Abstract—In this paper, a novel framework for optimizing the resource allocation in a millimeter-wave-non-orthogonal multiple access (mmWave-NOMA) communication for crowded venues is proposed. MmWave communications suffer from severe blockage caused by obstacles such as the human body, especially in a dense region. Thus, a detailed method for modeling the blockage events in the in-venue scenarios is introduced. Also, several mmWave access points are considered in different locations. The resource allocation problem in this network is formulated in the form of an optimization problem to maximize the network sum rate, which is NP-hard. Hence, a three-stage low-complex solution is proposed to solve the problem. At first, a user scheduling algorithm, i.e., modified worst connection swapping (MWCS), is proposed. Secondly, the antenna allocation problem is solved using the simulated annealing algorithm. Afterward, to maximize the network sum rate and guarantee the quality of service constraints, a non-convex power allocation optimization problem is solved by adopting the difference of convex programming approach. The simulation results show that, under the blockage effect, the proposed mmWave-NOMA scheme performs, on average, 12% better than the conventional mmWave-orthogonal multiple access (OMA) scheme. In addition, the proposed scheme considering blockage even outperforms the corresponding OMA system without blockage.

Index Terms— Millimeter-wave communications, non-orthogonal multiple access, resource allocation, human blockage, multi-AP

I. INTRODUCTION

Bandwidth shortage in the current wireless communications is one of the most critical barriers to meet the ever-increasing demand for high data rates in the 5G and beyond. Hence, a large amount of unoccupied bandwidth has recently been considered in the mmWave spectrum, i.e., the frequency range of 30 GHz to 300 GHz [1]. The regulatory agencies such as the federal communications commission (FCC) have authorized the use of various bands in the mmWave spectrum, including an unlicensed 60 GHz band with a 500 MHz bandwidth [2]. However, in the mmWave communication, there are many challenges including high path loss, sensitivity to blockage caused even by user’s bodies and objects, and lower path loss in line-of-sight (LoS) mode compared to non-LoS (NLoS) mode [1]. There are various techniques to alleviate these limitations such as dense access point (AP) deployment, i.e., small cells, applying beamforming, and the use of high-gain directional antennas [1].

Besides using mmWave in the next-generation wireless networks, massive connectivity and high spectrum efficiency are new requirements that conventional orthogonal multiple access techniques cannot satisfy them [3], [4]. In contrast to traditional OMA schemes, NOMA in the power domain can provide a high spectrum efficiency [3]–[5]. The main idea behind NOMA is sharing the same resources (e.g., frequency or time) by multiple users and separating them in the power domain using superposition coding (SC) at the transmitter and successive interference cancellation (SIC) at the receiver [3]. Given the above characteristics of the mmWave communication and NOMA, the use of the NOMA technique in the mmWave spectrum can bring many benefits [4], [5]. However, applying the NOMA scheme in a multiple-AP mmWave network requires solving new complex problems such as optimal AP placement, beam steering, user grouping, antenna, and power allocation.

A. Related works

Multiple AP deployment is one of the well-known effective methods to increase the availability of the LoS mmWave links and optimize network coverage [6], [7]. The work in [6] proposed a greedy algorithm to solve the AP deployment problem to maximize coverage. In [7], the authors have formulated a joint AP deployment and beam steering problem to minimize the number of required APs while the coverage constraints are met. However, the resource allocation problem has not been addressed in [6] and [7]. Recent works such as [8] and [9] have investigated the resource allocation problem and the user association for the multi-cell mmWave networks. Due to dense deployment in the mmWave cellular networks, the co-channel interference management is very challenging. In [8], the authors have introduced a load balancing user association scheme for mmWave cellular networks by defining an optimization problem to maximize the sum rate, and then they have proposed a heuristic algorithm to solve it. They have shown that the user association significantly affects network interference and users’ instantaneous rates. In [9], the authors have investigated a mmWave dense femtocell network. First, they have proposed a clustering scheme based on the LoS connectivity for co-channel interference management and then have applied the power and sub-channel allocation separately for users and femto APs in each cluster. It should be noted that none of the works in [7]–[9] apply the NOMA technique,
and all of them considered a simple probabilistic model for the availability of LoS mmWave links.

Recent studies have used the NOMA scheme in a single-cell mmWave communication \([10]–[15]\). When NOMA is applied to such a system, the issue of user grouping is raised. In this regard, in \([10]–[15]\), the resource allocation, and user grouping methods for the single-cell mmWave-NOMA communications have been studied. In \([10]\), the authors have used the random beamforming technique to steer beam toward users within each NOMA group. However, due to the narrow beams in mmWave communication, the random beamforming method cannot exploit all the potential of NOMA. In order to fix this problem, the authors in \([11]\) have proposed a beamwidth control method being appropriate for mmWave-NOMA communication and investigated its performance in terms of energy efficiency. The authors in \([12]\) have introduced the beam splitting technique for grouping the users located in different directions and then compared its performance with other existing methods in terms of sum rate. In \([13]\), the authors have firstly used a user grouping method based on channel correlation, and then they solved an analog beamforming problem with a boundary-compressed particle swarm algorithm to direct the beam to each user. In \([14]\), the authors have introduced a joint transmitter-receiver beamforming and power allocation design for a mmWave-NOMA communication based on pure analog beamforming. In \([15]\), the authors have adopted a clustering approach for NOMA implementation using hybrid beamforming. It should be noted that the blockage effect is not considered in \([10], [12]–[14], \) and \([15]\). In \([11]\), the blockage effect is modeled by a distance-dependent probabilistic model; however, the authors have not provided any solution to reduce this effect.

Although a multi-AP structure combined with the NOMA technique highly increases the performance of mmWave communication, the resource allocation problem is inherently complex in this network. This complexity increases more if the multi-AP mmWave-NOMA network is deployed to cover a dense indoor region. This is due to the fact that in the dense indoor scenarios, random blockage effects from a large number of users’ bodies should be considered in any user scheduling, antenna, and power allocation algorithms. Consequently, precise modeling of the user’s body blockage effect on the mmWave links and designing a resource allocation scheme under these random human blockages increase the fidelity of the model in the multi-AP mmWave-NOMA networks. In \([16]\), the authors have obtained a closed-form expression for the outage probability of multi-cell mmWave-NOMA networks. They have taken a random approach for grouping and assigning users to base stations (BSs) and have not addressed the power allocation. In \([17]\), the authors proposed an angle-domain NOMA scheme for the multi-cell mmWave networks, in which they have maximized the system sum rate by two different approaches: optimization of the precoders/decoders and cooperative NOMA. However, they did not take into account the blockage effect.

### B. Contributions

In this paper, we consider a multi-AP mmWave-NOMA scheme and propose a novel framework for resource allocation in the downlink communication for dense indoor venues. The proposed approach allows the network to adjust the transmission beams of the mmWave AP, user scheduling, and power allocation in the presence of the stochastic user’s body blockage over the mmWave links. In summary, our main contributions are as follows.

- We precisely model the effect of human blockage in three-dimensional (3D) dense indoor scenarios to evaluate the performance of multi-AP mmWave-NOMA networks. In these 3D dense indoor environments, a large number of users are seated close to one another. Our 3D blockage modeling is applicable to a variety of real dense indoor venues, such as a lecture hall, sports stadium, and theater.

- We formulate the joint user scheduling, antenna, and power allocation in mmWave-NOMA networks as an optimization problem to maximize the users’ sum rate under stochastic blockage effects and quality of service (QoS) constraints. Then, we propose a three-stage solution with low complexity to solve this optimization problem. In the first stage, the user scheduling is formulated as an optimization problem. Considering the channel correlation and channel gain difference, we develop a low complexity algorithmic solution, i.e., a modified worst connection swapping technique to assign users to mmWave APs and group them for NOMA implementation (i.e., user scheduling design). In the second stage, we apply a meta-heuristic algorithm, named simulated annealing (SA), to split the beams among users’ NOMA group. In the third stage, given the obtained user scheduling and antenna allocation strategy in the previous stages, the power allocation problem is formulated as a non-convex optimization problem to maximize the system sum rate and guarantee the QoS constraints.

- We transform the formulated problem as a canonical form and solve the new canonical form power allocation problem based on the difference of convex (DC) programming approach. Simulation results show that the spectral efficiency of the proposed mmWave-NOMA system under the blockage effect is, on average, 12% and 4% higher than the corresponding OMA system with and without blockage, respectively.

The rest of the paper is organized as follows. In Section \([11]\) the system model is presented. The resource allocation problem for mmWave-NOMA communication is formulated in Section \([11]\). Section \([11]\) offers the three-stage low-complex solution for the formulated resource allocation problem. In Section \([11]\) simulation results demonstrate the performance of the proposed solution, and a conclusion is presented in Section \([11]\)
NOMA groups in each mm-AP. We define a binary decision by one RF chain. Thus, the users grouping is done to form practically be obtained for a given venue [19].

\[ \Pr(\varphi) \]

\[ \text{variable, denoted by } k \]

\[ \text{in the orientation of the users within the horizon plane, the} \]

\[ \text{where each RF chain} \]

\[ \text{MDs use the microwave links to exchange control signals} \]

\[ \text{Since mmWave links are susceptible to the blockage and} \]

\[ \text{as a central control unit that sends the required control signals} \]

\[ \text{each RF chain in different directions [12]. Thus, to group the} \]

\[ \text{MDs} \]

\[ \text{mm-APs and MDs. Due to the power limitation, each MD} \]

\[ \text{has only one radio frequency (RF) chain, which is connected} \]

\[ \text{MMs} \]

\[ \text{antenna (ULA) array with} \]

\[ \text{Mm-AP and MMD antenna elements, respectively. A fully connected} \]

\[ \text{hybrid structure is used in the} \]

\[ \text{mm-APs and MDs. Due to the power limitation, each MD} \]

\[ \text{has only one radio frequency (RF) chain, which is connected} \]

\[ \text{to all of its MMD antenna elements via MMD phase shifters} \]

\[ \text{(PSs). Each mm-AP is equipped with a set} \]

\[ \text{of NRF chains} \]

\[ \text{where each RF chain} \]

\[ \text{is connected to all Mm-AP antenna elements through Mm-AP} \]

\[ \text{PSs. In this network, the} \]

\[ \text{μ-AP acts as a central control unit that sends the required control signals} \]

\[ \text{Since mmWave links are susceptible to the blockage and} \]

\[ \text{therefore are not reliable for exchanging the control signals,} \]

\[ \text{MDs use the microwave links to exchange control signals} \]

\[ \text{between themselves and μ-AP. In our model, mm-APs connect} \]

\[ \text{to the central controller for sending channel state information} \]

\[ \text{(CSI) and receiving control signals via the backhaul link. To} \]

\[ \text{reduce the probability of MDs being exposed to LoS link} \]

\[ \text{blockage, we deploy several mm-APs in different locations.} \]

\[ \text{In our model, we focus on in-venue scenarios in which a} \]

\[ \text{high number of active users sit on a fixed and densely-} \]

\[ \text{deployment seating chart such as in a sports stadium, a} \]

\[ \text{lecture hall, a concert venue, or a theater. Thus, based on the} \]

\[ \text{arrangement and direction of the seats inside a given region,} \]

\[ \text{we assume that K users sit on the seats and the location of} \]

\[ \text{each user} \]

\[ \text{in} \]

\[ \text{is given by} \]

\[ \text{Cartesian coordinates. Fig. 1 shows an illustrative example of our proposed model with 277 seats for the users. Due to the random changes} \]

\[ \text{in the orientation of the users within the horizon plane,} \]

\[ \text{the azimuthal angle of a given user} \]

\[ \text{is assumed to be a random variable, denoted by} \]

\[ \text{with a given probability distribution function} \]

\[ \text{where} \]

\[ \text{that can practically be obtained for a given venue [19].} \]

\[ \text{In our system, all users within a NOMA group are served} \]

\[ \text{by one RF chain. Thus, the users grouping is done to form} \]

\[ \text{NOMA groups in each mm-AP. We define a binary decision} \]

\[ \text{variable} \]

\[ \text{where} \]

\[ \text{if user} \]

\[ \text{is assigned to the} \]

\[ \text{mm-AP} \]

\[ \text{and RF chain} \]

\[ \text{due to the walls and nearby seats in an in-venue region is static. Due to the use of ULA (located in the horizontal plane) in the} \]
transceiver and since ULA propagation pattern only changes on the horizontal plane and is constant on the elevation plane [20], we consider 2D model here. In the following, we model the human body with a cylinder [21].

1) Self-body blockage: To determine the self-body blockage for user $k$ relative to mm-AP $b$, we define $A_{kb,1} \triangleq \{ \theta_E, 2\pi \} \cup [0, \theta_E]$. Consequently, $A_{kb,1} = [\theta_E, \theta_D]$.

2) Nearby-user blockage: We define set $A_{kb,2}$ to describe a set of angles corresponding to the red part of the user circle in Fig. 1. Then among the users within this circle, we only consider those who are taller than or equal to $H_{\text{person}}$ and their distance to mm-AP $b$ is less than $d$. 

Step 1: as it can be seen in Fig. 2, the effective distance $\bar{d}$ between user $k$ and a person with the height $H_{\text{person}}$ around it, is given by $\bar{d} = \frac{H_{\text{person}} - H_{\text{MD}}}{d}$, where $H_{\text{AP}}$ and $H_{\text{MD}}$ denote the heights of mm-AP and MD, respectively; note that if users' seats are on a platform above the ground level, the height of the platform must also be added to $H_{\text{person}}$ and $H_{\text{MD}}$.

Step 2: we draw a circle with center $(X_k, Y_k)$ and radius $\bar{d}$; then among the users within this circle, we only consider those who are taller than or equal to $H_{\text{person}}$ and their distance to mm-AP $b$ is less than $d$.

Step 3: for users holding step 2 conditions, we draw the tangent lines to each of their circles and consider only those lines that pass through the user circle.

Step 4: for the lines considered in step 3, we obtain the coordinates of the intersection of these lines with the user circle. We then calculate the angles corresponding to these points, such as $\theta_E$ and $\theta_D$ shown in Fig. 2.

Step 5: users corresponding to the lines considered in step 3 can create a shadow for the user. Consequently, complement of $A_{kb,2}$ is equal to the set of azimuthal angles that characterize the shaded areas. For example, for Fig. 2 $A_{kb,2} = [\theta_E, 2\pi] \cup [0, \theta_E]$.

After repeating the above steps and defining set $A_{kb} \triangleq A_{kb,1} \cap A_{kb,2}$, we can comment on the presence or absence LoS link between user $k$ and mm-AP $b$. For instance, the angles corresponding to the red part of the user circle in Fig. 2 are equivalent to $A_{kb}$ i.e., LOS angles.

Considering the blockage effect, we define a binary variable $e_{kb}$ where $e_{kb} = 0$ if the LoS link between user $k$ and mm-AP $b$ is blocked, otherwise $e_{kb} = 1$.

Fig. 2: Self-body blockage modeling.

![Fig. 2: Self-body blockage modeling.](image)

The channel matrix between the mm-AP $b$ and user $k$ can be represented as [22]

$$H_{kb} = e_{kb} \sqrt{\rho_{kb}} \alpha_{kb} \mathbf{a}_b \mathbf{H}_b \mathbf{a}_M^H(\theta_{kb})$$

LoS component

$$+ \sum_{l=1}^{L} \sqrt{\rho_{kb} c_{kb}} \mathbf{a}_M(\phi_{kb}) \mathbf{H}_b \mathbf{a}_M^H(\theta_{kb}),$$

Scattering components

where $\rho_{kb}$ and $\alpha_{kb} \sim \mathcal{C}(0,1)$ denote the average path loss and complex gain of the $l$-th path between user $k$ and mm-AP $b$, respectively. Here, $l = 0$ represents the LoS path and $l \in \{1, \ldots, L\}$ represents the $l$-th NLoS path. $\rho_{kb} = (\frac{c}{\pi f_c})^2 d_{kb}^{-\rho}$ in which $c = 3 \times 10^8$ m/s, $f_c$ represents the carrier frequency, $d_{kb}$ shows the distance of between user $k$ and mm-AP $b$, and $\gamma_{kb}$ denotes the path loss exponent of $l$-th path. $\mathbf{a}_M(\phi_{kb}) = [1, e^{j\phi_{kb}}, \ldots, e^{j(M_{\text{MD}}-1)\phi_{kb}}]^T \in \mathbb{C}^{M_{\text{MD}} \times 1}$ denotes the array response vector for the $l$-th path with AOD $\phi_{kb}$. $\mathbf{a}_M(\phi_{kb}) = [1, e^{j\phi_{kb}}, \ldots, e^{j(M_{\text{MD}}-1)\phi_{kb}}]^T \in \mathbb{C}^{M_{\text{MD}} \times 1}$ shows the array response vector for the $l$-th path with angle of arrival (AOA) $\phi_{kb}$ from mm-AP $b$ at user $k$.

We call the channel between mm-AP and MD RF chains as the effective channel. Consequently, the effective channel for the user $k$ allocated to $n$-th RF chain of mm-AP $b$ is defined as follows

$$\overline{\mathbf{h}}_{kb} = \mathbf{v}_k^H \mathbf{H}_{kb} \mathbf{w}_{bn},$$

where $\mathbf{v}_k \in \mathbb{C}^{M_{\text{MD}} \times 1}$ is the analog beamforming vector at user $k$ where $\mathbf{v}_k = 0$ if $e_{kb}=0$, otherwise $\mathbf{v}_k = \frac{1}{\sqrt{M_{\text{MD}}}} [1, \ldots, e^{j(M_{\text{MD}}-1)\phi_{kb}}]^T$. Likewise, the effective channel vector of user $k$ is defined as $\mathbf{\tilde{h}}_{kb} = [\mathbf{h}_{kb}(1), \ldots, \mathbf{h}_{kb}(N)]^T \in \mathbb{C}^{N \times 1}$ and given the set $K_b$ as the collection of $K_b$ users served by mm-AP $b$, the effective channel matrix for this mm-AP is defined as $\mathbf{H}_b = [\mathbf{h}_{kb}(1), \ldots, \mathbf{h}_{kb}(N)] \in \mathbb{C}^{N \times K_b}$. In our model, we modify the three-step low-complexity algorithm presented in [23] for channel estimation. In step 1, MDs transmit unique frequency tones

Fig. 3: Nearby-user blockage modeling.

![Fig. 3: Nearby-user blockage modeling.](image)
via one of the omnidirectional antennas in the antenna array to the mm-APs. Then, the mm-APs send estimated $\tilde{\theta}_{k,0}$ and $\tilde{\alpha}_{k,0}$ (LoS CSI) via backhaul link to the central controller. In step 2, the mm-APs send unique frequency tones to all the users using analog beamforming adjusted with $\tilde{\theta}_{k,0}$. Then, MDs transmit the estimated $\tilde{\alpha}_{k,0}$ over the microwave link to the central controller. In step 3, the effective channel is estimated after user scheduling and antenna allocation in the central controller and sent to the users and mm-APs. In addition, digital beamforming and power allocation are done by exploiting the effective channel. Following time division duplex and channel reciprocity, in our model, the estimated effective channel in uplink can be used for designing the digital precoder in the downlink.

**E. SIC decoding order**

Without loss of generality, we assume that the users are indexed in the descending order of their LoS gains, i.e.,

$$|\alpha_{1,0}|^2 \geq |\alpha_{2,0}|^2 \geq \ldots \geq |\alpha_{K,0}|^2,$$

where $b_k \in B$ denotes the mm-AP index assigned to user $k$. Note that if the LoS path is blocked, we will replace the NLoS noise at user $\nu_k$ indexed in the descending order of their LoS gains, i.e., $\nu_k \in \{b_1, \ldots, b_K\}$, to have a tractable resource allocation scheme, and to decrease the system overhead.

In our model the received signal vector at the users, $y = [y_1, \ldots, y_K] \in \mathbb{C}^{K \times 1}$, is given by

$$y = \tilde{H}^H G x + \nu,$$

where $\tilde{H} = \begin{bmatrix} \tilde{h}_{k,bn} \end{bmatrix} \in \mathbb{C}^{K \times BN}$, so that the vectors $\tilde{h}_{k,bn}$ form the $k$-th row of this matrix, $G \in \mathbb{C}^{BN \times BN}$ is a block diagonal matrix such that the matrices $G_1$ to $G_B$ are its main-diagonal blocks, $x = [x_1^T, \ldots, x_B^T]^T \in \mathbb{C}^{BN \times 1}$ and $\nu = [\nu_1, \ldots, \nu_K]^T \in \mathbb{C}^{K \times 1}$ is the noise vector. After some simplification, the received signal at user $k$, $y_k$, is given by

$$y_k = \begin{aligned}
  \tilde{h}_{k,bn}^H g_{bn} \sqrt{p_{bn} x_j} + &\sum_{j \in K, j \neq k} c_{jbn} \sqrt{p_{jbn} x_j} \quad \text{Desired signal} \\
  + \tilde{h}_{k,bn}^H g_{bn}^* \sum_{j \in K} c_{jbn}^* \sqrt{p_{jbn}^* x_j} \quad \text{Intra-group interference} \\
  + \sum_{b' \in B, b' \neq b} \tilde{h}_{k,b'}^H g_{b'} x_j + &\nu_k \quad \text{Inter-AP interference}
\end{aligned}$$

Per traditional downlink NOMA protocol, the individual rate for user $k$ received from $n$-th RF chain of mm-AP $b$ is given by

$$R_{k \rightarrow k}^{bn} = \log_2 \left( 1 + \frac{c_{kb,n} \tilde{h}_{k,bn}^H h_{k,bn}^*}{I_{k \rightarrow k}^{bn} + \sigma^2} \right),$$

where

$$I_{k \rightarrow k}^{bn} = \| \tilde{h}_{k,bn}^H g_{bn} \|^2 \sum_{j=1}^{k-1} c_{jbn} p_{jbn},$$

$$I_{k}^{bn} = \sum_{n' \in N, n' \neq n} \| \tilde{h}_{kin}^H g_{bn} \|^2 \sum_{j \in K} c_{jbn}^* p_{jbn},$$

and

$$I_{k}^{bn} = \sum_{b' \in B, b' \neq b} \sum_{n \in N} \| \tilde{h}_{k,bn}^H g_{b'n} \|^2 \sum_{j \in K} c_{jbn}^* p_{jbn},$$

where $\nu_k \sim \mathcal{C}\mathcal{N}(0, \sigma^2)$ denotes additive white Gaussian noise at user $k$ with the power of $\sigma^2$. In (6), the first term represents the desired signal of user $k$, the second term denotes intra-group interference caused by the other users within the NOMA group associated with $n$-th RF chain of mm-AP $b$, the third term called inter-group interference originates from all the other RF chains of mm-AP $b$, and the fourth term expresses the inter-AP interference induced by other mm-APs.
0 \leq p_{kbn}, \forall k \in \mathcal{K}, \forall b \in \mathcal{B}, \forall n \in \mathcal{N}, \quad (13f)

\sum_{k=1}^{K} \sum_{b=1}^{B} \sum_{n=1}^{N} c_{kbn} p_{kbn} \leq p_{\text{total}}, \quad (13g)

\begin{align*}
    c_{kbn} R_{\text{max}}^{bn} & \leq c_{kbn} R_{k-n}^{bn}, \forall k < i, \forall b \in \mathcal{B}, \forall n \in \mathcal{N}, \\
    c_{kbn} R_{\min} & \leq \sum_{b=1}^{B} \sum_{n=1}^{N} R_{k-b}^{bn}, \forall k \in \mathcal{K}, \\
    c_{kbn} & \in \{0, 1\}, \forall k \in \mathcal{K}, \forall b \in \mathcal{B}, \forall n \in \mathcal{N}, \quad (13i)
\end{align*}

\sum_{k=1}^{K} M_{kbn} \in \mathbb{N}. \quad (13k)

Constraint (13b) shows that the maximum capacity of each NOMA group is two users. Constraint (13c) indicates that each user can be assigned to at most one mm-AP and one RF chain. Constraint (13d) guarantees that the number of all allocated antenna elements on n-th RF chain of mm-AP b cannot be larger than \( M_{\text{AP}} \). Constraint (13e) is a lower bound on the number of antennas allocated to each user. Also, constraint (13f) is the total transmission power constraint in all of the mm-APs. Constraint (13h) guarantees successful SIC and constraint (13i) is the QoS constraint with the predefined threshold \( R_{\text{min}} \) as the minimum rate requirement of each user. Due to the existence of continuous and integer variables as well as nonlinear functions in the objective function and constraints of the problem (13), this problem is mixed-integer nonlinear programming (MINLP) and in general, NP-hard [24]. Consequently, directly solving problem (13) is challenging and intractable. To achieve an efficient solution with acceptable computational complexity, we propose a three-stage low-complex method, which is detailed in section IV.

IV. RESOURCE ALLOCATION

We decompose problem (13) into three sub-problems and propose a three-stage sub-optimal solution with low complexity to solve it. In the first stage, we set the user scheduling variable, \( c_{kbn} \), to maximize the system sum rate, \( R_{\text{sum,1}} \), based on LoS CSI (channel without scattering components), users’ channel correlation, users’ channel gain difference, and MWCS algorithm. At this stage, the power allocation and antenna allocation are considered to be uniform, for example if \( c_{kbn} = c_{kbn} = p_{kbn} = p_{\text{total}}/K \) also if \( c_{kbn} = c_{kbn} = 1 \), then \( M_{\text{AP}} = M_{kbn} \). We also assume that the digital precoder is unavailable, i.e., \( G_b = I_N, \forall b \in \mathcal{B} \). In the second stage, and for NOMA groups with more than one member, the antenna allocation is performed using a meta-heuristic algorithm called simulated annealing to maximize the system sum rate in this stage, \( R_{\text{sum,2}} \). After the first and the second stages, the effective channel is estimated to be used in the digital precoding in the next stage. In the third stage, we use ZF digital precoder to mitigate the effect of inter-group interference. We then determine the power allocation policy to maximize the system sum rate, \( R_{\text{sum,3}} \), while QoS constraints are met. For this purpose, we rewrite the optimization problem in (13) by taking into account the obtained user scheduling and antenna allocation strategy known as \( c_{kbn} \) and \( M_{kbn} \). This new optimization problem is still non-convex and difficult to be solved. The method used to solve this problem will be described in section IV.C.

A. First stage: user scheduling

1) Problem formulation: Assuming uniform power and equal antenna allocation as well as \( G_b = I, \forall b \in \mathcal{B} \), we define the following sub-problem to obtain the user scheduling strategy by

\begin{align*}
    \max_{k \in \mathcal{K}, b \in \mathcal{B}, n \in \mathcal{N}} R_{\text{sum,1}} \quad (14a) \\
    \text{s.t.} \quad \begin{cases}
        c_{kbn}, \\
        (13b), (13c), (13i).
    \end{cases} \quad (14b)
\end{align*}

The exhaustive search method can find the optimal solution of the problem (14), which for a fully loaded scenario, i.e., \( K = 2BN \), \( K_b = 2N, \forall b \in \mathcal{B} \), the total number of states to be searched is \( (K!/(2N)!^B) \times (2N)!/2N!N!^B \), where the first and second terms are the total number of user assignment and grouping states, respectively. As a result, this method is inefficient due to its high computational complexity. In this regard, inspired by the idea presented in [8], we propose a heuristic algorithm called MWCS algorithm to assign users to mm-APs, and then we group the users according to the difference and correlation between their channels which is the estimated channels that contain only the LoS path (LoS CSI).

2) The proposed user assignment algorithm: The MWCS algorithm is based on the fact that the suboptimality of a user assignment strategy may be due to the imposition of a weak link to an MD or its suffering from the high inter-AP interference. Consequently, swapping the worst connection will probably give a stronger link to the MD or reduce inter-AP interference that can lead to a better individual data rate.

To describe how the MWCS algorithm works, we define \( \Pi = \{\mathcal{K}_1, \ldots, \mathcal{K}_B\} \) as a partition of \( \mathcal{K} \), such that \( \forall b \neq b', \mathcal{K}_b \cap \mathcal{K}_{b'} = \emptyset \), and \( \bigcup_{b=1}^{B} \mathcal{K}_b = \mathcal{K} \), where \( \mathcal{K}_b \) denotes the set of the users assigned to mm-AP \( b \). In fact, \( \Pi \) represents the strategy of assigning users to mm-APs. The MWCS algorithm starts with an initial \( \Pi \), then users within each mm-AP are grouped by algorithm 2. Afterward, we can obtain the user scheduling strategy including user assignment and user grouping, i.e., \( \Pi \), where \( e \in \{0, 1\}^{KBN \times 1} \) denotes the collection of \( c_{kbn}, \forall k \in \mathcal{K}, \forall b \in \mathcal{B}, \forall n \in \mathcal{N} \). Then, the individual data rate of all users can be calculated by (7).

Definition 1. The corresponding connection with the lowest individual data rate is defined as the worst connection, and the user associated with this connection is known as the worst user.

Definition 2. The swap operation for user \( k \) in mm-AP \( b \) and user \( j \) in mm-AP \( b' \) such that \( b \neq b' \) is equivalent to

\[ K_b = (K_b \setminus \{k\}) \cup \{j\}, \text{ and } K_{b'} = (K_{b'} \setminus \{j\}) \cup \{k\}. \quad (15) \]
If $|K_b| < 2N$, let us consider a hole (empty capacity) instead of user $j$, then the swap operation is equivalent to

$$K_b = K_b \setminus \{k\}, \quad \text{and} \quad K_b' = K_b' \cup \{k\}. \quad (16)$$

The algorithm consists of two main steps.

First step: at $t$-th iteration to find a better user scheduling strategy, the worst user is swapped with the users and holes within other mm-APs, where for every swap such as $i$, a new partition such as $\Pi_t$ is obtained. Afterward, the users within the new mm-APs involved, $\mathbf{B}$, must be re-grouped by Algorithm 2 which results in a new user scheduling strategy, $c_t$. Among the obtained new strategies, $c_t, \forall t$, the strategy that achieves the highest sum rate is selected. Subsequently, if the selected strategy, $c_t$, leads to a higher sum rate than the previous strategy, $c_{t-1}$, the selected strategy $c_t$ will be replaced. This process repeats until no improvement in the sum rate is achieved. After this, the second step for further improvements is executed.

Second step: the first step deadlock can be overcome by removing the worst connection, $K_t \setminus \{k_{\text{worst}}\}$, and redefining it from the remaining connections and returning to the first step.

The algorithm terminates when no connection is left in the second step to redefine the worst connection. Finding a good initial partition has a great impact on the convergence speed and optimality of the heuristic algorithms. Therefore, instead of randomly generating the initial partition, among the mm-APs, we assign each user to the one that has at least one vacancy and can provide the strongest LoS link. In order to identify the strongest LoS link, we define $H_{kb,0} = a_{\text{MD}}(\phi_{kb})a^{t}_{\text{AP}}(\theta_{kb})$ as the LoS channel matrix between MD $k$ and mm-AP $b$, and $H_{kb,0} = \mathbf{v}_{\text{AP}}^{H}H_{kb,0} \in \mathbb{C}^{1 \times M_{\text{AP}}}$ as the LoS channel vector between mm-AP antennas $b$ and the RF chain of MD $k$. Accordingly, the mm-AP that can provide the strongest LoS link for MD $k$ as $b = \arg \max_{b \in \mathbf{B}} |H_{kb,0}|^2$.

3) The user grouping algorithm: According to the NOMA technique, to achieve better performance, two users with a significant difference between their channel gains (the near and far users) must be in the same group. On the other hand, given the spacial directivity of the mmWave channel, the users whose channels are highly correlated must be in the same group to exploit multiplexing gain [13]. In other words, the users whose channels are uncorrelated should be assigned to different groups to reduce the interference. Consequently, to select a user pair to form a group, we select the one with the highest channel correlation and channel gain difference. Given the LoS channels of users $i$ and $j$, $H_{kb,0}$ and $H_{jb,0}$. The channel gain difference, Diff$(i, j)$, and channel correlation, Corr$(i, j)$, between them are calculated as in [26].

Thus, we can formulate a multi-objective optimization problem with $[\text{Diff}(i, j), \text{Corr}(i, j)]$ as the vector of objective functions to select the desired user pair. The scalarization methods can be used to solve such optimization problems, one of which is the weighted sum method [27]. This method replaces the vector of objective functions with $w_1 \text{Corr}(i, j) + w_2 \text{Diff}(i, j)$, where $\text{Corr}(i, j)$ and $\text{Diff}(i, j)$ are the normalized channel correlation and channel gain difference by the min-max normalization method [28]. Consequently, the scalarized multi-objective optimization problem is as follows.

$$\max_{(i,j)} \quad w_1 \text{Corr}(i, j) + w_2 \text{Diff}(i, j) \quad (17a)$$
$$\text{s.t.} \quad i < j, \forall i \in \mathcal{K}, \forall j \in \mathcal{K}, \quad (17b)$$

where $\mathcal{K}$ denotes the collection of non-grouped users and $w_1, w_2 \in (0, 1)$ are the weights of $\text{Corr}(i, j)$ and $\text{Diff}(i, j)$, respectively, that must confirm $w_1 + w_2 = 1$.

After specifying $\Pi = \{K_1, \ldots, K_{|\mathcal{B}|}\}$, the mm-APs whose users have been changed, i.e., $\mathbf{B}$, need to be re-grouped. For any member of $\mathbf{B}$ such as mm-AP $b$, if its number of users exceeds its number of RF chains, namely $|K_b| > N$, we must form $|K_b| - N$ two-user groups. Each time problem (17) is solved, a group is formed, and members of that group are exited from the set $\mathcal{K}$, so if we repeat this process $|K_b|$ times for mm-AP $b$ the grouping of users in this mm-AP ends. It should be noted that we use the exhaustive search method to solve problem (17), which requires investigating all $\binom{|\mathcal{K}|-1}{2}$ possible states. The method of grouping users is summarized in Algorithm 2.

4) Convergence: The system sum rate is bounded due to the limited resources such as power. On the other hand, in Algorithm 1 we only accept changes that lead to a strict increase in the system sum rate. Thus, due to the limitation of the system sum rate, Algorithm 1 converges after a few iterations.
Algorithm 2: User grouping Algorithm

Initialization: Initialize the iteration index \( t = 1 \). Specify the strategy of assigning users to mm-APs \( (\Pi_1 \Pi_2) \) and set \( B \) according to Algorithm 1. Sort the indexes of the users based on \( \Pi_2 \).

for \( \forall b \in B \) do

if \( |K_b| > N \) then

   Initialize the collection of non-grouped users \( K = K_b \).

   for \( n = 1 \) to \( |K_b| - N \) do

      Solve problem (2) and obtain \( \{i^*, j^*\} \). Put users \( i^* \) and \( j^* \) in a group, \( c_{bn} = c_{b_{\text{new}}} = 1 \). Update the collection of non-grouped users, \( K = K \setminus \{i^*, j^*\} \).

   end

else

   Form \( |K_b| \) single-user groups.

end

end

Output: \( c \)

5) Complexity analysis: Given the definition of floating-point operations (flops) in [29] and ignoring all terms except the dominant term and only considering non-zero entries of vectors and matrices [30], we evaluate the complexity of the proposed algorithm by counting the number of flops. In the \( t \)-th iteration of Algorithm 1 in a fully loaded scenario, i.e., \( K = 2BN \) \( (|K_b| = 2N, \forall b \in B) \), there are \( |K_b|(B - 1) \) swaps for the worst user, in addition, with each swap, the users within the two mm-APs involved must be re-grouped by Algorithm 2. By ignoring the ineffective terms versus the dominant term and given the complexity of sorting operation \( (\arg\max \text{ or } \arg\min) \) [31], the complexity of Algorithm 2 is \( O(8NMA_\text{P}(M_{\text{MD}} + N)) \). Also, the complexity of computing the system sum rate is \( O(2KN^2A_\text{P}M_{\text{MD}}) \). Based on these complexities, among the different parts of Algorithm 1 line 12 has the highest complexity so that the complexity of the other parts can be ignored. Accordingly, the computational complexity of Algorithm 1 is \( O(4KB^2N^3A_\text{P}M_{\text{MD}}) \) per iteration. As a result, given the total number of iterations in the worst case, the computational complexity of the MWCS Algorithm at the worst case and for fully loaded scenario is given by \( O \left( \left( K'/(2N)^B \right)(4KB^2N^3A_\text{P}M_{\text{MD}}) \right) \) which is much lower than the complexity of the exhaustive search method, i.e., \( O\left( \left( (2N)!/2^N N! \right)^B \left( K'/(2N)^B \right)(2KN^2A_\text{P}M_{\text{MD}}) \right) \). In particular, for \( B = 3, N = 3 \) and \( K = 2BN \), the MWCS Algorithm, at worst case, can reduce the number of flops by 99.4% compared to the exhaustive search method.

B. Second stage: antenna allocation

1) Problem formulation: After obtaining the user scheduling strategy, \( c^* \), we formulate the antenna allocation problem with the assumption of uniform power allocation between users and without using the digital precoder as follows.

\[
\max_m \quad R_{\text{sum},2} \tag{18a}
\]

s.t. \( m_q \in \{M_{\text{min}}, \ldots, A_\text{P} - M_{\text{min}}\}, \forall q \in Q \), \( \tag{18b} \)

where \( Q \) is a set of \( Q \) two-user NOMA groups across the system. We also define vector \( m \in \mathbb{N}^{1 \times Q} \) for simplicity, where each element of vector \( m \), \( m_q \), determines the number of antennas assigned to both users of the NOMA group associated with this element, i.e., \( m_q \) and \( A_\text{P} - m_q \).

The exhaustive search can be used to find the optimal solution of problem (18), which at the fully loaded scenario, \( Q = BN \), has a computational complexity equal to \( O \left( (M_{\text{AP}} - 2M_{\text{min}})^{BN}(2KB^2N^2A_\text{P}M_{\text{MD}}) \right) \). However, due to its high computational complexity, it imposes a high overhead on the system. To address this problem, we use the SA algorithm.

2) The antenna allocation algorithm: The mechanism presented in the SA algorithm is very suitable for avoiding the local maximum. The proposed SA algorithm for solving the problem (18) is summarized in Algorithm 3. To create the neighbor solution in Algorithm 3, we randomly choose one of the numbers 1 to \( Q - 1 \) as the number of vector elements \( m \) that must be changed, such as \( q' \in \{1, \ldots, Q - 1\} \). Then, among the elements of \( m \), we randomly select \( q' \) elements and determine the new value of each element such as \( m_q \) by randomly selecting from \( \{M_{\text{min}}, \ldots, A_\text{P} - M_{\text{min}}\} \).

Algorithm 3: Antenna Allocation Algorithm

Initialization: Set the temperature to \( T = T_0 \). Consider \( m = [M_{\text{AP}}, \ldots, M_{\text{AP}}] \times \ast \) as the current solution. Compute the current sum rate for \( m \), \( R_{\text{sum},2} \). Moreover \( m^* = m \) and \( R_{\text{sum},2}^* = R_{\text{sum},2} \).

while \( T \geq T_1 \) do

for \( t = 1 \) to \( t_{\text{max}} \) do

Create a neighbor solution, \( m^{\text{new}} \) and Compute sum rate for \( m^{\text{new}}, \quad R_{\text{sum},2}^{\text{new}} \). Accept \( m^{\text{new}} \) with probability of \( P \), \( R_{\text{sum},2}^{\text{new}} \) and \( m^* \) as the current sum rate, \( R_{\text{sum},2} \), and the current solution, \( m \), respectively.

end

end

if \( R_{\text{sum},2}^* \geq R_{\text{sum},2} \) then

end

end

Output: \( m^* \)

C. Third stage: digital precoder and power allocation

1) ZF digital precoder: We use a ZF digital precoder to reduce interference between the NOMA groups within each mm-AP. Since each NOMA group can contain two users, we perform singular value decomposition (SVD) on the equivalent channel for each NOMA group , i.e., \( H_{bn} \in \mathbb{C}^{N \times Q}, \forall b \in B \) and \( \forall n \in N \), which denotes the equivalent channel matrix corresponding to the NOMA group served by RF chain \( n \) in mm-AP \( b \) and \( Q_{bn} \) is a set that includes all users of the NOMA group associated with RF chain \( n \) in mm-AP \( b \). Now taking SVD of \( H_{bn} \) we have

\[
H_{bn}^H = U_{bn} \Sigma_{bn} V_{bn}^H, \tag{19}
\]

where \( U_{bn} = [u_{bn1}, \ldots, u_{bnn}] \in \mathbb{C}^{Q_{bn} \times |Q_{bn}|} \) is the left singular matrix, \( \Sigma_{bn} \) is singular value matrix that its diagonal entries are known as singular values of \( H_{bn}^H \), and \( V_{bn} \) is the
right singular matrix. Thus, the equivalent channel vector of NOMA group associated with RF chain \( n \) in mm-AP \( b \) is given by
\[
\hat{\mathbf{h}}_{bn} = \mathbf{H}_{bn} \mathbf{u}_{bn1} \in \mathbb{C}^{N \times 1}.
\] (20)

Note that, the equivalent channel vector for single-user NOMA group, \( |Q_{bn}| = 1 \), can be directly delivered with its effective channel vector, i.e., \( \hat{\mathbf{h}}_{bn} = \mathbf{h}_{bn} \). Now, the equivalent channel matrix for all the NOMA groups on all the RF chains of mm-AP \( b \) is given by
\[
\hat{\mathbf{H}}_b = [\hat{\mathbf{h}}_{b1}, \ldots, \hat{\mathbf{h}}_{bN}] \in \mathbb{C}^{N \times N},
\] (21)
where \( \hat{\mathbf{H}}_b = \mathbf{I}_N \) if no user is assigned to mm-AP \( b \). Consequently, the ZF digital precoder is given by
\[
\mathbf{G}_b = \hat{\mathbf{H}}_b^H \left( \hat{\mathbf{H}}_b \hat{\mathbf{H}}_b^H \right)^{-1} \in \mathbb{C}^{N \times N}.
\] (22)

2) Power allocation design: Given the effective channel matrix \( \hat{\mathbf{H}}_b \) and the digital precoder \( \mathbf{G}_b \), we formulate the power allocation problem in the following optimization problem as
\[
\max_{\mathbf{p}_{bn} \in \mathbb{R}^{KN}} \quad R_{\text{sum}} \quad \text{s.t.} \quad \sum_{k \in \mathcal{K}, b \in \mathcal{B}, n \in \mathcal{N}} p_{bn} \leq \mathbf{p}_n^0, \quad p_{bn} \geq 0.
\] (23a)

In problem (23), \( c_{kbn}^*, M_{kbn} \) and \( m_{kbn} \) are replaced by the obtained \( c_{kbn}^*, M_{kbn} \), and \( m_{kbn} \) in the first and the second stages. Problem (23) is non-convex optimization problem, however, it can be equivalently converted to a canonical form of a DC programming technique [33] as follows
\[
\min_{\mathbf{p}} \quad F_1(\mathbf{p}) - F_2(\mathbf{p})
\] (24a)
\[
\text{s.t.} \quad 0 \leq \mathbf{p}, \quad \mathbf{c}^T \mathbf{p} \leq \mathbf{p}_{\text{total}}, \quad c_{kbn}^* M_{kbn}^2 D_{k-b}^{bn} \leq c_{kbn}^* M_{kbn}^2 D_{k-b}^{bn}, \quad \forall k \leq i, \forall b \in \mathcal{B}, \forall n \in \mathcal{N},
\] (24b)
\[
\left( 2c_{kbn}^* M_{kbn}^2 \right) D_{k-b}^{bn} \leq c_{kbn}^* M_{kbn}^2 |\mathbf{h}_k^H \mathbf{g}_{bn}|^2,
\] (24c)
\[
\forall k \in \mathcal{K}, \forall b \in \mathcal{B}, \forall n \in \mathcal{N},
\] (24d)
where \( \mathbf{p} \in \mathbb{R}^{KN \times 1} \) denotes the collection of \( \mathbf{p}_{bn} \), also \( F_1(\mathbf{p}) \) and \( F_2(\mathbf{p}) \) are given by
\[
F_1(\mathbf{p}) = - \sum_{k \in \mathcal{K}} \sum_{b \in \mathcal{B}} \sum_{n \in \mathcal{N}} \log_2 \left( D_{k-b}^{bn} + 1 \right) \left( \mathbf{p}_{bn} \right),
\] (25)
and
\[
F_2(\mathbf{p}) = - \sum_{k \in \mathcal{K}} \sum_{b \in \mathcal{B}} \sum_{n \in \mathcal{N}} \log_2 \left( D_{k-b}^{bn} \right),
\] (25)
respectively, in which
\[
D_{k-b}^{bn} = |\mathbf{h}_k^H \mathbf{g}_{bn}|^2 c_{i-b} p_{bn} + i_{k-b}^{bn} + \Pi_k^b + \Pi_i^b + \sigma^2,
\] (26)
and
\[
D_{k-b}^{bn} = i_{k-b}^{bn} + \Pi_k^b + \Pi_i^b + \sigma^2.
\] (27)

Since the functions (26) and (27) are affine with respect to \( \mathbf{p} \), therefore, \( F_1(\mathbf{p}) \) and \( F_2(\mathbf{p}) \) are differentiable convex functions with respect to \( \mathbf{p} \). Consequently, according to the first-order condition for the convex functions [30] we have
\[
F_2(\mathbf{p}) \geq F_2(\mathbf{p}_t) + \nabla \mathbf{p} F_2(\mathbf{p}_t)^T (\mathbf{p} - \mathbf{p}_t),
\] (28)
where \( \nabla \mathbf{p} F_2(\mathbf{p}_t) \) is the gradient of \( F_2(\mathbf{p}) \) with respect to \( \mathbf{p} \) and is given by
\[
\frac{\partial F_2(\mathbf{p})}{\partial p_{bn}} |_{\mathbf{p}_t} = - \frac{1}{\ln(2)} \sum_{k'=k+1}^K \left| \frac{c_{kbn}^*}{c_{kbn}^*} \right| D_{k-b}^{bn} D_{k-b}^{bn} p_{bn}^0, \quad k \leq i, \forall b \in \mathcal{B}, \forall n \in \mathcal{N}.
\] (29)

Now, we can obtain an upper bound for the minimization problem (24) by solving the following convex optimization problem
\[
\min_{\mathbf{p}} \quad F_1(\mathbf{p}) - \tilde{F}_2(\mathbf{p})
\] (30a)
\[
\text{s.t.} \quad (24b), (24c), (24d), (24f).
\] (30b)

To find a tight upper bound in (30), we use the algorithm introduced by the DC programming method, [32], [33]. Since the functions \( F_1(\mathbf{p}) \) and \( F_2(\mathbf{p}) \) are differentiable convex, the DC-based algorithm converges to a stationary point with a polynomial time computational complexity [33].

V. SIMULATION RESULTS

To perform the simulation, we have selected Shahid Fotouhi Hall of Isfahan University of Technology, which is shown in Fig. 4. The coordinates of the seats and mm-APs for this scenario are available in [34]. Note that the locations of mm-APs are numerically optimized to reduce the average blockage probability for all seats [6], [7]. All \( K \) users are randomly selected from the people sitting on the seats. Since
most of the time, the users look toward the center of the front stage, for the user’s orientation angle, \( \Psi_k \), we consider a probability distribution based on Fig. 5, in which \( \theta_k \) is a location-dependent parameter that is calculated based on the coordinates of the user, \((X_k, Y_k)\), and the center of the front stage, \((0,0)\). We compare the simulation results of the proposed approach with the TDMA scheme, which is one of the mmWave-OMA conventional techniques. All statistical results are averaged over a large number of independent runs, i.e., 2000. The simulation parameters are shown in Table I.

**Table I: Simulation parameters**

| Parameter | Value | Parameter | Value |
|-----------|-------|-----------|-------|
| \( f_c \) | 60 GHz | \( H_{\text{mm}} \) | 125 cm |
| \( L \) | 2 | \( R_{\text{MD}} \) | 70 cm |
| \( \gamma_0 \) | 2.25 | \( t_{\text{max}} \) | 12 |
| \( \sigma^2 \) | -80 dBm | \( \beta \) | 0.95 |
| \( M_{\text{MD}} \) | 15 | \( T_0 \) | 10 |
| \( M_{\text{min}} \) | \( M_{\text{MD}}/6 \) | \( \epsilon_1 \) | \( 7 \times 10^{-11} \) |
| \( R_{\text{min}} \) | 0.25 bit/sec/Hz | \( w_1 \) | 0.6 |

Fig. 6 illustrates the average sum rate in the first stage, \( R_{\text{sum,1}} \), to compare the performance of the MWCS with the exhaustive search method. Given the high computational complexity of the exhaustive search method, we have simulated only two cases with a small number of users, mm-APs, and RF chains. Note that the proposed MWCS algorithm is applicable to the cases with more number of users, mm-APs, and RF chains. Fig. 6 shows that the MWCS algorithm reaches 97.7% of the optimal value within only 13 iterations on average, which demonstrates the fast convergence and the effectiveness of our proposed algorithm.

Fig. 7 shows the convergence of the antenna allocation algorithm versus the exhaustive search method. As it can be seen, this algorithm achieves the globally optimal solution while its computational complexity is very lower than the exhaustive search method.

Fig. 8 illustrates the average sum rate in the third stage, \( R_{\text{sum}} \), versus the total power budget. In this figure, in addition to the performance of the proposed system, the performance of the mmWave-OMA system is also shown as a performance benchmark for two cases with different values for \( K,B \), and \( N \). The simulation results in Fig. 8 show that the proposed mmWave-NOMA system outperforms the mmWave-OMA system, and this performance improvement becomes more evident by increasing the number of surplus users over the number of RF chains. In particular, for \( K = 21, B = 3, N = 5 \), under blockage effect, the proposed mmWave-NOMA system performs, on average, 12% better than the mmWave-OMA system, while this value is 15.8% without considering the blockage effect. This is due to the fact that using NOMA, the intra-group interference in the user with better channel condition is controlled by SIC, and in the user with bad channel condition is negligible owing to applying the proposed resource allocation method. Besides, the inter-group interference decreases using the ZF precoder, and by appropriately assigning users to mm-APs, we can reduce inter-AP interference. Moreover, we observe that the average sum rate increases with the total power budget. However, due to the residual interference in the system, the sum rate uptrend is slowing. In this figure, the performance degradation due to the blockage effect is observable, which is not negligible for dense indoor scenarios. Additionally, the more RF chains, the more directions each mm-AP can send data, and the more users are covered simultaneously, resulting in the higher the sum rate.

Fig. 9 shows the average sum rate in the third stage, \( R_{\text{sum}} \), versus the number of antennas equipped at each mm-AP.
the number of antennas per mm-AP increases, the array gain goes up, and beams with lower sidelobes are formed. Thus, the inter-AP interference decreases, and the average sum rate monotonically increases with the number of antennas at each mm-AP. Fig. 9 also confirms that under the blockage effect and without blockage effect, for $K = 21$, $B = 3$, $N = 5$, the proposed mmWave-NOMA system performs, on average, 8.6% and 10.4% better than the mmWave-OMA system, respectively. With the increasing number of users, a higher number of them may expose to the blockage; thus, considering this blockage effect, the more number of users leads to a higher decrease in the system sum rate. Also, note that Figures 8 and 9 show that the mmWave-NOMA system under blockage even performs 4% better than the corresponding OMA system without blockage.

Fig. 10 shows the average sum rate in the third stage, $R_{\text{sum}}$, versus the number of mm-APs for the proposed mmWave-NOMA system. We have simulated three different setups for two blockage scenarios consisting of a low blockage scenario (MD in hand) with $H_{\text{MD}} = 70$ cm and a high blockage scenario (MD in pocket) with $H_{\text{MD}} = 50$ cm. As it can be seen in Fig. 10, for all setups and scenarios, the compromise is made at $B = 3$, while for the high interference setup in low blockage conditions, $B = 1$ is the best. In fact, for the high interference setup in low blockage conditions, the inter-AP interference has a predominant effect compared to the blockage probability. Consequently, by increasing the number of mm-APs, the system sum rate decreases. Consequently, the number of required mm-APs needs to be chosen based on a tradeoff between the blockage probability and co-channel interference.

VI. CONCLUSIONS

In this paper, we have proposed a novel framework for resource allocation in the downlink of mmWave-NOMA communication through multi-AP for dense venues. The resource allocation in this network is very challenging due to the high co-channel interference caused by dense deployment and the inherent complexity of the system stemming from the combination of multi-AP structure, mmWave communication, and NOMA technique. To solve the highly computational complex resource allocation problem of the desired network, we have broken the main problem into three sub-problems. Then, we have proposed an algorithm for each sub-problem and evaluated the performance of each of them in terms of optimality, complexity, and convergence. Simulation results show that the sum rate of the proposed mmWave-NOMA scheme under the blockage effect is still 4% higher than the corresponding OMA scheme without blockage. Also, we have found that a tradeoff between the amount of co-channel interference and the availability of LoS mmWave links should be considered to find the required number of AP in a mmWave-NOMA scheme.

REFERENCES

[1] I. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, “Millimeter-wave communications: physical channel models, design considerations, antenna construction and link-budget,” IEEE Communications Surveys & Tutorials, vol. 20, no. 2, pp. 870–913, Secondquarter 2018.
[1] B. Wang, L. Dai, Z. Wang, N. Ge, and S. Zhou, “Spectrum and energy-efficient beamspace MIMO-NOMA for millimeter-wave communication systems using lens antenna array,” IEEE Journal on Selected Areas in Communications, vol. 35, no. 10, pp. 2570–2382, Oct. 2017.

[2] L. Zhao, D. W. K. Ng, and J. Yuan, “Multi-user precoding and channel estimation for hybrid millimeter wave systems,” IEEE Journal on Selected Areas in Communications, vol. 35, no. 7, pp. 1576–1590, July 2017.

[3] C. A. Floudas, Nonlinear and mixed-integer optimization: fundamentals and applications. Oxford University Press, 1995.

[4] P. J. Van Laarhoven and E. H. Aarts, Simulated annealing: Theory and applications. Springer, 1987.

[5] Y. Zhao, W. Xu, and S. Jin, “An minorization-maximization based hybrid precoding in NOMA-mMIMO,” in International Conference on Wireless Communications and Signal Processing (WCSP), Nanjing, China, Oct 2017.

[6] R. T. Marler and J. S. Arora, “Survey of multi-objective optimization methods for engineering,” Structural and multidisciplinary optimization, vol. 26, no. 6, pp. 369–395, Mar. 2004.

[7] Z. Mustaffa and Y. Yusof, “A comparison of normalization techniques in predicting dengue outbreak,” in International Conference on Business and Economics Research, vol. 1, Kuala Lumpur, Malaysia, 2011, pp. 345–349.

[8] R. Hunger, Floating point operations in matrix-vector calculus. Munich University of Technology, Inst. for Circuit Theory and Signal Processing, 2005.

[9] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge university press, 2004.

[10] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to algorithms. MIT press, 2009.

[11] N. Vucic, S. Shi, and M. Schubert, “DC programming approach for resource allocation in wireless networks,” in International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks, Avignon, France, May 2010, pp. 380–386.

[12] Z. Wei, D. W. K. Ng, J. Yuan, and H. Wang, “Optimal resource allocation for power-efficient MC-NOMA with imperfect channel state information,” IEEE Transactions on Communications, vol. 65, no. 9, pp. 3944–3961.

[13] The coordinates of the seats and mm-APs. [Online]. Available: https://iutbox.iut.ac.ir/index.php/s/YSxXXgnT5H9Jetn