User Grouping and Resource Allocation for Joint Communication and Positioning in mmWave Multi-cell Networks

Xueni Luo, Xiaofeng Lu, Boyu Jin, Benquan Yin and Kun Yang, *Follow, IEEE*

***Abstract*—B5G/6G expands a new application scenario of joint communication and positioning which can simultaneously provide high-quality communication and positioning services. Millimeter wave (mmWave) and massive Multiple Input Multiple Output (MIMO) can help systems to achieve high-quality communication and generate high-directional beams to assist positioning. In this paper, the proposed Structured Perturbed Orthogonal Matching Pursuit (SPOMP) could alleviate the pilot pollution in massive MIMO systems and break the resolution of angular estimation. Based on the above super-resolution estimation, we develop a dynamic two-stage multi-cell user grouping scheme to reduce interference and improve resource utilization. Combined with the grouping results and our derived performance metric, a joint optimization problem for power and bandwidth allocation is proposed to maximize the comprehensive performance of joint communication and positioning while guaranteeing performance bounds. An effective cyclic iteration algorithm based on Alternating Direction Method of Multipliers (ADMM) is present to solve the proposed optimization problem. Numerical results show that the proposed joint user grouping and resource allocation scheme achieves a larger rate-accuracy region and a better multi-cell service performance balance compared to conventional schemes.**

***Index Terms*—Millimeter wave, massive MIMO, joint communication and positioning, user grouping, resource allocation.**

I. Introduction

UNDER B5G/6G networks, there have been a large number of emerging applications, such as unmanned driving technology [1], [2], multi-device Internet of Things [3], [4] and crowd sensing technology [5], which put forward higher requirements for the improvement of communication systems, including not only high-quality wireless communication links, but also positioning services with high robustness and accuracy [6], [7]. As an important technology in 5G standards, mmWave has a broad spectrum resource to provide larger channel bandwidth. Meanwhile, massive MIMO can obtain the high directional gain and higher data rate through hybrid analog/digital beamforming (HADB) technology and multiple data streams [8], [9], to deal with the serious path loss of mmWave and broaden the application range of mmWave [10]–[12]. So massive MIMO communication networks combined with mmWave is widely considered as the key technology to support the joint communication and positioning systems [13].

In multi-user scenarios, the practical channel matrices of different users are commonly correlated with each other due to their limited antenna number and dense distribution area [14]. The authors in [15] isolate the users whose correlation coefficient is larger than a certain threshold in separate groups which are served in different time to reduce the inter-user interference which is a main influencing factor to the multi-user systems' performance. However, they do not consider to reuse the precious spectrum resource. Spatial correlation between different users is a key point for user grouping. In [16], the proposed user grouping scheme assigns the users with low correlation to different groups to achieve spatial and frequency multiplexing gain.

Meanwhile, there is a resource competition between the communication and positioning services due to the limited resources in multi-user scenarios. Some researchers obtain the trade-off between data rate and positioning accuracy through the beam optimization and resource allocation schemes in single-user and multi-user scenarios [17], [18]. However, above researches do not pay enough attention to the vital spectrum resources. The authors in [19]–[22] propose the bandwidth allocation schemes by allocating a single sub-carrier or multiple sub-carriers to the users. However, the low energy efficiency of such schemes would violate the concept of green communication. In [23], the authors propose a dual objective optimization problem of power and bandwidth allocation while not based on joint communication and positioning. In [16], the authors research the multi-dimensional resource allocation, including power, beams, and sub-carriers, in multi-user MIMO systems. However, they assume that the CSI is idealized and known. Moreover, the authors in [16], [24] combine resource allocation schemes with user grouping strategies in order that the different groups can use the same resources to improve the resource utilization.

Effective user grouping and resource allocation strategies are necessary for all users to enjoy two high-quality ser-

This work was supported in part by the National Natural Science Foundation of China under Grants 62071364, 62132004, and 62231027, and in part by the MOST Major Research and Development Project under Grant 2021YFB2900204. *(Corresponding Author: Xiaofeng Lu.)*

Manuscript received xxxx 00, 0000; revised xxxx 00, 0000.

Xueni Luo, Xiaofeng Lu, Boyu Jin and Benquan Yin are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xian 710071, China.(email: lxn@stu.xidian.edu.cn; luxf@xidian.edu.cn; 22011210625@stu.xidian.edu.cn; bqyin@stu.xidian.edu.cn)

Kun Yang is with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Essex, U.K. (e-mail: kyang@ieee.org).

vices simultaneously. Meanwhile, knowing the accurate CSI is also the basis of reasonable grouping criterias and effective resource allocation schemes. Considering the characteristics of mmWave and the massive access terminals, the direct consequence of the estimation error of CSI is the loss of the expected signal and increasing interference, which are fatal to the performance of joint communication and positioning systems [25]. The traditional sub-space algorithms, including Estimating Signal Parameters Via Rotational Invariance Techniques (ESPRIT) [26] and Multiple Signal Classification (MUSIC) [27], need a large number of snapshots and consume a lot of training resources. The authors of [28] proof the sparsity of mmWave massive MIMO channel in angular domain, and convert the high-dimensional channel to the low-dimensional channel in spatial domain. In [29], the authors propose the improved Orthogonal Matching Pursuit (OMP) to solve the power leakage around the grid points due to the grid quantization. And the authors of [30] extend the above method to mmWave networks. However, all of these algorithms need the known channel sparsity, and whether they can be applied under mmWave massive MIMO systems is unknown.

In recent years, to obtain further system performance improvement and enhance the utilization of limited resource, the research focus has shifted from single-cell systems to more practical multi-cell scenarios and the inter-cell interference (ICI) becomes a main limiting factor of system performance. It is clear that realizing the multi-cell cooperation could reduce ICI and improve the multi-cell system performance [31]. Motivated by the above and considering multi-cell systems, we investigate a joint user grouping and resource allocation scheme to get a larger rate-accuracy region in joint communication and positioning systems. The proposed system uses the positioning service to obtain the information of the users and the surrounding environment. By optimizing the resource scheduling strategy and using the beamforming technology to ensure the quality of communication service and improve the resource utilization, so as to follow the concept of green communication. Against the existing approaches, the key contributions of this paper are as follows:

- To improve the two-stage Structured Perturbed Orthogonal Matching Pursuit (SPOMP) based on iteratively updating the tradeoff weight between the sparsity of channel and the measurement vector's fitting error. So we can obtain the structured perturbed parameters to reduce the resolution loss of channel parameters caused by the grid quantification in sparse recovery methods and pilot pollution in massive MIMO systems.
- To propose the dynamic two-stage multi-cell user grouping scheme based on HADB technology and the estimated super-resolution CSI to reduce the inter-cell and intracell interference and improve the resource utilization, simultaneously.
- To develop two new information metrics, data rate and positioning estimation rate. Based on them, the joint communication and positioning optimation question can be present under the similar physical concepts of two services. By quantifying the inter-cell interference and relaxing the intra-group interference to a constant, the proposed NP-hard problem can be changed to the proper reformulation which is able to applying the ADMM and Coordinate Descent and we can obtain the suboptimal resource allocation scheme easily.

The rest of this paper is organized as follows: Section II describes the mmWave sparse channel model and comprehensive multi-cell system model. Section III introduces the performance metrics. Section IV presents the optimization problem and proposed joint transmission scheme. Section V shows the simulation results. Finally, the conclusions are stated in Section VI.

Here are some notations will be used in this paper:

- $(\cdot)^{\mathrm{T}}$, $(\cdot)^{\mathrm{H}}$, $(\cdot)^{*}$, $(\cdot)^{-1}$, $(\cdot)^{\dagger}$: transpose, Hermitian, conjugate, inverse, pseudo-inverse, respectively.
- $\otimes$, $\odot$: Kronecker product, Khatri-Rao product, respectively.
- $\mathrm{E}\{\cdot\}$, $\mathrm{vec}(\cdot)$, $H(\cdot)$: expectation operation, vectorization operation, entropy operation, respectively.
- $\|\cdot\|_0$, $\|\cdot\|_F$, $[T]$: l_0-norm, Frobenius norm, the set of non-negative integers $\{1, 2, ..., T\}$, respectively.

Throughout this paper, matrices and vectors are denoted by upper-case and lower-case boldface letters, respectively.

II. System model

A. Transmission Model

Consider a HADB mmWave MIMO network where the locations of the static user equipments (UEs) are unknown and the locations of the base stations (BSs) are known, all equipped with a uniform linear array (ULA) containing an odd number of antennas. To guarantee multi-stream data transmissions, assume that each BS is equipped with M antennas, M_{RF} RF chains, and $M_{\mathrm{DS}} > 1$ data streams. Generally, it is assumed that $M_{\mathrm{DS}} \leq M_{\mathrm{RF}} \leq M$, $N_{\mathrm{DS}} \leq N_{\mathrm{RF}} \leq N$, and $M_{\mathrm{DS}} = N_{\mathrm{DS}} \leq \min\{M_{\mathrm{RF}}, N_{\mathrm{RF}}\}$, to ensure the rationality of the system.

Each symbol is sent through its specialized RF chain, so that the BS in our OFDM system transmits the continuous M_{RF} symbols at frame t. When the symbols in frame t transmitting, the channel is assumed to be time-invariant, while the channel will be considered to be time-varying in different frames. For symbol s at frame t, the pilot is sent through a dedicated RF chain expressed by the precoder operation as $\left\{\mathbf{f}_{t,s} \in \mathbb{C}^{M\times 1}; \|\mathbf{f}_{t,s}\|_2^2 = 1, \forall t \in [T], \forall s \in [M_{\mathrm{RF}}]\right\}$, where T is the number of snapshots. Similarly, each UE uses N_{RF} RF chains to complete the combiner operation denoted as $\left\{\mathbf{W}_{t,s} \in \mathbb{C}^{N\times N_{RF}}; \|\mathbf{W}_{t,s}\|_2^2 = 1, \forall t \in [T], \forall s \in [M_{\mathrm{RF}}]\right\}$. The received signal during the symbol s at frame t can be expressed by

$$\mathbf{y}_{t,s} = \mathbf{W}_{t,s}^{\mathrm{H}}\mathbf{H}_t\mathbf{f}_{t,s}\mathbf{x}_t + \mathbf{W}_{t.s}^{\mathrm{H}}\mathbf{n}_{t,s}, t \in [T], s \in [M_{\mathrm{RF}}], \quad (1)$$

where $\mathbf{H}_t \in \mathbb{C}^{N\times M}$ denotes the channel matrix in frame t from the BS to the UE, $\mathbf{x}_t$ represents the transmitted pilot signal known to the BS and the UE, $\mathbf{n}_{t,s} \in \mathbb{C}^{N\times 1}$ denotes the noise assumed to be i.i.d Gaussian distributed, $\mathbf{n}_{t,s} \sim \mathcal{CN}\left(0, \sigma_n^2 \mathrm{I}_N\right)$, with the known noise variance σ_n^2.

The received signal corresponding to all the symbols in frame t can be combined in columns as $\tilde{\mathbf{y}}_t = \left[\mathbf{y}_{t,1}^{\mathrm{T}}, ..., \mathbf{y}_{t,s}^{\mathrm{T}}\right] \in \mathbb{C}^{M_{\mathrm{RF}} N_{\mathrm{RF}} \times 1}, \forall s \in [M_{\mathrm{RF}}], \forall t \in [T]$. Using the vectorized form of the matrices, it can be mathematically expressed as

$$\tilde{\mathbf{y}}_t = \left(\mathbf{F}_t \odot \mathbf{W}_t^*\right)^{\mathrm{T}} \mathrm{vec}\left(\mathbf{H}_t\right) \tilde{\mathbf{x}}_t + \mathbf{W}_t \tilde{\mathbf{n}}_t, \tag{2}$$

where $\mathbf{W}_t = \left[\mathbf{W}_{t,1}^{\mathrm{T}}, ..., \mathbf{W}_{t,s}^{\mathrm{T}}\right]^{\mathrm{T}}$, $\mathbf{F}_t = [\mathbf{f}_{t,1}, ..., \mathbf{f}_{t,s}]$, and $\tilde{\mathbf{n}}_t = [\mathbf{n}_{t,1}, ..., \mathbf{n}_{t,s}]$. ${\mathbf{W}_t}^{\mathrm{H}} \mathbf{W}_t = {\mathbf{F}_t}^{\mathrm{H}} \mathbf{F}_t = \mathbf{I}$ is apparently true. Then assuming $\mathbf{\Phi}_t = \left(\mathbf{F}_t \odot \mathbf{W}_t^*\right)^{\mathrm{T}} \in \mathbb{C}^{M_{\mathrm{RF}} N_{\mathrm{RF}} \times MN}$, $\tilde{\mathbf{n}}_{t,agg} = \tilde{\mathbf{W}}_t \tilde{\mathbf{n}}_t \in \mathbb{C}^{M_{\mathrm{RF}} N_{\mathrm{RF}} \times 1}$ and $\mathbf{h}_t = \mathrm{vec}\left(\mathbf{H}_t\right) \in \mathbb{C}^{NM \times 1}$, the signal in (2) can be simplified to:

$$\tilde{\mathbf{y}}_t = \mathbf{\Phi}_t \mathbf{h}_t \tilde{\mathbf{x}}_t + \tilde{\mathbf{n}}_{t,agg}. \tag{3}$$

Assuming that the channel contains K spatial clusters and each cluster is composed of L multipath components (MPCs). For simplicity, we assume that K and L remain fixed within the duration of the channel parameter estimation, then the total number of MPCs is KL.

Due to the rich bandwidth in the mmWave frequency, the coherence time of time-varying fading coefficients is much less than the angular coherence time, which means that the spatial characteristics of mmWave signals can be regarded as constant during the coherent time of the small scale fading coefficients. So the channel matrix can be denoted as

$$\mathbf{H}_t = \frac{1}{Loss} \sum_{k=1}^{K} \sum_{l=1}^{L} \alpha_{k,l,t} \mathbf{a}_{\mathrm{UE}}\left(\theta_{k,l}^{rx}\right) \mathbf{a}_{\mathrm{BS}}\left(\theta_{k,l}^{tx}\right)^{\mathrm{H}}, \tag{4}$$

where $Loss$ is the average path loss, $\alpha_{k,l,t}$ is the small scale fading coefficient at frame t of the l-th MPC in the k-th cluster and assumed to be i.i.d complex Gaussian distributed. $\theta_{k,l}^{rx} \in [0, \pi)$ and $\theta_{k,l}^{tx} \in [0, \pi)$ denote the AoA and AoD of the l-th MPC in the k-th cluster, respectively, which remain constant over the snapshot. $\mathbf{a}_{\mathrm{BS}}\left(\theta_{k,l}^{tx}\right) \in \mathbb{C}^{M \times 1}$ and $\mathbf{a}_{\mathrm{UE}}\left(\theta_{k,l}^{rx}\right) \in \mathbb{C}^{N \times 1}$ are the normalized array response of the BS ULA and UE ULA, respectively, whose elements are expressed as

$$\left[\mathbf{a}_{\mathrm{BS}}\left(\theta_{k,l}^{tx}\right)\right]_m = \frac{1}{\sqrt{M}} e^{j \frac{2\pi}{\lambda} d_{\mathrm{BS}} (m-1) \cos\left(\theta_{k,l}^{tx}\right)}, \forall m \in [M], \tag{5a}$$

$$\left[\mathbf{a}_{\mathrm{UE}}\left(\theta_{k,l}^{rx}\right)\right]_n = \frac{1}{\sqrt{N}} e^{j \frac{2\pi}{\lambda} d_{\mathrm{UE}} (n-1) \cos\left(\theta_{k,l}^{rx}\right)}, \forall n \in [N], \tag{5b}$$

where $d_{\mathrm{BS}} = d_{\mathrm{UE}} = \lambda/2$ indicate the inter-antenna spacing in the BS ULA and UE ULA, respectively, and λ is the carrier wavelength.

B. Sparse Representation of the mmWave Channel

The finite scattering characteristic of the mmWave causes the channel to show sparsity in AoA-AoD domain, and different AoA-AoDs denote different paths, so the mmWave channel can be estimated through sparse recovery methods. Ignoring the height gap between BSs and UEs, we discretize the AoA-AoD tuple $(\theta^{rx}, \theta^{tx})$ in the continuous parameter space $([0, \pi) \times [0, \pi))$ into a set of finite grid points to establish a grid in a two-dimensional space. When the actual AoA-AoD coincides with a grid point, the actual channel can be estimated ideally.

To ensure the accuracy of estimating the line-of-sight (LoS) path and improve the resolution of the estimated parameters, we choose the uniform sampling scheme in the virtual $\cos(\theta)$ domain. In this way, the AOA/AOD are non-uniformly quantized so that the spacing of adjacent grid points in the grid center is significantly smaller than that at the grid edges. Compared with the uniform sampling scheme in physical angle domain, it would reduce the correlation of the redundant dictionary and preserve orthogonality over the virtual domain [32]. The set containing G_{UE} and G_{BS} discretized grid points in AOA and AOD domains can be respectively denoted as

$$\Theta^{rx} = \left\{\theta_{G_i^{\mathrm{UE}}}^{rx} : \cos\left(\theta_{G_i^{\mathrm{UE}}}^{rx}\right) = 1 - \frac{2\left(G_i^{\mathrm{UE}} - 1\right)}{G_{\mathrm{UE}}}, G_i^{\mathrm{UE}} = 1, ..., G_{\mathrm{UE}}\right\},$$
$$\Theta^{tx} = \left\{\theta_{G_i^{\mathrm{BS}}}^{tx} : \cos\left(\theta_{G_i^{\mathrm{BS}}}^{tx}\right) = 1 - \frac{2\left(G_i^{\mathrm{BS}} - 1\right)}{G_{\mathrm{BS}}}, G_i^{\mathrm{BS}} = 1, ..., G_{\mathrm{BS}}\right\}, \tag{6}$$

where $\theta_{G_i^{\mathrm{UE}}}^{rx} \in [0, \pi)$ and $\theta_{G_i^{\mathrm{BS}}}^{tx} \in [0, \pi)$ indicate the AoA and AoD in the physical domain corresponding to the grid points in the virtual domain, respectively.

The array responses corresponding to all of grid points in the form of column vectors are composed into matrices as $\mathbf{A}_{\mathrm{UE}} = \left[\mathbf{a}_{\mathrm{UE}}\left(\theta_1^{rx}\right), ..., \mathbf{a}_{\mathrm{UE}}\left(\theta_{G_{\mathrm{UE}}}^{rx}\right)\right]$ and $\mathbf{A}_{\mathrm{BS}} = \left[\mathbf{a}_{\mathrm{BS}}(\theta_1^{tx})^{\mathrm{T}}, .., \mathbf{a}_{\mathrm{BS}}\left(\theta_{G_{\mathrm{BS}}}^{tx}\right)^{\mathrm{T}}\right]$. $\mathbf{A}_{\mathrm{UE}} \in \mathbb{C}^{N \times G_{\mathrm{UE}}}$ and $\mathbf{A}_{\mathrm{BS}} \in \mathbb{C}^{G_{\mathrm{BS}} \times M}$ represent the array response matrices built on the sets Θ^{rx} and Θ^{tx} , respectively. We can convert the channel matrix in (4) into

$$\mathbf{H}_t = \mathbf{A}_{\mathrm{UE}} \mathbf{H}_t^G \mathbf{A}_{\mathrm{BS}}^{\mathrm{H}}, \tag{7}$$

where $\mathbf{H}_t^G \in \mathbb{C}^{G_{\mathrm{UE}} \times G_{\mathrm{BS}}}$ is the virtual channel matrix obtained after the discretization in angular domain. Using the vectorized form of the matrices, (7) can be changed again into

$$\mathbf{h}_t = \mathrm{vec}\left(\mathbf{H}_t\right) = \mathbf{\Psi} \mathbf{h}_t^G, \tag{8}$$

where $\mathbf{h}_t^G = \mathrm{vec}\left(\mathbf{H}_t^G\right)$, and $\mathbf{\Psi} = \left(\mathbf{A}_{\mathrm{BS}}^* \otimes \mathbf{A}_{\mathrm{UE}}\right)$ represents the dictionary matrix in angular domain, each of whose column represents the vector form of the array response corresponding to the AoA-AoD at each grid point.

Based on (8), the received signal model in (3) can be converted to

$$\tilde{\mathbf{y}}_t = \mathbf{\Phi}_t \mathbf{\Psi} \mathbf{h}_t^G \tilde{\mathbf{x}}_t + \tilde{\mathbf{n}}_{t,agg}. \tag{9}$$

When obtaining the received signal, the virtual channel can be recovered through (9), and the actual channel can be recovered through the sparsity of the channel.

C. Multi-cell System Model

Based on above models, we consider a multi-cell downlink system including C cells and U users randomly distributing in a two-dimensional space, each cell whose radius is R contains one BS in its center. The total power of a BS used and the bandwidth of the system used for communication and positioning services are $P_{\max}$ and $B_{\max}$, respectively.

The received signal transmitted from the b_i-th BS to the u_j-th UE is expressed as

$$\begin{aligned}\tilde{\mathbf{y}}_{b_i,u_j} &= \mathbf{W}^{\mathrm{H}}_{b_i,u_j}\mathbf{H}_{b_i,u_j}\mathbf{F}_{b_i,u_j}\mathbf{x}^{b_i,u_j} \\ &+ \mathbf{W}^{\mathrm{H}}_{b_i,u_j}\mathbf{H}_{b_i,u_j}\sum_{m=1,m\neq j}^{U}\mathbf{F}_{b_i,u_m}\mathbf{x}^{b_i,u_m} \\ &+ \mathbf{W}^{\mathrm{H}}_{b_i,u_j}\sum_{n=1,n\neq i}^{B}\mathbf{H}_{b_n,u_j}\mathbf{F}_{b_n,u_j}\mathbf{x}^{b_n,u_j} + \mathbf{W}^{\mathrm{H}}_{b_i,u_j}\mathbf{n}.\end{aligned} \tag{10}$$

The first item in (10) denotes the desired receiving signal; the second item represents the intra-cell interference including the inter-symbol interference; the third term expresses the inter-cell interference; and the last item is the noise.

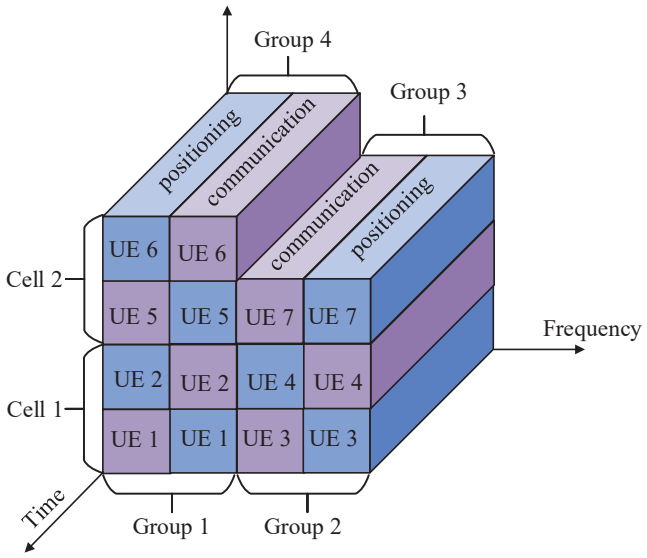

Fig. 1. Frequency division multiplexing model of joint communication and positioning with user grouping in multi-cell systems.

Based on [33], we use a frequency division multiplexing model with user grouping in our multi-cell system to improve resource utilization and ensure the hours of communication and positioning services in Fig.1. In a fixed time slot, the BSs provide UEs with communication service and positioning service simultaneously. After dividing the UEs into groups, the positioning and communication services for one UE occupy different sub-carriers, and the UEs in different groups use orthogonal sub-carriers. Meanwhile, the UEs in a group or different cells could use the same sub-carriers to enhance the frequency resource utilization.

Due to the limits of total power and bandwidth, each BS must allocate the power and the time-frequency resources reasonably to different UEs in its cell while transmitting the signal. Assumed that P_{b_i,u_j} is the signal power assigned by the b_i-th BS to the u_j-th UE, respectively. The u_j-th UE is assigned to a cell whose user set can be expressed as $\mathbf{q}_{b_i}$, which contains N_{b_i} UEs, meanwhile the UE is also assigned to a group whose user set can be expressed as $\mathbf{z}_{b_i,g}$, which contains $N_{b_i,g}$ UEs, it is clear that the bandwidth assigned to everyone of the UEs in g-th group can be expressed by $B_{b_i,g}$. And the b_i-th cell contains g_{b_i} groups. According to the resource multiplication strategy mentioned above, the assigned resource need to satisfy

$$\sum_{\forall u_j\in\mathbf{q}_{b_i}} P_{b_i,u_j}\leq P_{\max}, \forall b_i\in[C]\,; \tag{11a}$$

$$\sum_{\forall g\in[g_{b_i}]} B_{b_i,g}\leq B_{\max}, \forall b_i\in[C]\,; \tag{11b}$$

(11a) means that the sum of the total power that each BS can schedule to its UEs can not exceed $P_{\max}$, and (11b) means that the sum of the bandwidth that each BS can schedule to its groups can not exceed $B_{\max}$.

III. Performance Metric

A. HADB Technology

Multiple MPCs will make the sub-carriers of the received signal no longer orthogonal to each other, resulting in co-channel interference. Moreover, the multipath fading effect can greatly disperse the SINR of the non-line-of-sight (NLoS) MPCs. Therefore, we use the Maximum Ratio Transmission (MRT) and Zero Forcing (ZF) precoding technology to minimize the different types of interference and optimize the recieved SINR by setting the precoder and combiner matrices to match the multi-user CSI.

The BSs use the ZF algorithm to eliminate intra-cell interference, the precoder matrix of u_j-th UE of b_i-th cell based on its corresponding combiner matrix is represented as

$$\mathbf{F}_{b_i,u_j} = \frac{\mathbf{H}_{b_i,u_j}}{\gamma_{b_i,u_j}}, \tag{12}$$

where $\gamma_{b_i,u_j} = \sqrt{\mathrm{Tr}\left(\mathbf{H}_{b_i}(\mathbf{H}_{b_i})^{\mathrm{H}}\right)}$ denotes the normalization parameter of power and $\mathbf{H}_{b_i} = \left[\mathbf{H}_{b_i,u_j}\right]^{\mathrm{H}}, \forall u_j \in [N_{b_i}]$. According to the definition of SINR, the inter-cell interference is denoted as

$$\sigma^2_{b_i,u_j} = \sum_{n=1,n\neq i}^{B}\left\|\mathbf{W}^{\mathrm{H}}_{b_i,u_j}\mathbf{H}_{b_n,u_j}\mathbf{F}_{b_n,u_j}\mathbf{x}^{b_n,u_j}\right\|_2^2. \tag{13}$$

In fact, due to the size of antenna and device, there are not enough antennas deployed in UEs. Assumed that $(B-1)M \geq N$, we make $\mathbf{W}_{b_i,u_j}$ equal the right singular matrix corresponding to the smallest singular value λ_k of $\mathbf{T}_{b_i,u_j}$, so that $\left\|\mathbf{T}_{b_i,u_j}\mathbf{W}_{b_i,u_j}\right\|_2^2 = \lambda_k^2$. Then $\lambda_k^2 P_{\max}$ denotes the reduced inter-cell interference from other BS, which is not completely eliminated, but reduced to achieve our goal.

B. Data Rate and Positioning Estimation Rate

The quality of communication services can be described by communication SINR and channel capacity. From the received

signal model in (10), the SINR of the u_j-th UE of b_i-th cell can be expressed as

$$\mathrm{SINR}_{b_i,u_j}^{\mathrm{C}} = \frac{P_{b_i,u_j}^{\mathrm{C}} \left\| \mathbf{H}_{b_i,u_j} \frac{\mathbf{H}_{b_i,u_j}}{\gamma_{b_i,u_j}} \right\|_2^2}{(C-1)\lambda_k^2 P_{\max} + \sigma_n^2 + \sum\limits_{\forall u_k \in [\mathbf{z}_{b_i,g}], k \neq j} P_{b_i,u_k}^{\mathrm{C}} \left\| \mathbf{H}_{b_i,u_j} \frac{\mathbf{H}_{b_i,u_k}}{\gamma_{b_i,u_k}} \right\|_2^2}. \tag{14}$$

The data rate can be expressed based on (14) as:

$$R_{b_i,u_j}^{\mathrm{C}} = B_{b_i,g}^{\mathrm{C}} \log_2 \left(1 + \mathrm{SINR}_{b_i,u_j}^{\mathrm{C}}\right). \tag{15}$$

From (14) and (15), it is clear that the data rate mainly depends on the signal power and bandwidth assigned to the UE.

To find a positioning metric of the same magnitude as the Shannon capacity and motivated by the radar estimation rate [34], we develop a new metric, the positioning estimation rate R_{b_i,u_j}^{O} which is analogous to the data information rate. The physical meaning of the positioning estimation rate is the maximum rate of the information about the channel parameters transmitted by the received signal. And R_{b_i,u_j}^{O} can be expressed by the estimation entropy and the reveived signal extropy as:

$$R_{b_i,u_j}^{\mathrm{O}} = \frac{h_r - h_{est}}{T_{b_i,g}^{\mathrm{O}}}, \tag{16}$$

where $T_{b_i,g}^{\mathrm{O}} = \frac{1}{B_{b_i,g}^{\mathrm{O}}}$ denotes the symbol length. h_{est} represents the estimation entropy, namely, the estimation uncertainty. Under the assumption of Gaussian estimation error, h_{est} can be expressed as

$$h_{est} = \frac{1}{2}\log_2\left(2e\pi\sigma_{est}^2\right) = \frac{1}{2}\log_2\left(\frac{2e\pi}{2M^3 v^2\,\mathrm{SINR}_{b_i,u_j}^{\mathrm{O}}}\right), \tag{17}$$

where σ_{est}^2 is the Cramer-Rao lower bound for angle estimation [35], with $v^2 = \pi^2 d^2 \cos^2\theta_0 (M-1)/3\lambda^2 = \pi^2 \cos^2\theta_0 (M-1)/12$, where θ_0 is the beam pointing-angle of the array. Assuming the channel parameter is Gaussian distributed with variance σ_{para}^2, the reveived signal extropy h_r is given by

$$h_r = \frac{1}{2}\log_2\left[2e\pi\left(\sigma_{est}^2 + \sigma_{para}^2\right)\right]. \tag{18}$$

(16) can be written as

$$\begin{aligned} R_{b_i,u_j}^{\mathrm{O}} &= \frac{B_{b_i,g}^{\mathrm{O}}}{2}\log_2\left(1 + \frac{\sigma_{para}^2}{\sigma_{est}^2}\right) \\ &= \frac{B_{b_i,g}^{\mathrm{O}}}{2}\log_2\left(1 + \sigma_{para}^2 2M^3 v^2\,\mathrm{SINR}_{b_i,u_j}^{\mathrm{O}}\right). \end{aligned} \tag{19}$$

Above proposed metrics allow communication and positioning services to be jointly optimal and compromised under the similar physical concepts. Both of date rate and positioning estimation rate will reflect the efficiency of transferring the transmitting information and channel information through the joint communication and positioning system, which can provide guidance for our system designers.

IV. JOINT DYNAMIC USER GROUPING AND RESOURCE ALLOCATION SCHEME

In this section, we first present our two-stage multi-cell user grouping based on super-resolution estimation. Then, we formulate the joint user grouping and resource allocation problem and its corresponding solution algorithm to optimize the system performance.

A. Super-resolution Channel Parameters Estimation

This traditional sparse recovery algorithms are suitable only when the actual AoA-AoD is aligned with a grid point which is an ideal on-grid case. Thus, the grid-dependent method still has some defects in the resolution due to the limits of the distance between the neighboring grid points. And when KL is greater than G_{UE} or G_{BS}, the AOA-AOD of at least two paths will be mapped to the same grid point, which greatly affects the final estimation result.

The limited number of grids implies that the estimation resolution is related to the grid spacing due to the mismatch between the pre-designed grid and the actual sparse basis, i.e., the basis mismatch phenomenon. Increasing the grid number will increase the estimation resolution while increasing the complexity of algorithm dramatically. In contrast, we introduce a new parameter to represent the deviation between the actual sparse basis and the pre-designed grid, thus eliminating the off-grid effect while guaranteeing a lower complexity, and breaking the limit of resolution.

To eliminate the dictionary matrix mismatch due to the off-grid effect to ensure (8) and (9) valid, the actual AoA-AoD of the l-th MPC is represented as $\tilde{\theta}_l^{rx} = \theta_l^{rx} + \delta_l^{rx}$ and $\tilde{\theta}_l^{tx} = \theta_l^{tx} + \delta_l^{tx}$. θ_l^{rx} is the grid point nearest to $\tilde{\theta}_l^{rx}$ in the set Θ^{rx}, and $\delta_l^{rx} \in [0, \Delta\theta^{rx}/2]$ is the perturbation parameter of $\tilde{\theta}_l^{rx}$ corresponding to θ_l^{rx}. It is similar to the AOD domain. Based on the more accurate $\left(\tilde{\theta}_l^{rx}, \tilde{\theta}_l^{tx}\right)$, we can update $\mathbf{\Psi}_l$, the column corresponding to the l-th MPC, as

$$\begin{aligned} &\mathbf{\Psi}_l = \mathbf{a}_{\mathrm{UE}}\left(\theta_l^{rx}\right)\mathbf{a}_{\mathrm{BS}}\left(\theta_l^{tx}\right)^{\mathrm{H}} \\ &\rightarrow \mathbf{\Psi'}_l = \mathbf{a}_{\mathrm{UE}}\left(\theta_l^{rx} + \delta_l^{rx}\right)\mathbf{a}_{\mathrm{BS}}\left(\theta_l^{tx} + \delta_l^{tx}\right)^{\mathrm{H}}. \end{aligned} \tag{20}$$

Considering all of the KL MPCs, we can adapt $\mathbf{\Psi}$ to the new virtual channel matrix $\mathbf{\Psi'}$ by (20) and greatly reduce the off-grid effect. It is necessary to capture the perturbation parameters occurring in angle domain accurately in order to complete super-resolution estimations.

Based on (9), the tranditional sparse recovery problem can be formulated as

$$\min_{\mathbf{h}_t^G, \boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}} \|\mathbf{h}_t^G\|_0, \text{ s.t. } \left\|\tilde{\mathbf{y}}_t - \mathbf{\Phi}_t \mathbf{\Psi'} \mathbf{h}_t^G \tilde{\mathbf{x}}_t\right\|_{\mathrm{F}} \leq \varepsilon, \tag{21}$$

where ε is the error tolerance parameter. Refer to [36], we can replace the l_0-norm by a log-sum function and translate (21) to a unconstrained optimization problem, and the iterative objective function is

$$\begin{aligned} &\min_{\mathbf{h}_t^G, \boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}} S^{(i)}\left(\mathbf{h}_t^G, \boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}\right) = \\ &\lambda^{-1}\left(\mathbf{h}_t^G\right)^{\mathrm{H}} \mathbf{D}^{(\mathbf{i})} \mathbf{h}_{\mathbf{t}}^{\mathbf{G}} + \left\|\tilde{\mathbf{y}}_{\mathbf{t}} - \mathbf{\Phi}_{\mathbf{t}} \mathbf{\Psi'} \mathbf{h}_{\mathbf{t}}^{\mathbf{G}} \tilde{\mathbf{x}}_{\mathbf{t}}\right\|_{\mathrm{F}}. \end{aligned} \tag{22}$$

In i-th iteration, $\mathbf{D}^{(\mathbf{i})}$ can be expressed as

$$\mathbf{D}^{(i)} = \operatorname{diag}\left(\frac{1}{\left|\left(h_t^G\right)_1^{(i)}\right|^2+\delta}, \cdots, \frac{1}{\left|\left(h_t^G\right)_{G_{\mathrm{UE}}G_{\mathrm{BS}}}^{(i)}\right|^2+\delta}\right), \tag{23}$$

where $\delta > 0$ is a constant. The first item in $S^{(i)}$ denotes the sparsity of the estimated $\mathbf{h}_t^G$, the second one represents the residue vector. $\lambda > 0$ is the regularization parameter which controls the tradeoff between the channel sparsity and the matching error.

After get the on-grid coarse results by OMP, in the $i+1$-th iteration, we can update $\lambda, \mathbf{h}_t^G, \boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}$ alternately. Firstly, if the result of i-th iteration is matched with the measurement vector poorly, we will decrease λ to improve the sparsity. While if the result iteration is matched well, we will increase λ to pursue a better fitting. To be specific, λ can be updated by

$$\lambda = \min\left(d/r^{(i)}, \lambda_{max}\right), \tag{24}$$

where $r^{(i)} = \left\|\tilde{\mathbf{y}}_t - \boldsymbol{\Phi}_t(\boldsymbol{\Psi}')^{(i)}\left(\mathbf{h}_t^G\right)^{(i)}\tilde{\mathbf{x}}_t\right\|_{\mathrm{F}}$ is the residue. d and λ_{max} are two constants.

In the i-th iteration, if we make the partial derivative of (22) equal to zero, the optimal $\left(\mathbf{h}_t^G\right)^{(i)}$ can be obtained by

$$\begin{aligned}\left(\mathbf{h}_t^G\right)^{(i)}\left(\boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}\right) &= \arg\min_{\mathbf{h}_t^G} S^{(i)}\left(\mathbf{h}_t^G, \boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}\right) \\ &= \left(\lambda^{-1}\mathbf{D}^{(i)} + \sum_{p=1}^{N_{\mathrm{X}}}\mathbf{K}^H\mathbf{K}\right)^{-1}\left(\sum_{p=1}^{N_{\mathrm{X}}}\mathbf{K}^H\tilde{\mathbf{y}}_t\right),\end{aligned} \tag{25}$$

Then we also can get the corresponding optimal $S^{(i)}$ which can be expressed by a function of $\boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}$ as:

$$\begin{aligned}S_{\mathrm{opt}}^{(i)}\left(\boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}\right) &= \min_{\mathbf{h}_t^G} S^{(i)}\left(\mathbf{h}_t^G, \boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}\right) \\ &= -\left(\sum_{p=1}^{N_{\mathrm{X}}}\tilde{\mathbf{y}}_t^H\mathbf{K}\right)\cdot\left(\lambda^{-1}\mathbf{D}^{(i)} + \sum_{p=1}^{N_{\mathrm{X}}}\mathbf{K}^H\mathbf{K}\right)^{-1} \\ &\quad\cdot\left(\sum_{p=1}^{N_{\mathrm{X}}}\mathbf{K}^H\tilde{\mathbf{y}}_t\right) + \sum_{p=1}^{N_{\mathrm{X}}}\tilde{\mathbf{y}}_t^H\tilde{\mathbf{y}}_t,\end{aligned} \tag{26}$$

where $\mathbf{K} = \mathbf{A}_{\mathrm{UE}} diag\left(\mathbf{A}_{\mathrm{BS}}^{\mathrm{H}}\tilde{\mathbf{x}}_t\right)$. Then, we can use the gradient descent method to update $\boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}$:

$$\begin{aligned}\left(\boldsymbol{\delta}^{rx}\right)^{(i+1)} &= -\eta\cdot\nabla_{\boldsymbol{\delta}^{rx}} S_{\mathrm{opt}}^{(i)}\left(\left(\boldsymbol{\delta}^{tx}\right)^{(i)}, \left(\boldsymbol{\delta}^{rx}\right)^{(i)}\right), \\ \left(\boldsymbol{\delta}^{tx}\right)^{(i+1)} &= -\eta\cdot\nabla_{\boldsymbol{\delta}^{tx}} S_{\mathrm{opt}}^{(i)}\left(\left(\boldsymbol{\delta}^{tx}\right)^{(i)}, \left(\boldsymbol{\delta}^{rx}\right)^{(i)}\right),\end{aligned} \tag{27}$$

where η is the step-length.

The proposed algorithm can achieve super-resolution channel parameter estimations whose steps are detailed in Algorithm 1. Because we left out the delay part in (4), our proposed Algorithm 1 can only obtain the AOA and AOD of every MPC, which will rely on the necessity of the existence of a LoS path between the BS and the UE, in which case it is possible to perform the three-point positioning strategy to the UE with three BSs in the presence of only the angles.

Algorithm 1 is two-stage iterative process, for each MPC, Algorithm 1 need to obtain the coarse on-grid point and the off-grid parameter. The last period determines the overall computational complexity. The computational complexity in each iteration mainly depends on the gradient descent method step 12, which can be expressed as $O\left((M+N)(KL)^2\right)$.

Algorithm 1 Structured Perturbed Orthogonal Matching Pursuit

Input : $\tilde{\mathbf{y}}_t, \boldsymbol{\Phi}_t, \forall t \in [T], \varepsilon, \boldsymbol{\Psi}, KL, \tilde{\mathbf{x}}_t, h_{th}$;
Initialization : $\tilde{\mathbf{y}}_{t_\perp} = \tilde{\mathbf{y}}_t, k = 1, i = 1, \boldsymbol{\Lambda}_0 = \{\}, \mathbf{S}_0 = \{\}$;
1: **while** $k \le KL$ **do**
2: $\quad j' = \arg\max_j \sum_{t=1}^{T}\left(\boldsymbol{\Phi}_t\boldsymbol{\Psi}\right)_j^{\mathrm{T}}\left(\tilde{\mathbf{y}}_t - \tilde{\mathbf{h}}_t^G\mathbf{S}_k\right)$;
3: $\quad \mathbf{S}_k = \mathbf{S}_{k-1} \cup \left(\boldsymbol{\Phi}_t\boldsymbol{\Psi}\right)_{j'}$, $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}_{k-1} \cup \{j'\}$;
4: $\quad \tilde{\mathbf{h}}_t^G = \mathbf{S}_k^{\dagger}\tilde{\mathbf{y}}_t$;
5: $\quad k = k+1$
6: **end while**
7: **Output**: Coarse AoAs/AoDs results $\boldsymbol{\theta}^{rx}, \boldsymbol{\theta}^{tx} \leftarrow \boldsymbol{\Lambda}_k$;
8: Initialization $\left(\mathbf{h}_t^G\right)^{(0)} \leftarrow$ (21);
9: **while** $\left\|\left(\mathbf{h}_t^G\right)^{(i)} - \left(\mathbf{h}_t^G\right)^{(i-1)}\right\|_2 \ge \varepsilon$ **do**
10: $\quad$ update $\lambda \leftarrow$ (20);
11: $\quad$ construct $S_{\mathrm{opt}}^{(i)}\left(\boldsymbol{\delta}^{tx}, \boldsymbol{\delta}^{rx}\right) \leftarrow$ (22);
12: $\quad$ update $\left(\boldsymbol{\delta}^{rx}\right)^{(i+1)}, \left(\boldsymbol{\delta}^{tx}\right)^{(i+1)} \leftarrow$ (23);
13: $\quad$ update $\left(\mathbf{h}_t^G\right)^{(i+1)} \leftarrow$ (21);
14: $\quad$ Prune l-th path if $\left(h_{t,l}^G\right)^{(i+1)} < h_{th}$;
15: $\quad i = i+1$
16: **end while**
17: **Output** : $\mathbf{H}_t = \sum_l h_{t,l}^G\mathbf{a}_{\mathrm{UE}}\left(\theta_l^{rx} + \delta_l^{rx}\right)\mathbf{a}_{\mathrm{BS}}(\theta_l^{tx} + \delta_l^{tx})^{\mathrm{H}}$

B. Dynamic Two-stage Multi-cell user Grouping based on Super-resolution Estimation

In the first stage, every UE suppresses the inter-cell interference through the HADB technology, so that the UE's received SINR depends on the squared value of the minimum singular value of its joint interference channel matrix. Therefore, to maximize the received SINR, we use the grouping method based on the users' positions according to the fixed grouping threshold α. To be brief, we assign each UE into its suitable cell in terms of the interference caused by this UE to other cells if it is allocated to this cell. Assumed that the resulting user set of the b_i-th cell can be expressed as $\mathbf{C}^{b_i}$, which contains N_{b_i} UEs after the first stage.

In the second stage, we complete user grouping in each cell. For a single cell, it is necessary to improve the resource utilization and try to minimize the inter-user interference. Therefore, we group the UEs in each cell according to the channel correlation between different UEs. The UEs with low correlation will be assigned to the same group. In this way, UEs in a group can use the orthogonal sub-carriers to complete communication and positioning simultaneously and avoid the intra-group interference, while UEs in different groups can use the same sub-carriers to enhance the resource utilization with less inter-group interference.

Algorithm 2 Dynamic Two-stage Multi-cell user Grouping Algorithm based on Super-resolution Estimation

Input : B, U, R, α

Initialization : $g_{b_i}=1, N_{b_i}=0, \mathbf{C}^{b_i}=\{\}, \forall i \in [B]$,

for all $j, 1 \le j \le U$ **do**
  $\mathbf{T}_{b_i,u_j}=\mathbf{U}_{b_i,u_j}\mathbf{\Sigma}_{b_i,u_j}\mathbf{V}^{\mathrm{H}}_{b_i,u_j} \leftarrow \mathbf{H}_{b_i,u_j}, \forall i \in [B]$;
  $b_k \leftarrow$ the smallest singular value of $\mathbf{\Sigma}_{b_i,u_j}$;
  $\mathbf{C}^{b_k}=\mathbf{C}^{b_k} \cup \{u_j\}$;
  $N_{b_k}=N_{b_k}+1$;
  $u^{b_k}_{N_{b_k}}=u_j$;
end for
for all $i, 1 \le i \le B$ **do**
  $\mathbf{z}_{b_i,g_{b_i}}=\{u^{b_i}_1\}$, $N_{b_i,g_{b_i}}=1$;
  for all $l, 2 \le l \le N_{b_i}$ **do**
    $\rho\left(u^{b_i}_1, u^{b_i}_l\right) \leftarrow (34)$;
    if $\rho < \alpha$ **then**
      $\mathbf{z}_{b_i,g_{b_i}}=\mathbf{z}_{b_i,g_{b_i}} \cup u^{b_i}_l$;
      $N_{b_i,g_{b_i}}=N_{b_i,g_{b_i}}+1$;
    end if
    repeat
      if $\{\exists u^{b_i}_{n1}, u^{b_i}_{n2}, u^{b_i}_m \in \mathbf{z}_{b_i,g_{b_i}}, n_1 \ne n_2 \ne m\}$ & $\{\rho\left(u^{b_i}_m, u^{b_i}_{n1}\right) > \alpha\}$ & $\{\rho\left(u^{b_i}_m, u^{b_i}_{n2}\right) > \alpha\}$ **then**
        $\mathbf{z}_{b_i,g_{b_i}}=\mathbf{z}_{b_i,g_{b_i}} \backslash \{u^{b_i}_m\}$;
        $N_{b_i,g_{b_i}}=N_{b_i,g_{b_i}}-1$;
      end if
    until $\mathbf{z}_{b_i,g_{b_i}}$ is not changed
  end for
  $g_{b_i}=g_{b_i}+1$;
end for
Output : $\left[\mathbf{z}_{b_i,1}, ..., \mathbf{z}_{b_i,g_{b_i}}\right], \forall i \in [B]$;

After using the HADB technology to eliminate the co-channel interference in $\mathbf{H}^{\mathrm{H}}_{u_i}\mathbf{H}_{u_i} \in \mathbb{C}^{M\times M}$, we denote the CSI of the u_i-th UE as $\mathbf{\Gamma}_{u_i}$, which is obtained by

$$\mathbf{\Gamma}_{u_i}=\sum_{j=1}^{M}\lambda_{u_i,j}\boldsymbol{\xi}_{u_i,j}, \tag{28}$$

where $\lambda_{u_i,j}$ and $\boldsymbol{\xi}_{u_i,j}$ are the j-th eigenvalue and its corresponding eigenvector of $\mathbf{H}^{\mathrm{H}}_{u_i}\mathbf{H}_{u_i}$, respectively. The channel correlation between the u_j-th UE and u_k-th UE can be calculated by

$$\rho\left(u_j,u_k\right)=\frac{\left\|\left(\mathbf{\Gamma}^{\mathrm{H}}_{u_j}\mathbf{\Gamma}_{u_j}\right)^{\mathrm{H}}\left(\mathbf{\Gamma}^{\mathrm{H}}_{u_k}\mathbf{\Gamma}_{u_k}\right)\right\|_2}{\left\|\mathbf{\Gamma}^{\mathrm{H}}_{u_j}\mathbf{\Gamma}_{u_j}\right\|_2\left\|\mathbf{\Gamma}^{\mathrm{H}}_{u_k}\mathbf{\Gamma}_{u_k}\right\|_2}. \tag{29}$$

The steps of the proposed two-stage multi-cell user grouping algorithm are detailed in Algorithm 2.

According to Algorithm 2, in the first stage, each user is traversed once to calculate its channel interference matrix, in the second stage, each user is traversed once to calculate its channel correlation matrix. So the computational complexity of Algorithm 2 is $O\left(U^2M^2N^2\right)$.

C. Joint Resource Allocation Problem in Multi-cell Networks

In this section, based on the above proposed transmission scheme and our derived performance metrics for communication and positioning services. If only considering the communication services, the optimization sub-problems to allocate power and bandwidth to the u_j-th UE in b_i-th cell can be expressed as

$$\underset{B^{\mathrm{C}}_{b_i,g},P^{\mathrm{C}}_{b_i,u_j}}{\arg\max}\ R^{\mathrm{C}}_{b_i,u_j}, \tag{30a}$$

$$s.t. B^{\mathrm{C}}_{b_i,g} \ge B_{\mathrm{thr}}, P^{\mathrm{C}}_{b_i,u_j} \ge P_{\mathrm{thr}}, \tag{30b}$$

$$\forall b_i \in [C], \forall u_j \in [N_{b_i}], \forall g \in [g_{b_i}]; \tag{30c}$$

(30b) denote the lower bounds of communication resource allocated to each UE; the resource allocation optimization model for joint communication and positioning can be established based on the multi-cell user grouping results.

We develop the resource allocation scheme for a single cell, which can be directly applied to each cell in multi-cell systems, because UEs in different cells can use the same subcarriers to realize resource multiplexing. Considering a single group in the b_i-th cell, the grouping and allocating vectors can be expressed as

$$\mathbf{Z}=[\mathbf{z}_g], \forall g \in [g_{b_i}]; \tag{31a}$$

$$\mathbf{P}_{b_i}=\left[\mathbf{P}^{\mathrm{C}}_{b_i}, \mathbf{P}^{\mathrm{O}}_{b_i}\right] \tag{31b}$$

$$=\left[P^{\mathrm{C}}_{b_i,1}, ..., P^{\mathrm{C}}_{b_i,u_j}, P^{\mathrm{O}}_{b_i,1}, ..., P^{\mathrm{O}}_{b_i,u_j}\right], \forall u_j \in [N_{b_i}]; \tag{31c}$$

$$\mathbf{B}_{b_i}=\left[\mathbf{B}^{\mathrm{C}}_{b_i}, \mathbf{B}^{\mathrm{O}}_{b_i}\right] \tag{31d}$$

$$=\left[B^{\mathrm{C}}_{b_i,1}, ..., B^{\mathrm{C}}_{b_i,g}, B^{\mathrm{O}}_{b_i,1}, ..., B^{\mathrm{O}}_{b_i,g}\right], \forall g \in [g_{b_i}]; \tag{31e}$$

where (31a) denotes the grouping result of each cell obtained by Algorithm 2; (31c) is the vector of allocating the power for all UEs' different services; and (31d) is the vector of allocating the bandwidth for different groups.

In order to improve the cell's total comprehensive performance of communications and positioning services, we construct the joint optimization problem whose performance target is the sum of all UEs' received performance metrics in the two services by maxmizing the comprehensive performance of the links as

$$\arg\max_{\mathbf{P}_{b_i},\mathbf{B}_{b_i}} \sum_{\forall \mathbf{z_g} \in \mathbf{Z}} \sum_{\forall u_j \in \mathbf{z_g}} \left[a_{u_j}R^{\mathrm{C}}_{b_i,u_j}+a_{u_j}b^{O}_{u_j}R^{\mathrm{O}}_{b_i,u_j}\right], \tag{32a}$$

$$s.t. \sum_{\forall g \in [g_{b_i}]} B^{\mathrm{O}}_{b_i,g} \le b^{O}_{u_j,B}B_{\max}, \sum_{\forall g \in [g_{b_i}]} B^{\mathrm{C}}_{b_i,g} \le b^{C}_{u_j,B}B_{\max}; \tag{32b}$$

$$B^{\mathrm{C}}_{b_i,g} \ge B_{\mathrm{thr}}, B^{\mathrm{O}}_{b_i,g} \ge B_{\mathrm{thr}}, \forall g \in [g_{b_i}]; \tag{32c}$$

$$\sum_{\forall u_j \in [N_{b_i}]} P^{\mathrm{C}}_{b_i,u_j} \le b^{C}_{u_j,P}P_{\max}; \tag{32d}$$

$$\sum_{\forall u_j \in [N_{b_i}]} P^{\mathrm{O}}_{b_i,u_j} \le b^{O}_{u_j,P}P_{\max}; \tag{32e}$$

$$P^{\mathrm{C}}_{b_i,u_j} \ge P_{\mathrm{thr}}, P^{\mathrm{O}}_{b_i,u_j} \ge P_{\mathrm{thr}}, \forall u_j \in [N_{b_i}]; \tag{32f}$$

where $b^C_{u_j,P}$, $b^O_{u_j,P}$ and $b^C_{u_j,B}$, $b^O_{u_j,B}$ denote the weights of communication and positioning resource allocation for u_j-th user, respectively. Meanwhile, a_{u_j} is the weight of comprehensive performance for u_j-th user. It is clear that $b^C_{u_j,P}+b^O_{u_j,P}=1$, $b^C_{u_j,B}+b^O_{u_j,B}=1$ and $\sum_{\forall u_j\in[N_{b_i}]} a_{u_j}=1$. (32b), (32d) and (32e) mean that the power assigned to each UE in this cell for the two services should not be greater than the upper limit of power and bandwidth that can be allocated.

While guaranteeing the constraints containing (32b), (32d) and (32e) to be established, it is clear that the optimal solution of (32a) can be obtained when all usable power and bandwidth are used. λ_k^2 is related to the user's joint interference channel matrix which can be obtained by Algorithm 1 and the channel noise is a uniform modeling parameter in this section, hence assuming that both of λ_k^2 and σ^2 are known to us. So (32b), (32d) and (32e) can be exchanged to

$$\sum_{\forall g\in[g_{b_i}]} B^{\mathrm{O}}_{b_i,g}=b^O_{u_j,B}B_{\max},\ \sum_{\forall g\in[g_{b_i}]} B^{\mathrm{C}}_{b_i,g}=b^C_{u_j,B}B_{\max}; \tag{33a}$$

$$\sum_{\forall u_j\in[N_{b_i}]} P^{\mathrm{C}}_{b_i,u_j}=b^C_{u_j,P}P_{\max}; \tag{33b}$$

$$\sum_{\forall u_j\in[N_{b_i}]} P^{\mathrm{O}}_{b_i,u_j}=b^O_{u_j,P}P_{\max}; \tag{33c}$$

D. Solution Algorithm to the Joint Resource Allocation Problem

Solving (32) is equivalent to finding the optimal (suboptimal) solution of matrices' product under equation constraints. It belongs to linear 2-block variable product optimization problems, which is an NP-Hard problem and difficult to be solved by conventional optimization algorithms. In the optimization theory, ADMM algorithm is suitable in such a high dimension with the excellent convergence accuracy [37].

There are two groups of variables that are $\{\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{\mathrm{C}}_{b_i}\}$ and $\{\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{\mathrm{O}}_{b_i}\}$ in (32a). So the original problem can be directly solved by ADMM. At first, in the case of determining the two groups of variables, we can rewrite (32) in a standard form according to ADMM as

$$\arg\min_{\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{\mathrm{C}}_{b_i},\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{\mathrm{O}}_{b_i}} \quad -f\left(\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{\mathrm{C}}_{b_i}\right)-g\left(\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{\mathrm{O}}_{b_i}\right), \tag{34a}$$

$$s.t.\ \mathbf{1}^T\mathbf{P}^{\mathrm{C}}_{b_i}-b^C_{u_j,P}P_{\max}=0; \tag{34b}$$

$$\mathbf{1}^T\mathbf{P}^{O}_{b_i}-b^O_{u_j,P}P_{\max}=0; \tag{34c}$$

$$\mathbf{1}^T\mathbf{B}^{\mathrm{C}}_{b_i}-b^C_{u_j,B}B_{\max}=0; \tag{34d}$$

$$\mathbf{1}^T\mathbf{B}^{\mathrm{O}}_{b_i}-b^O_{u_j,B}B_{\max}=0; \tag{34e}$$

$$\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{C}_{b_i}\in S_C;\ \mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{O}_{b_i}\in S_O; \tag{34f}$$

where the feasible sets S_C, S_P for the two groups of variables can be expressed as

$$S_C=\{(\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{C}_{b_i})|\mathbf{P}^{\mathrm{C}}_{b_i}\geq P_{\mathrm{thr}}\cdot\mathbf{1},\mathbf{B}^{C}_{b_i}\geq B_{\mathrm{thr}}\cdot\mathbf{1}\}; \tag{35a}$$

$$S_O=\{(\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{O}_{b_i})|\mathbf{P}^{\mathrm{O}}_{b_i}\geq P_{\mathrm{thr}}\cdot\mathbf{1},\mathbf{B}^{O}_{b_i}\geq B_{\mathrm{thr}}\cdot\mathbf{1}\}; \tag{35b}$$

The augmented Lagrangian function corresponding to (30) is

$$\begin{aligned}
&L_A\left(\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{\mathrm{C}}_{b_i},\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{\mathrm{O}}_{b_i},\lambda_1,\lambda_2,\lambda_3,\lambda_4\right)=\\
&-f\left(\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{\mathrm{C}}_{b_i}\right)-g\left(\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{\mathrm{O}}_{b_i}\right)+\lambda_1\left(\mathbf{1}^T\mathbf{P}^{O}_{b_i}-b^O_{u_j,P}P_{\max}\right)\\
&+\lambda_2\left(\mathbf{1}^T\mathbf{P}^{C}_{b_i}-b^C_{u_j,P}P_{\max}\right)+\lambda_3\left(\mathbf{1}^T\mathbf{B}^{\mathrm{O}}_{b_i}-b^O_{u_j,B}B_{\max}\right)\\
&+\lambda_4\left(\mathbf{1}^T\mathbf{B}^{\mathrm{C}}_{b_i}-b^C_{u_j,B}B_{\max}\right)+\frac{\mu}{2}\left|\mathbf{1}^T\mathbf{P}^{\mathrm{C}}_{b_i}-b^C_{u_j,P}P_{\max}\right|^2\\
&+\frac{\mu}{2}\left|\mathbf{1}^T\mathbf{B}^{\mathrm{C}}_{b_i}-b^C_{u_j,B}B_{\max}\right|^2+\frac{\mu}{2}\left|\mathbf{1}^T\mathbf{B}^{\mathrm{O}}_{b_i}-b^O_{u_j,B}B_{\max}\right|^2\\
&+\frac{\mu}{2}\left|\mathbf{1}^T\mathbf{B}^{\mathrm{O}}_{b_i}-b^O_{u_j,B}B_{\max}\right|^2,
\end{aligned} \tag{36}$$

where λ_1, λ_2, λ_3 and λ_4 are the lagrange multipliers, $\mu>0$ is the penalty parameter. The last two items of (36) is the square regular term to ensure the algorithm convergence. We update the variables at the $(k+1)$-th iteration based on the framework of the ADMM to solve (34) as

$$\begin{aligned}
&\left(\mathbf{P}^{\mathrm{C}\,(k+1)}_{b_i},\mathbf{B}^{C\,(k+1)}_{b_i}\right)\\
&=\arg\min_{\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{C}_{b_i}\in S_C} L_A\left(\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{\mathrm{C}}_{b_i},\mathbf{P}^{\mathrm{O}\,(k)}_{b_i},\mathbf{B}^{\mathrm{O}\,(k)}_{b_i},\lambda_1^k,\lambda_2^k,\lambda_3^k,\lambda_4^k\right)\\
&\left(\mathbf{P}^{\mathrm{O}\,(k+1)}_{b_i},\mathbf{B}^{O\,(k+1)}_{b_i}\right)\\
&=\arg\min_{\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{O}_{b_i}\in S_O} L_A\left(\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{\mathrm{O}}_{b_i},\mathbf{P}^{\mathrm{C}\,(k+1)}_{b_i},\mathbf{B}^{\mathrm{C}\,(k+1)}_{b_i},\lambda_1^k,\lambda_2^k,\lambda_3^k,\lambda_4^k\right)
\end{aligned} \tag{37}$$

To obtain the suboptimal resource allocation scheme easilier, we can use Coordinate Descent to solve (37). To be specific, (37) can be exchanged to

$$\begin{aligned}
&\mathbf{P}^{\mathrm{C}\,(k+1)}_{b_i}=\\
&\arg\min_{\mathbf{P}^{\mathrm{C}}_{b_i}\in S_C} L_A\left(\mathbf{P}^{\mathrm{C}}_{b_i},\mathbf{B}^{\mathrm{C}\,(k)}_{b_i},\mathbf{P}^{\mathrm{O}\,(k)}_{b_i},\mathbf{B}^{\mathrm{O}\,(k)}_{b_i},\lambda_1^k,\lambda_2^k,\lambda_3^k,\lambda_4^k\right)\\
&\mathbf{B}^{C\,(k+1)}_{b_i}=\\
&\arg\min_{\mathbf{B}^{C}_{b_i}\in S_C} L_A\left(\mathbf{P}^{\mathrm{C}\,(k+1)}_{b_i},\mathbf{B}^{\mathrm{C}}_{b_i},\mathbf{P}^{\mathrm{O}\,(k)}_{b_i},\mathbf{B}^{\mathrm{O}\,(k)}_{b_i},\lambda_1^k,\lambda_2^k,\lambda_3^k,\lambda_4^k\right)\\
&\mathbf{P}^{\mathrm{O}\,(k+1)}_{b_i}=\\
&\arg\min_{\mathbf{P}^{\mathrm{O}}_{b_i}\in S_O} L_A\left(\mathbf{P}^{\mathrm{O}}_{b_i},\mathbf{B}^{\mathrm{O}\,(k)}_{b_i},\mathbf{P}^{\mathrm{C}\,(k+1)}_{b_i},\mathbf{B}^{\mathrm{C}\,(k+1)}_{b_i},\lambda_1^k,\lambda_2^k,\lambda_3^k,\lambda_4^k\right)\\
&\mathbf{B}^{O\,(k+1)}_{b_i}=\\
&\arg\min_{\mathbf{B}^{\mathrm{O}}_{b_i}\in S_O} L_A\left(\mathbf{P}^{\mathrm{O}\,(k+1)}_{b_i},\mathbf{B}^{\mathrm{O}}_{b_i},\mathbf{P}^{\mathrm{C}\,(k+1)}_{b_i},\mathbf{B}^{\mathrm{C}\,(k+1)}_{b_i},\lambda_1^k,\lambda_2^k,\lambda_3^k,\lambda_4^k\right)
\end{aligned} \tag{38}$$

If we relax the intra-group interference in the expression of SINR to a constant, it is clear that (38) are convex programs which can be solved efficiently by off-the-shelf solutions, such as CVX. If the termination condition of the algorithm which is expressed as

$$\begin{aligned}
&\left|f\left(\mathbf{P}^{\mathrm{C}\,(k+1)}_{b_i},\mathbf{B}^{\mathrm{C}\,(k+1)}_{b_i}\right)-f\left(\mathbf{P}^{\mathrm{C}\,(k)}_{b_i},\mathbf{B}^{\mathrm{C}\,(k)}_{b_i}\right)\right|\\
&+\left|g\left(\mathbf{P}^{\mathrm{O}\,(k+1)}_{b_i},\mathbf{B}^{\mathrm{O}\,(k+1)}_{b_i}\right)-g\left(\mathbf{P}^{\mathrm{O}\,(k)}_{b_i},\mathbf{B}^{\mathrm{O}\,(k)}_{b_i}\right)\right|\leq\varepsilon
\end{aligned} \tag{39}$$

is meet, the final solution can be obtained, where ε is the stop-

ping residual. Otherwise, we update the Lagrange multipliers after each iteration as:

$$\lambda_1^{k+1} = \lambda_1^k + \mu\left(\mathbf{1}^T\mathbf{P}_{b_i}^{\mathrm{O}} - b_{u_j,P}^{O}P_{\max}\right); \tag{40a}$$

$$\lambda_2^{k+1} = \lambda_2^k + \mu\left(\mathbf{1}^T\mathbf{P}_{b_i}^{\mathrm{C}} - b_{u_j,P}^{C}P_{\max}\right); \tag{40b}$$

$$\lambda_3^{k+1} = \lambda_3^k + \mu\left(\mathbf{1}^T\mathbf{B}_{b_i}^{\mathrm{O}} - b_{u_j,B}^{O}B_{\max}\right); \tag{40c}$$

$$\lambda_4^{k+1} = \lambda_4^k + \mu\left(\mathbf{1}^T\mathbf{B}_{b_i}^{\mathrm{C}} - b_{u_j,B}^{C}B_{\max}\right); \tag{40d}$$

The number of UEs in each group is random, to avoid the appearance of the case that the lower performance bounds are not met because of the insufficient preset power in the process of iteration, we additionally introduce a power revision operation as $P_{\max} = P_{\max} + V$, where V is an arbitrary constant. When each iteration ends, we reset the value of V and revise the total power to avoid insufficient performance until the solution of the joint optimization problem no longer changes or V is set to zero in multiple iterations.

Algorithm 3 Cyclic Iteration Algorithm based on ADMM to solve Joint Resouce Allocation Problem

Input : $U, C, \mathbf{Z}, P_{\max}, B_{\max}, P_{\mathrm{thr}}, B_{\mathrm{thr}}, \sigma_n^2, \lambda_k^2, \varepsilon, \mu, b_{u_j,B}^{C}, b_{u_j,B}^{O}, b_{u_j,P}^{C}, b_{u_j,P}^{O}, \mathbf{H}_{b_i,u_j}, \gamma_{b_i,u_j}, a_{u_j}, \forall b_i \in [C], \forall u_j \in [N_{b_i}]$

Initialization : $\mathbf{P}, \mathbf{B}, V,$

1: **for** each cell **do**
2: **repeat**
3: $L_A\left(\mathbf{P}_{b_i}^{\mathrm{C}}, \mathbf{B}_{b_i}^{\mathrm{C}}, \mathbf{P}_{b_i}^{\mathrm{O}}, \mathbf{B}_{b_i}^{\mathrm{O}}, \lambda_1, \lambda_2, \lambda_3, \lambda_4\right) \leftarrow$(36);
4: $\{\mathbf{P}_{b_i}^{\mathrm{C}}{}^{(k+1)}, \mathbf{B}_{b_i}^{\mathrm{C}}{}^{(k+1)}, \mathbf{P}_{b_i}^{\mathrm{O}}{}^{(k+1)}, \mathbf{B}_{b_i}^{O}{}^{(k+1)}\} \leftarrow$(38);
5: Update $\lambda_1^{k+1}, \lambda_2^{k+1}, \lambda_3^{k+1}, \lambda_4^{k+1} \leftarrow$(40);
6: **until** (39);
7: **end for**
8: **Output** : $[\mathbf{P}_{b_i}], [\mathbf{X}_{b_i}], \forall b_i \in [C]$;

Therefore, the solution in the above iteration process will approach the optimal solution of (30) gradually. The steps of the solution algorithm are detailed in Algorithm 3.

Assuming that Algorithm 3 get its convergence requires K iterations. In each iteration, it should 4 CVX solution processes. If the CVX solution uses the Interior Point method, the total computational complexity of Algorithm 3 is $O\left(U^{3.5}log(1/\varepsilon)K\right)$.

V. Simulation Results

To evaluate the performance of the proposed algorithms, we conduct the simulations based on the mmWave massive MIMO multi-cell systems with 3 adjacent cells whose radius are 30m deployed at a two-dimensional area of 120m×120m. There are 12 UEs with random distribution and unknown location and 1 BS with known position in each cell. The location of all equipments is shown in Fig.2, where the asterisk indicates the BS, the hexagonal area is the coverage of the corresponding BS with same color.

The simulation parameters are illustrated in Table 1. Meanwhile, the actual AOA-AOD is randomly generated in every Monte Carlo simulation. Considering a simplified system containing a BS and a UE, we use Algorithm 1 to obtain

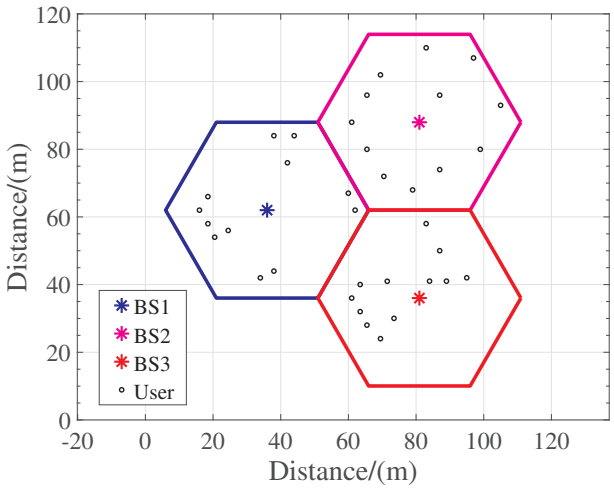

Fig. 2. Initial network geometry with 3 known BSs and 36 unknown UEs.

the multi-user CSI, and compare its performance against the traditional channel estimation algorithms, including the Minimum Mean Square Error Estimation (MMSE) and OMP. The channel estimation algorithms are evaluated by the Normalized Mean Square Error (NMSE), as $E\left(\frac{\|\mathbf{H}-\tilde{\mathbf{H}}\|_F^2}{\|\mathbf{H}\|_F^2}\right)$, where $\tilde{\mathbf{H}}$ and $\mathbf{H}$ denote the estimated channel and the actual channel, respectively.

TABLE I
SIMULTION PARAMETERS SETTINGS

Parameter	Configuration Value
Center Frequency	60GHz
Bandwidth	100MHz
Subcarrier Interval	150kHz
Number of BSs	3
Number of UEs	36
Antennas Number in BS	16, 18, 20
Antennas Number in UE	1,2
Number of Grid Points	10, 80, 300
Transmit Power	5:10:55W
Number of MPCs	16:2:30
Number of Snapshots	6
Stopping Residual	0.001

In Fig.3, we vary the Signal Noise Ratio (SNR) and evaluate the NMSE of the proposed algorithm. In addition to the above two traditional channel estimation algorithms, we also compare our proposed algorithm with a current state-of-the-art parameter estimation algorithm, ESPRIT. After we get ESPRIT's estimated AOAs, we can use Least Squares (LS) to obtain the estimated channel. We set $N = 1$ in this simulation because ESPRIT can only estimate the AOA. It is clear that Algorithm 1 outperforms other algorithms and the performance of the 4 algorithms always gradually improved with the increasing SNR. Compared with OMP, Algorithm 1 breaks the limit of resolution and achieves a large performance improvement by introducing a new parameter to represent the deviation between the actual sparse basis and the pre-designed

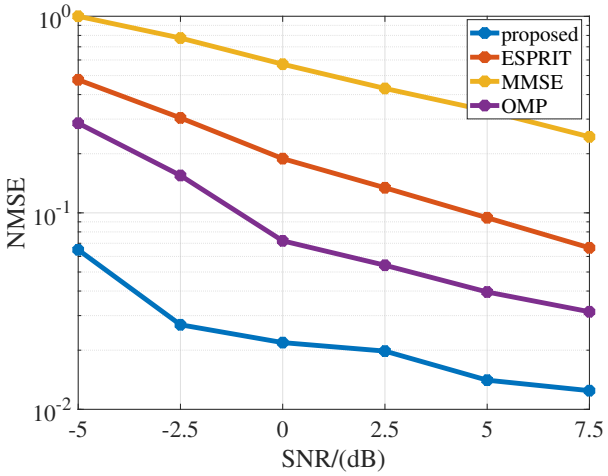

Fig. 3. Comparison of NMSE between different methods versus SNR.

grid.

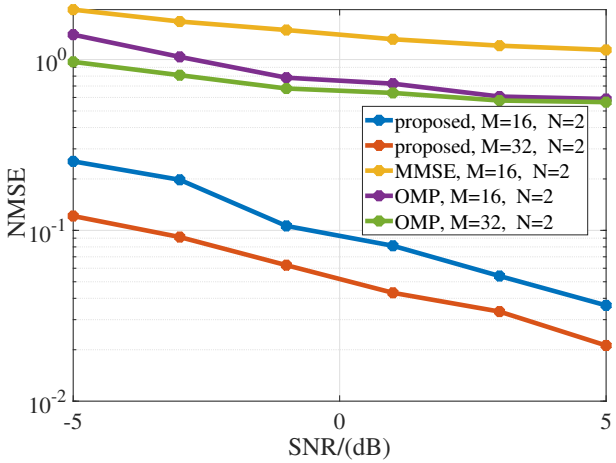

Fig. 4. Comparison of NMSE between different methods versus SNR and the number of antennas.

In the subsequent simulation, we all adopt the system with multiple antennas at the UEs' end. In Fig.4, we vary the SNR and evaluate the NMSE of the 3 algorithms with different numbers of antennas. With the same number of antennas, Algorithm 1 outperforms MMSE and OMP as shown in Fig.3. The increasing number of antennas can improve the performance of Algorithm 1 and OMP because of the orthogonality due to the multi-antenna gain.

In Fig.5, we vary the number of MPCs and grids and evaluate the NMSE of 3 algorithms with $M = 16$, $N = 2$ and $SNR = 5\text{dB}$. No matter what number of grids we set, Algorithm 1 outperforms MMSE and OMP and the performance of the other two algorithms always fluctuate slightly with the increasing number of MPCs. Because MMSE directly obtains the overall channel matrix which is independent to the number of MPCs. And OMP obtains the CSI of each MPC during multiple iterations, the number of MPCs will affect the number of iterations with less impact on final performance. The increasing number of grids can improve the performance of OMP at the very beginning because of the increasing grid resolution, however the off-grid effect would limit the performance gains. Algorithm 1's performance has not been improved significantly with the increasing number of grids because it breaks the limit of grid resolution. To reduce the overhead, the number of MPCs and grids ares set to 3 and 100 in our subsequent simulations, respectively.

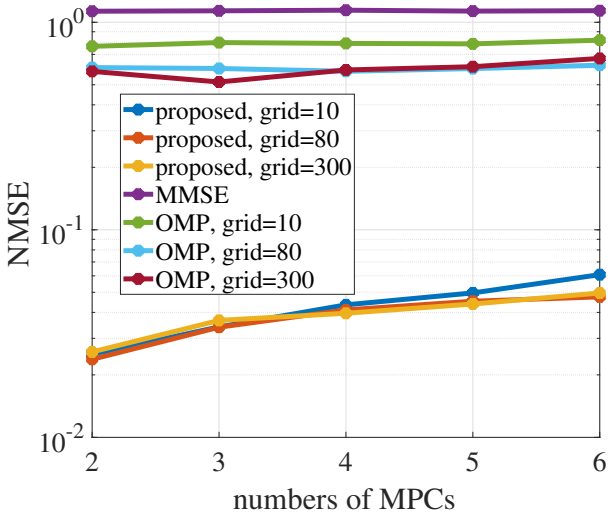

Fig. 5. Comparison of NMSE between different methods versus the number of grids and MPCs.

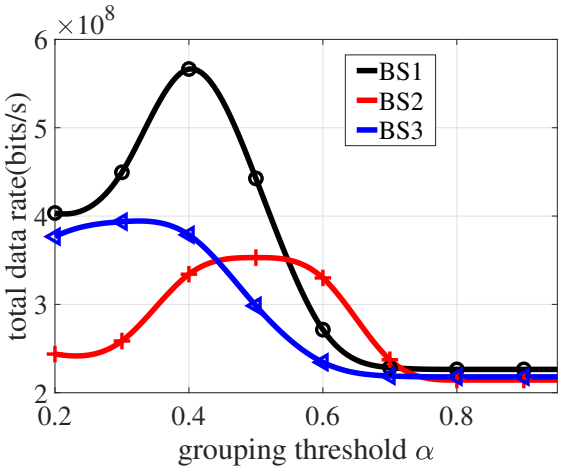

Fig. 6. Graph about total data rate of the proposed joint algorithm in different α.

Based on the different grouping results of Algorithm 2 with different grouping thresholds, we can use Algorithm 3 to allocate resource to each UE in each cell.

To study the influence of different grouping thresholds on the services, Fig.6 shows the communication performance while varying the grouping threshold α with $M = 32$, $N = 16$ and $Power = 5W$, because the expression of the positioning performance is similar to that of communication performance. All resources are equally allocated to every UE. In Fig.6, the six curves have the same trends. Regardless of the number of UEs in different cells, the optimal α is usually around 0.4, the system performance tends to increase and then decrease, while the grouping threshold α increasing. Because if α is too large or too small, the grouping result is that all UEs in one group or each UE in a separate group, neither of which is

optimal. In each cell, the increasing α causes the decreasing number of user groups. Before reaching the best performance, Algorithm 2 assigns the UEs with low channel correlation into the same group, thus reducing the intra-cell interference. After reaching the peak, the number of user groups decreases sharply with the increasing α and the performance decreases again under the condition of high intra-cell interference. Subsequent simulations set α in cells corresponding to BS1, BS2, and BS3 as 0.4, 0.3, and 0.5, respectively.

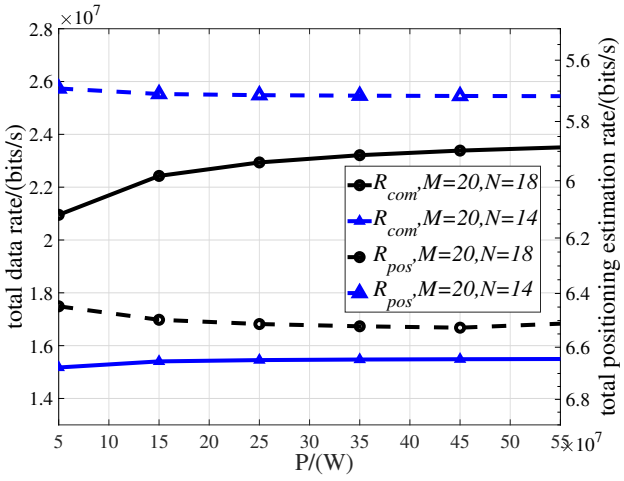

Fig. 7. Graph about total data rate and total positioning rate of the proposed joint algorithm in different transmit power and number of antennas.

After determining the threshold, in Fig.7, we vary the transmit power and evaluate the total communication and positioning performance at different numbers of antennas. With the increasing transmit power, the total data rate shows an upward trend and the total positioning rate shows a downward trend corresponding to improved positioning performance. With the increasing number of receiving antennas, the communication performance is enhanced by about from 10% to 20% under different transmit power and the positioning performance increases by from 8% to 10%. All of above phenomenas are due to the gain generated by multiple antenna arrays. Similar to Fig.4, the multi-antenna gain could enhance not only the accuracy of channel estimation, but also the joint communication and positioning performance.

To verify the validity of Algorithm 2 and Algorithm 3, Fig.8 shows the performance trade-off between communication and positioning of the proposed joint algorithm and two conventional resource allocation algorithms, including the Equal Bandwidth Allocation (EBA) + proposed Power Allocation results and Equal Bandwidth and Power Allocation (EBPA), under different user grouping thresholds and performance weights, $b_{u_j}^C$ and $b_{u_j}^O$, with $M = 20$, $N = 18$ and $P = 10W$, $B = 10MHz$. With the increasing rate, the corresponding positioning performance shows a decreasing trend due to the resource competition between the two services. With the same grouping threshold, the proposed joint algorithm obtains the best performance because its resource allocation scheme is more suitable to get the multi-user performance balance than other two resource allocation algorithm. And with the same resource allocation algorithm, the performance in $\alpha = 0.5$ is

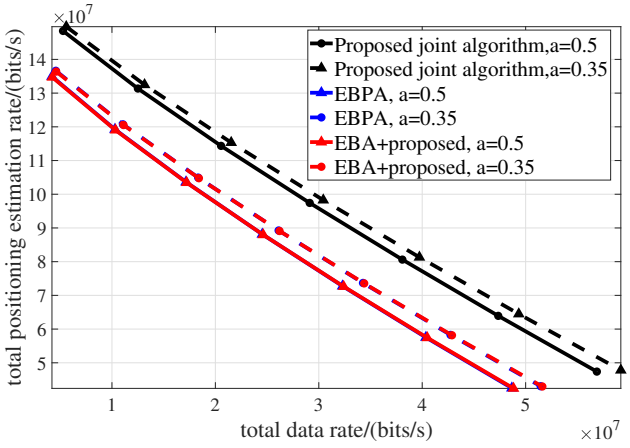

Fig. 8. Joint communication and positioning performance in different resource allocation and user grouping algorithms.

better than the performance in $\alpha = 0.35$ which is same as the phenomenon in Fig.6.

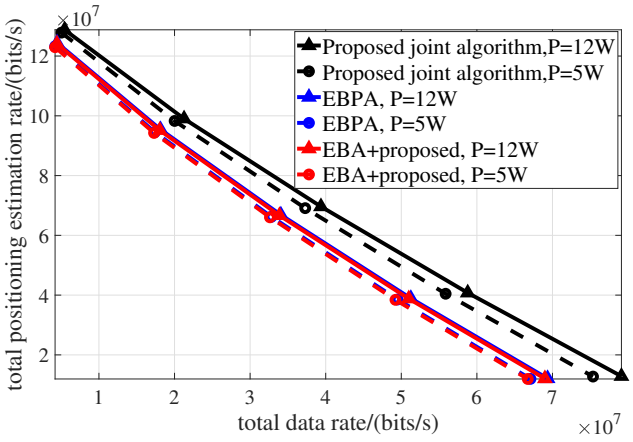

Fig. 9. Graph about the performance trade-off in different resource allocation algorithms and transmit power.

To prove the superiority of Algorithm 3, Fig.9 compares the performance trade-off of the 3 resource allocation schemes with Algorithm 2, at 2 different transmit power and performance weights with $M = 20$, $N = 18$. Compared with the other two algorithms, with the same data rate or positioning rate, the proposed joint algorithm has better positioning or communication performance, and with the growing data rate or positioning rate, the decline speed of its positioning or communication performance is also significantly slower. Because Algorithm 3 efficiently allocates power and bandwidth under performance constraints, instead of assigning the average resource to each user, thus performing well in both services. When the power changes from 5W to 12W, the joint performance of all algorithms are improved to a certain extent due to the sufficient power.

VI. Conclusion

In this paper, a joint dynamic two-stage user grouping and multi-dimensional resource allocation scheme based on super-resolution estimation is proposed, in mmWave massive MIMO systems considering multi-cell scenarios. Through the

proposed scheme and HADB technology, all users could share the same time-frequency resource blocks with the minimum interference and performance loss. Then we transfer the original problem into a joint optimization problem with multiple separable control variables, including sub-carriers and power. We apply ADMM which is suitable in high dimension conditions to solve such linear 2-block variable product optimization problem. The numeric simulation results demonstrate that the proposed joint algorithm obtains better comprehensive system overall throughput by efficiently exploiting wireless resources and multi-cell cooperation than the traditional joint algorithms. Future works might concern that the extension to the uniform planar array (UPA) equipments and the scalability of proposed strategy to multi-cell systems with multiple base stations in a single cell. In addition, the research of estimating AOA-AOD-delay makes it possible to use a single BS for positioning without los path which could be taken into account.

REFERENCES

[1] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 2, no. 3, pp. 194–220, 2017.

[2] J. Thomas, J. Welde, G. Loianno, K. Daniilidis, and V. Kumar, "Autonomous flight for detection, localization, and tracking of moving targets with a small quadrotor," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1762–1769, 2017.

[3] S. G. Nagarajan, P. Zhang, and I. Nevat, "Geo-spatial location estimation for internet of things (iot) networks with one-way time-of-arrival via stochastic censoring," *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 205–214, 2016.

[4] M. Z. Win, F. Meyer, Z. Liu, W. Dai, S. Bartoletti, and A. Conti, "Efficient multisensor localization for the internet of things: Exploring a new class of scalable localization algorithms," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 153–167, 2018.

[5] R. Estrada, R. Mizouni, H. Otrok, A. Ouali, and J. Bentahar, "A crowdsensing framework for allocation of time-constrained and location-based tasks," *IEEE Transactions on Services Computing*, vol. 13, no. 5, pp. 769–785, 2017.

[6] F. Liu, Y. Cui, C. Masouros, J. Xu, T. X. Han, Y. C. Eldar, and S. Buzzi, "Integrated sensing and communications: Towards dual-functional wireless networks for 6g and beyond," *IEEE journal on selected areas in communications*, 2022.

[7] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," *IEEE network*, vol. 34, no. 3, pp. 134–142, 2019.

[8] M. Feng and S. Mao, "Harvest the potential of massive mimo with multi-layer techniques," *IEEE Network*, vol. 30, no. 5, pp. 40–45, 2016.

[9] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[10] M. Xiao, S. Mumtaz, Y. Huang, L. Dai, Y. Li, M. Matthaiou, G. K. Karagiannidis, E. Björnson, K. Yang, I. Chih-Lin *et al.*, "Millimeter wave communications for future mobile networks (guest editorial), part i," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 7, pp. 1425–1431, 2017.

[11] E. G. Larsson, T. Marzetta, F. Rusek *et al.*, "Scaling up mimo: Opportunities and challenges with very large arrays," *Sig Process Mag IEEE*, vol. 30, no. 1, pp. 40–60, 2013.

[12] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive mimo: Benefits and challenges," *IEEE journal of selected topics in signal processing*, vol. 8, no. 5, pp. 742–758, 2014.

[13] A. Sakhnini, M. Guenach, A. Bourdoux, and S. Pollin, "A cramér-rao lower bound for analyzing the localization performance of a multistatic joint radar-communication system," in *2021 1st IEEE International Online Symposium on Joint Communications & Sensing (JC&S)*. IEEE, 2021, pp. 1–5.

[14] X. Yang, S. Zhang, B. Gao, and J. Cao, "A low complexity joint user grouping and resource allocation algorithm in massive mimo systems," in *2019 IEEE 19th International Conference on Communication Technology (ICCT)*. IEEE, 2019, pp. 914–919.

[15] M. Alkhaled, E. Alsusa, and W. Pramudito, "Adaptive user grouping algorithm for the downlink massive mimo systems," in *2016 IEEE Wireless Communications and Networking Conference*. IEEE, 2016, pp. 1–6.

[16] Y. Bao, B. Yu, B. Yin, X. Luo, and X. Lu, "Resource allocation for joint communication and localization systems with mu-mimo," *IEEE Access*, vol. 10, pp. 124 649–124 662, 2022.

[17] G. Destino and H. Wymeersch, "On the trade-off between positioning and data rate for mm-wave communication," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2017, pp. 797–802.

[18] D. Kumar, J. Saloranta, G. Destino, and A. Tölli, "On trade-off between 5g positioning and mmwave communication in a multi-user scenario," in *2018 8th International Conference on Localization and GNSS (ICL-GNSS)*. IEEE, 2018, pp. 1–5.

[19] X. Lu, K. Yang, N. Fan, H. Guo, and H. Zhang, "Joint clustering of users and resources for multi-cell vmimo-sc-fdma uplink systems," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1417–1430, 2018.

[20] B. Yu, Y. Bao, K. Cheng, R. Chen, and X. Lu, "Resource virtualization and allocation for tdd-f-ofdm systems with mu-mimo," *IEEE Access*, vol. 8, pp. 219 047–219 061, 2020.

[21] S.-M. Tseng and Y.-F. Chen, "Average psnr optimized cross layer user grouping and resource allocation for uplink mu-mimo ofdma video communications," *IEEE Access*, vol. 6, pp. 50 559–50 571, 2018.

[22] Z. Wang, K. Han, X. Shen, W. Yuan, and F. Liu, "Achieving the performance bounds for sensing and communications in perceptive networks: Optimal bandwidth allocation," *IEEE Wireless Communications Letters*, vol. 11, no. 9, pp. 1835–1839, 2022.

[23] S. Pan, Y. Yan, K. A. Bonsu, and W. Zhou, "Resource allocation algorithm for mu-mimo systems with double-objective optimization under the existence of the rank deficient channel matrix," *IEEE Access*, vol. 7, pp. 61 307–61 319, 2019.

[24] X. Lu, Q. Ni, W. Li, and H. Zhang, "Dynamic user grouping and joint resource allocation with multi-cell cooperation for uplink virtual mimo systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3854–3869, 2017.

[25] C. Hu, L. Dai, T. Mir, Z. Gao, and J. Fang, "Super-resolution channel estimation for mmwave massive mimo with hybrid precoding," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8954–8958, 2018.

[26] A. Liao and Z. Gao, "Super-resolution channel estimation for mmwave massive mimo," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–5.

[27] Z. Guo, X. Wang, and W. Heng, "Millimeter-wave channel estimation based on 2-d beamspace music method," *IEEE transactions on wireless communications*, vol. 16, no. 8, pp. 5384–5394, 2017.

[28] K. Yu, M. Shen, R. Wang, and Y. He, "Joint nuclear norm and l1–2-regularization sparse channel estimation for mmwave massive mimo systems," *IEEE Access*, vol. 8, pp. 155 409–155 416, 2020.

[29] O. Teke, A. C. Gurbuz, and O. Arikan, "Perturbed orthogonal matching pursuit," *IEEE Transactions on Signal Processing*, vol. 61, no. 24, pp. 6220–6231, 2013.

[30] C. K. Anjinappa, A. C. Gurbuz, Y. Yapici, and I. Guvenc, "Off-grid aware channel and covariance estimation in mmwave networks," *IEEE Transactions on Communications*, vol. 68, no. 6, pp. 3908–3921, 2020.

[31] C. Qin, S. Zeng, C. Wang, D. Pan, W. Wang, and Y. Zhang, "A distributed interference alignment approach based on grouping in heterogeneous network," *IEEE Access*, vol. 6, pp. 2484–2495, 2017.

[32] J. Lee, G.-T. Gil, and Y. H. Lee, "Channel estimation via orthogonal matching pursuit for hybrid mimo systems in millimeter wave communications," *IEEE Transactions on Communications*, vol. 64, no. 6, pp. 2370–2386, 2016.

[33] R. Koirala, B. Denis, B. Uguen, D. Dardari, and H. Wymeersch, "Localization and throughput trade-off in a multi-user multi-carrier mmwave system," *IEEE Access*, vol. 7, pp. 167 099–167 112, 2019.

[34] A. R. Chiriyath, B. Paul, G. M. Jacyna, and D. W. Bliss, "Inner bounds on performance of radar and communications co-existence," *IEEE Transactions on Signal Processing*, vol. 64, no. 2, pp. 464–474, 2016.

[35] C. Shi, D. Xu, Y. Zhou, and W. Tu, "Range-doa information and scattering information in phased-array radar," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, 2019, pp. 747–752.

[36] C. Hu, L. Dai, T. Mir, Z. Gao, and J. Fang, "Super-resolution channel estimation for mmwave massive mimo with hybrid precoding," *IEEE*

Transactions on Vehicular Technology, vol. 67, no. 9, pp. 8954–8958, 2018.

[37] L. Chen, D. Sun, and K. C. Toh, "A note on the convergence of admm for linearly constrained convex optimization problems," *Computational Optimization and Applications*, vol. 66, pp. 327–343, 2017.

Xueni Luo was born in Hunan, China. She received the B.Sc. degree in communication engineering from Xidian University, Xi'an, China, in 2021. She is currently pursuing the M.Sc. degree in communication and information system with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an. Her research interests include OFDM, mmWave, and resource allocation of wireless communications.

Xiaofeng Lu received the B.Sc. degree from Sichuan University, Chengdu, China, in 1996, the M.Sc. degree from Hunan University, Changsha, China, in 1999, and the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2006. From 1999 to 2003, he was a Research and Development Engineer with the Wuhan Research Institute of Post and Telecommunications. He is currently a Full Professor with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, China. His main research interests lie in the area of broadband wireless communications, covering topics, such as joint communication and positioning and MU-MIMO.

Boyu Jin was born in Henan, China. He received the B.Sc. degree in communication engineering from Zhengzhou University, Zhengzhou, China, in 2022. He is currently pursuing the M.Sc. degree in communication and information system with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an. His research interests include channel estimation and resource allocation of wireless communications.

Benquan Yin was born in Anhui, China. He received the B.Sc. degree in optoelectronic information science and technology from Xidian University, Xi'an, China, in 2020. He received the M.Sc. degree in communication and information system with the State Key Laboratory of Integrated Service Networks, Xidian University, Xi'an, China, in 2023. His research interests include millimeter wave, virtual MIMO, and resource allocation of wireless communications.

Kun Yang (F'23) received his PhD from the Department of Electronic & Electrical Engineering of University College London (UCL), UK. He is currently a Chair Professor in the School of Computer Science & Electronic Engineering, University of Essex, UK, leading the Network Convergence Laboratory (NCL). He is also an affiliated professor of Nanjing University, China. His main research interests include wireless networks and communications, future Internet and edge computing. In particular he is interested in energy aspects of future communication systems such as 6G and new AI (artificial intelligence) technique for wireless. He has managed research projects funded by UK EPSRC, EU FP7/H2020, and industries. He has published 400+ papers and filed 30 patents. He serves on the editorial boards of a number of IEEE journals (e.g., IEEE TNSE, TVT, WCL). He is a Deputy Editor-in-Chief of IET Smart Cities Journal. He has been a Judge of GSMA GLOMO Award at World Mobile Congress – Barcelona since 2019. He was a Distinguished Lecturer of IEEE ComSoc (2020-2021). He is a Member of Academia Europaea (MAE), a Fellow of IEEE, a Fellow of IET and a Distinguished Member of ACM.