( \beta_1^{(2)} = 10^{-2}, \beta_2 \in \{0.08, 0.05, 0.01\} \).

| \( \gamma \) (dB) | 10 dB | 30 dB |
|-----------------|-------|-------|
| \( \beta_2 \)    | \( T_{FL} \) | \( T_{PL} \) | \( R_1 \) | \( T_{FL} \) | \( T_{PL} \) | \( R_1 \) |
| 10^{-2}, 1      | 441.4 | 255.4 | 58%   | 132.6 | 92.8   | 69%   |
| 10^{-3}, 1      | 452.4 | 276.3 | 61%   | 291.4 | 226.8 | 77%   |
| 10^{-2}, 2      | 350.4 | 179.9 | 51%   | 120.5 | 63.3  | 52%   |
| 10^{-3}, 2      | 351.7 | 201.2 | 57%   | 277.5 | 167.7 | 60%   |

The same approach can be followed to evaluate the system performance in the case that the bias is random, or when the channel gains have estimation errors.

VI. CONCLUSION AND FUTURE WORK

This work considered the adaptive bit loading problem for a two-user OFDM-NOMA, where the objective is to maximize the total throughput while satisfying the users’ individual BER constraints. Because the complexity of bit loading for OFDM-NOMA is significantly higher than OFDM-OMA, we proposed an efficient heuristic algorithm that converts the two-dimensional optimization process into two one-dimensional processes, and thus significantly reduce the complexity of the bit loading process. The proposed algorithm is based on the incremental allocation concept, however, several fundamental...
modifications are preformed to accommodate the difference between OMA and NOMA. For example, the need for frequent search for the maximum BER was eliminated by performing the allocation based on the channel gain instead of the BER. Moreover, different initial starting vectors and loading directions were applied to handle the multi-user interference. The obtained numerical results show that bit loading for NOMA can significantly improve the total throughput because the NOMA, in this case, has an inherent spectrum pooling feature that enables maximizing the available spectrum utilization. More specifically, it has been shown that the bit loading may provide spectral efficiency advantage of more than 300% for NOMA over OMA at low and moderate SNRs, and about 100% at high SNRs. The computational complexity of the proposed algorithm was evaluated for various operating conditions, and the obtained results showed that the proposed PL approach can reduce the complexity by about 50% for several cases of interest as compared to typically used FL solution However, the bit loading complexity for OFDM-NOMA is generally higher than OFDM-OMA due to the heavy computations of the BER.

Although this work considered the bit loading for NOMA, the number of users is limited to two. Consequently, extending the proposed algorithms to three or more users cases is crucial. Generally speaking, the proposed algorithm can be applied to any number of users by following the same approach used in this work. Nevertheless, the computational complexity may increase nonlinearly versus the number of users. More specifically, computing the BER for more than two users with high modulation orders requires extensive computational power [30]. Moreover, the power assignment is another challenge because using fixed power for all modulation schemes may degrade the performance for certain users. For example, for the three-user case with 16-QAM, the third user maximum power is about $6.5 \times 10^{-3}$ while if the modulation is switched to 4-QAM, then that user can be allocated up to 0.18 [29]. Therefore, extending this work to more than two users is an interesting research problem that we will consider in our future work.

It is worth noting adaptive power allocation for multicarrier systems with QoS constraints is another interesting problem that should be tackled. The highly nonlinear BER formulas for NOMA complicates the power computation efforts. Therefore, an efficient design is needed to solve this problem efficiently.

**APPENDIX I: LIST OF ACRONYMS**

- **5G** fifth generation.
- **AWGN** additive white Gaussian noise.
- **BER** bit error rate.
- **BS** base station.
- **CP** cyclic prefix.
Fig. 4. The total system throughput, where $P_{Th}^{(1)} = 10^{-2}$, $P_{Th}^{(2)} = 10^{-3}$, and $\beta_1 \in \{0.92, 0.95, 0.99\}$.

Fig. 5. Modulation distributions for $U_1$ and $U_2$, $\bar{\gamma} = 10$ dB, $\beta_1 = 0.95$ and $P_{Th}^{(1)} = P_{Th}^{(2)} = 10^{-2}$.

Fig. 6. Modulation distribution for $U_1$ and $U_2$, $\bar{\gamma} = 30$ dB, $\beta_1 = 0.95$, and $P_{Th}^{(1)} = P_{Th}^{(2)} = 10^{-2}$. 
Fig. 7. Modulation distribution for $U_1$ and $U_2$, $\bar{\gamma} = 10$ dB, $\beta_1 = 0.95$, $P_{Th}^{(1)} = 10^{-2}$ and $P_{Th}^{(2)} = 10^{-3}$.

Fig. 8. Modulation distribution for $U_1$ and $U_2$, $\bar{\gamma} = 30$ dB, $\beta_1 = 0.95$, $P_{Th}^{(1)} = 10^{-2}$ and $P_{Th}^{(2)} = 10^{-3}$.

Fig. 9. Throughput of Algorithm 4 using IAA and MIAA, where $\beta_1 = 0.95$ and $P_{Th}^{(1)} = P_{Th}^{(2)} = 10^{-2}$.

Fig. 10. Throughput of $U_1$, $U_2$ and total throughput with imperfect SNR estimation, $\beta_1 = 0.95$ and $P_{Th}^{(1)} = P_{Th}^{(2)} = 10^{-2}$.

| Acronym | Description |
|---------|-------------|
| CR      | cognitive radio |
| CSI     | channel state information |
| DFT     | discrete Fourier transform |
| FL      | fully loaded |
| I-MIAA  | iterative MIAA |
| IAA     | incremental allocation algorithm |
| IDFT    | inverse discrete Fourier transform |
| IoT     | Internet of things |
| ISI     | inter-symbol interference |
| IUI     | inter-user interference |
| JMuD    | joint multi-user detection |
| MIAA    | modified IAA |
| NOMA    | non-orthogonal multiple access |
| NP      | non-deterministic polynomial-time |
| OFDM    | orthogonal frequency-division multiplexing |
| OMA     | orthogonal multiple access |
| PL      | partially loaded |
| QAM     | quadrature amplitude modulation |
| QoS     | quality of service |
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2021.3125448, IEEE Transactions on Vehicular Technology

APENDIX II: LIST OF SYMBOLS

In the following list of symbols, the indices \( k \) and \( n \) correspond to the subcarrier and user index, respectively. Both indices will be dropped unless it is necessary to state one or both of them explicitly.

\( \mathcal{M} \): Number of possible modulation orders  
\( K \): Number of subcarriers  
\( \mathcal{N} \): Number of users  
\( U_n \): \( n \)th user  
\( y \): NOMA data symbol  
\( P_T \): Total transmit power  
\( \beta \): Power allocation coefficient  
\( x \): User data symbol  
\( \mathbb{E} [\cdot] \): Statistical expectation  
\( b_k \): Number of bits for the \( k \)th subcarrier  
\( M \): Modulation order  
\( G \): Channel frequency response  
\( g \): Multipath component gain  
\( Q + 1 \): Number of multipath components  
\( w \): AWGN  
\( \xi \): Path loss exponent  
\( d \): Distance to the BS  
\( \alpha \): Channel envelope  
\( \bar{P}_{TH} \): Average BER threshold  
\( \theta_{\text{max}} \): Maximum number of bits  
\( P_k \): Instantaneous error probability for subcarrier \( k \)  
\( \bar{P} \): Average BER per OFDM symbol  
\( \bar{\gamma} \): Average SNR  
\( T_{\text{T}} \): Normalized total throughput per subcarrier  
\( \Gamma \): Operation mode indicator  
\( v_k^{(n)} \triangleq \log_2 M_k^{(n)} \)  
\( v_k^{(n)} \triangleq \sqrt{ M_k^{(n)} } \)  
\( I_k^{(n)} \triangleq v_k^{(n)} - 1 \)  
\( \lambda_{1,k}^{(n)} \triangleq \left[ 1 + x_k^{(n)} (2)^{\frac{1}{2}} \right] + x_k^{(n)} \)  
\( S_k \triangleq (-1)^{1-x_k^{(n)}} \)  
\( L_{1,k}^{(n)} \triangleq (1 - 2^{-i})v_k^{(n)} - 1 \)  
\( g_k^{(n)} (a, b) \triangleq \alpha_k^{(n)} (a \sqrt{\gamma_1} \pm b \sqrt{\gamma_2}) \)  
\( D_{1,k}^{(n)} \triangleq (-1)^{1-x_k^{(n)}} \lambda_{2,k}^{(n)} \)  
\( D_{2,k}^{(n)} \triangleq 2^{1-x_k^{(n)}} \left[ \frac{1}{2} \right] \)  
\( D_{3,k}^{(n)} \triangleq 2^{x_k^{(n)}} - 1 \left[ l_2^{(1-x_k^{(n)})} \log_2 (v_k^{(n)} - 1) + 1 \right] \)  
\( c_i \triangleq 2 - \beta n - 3i,1 \)  

REFERENCES

[1] S. Mattisson, “An overview of 5G requirements and future wireless networks: Accommodating scaling technology,” IEEE Solid-State Circuits Mag., vol. 10, no. 3, pp. 54–60, May 2018.
[2] L. Bariah, S. Muhaïdât, and A. Al-Dweïk, “Error probability analysis of NOMA-based relay networks with SWIPT,” IEEE Commun. Lett., vol. 23, no. 7, pp. 1223–1226, Jul. 2019.
[3] E. Hossain and M. H. Hasan, “5G cellular: Key enabling technologies and research challenges,” IEEE Instrum. Meas. Mag., vol. 18, no. 3, pp. 11–21, Jun. 2015.
[4] F. Zhou, Y. Wu, R. Q. Hu, Y. Wang, and K. K. Wong, “Energy-efficient NOMA enabled heterogeneous cloud radio access networks,” IEEE Netw., vol. 32, no. 2, pp. 152–160, Mar. 2018.
[5] S. Rezvani, N. M. Yamchi, M. R. Javan, and E. A. Jorswieck, “Resource allocation in virtualized CoMP-NOMA HetNets: Multi-connectivity for joint transmission,” IEEE Trans. Commun., vol. 69, no. 6, pp. 4172–4185, Mar. 2021.
[6] Q. Nadeem, A. Kamoun, M. Debbah, and M. Alouini, “Design of 5G full dimension massive MIMO systems,” IEEE Trans. Commun., vol. 66, no. 2, pp. 726–740, Feb. 2018.
[7] W. Pramudito et al., “Load-aware energy efficient adaptive large scale antenna system,” IEEE Access, vol. 8, pp. 82 592–82 606, Apr. 2020.
[8] H. Mukhtar, A. Al-Dweïk, and A. Shami, “Turbo product codes: Applications, challenges, and future directions,” IEEE Commun. Surveys Tutts., vol. 18, no. 4, pp. 3052–3069, 4th Quart. 2016.
[9] A. Al-Ismawi, M. A. Imran, R. Tafazooli, and H. Chen, “Subcarrier and power allocation for LDS-OFDM system,” in Proc. IEEE Veh. Technol. Conf. (VTC), May 2011, pp. 1–5.
[10] Z. Liu and L.-L. Yang, “Sparse or dense: A comparative study of code-domain NOMA systems,” IEEE Trans. Wireless Commun., vol. 20, no. 8, pp. 4768–4780, Mar. 2021.
[11] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. Kwak, “Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges,” IEEE Commun. Surveys Tutts., vol. 19, no. 2, pp. 721–742, 2nd Quart. 2017.
[12] A. Tusha, S. Dogan, and H. Arslan, “A hybrid downlink NOMA with OFDM and OFDM-IM for beyond 5G wireless networks,” IEEE Signal Process. Lett., vol. 27, pp. 491–495, Mar. 2020.
[13] A. Adnan, Y. Liu, C. Chow, and C. Yeh, “Demonstration of non-Hermitian symmetry (NHS) IFFT/FFT size efficient OFDM non-orthogonal multiple access (NOMA) for visible light communication,” IEEE Photon. J., vol. 12, no. 3, pp. 1–5, Apr. 2020.
[14] A. M. Wyglinski, F. Labeau, and P. Kabal, “Bit loading with BER-constraint for multicarrier systems,” IEEE Trans. Wireless Commun., vol. 4, no. 4, pp. 1383–1387, Sep. 2005.
[15] F. Kalbat et al., “Direct bit loading with reduced complexity and overhead for precoded OFDM systems,” IEEE Trans. Veh. Technol., vol. 68, no. 7, pp. 7169–7173, July 2019.
[16] Q. Sun, L. Li, A. Toli, M. Juntti, and J. Mao, “Optimal energy efficient bit and power loading for multicarrier systems,” IEEE Commun. Lett., vol. 18, no. 7, pp. 1194–1197, Apr. 2014.
[17] D. Wang, Y. Cao, L. Zheng, and Z. Du, “Iterative group-by-group bit-loading algorithms for OFDM systems,” IEEE Trans. Veh. Technol., vol. 62, no. 8, pp. 4311–4315, Oct. 2013.
[18] Z. Song, Q. Ni, and X. Sun, “Spectrum and energy efficient resource allocation with QoS requirements for hybrid MC-NOMA 5G systems,” IEEE Access, vol. 6, pp. 37 055–37 069, July 2018.
[19] 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, Mar. 2020.
[20] S. Luo and K. C. Teh, “Adaptive transmission for cooperative NOMA system with buffer-aided relaying,” IEEE Commun. Lett., vol. 21, no. 4, pp. 937–940, Jan. 2017.
[21] D. J. Dechene and A. Shami, “Energy-aware resource allocation strategies for LTE uplink with synchronous HARQ constraints,” IEEE Trans. Mobile Comput., vol. 13, no. 2, pp. 422–433, Dec. 2014.
[22] Y. Iqaili and A. Al-Dweïk, “Adaptive bit loading with reduced computational time and complexity for multicarrier wireless communications,”
IEEE Trans. Aerosp. Electron. Syst., vol. 56, no. 3, pp. 2497–2506, June 2020.

[23] I. Lee and J. Kim, “Average symbol error rate analysis for non-orthogonal multiple access with M-ary QAM signals in Rayleigh fading channels,” IEEE Commun. Lett., vol. 23, no. 8, pp. 1328–1331, Aug. 2019.

[24] T. Assaf, A. Al-Dweik, M. S. E. Moursi, H. Zeineldin, and M. Al-Jarrah, “NOMA receiver design for delay-sensitive systems,” IEEE JST., pp. 1–12, Nov. 2020.

[25] T. Assaf, A. Al-Dweik, M. E. Moursi, and H. Zeineldin, “Exact BER performance analysis for downlink NOMA systems over Nakagami-m fading channels,” IEEE Access, vol. 7, pp. 134 539–134 555, Sep. 2019.

[26] T. Assaf, A. J. Al-Dweik, M. S. E. Moursi, H. Zeineldin, and M. Al-Jarrah, “Exact bit error-rate analysis of two-user NOMA using QAM with arbitrary modulation orders,” IEEE Commun. Lett., vol. 24, no. 12, pp. 2705–2709, Aug. 2020.

[27] Kyongkuk Cho and Dongweon Yoon, “On the general BER expression of one- and two-dimensional amplitude modulations,” IEEE Trans. Commun., vol. 50, no. 7, pp. 1074–1080, July 2002.

[28] M. Tariq, A. Al-Dweik, B. Mohammad, H. Saleh, and T. Stouraitis, “Computational power evaluation for energy-constrained wireless communications systems,” IEEE Open J. Commun. Soc., vol. 1, pp. 308–319, Mar. 2020.

[29] Y. Iraqi and A. Al-Dweik, “Power allocation for reliable SIC detection of rectangular QAM-based NOMA systems,” IEEE Trans. Veh. Technol., vol. 70, no. 8, pp. 8355–8360, July 2021.

[30] H. Yahya, E. Alsusa, and A. Al-Dweik, “Exact BER analysis of NOMA with arbitrary number of users and modulation orders,” IEEE Trans. Commun., vol. 69, no. 9, pp. 6330–6344, June 2021.

Hatem Zeineldin (M’06–SM’13) received the B.Sc. and M.Sc. degrees in electrical engineering from Cairo University, Giza, Egypt, in 1999 and 2002, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2006. He was with Smith and Andersen Electrical Engineering, Inc., North York, ON, USA, where he was involved in projects involving distribution system designs, protection, and distributed generation. He was a Visiting Professor with the Massachusetts Institute of Technology, Cambridge, MA, USA. He is currently a Professor with the Khalifa University of Science and Technology, Abu Dhabi, UAE, and is currently on leave from the Faculty of Engineering, Cairo University, Giza. His current research interests include distribution system protection, distributed generation, and micro grids. He is currently an Editor for the IEEE Transactions on Energy Conversion and the IEEE Transactions on Smart Grid.

Mohamed El Moursi (M’12, SM15) received the B.Sc. and M.Sc. degrees from Mansoura University, Mansoura, Egypt, in 1997 and 2002, respectively, and the Ph.D. degree from the University of New Brunswick (UNB), Fredericton, NB, Canada, in 2005, all in electrical engineering. He was a Research and Teaching Assistant in the Department of Electrical and Computer Engineering, UNB, from 2002 to 2005. He joined McGill University as a Post-doctoral Fellow with the Power Electronics Group.

Taseem Assaf (S’14) received the B.Sc. degree in communications engineering from Khalifa University (KU), Abu Dhabi, UAE, in 2014, and the master’s degree in electrical engineering from the American University of Sharjah (AUS), Sharjah, UAE, in 2016. She is currently pursuing the Ph.D. degree from KU. Her research interests include smart grids, optimization, wireless communications, and machine learning.

Arafat Al-Dweik (S’97–M’01–SM’04) received the M.S. degree (summa cum laude) and Ph.D. degree (magna cum laude) in electrical engineering from Cleveland State University, Cleveland, OH, USA. He served with Efficient Channel Coding, Inc., Cleveland, USA, the Department of Information Technology, Arab American University in Palestine and the University of Guelph, ON, Canada. Since 2003, he has been with the Department of Electrical Engineering, Khalifa University in UAE.

He is currently an Adjunct Research Professor with Western University and the University of Guelph in Canada, and University of Manchester in UK. Dr. Al-Dweik is a recipient of the Fulbright Scholarship, the Hijjawi Award for Applied Sciences, the Fulbright Alumni Development Grant, Dubai Award for Sustainable Transportation and the Leader-Founder Award in UAE. He serves as an Editor for the IEEE Transactions on Vehicular Technology and IET Communications. He is a registered Professional Engineer in the Province of Ontario, Canada.