Correspondence

Scheduling for Downlink OFDMA With IRS Reconfiguration Constraints

Alberto Rech, *Member, IEEE*,
Leonardo Badia, *Senior Member, IEEE*,
and Stefano Tomasin, *Senior Member, IEEE*

***Abstract*—The technical limitations of the intelligent reflecting surface (IRS) (re)configurations in terms of both communication overhead and energy efficiency must be considered when IRSs are used in cellular networks. In this paper, we investigate the downlink time-frequency scheduling of an IRS-assisted multi-user system in the orthogonal frequency-division multiple access (OFDMA) framework wherein both the set of possible IRS configurations and the number of IRS reconfigurations within a time frame are limited. We formulate the sum rate maximization problem as a non-polynomial (NP)-complete generalized multi-knapsack problem. A heuristic greedy algorithm for the joint IRS configuration and time-frequency scheduling is also proposed. Numerical simulations prove the effectiveness of our greedy solution.**

***Index Terms*—Intelligent reflecting surfaces; millimeter wave communication; orthogonal frequency-division multiple access.**

I. Introduction

Intelligent reflecting surfaces (IRSs) consist of meta-surfaces with radiating elements that can passively tune the phase shift of incoming signals to collectively reflect them in the desired propagation direction without active amplification [1]. They are a promising solution to enhance network coverage, especially in the context of millimeter wave (mmWave) communications.

Downlink scheduling solutions for IRS-assisted communications have been extensively studied for cellular networks under several implementation constraints. Dynamic optimization schemes adjusting IRS configurations over each time slot have been explored in [2], [3]. Instead, the authors of [4] consider a 2-user downlink transmission in a IRS-aided scenario over fading channels, comparing results of different basic orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA) schemes. The study reveals that while NOMA appears to be the best solution, time division multiple access (TDMA) outperforms frequency division multiple access due to the lack of frequency selective beamforming capabilities at the IRS. Additionally, the performance of NOMA scheduling solutions, including rate-splitting multiple access, has been evaluated in [5], [6].

Nevertheless, the majority of the literature on IRSs relies on problematic premises. Specifically, the assumption of an ideal control channel with the base station is prevalent in the literature, while actual deployments are expected to have wireless, error-prone IRS control channels, possibly implemented with low-cost technologies [7]. This introduces constraints on the IRS reconfiguration period, which results in synchronization issues and increased power consumption [8]. Indeed, early IRS prototypes display non-negligible phase-shift reconfiguration times [9], [10]. Such overhead increases with the size of the IRS, and it is expected to become a serious issue with the extremely large IRSs needed to overcome channel losses in harsh propagation environments [11], [12]. In this context, it is crucial to design resource allocation algorithms that mitigate the limitations imposed by the constrained IRS reconfigurations. This kind of constraint has been studied in [13], with a characterization of both OMA and NOMA schemes in a 2-user IRS-aided single input single output system with Rayleigh fading channels.

However, the main issue of NOMA is complex signal processing to perform interference cancellation at the receiver side. Therefore, NOMA is considered more appealing in uplink than downlink. Consequently, and in line with the current 3rd Generation Partnership Project (3GPP) frame structure [14], a TDMA scheduler for multi-user multiple input multiple output (MIMO) combined with user aggregation (i.e., clustering) techniques can optimize the system sum rate or the user scheduling fairness [15].

In this paper, we propose a novel OMA scheduling policy for an IRS-aided downlink communication system. For downlink orthogonal frequency-division multiple access (OFDMA) cellular transmissions, we adopt a joint resource allocation and IRS configuration to maximize the system sum rate. The communication and energy overhead of IRS reconfiguration is constrained by limiting the number of reconfigurations within each scheduling period. This forces the reuse of the same configurations for multiple users [16]. Moreover, we consider the case where the IRS configuration can only be chosen within a *codebook* of configurations to further reduce the control overhead [17]. To the best of our knowledge, this is the first study to address OFDMA scheduling with IRS while incorporating such configuration constraints, thereby offering a comprehensive and scalable solution to the challenges posed by multi-user environments in IRS-assisted communication systems. We formulate the sum rate maximization as a non-polynomial (NP) complete generalized multi-knapsack problem and propose a greedy algorithm for the joint IRS configuration and time-frequency scheduling. Numerical simulations prove the effectiveness of our solution. [1]

II. System Model

We consider the downlink transmission of a cellular system shown in Fig. 1, where the transmission from the next generation Node Base

Manuscript received 6 January 2024; revised 15 March 2024 and 21 May 2024; accepted 8 July 2024. This work was supported in part through the Project funded under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.3-Call for tender Number 341 of March 15, 2022, of Italian Ministry of University and Research, and in part by the European Union – NextGenerationEU Project code MUR: PE_0000001 Concession Decree Number 1549 of 11/10/2022 adopted by the Italian Ministry of University and Research, CUP C93C22005250001 Project title RESearch and innovation on future Telecommunications systems and networks, to make Italy more smART (RESTART). The review of this article was coordinated by Prof. Yiqing Zhou. *(Corresponding author: Alberto Rech.)*

Alberto Rech is with the Department of Information Engineering, University of Padova, 35122 Padova PD, Italy, and also with Paris Research Center, 92100 Huawei, France (e-mail: alberto.rech@huawei.com).

Leonardo Badia and Stefano Tomasin are with the Department of Information Engineering, University of Padova, 35122 Padova PD, Italy.

Digital Object Identifier 10.1109/TVT.2024.3426621

[1] *Notation.* Scalars are denoted by italic letters; vectors and matrices by boldface lowercase and uppercase letters, respectively; sets are denoted by calligraphic uppercase letters. $\boldsymbol{A}^{\mathrm{T}}$ and $\boldsymbol{A}^{\dagger}$ denote the transpose and the conjugate transpose of matrix $\boldsymbol{A}$, respectively. $\mathrm{diag}(\boldsymbol{a})$ indicates a square diagonal matrix with the elements of $\boldsymbol{a}$ on the principal diagonal. The imaginary unit is $j = \sqrt{-1}$. Finally, $\mathbb{E}[\cdot]$ denotes statistical expectation.

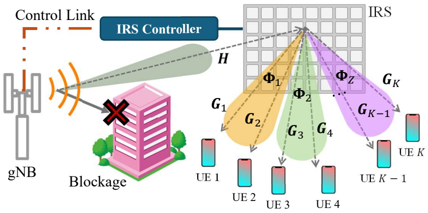


Fig. 1. OFDMA scheduling for IRS-assisted multi-UE communication.

(gNB) to K UEs is assisted by an IRS. The gNB and each UEs are equipped with N_g and N_U antennas, respectively. We assume that the direct links between the gNB and the UEs are unavailable due to a deep blockage, therefore the gNB transmits signals to the UEs by only exploiting the IRS cascade channels. The IRS configuration is managed by the gNB though the IRS controller, by exploiting a dedicated link between the gNB and the IRS.

The gNB schedules the UEs in the time-frequency domain by allocating resource blocks (RBs) from a grid of K RBs. We consider K as an integer multiple of F, with one UE scheduled per RB, as in our scenario we expect that the gNB-IRS channel has rank one, i.e., it has a single dominant path: this makes spatial multiplexing unfeasible. However, when higher-rank channels are available between the IRS and the gNB, spatial multiplexing can be used and our approach can be suitably modified to accommodate this scenario. The investigation of this point is left for future work.

Let f_i be the carrier frequency of the RB identified by frequency index i and an arbitrary time slot, and let $\mathcal{F} = \{f_i, i = 1, \ldots, F\}$ be the set of all carrier frequencies. We assume that UEs are either static or moving slowly, which is the most typical application scenario for IRS-aided networks. Therefore, we assume that once the perfect channel estimation of all UEs is acquired at the gNB at the beginning of each frame, the channels remain constant for its duration. We assume the gNB knows the cascade channel to the UEs for any IRS configuration.

IRS Model & Beamforming Codebook. Each element of the IRS acts as an omnidirectional antenna element that reflects the impinging EM field, introducing a tunable phase shift on the baseband-equivalent signal. We denote with $\phi_n = e^{j\theta_n}$ the reflection coefficient of the n-th IRS element, where $\theta_n \in \left\{0, \frac{2\pi}{2^{b_\text{I}}}, \ldots, \frac{2\pi(2^{b_\text{I}}-1)}{2^{b_\text{I}}}\right\}$ is the induced phase shift, with b_I-bits quantization. The *IRS configuration* is then defined as $\boldsymbol{\Phi} = \text{diag}(\phi_1, \ldots, \phi_{N_\text{I}})$.

To further reduce the overhead of the IRS configuration, we consider a codebook $\mathcal{C}_{\boldsymbol{\Phi}}$ from which matrix $\boldsymbol{\Phi}$ is chosen. This discrete design is compliant with the currently standardized initial access framework [18]. A large variety of codebooks for both near-field and far-field communication have been discussed in the literature, and the evaluation of their impact on system performance is out of the scope of this work. In this paper, we considered a simple design of *cell-specific codebook*, derived from the channel measurements in the cell (more details are given in Section III-C).

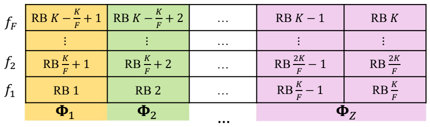
Transmission model. With reference to carrier i, we denote with $\boldsymbol{H}(f_i) \in \mathbb{C}^{N_\text{I}\times N_\text{g}}$ the gNB-IRS channel matrix and with $\boldsymbol{G}_k(f_i) \in \mathbb{C}^{N_\text{U}\times N_\text{I}}$ the channel matrix of the link between the IRS and UE k. We consider single-stream transmissions, where the gNB uses the beamforming vector $\boldsymbol{w}_\text{g} \in \mathbb{C}^{N_\text{g}\times 1}$. Note that this assumption matches the IRS-aided mmWave scenario, where the cascade channel rank is insufficient to perform multi-stream transmissions [19], [20]. Let $\boldsymbol{s}_k$ be the signal transmitted by the gNB to UE k assigned to carrier i, the received signal can be expressed as

$$z_k = \boldsymbol{v}_k^\text{T} \boldsymbol{G}_k(f_i)\boldsymbol{\Phi}\boldsymbol{H}(f_i)\boldsymbol{w}_\text{g}\boldsymbol{s}_k + \boldsymbol{v}_k^\text{T}\boldsymbol{n}_k, \tag{1}$$

where $\boldsymbol{v}_k \in \mathbb{C}^{N_\text{U}\times 1}$ is the beamforming vector at UE k, $\boldsymbol{\Phi}$ is the IRS configuration, and $\boldsymbol{n}_k$ is the circularly symmetric complex Gaussian noise vector with independent entries having zero-mean and variance σ_n^2. We assume that the gNB beamforms the signal toward the gNB-IRS line-of-sight (LoS) angle and, once the IRS configuration is fixed. Then, the beamformer at the UE matches its cascade channel, i.e., $\boldsymbol{v}_k$ is the singular vector corresponding to the largest singular value of $[\boldsymbol{G}_k(f_i)\boldsymbol{\Phi}\boldsymbol{H}(f_i)]^\dagger$. UE k attains the achievable rate

$$R_k(\boldsymbol{\Phi}, f_i) = \log_2\left(1 + \frac{|\boldsymbol{v}_k^\text{T}\boldsymbol{G}_k(f_i)\boldsymbol{\Phi}\boldsymbol{H}(f_i)\boldsymbol{w}_\text{g}|^2\sigma_s^2}{|\boldsymbol{v}_k^\text{T}|^2\sigma_n^2}\right), \tag{2}$$

where σ_s^2 and σ_n^2 are the transmit and noise power, respectively. Note that for the sake of a simpler explanation, we assume that the IRS introduces the same phase shift at all frequencies, thus $\boldsymbol{\Phi}$ does not depend on f_i. However, in the following, we will only use the value of achievable rates $R_k(\boldsymbol{\Phi}, f_i)$ for different values of $\boldsymbol{\Phi}$, k, and i, and a frequency-dependent behavior of the IRS can be accommodated in our model by changing (2).



f_F	RB $K-\frac{K}{F}+1$	RB $K-\frac{K}{F}+2$	…	RB $K-1$	RB K
	⋮	⋮	…	⋮	⋮
f_2	RB $\frac{K}{F}+1$	RB $\frac{K}{F}+2$	…	RB $\frac{2K}{F}-1$	RB $\frac{2K}{F}$
f_1	RB 1	RB 2	…	RB $\frac{K}{F}-1$	RB $\frac{K}{F}$
	$\boldsymbol{\Phi}_1$	$\boldsymbol{\Phi}_2$	…	$\boldsymbol{\Phi}_Z$	

Fig. 2. Example of resource grid with cluster configuration assignment.

III. Configuration and User Scheduling

In general, different IRS configurations should be adopted for each UE to maximize its achievable rate (2) based on its position in the cell and on the channel conditions. However, $\boldsymbol{\Phi}$ is not frequency-selective, i.e., its configuration is the same at each RB in the same time slot. Moreover, we here impose a constraint on the number of IRS reconfigurations per time frame, to limit the number of reconfigurations and reduce the communication overhead. This may also account for practical limitations that might arise in realistic deployments. However, such constraints usually lead to an achievable rate degradation as sub-optimal IRS configurations could be adopted to serve some UEs.

In detail, we formulate a constrained discrete optimization problem limiting the re-configurations within a time frame to a maximum number of $Z \leq K/F$. Within this time frame the gNB serves the K UEs by splitting them into Z disjoint subsets (or clusters) $\mathcal{U}_1, \ldots, \mathcal{U}_Z$, each with cardinality $\alpha_z F$, with $\alpha_z \in \mathbb{N}$, for $z = 1, \ldots, Z$. While serving all the UEs in subset $\mathcal{U}_z$ the IRS configuration is kept fixed to $\boldsymbol{\Phi}_z$. Fig. 2 displays the considered resource grid with an example of IRS configuration assignment for for $Z < K/F$, $|\mathcal{U}_1| = |\mathcal{U}_2| = F$ and $|\mathcal{U}_Z| = 2F$. Note that, if the codebook is small, several sets of UEs could share the optimal configuration; in such a case, the clusters are merged, and the number of reconfigurations is further reduced.

Now, we must decide a) which IRS configuration is used for each of the Z clusters (choice of $\boldsymbol{\Phi}_z$), b) for how many slots each configuration is used (α_z), and c) how UEs are assigned to RBs (choice of $x_{k,z,i}$). In formulas, let $\mathcal{S}_i$, $i = 1, \ldots, F$, be the set of UEs assigned to carrier i. Also, define the assignment variables

$$x_{k,z,i} = \begin{cases} 1 & \text{if } k \in \mathcal{U}_z \cap \mathcal{S}_i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The joint resource allocation and configuration optimization problem can be stated as the following generalized assignment

$$\max_{\substack{\boldsymbol{\Phi}_z, \alpha_z \\ x_{k,z,i}}} \sum_{z=1}^{Z}\sum_{k=1}^{K}\sum_{i=1}^{F} R_k(\boldsymbol{\Phi}_z, f_i)x_{k,z,i} \tag{4a}$$

$$\text{s.t.}\boldsymbol{\Phi}_z \in \mathcal{C}_{\boldsymbol{\Phi}}, \tag{4b}$$

$$x_{k,z,i} \in \{0,1\}, \quad \forall k,z,i, \tag{4c}$$

$$\textstyle\sum_{z=1}^{Z}\sum_{i=1}^{F} x_{k,z,i} = 1 \quad \forall k, \tag{4d}$$

$$\textstyle\sum_{k=1}^{K}\sum_{i=1}^{F} x_{k,z,i} = \alpha_z F \quad \forall z, \tag{4e}$$

where constraint (4b) imposes the IRS configurations to be chosen within codebook $\mathcal{C}_{\boldsymbol{\Phi}}$, and (4c)-(4d) denote the assignment of each UE to a unique RB. Instead, constraint (4e) imposes the cluster cardinalities as an integer multiple of F, reflecting the frequency non-selectivity of the IRS configuration. Due to (4d)-(4e), $\sum_{z=1}^{Z}\alpha_z = K/F$, and α_z is the number of time slots for which $\boldsymbol{\Phi}_z$ is kept.

Note that (4) belongs to the class of generalized multi-knapsack problems, well-known as NP-complete. Its solution requires an exhaustive search over all the discrete parameters, therefore, a heuristic approach, which splits (4) into two sub-problems, is convenient. Moreover, we remark that the size of codebook $\mathcal{C}_{\boldsymbol{\Phi}}$ may be extremely large, up to the case $|\mathcal{C}_{\boldsymbol{\Phi}}| = 2^{b_{\mathrm{I}}N_{\mathrm{I}}}$, i.e., all the combinations of phase shifts, thus exacerbating the problem complexity.

A. *Optimization Problem Decomposition*

To simplify (4), we first decompose the joint resource allocation and configuration assignment into two sub-problems named the *configurations assignment* and the *RB assignment*, respectively.

The configuration assignment sub-problem assigns one UE per cluster, leaving $K-Z$ UEs unassigned, and sets the IRS configuration for each cluster. The problem can be written as

$$\max_{\substack{\boldsymbol{\Phi}_z \\ x_{k,z,i}}} \sum_{z=1}^{Z}\sum_{i=1}^{F} R_k(\boldsymbol{\Phi}_z, f_i)x_{k,z,i} \tag{5a}$$

$$\text{s.t.} \;\; \text{(4b)}, \text{(4c)}, \; \textstyle\sum_{k=1}^{K}\sum_{i=1}^{F} x_{k,z,i} = 1 \quad \forall z. \tag{5b}$$

In the RB assignment sub-problem, instead, the remaining UEs are assigned to the clusters defined with (5) as

$$\max_{\substack{\alpha_z \\ x_{k,z,i}}} \sum_{z=1}^{Z}\sum_{k=1}^{K}\sum_{i=1}^{F} R_k(\boldsymbol{\Phi}_z, f_i)x_{k,z,i}, \;\; \text{s.t. (4c), (4d), (4e).} \tag{6}$$

Note that both the RB assignments and the cluster cardinality constraint (4e), are assessed in this second step, as α_z, $z = 1,\ldots,Z$ are optimization variables. Moreover, (6) is still a multi-knapsack assignment problem with variable knapsack capacities, therefore belonging to the class of NP-complete problems.

B. *Greedy Maximum-Rate Scheduler (GMAX)*

To reduce the complexity of (4), we resort to a greedy approach, denoted as Greedy MAXimum-rate scheduler (GMAX) algorithm, summarized in Algorithm 1.

First, we observe that (5) can be solved by exhaustively computing $R_k(\boldsymbol{\Phi}, f_i)$ for all $i = 1,\ldots,F$, $\boldsymbol{\Phi} \in \mathcal{C}_{\boldsymbol{\Phi}}$, and $k = 1,\ldots,K$, and then selecting the Z UEs (with their IRS configuration) providing the highest rate. Each of the selected UEs is assigned to the RB maximizing (2), respectively.

To handle the remaining UEs and solve (6), instead, we resort to a greedy approach. Let $\mathcal{P} = \{\boldsymbol{\Phi}_1, \boldsymbol{\Phi}_2, \ldots, \boldsymbol{\Phi}_Z\}$ be the set of IRS configurations of each cluster, GMAX solves

$$(k,z,i) = \operatorname*{argmax}_{k,z,i} R_k(\boldsymbol{\Phi}_z, f_i) \tag{7a}$$

$$\text{s.t.} \quad \text{(4e)}, \; \boldsymbol{\Phi}_z \in \mathcal{P}, \tag{7b}$$

$$k \in \{k : x_{k,z,i} = 0 \; \forall\, z, i\}, \tag{7c}$$

$$(z,i) \in \{(z,i) : x_{k,z,i} = 0 \; \forall\, k\}. \tag{7d}$$

Algorithm 1 Greedy Maximum-Rate Scheduler

```
Input: $R_k(\boldsymbol{\Phi}, f_i)$ for all $k, i, \boldsymbol{\Phi} \in \mathcal{C}_{\boldsymbol{\Phi}}$
Output: $\mathcal{P}, x_{k,z,i}$, for all $k, z, i$
$\alpha_z \leftarrow 1$,    for all $z$
$x_{k,z,i} \leftarrow 0$,    for all $k, z, i$
$(x_{k,z,i}, \boldsymbol{\Phi}_z) \leftarrow$ solve (5) exhaustively
while $\sum_{z=1}^{Z}\sum_{k=1}^{K}\sum_{i=1}^{F} x_{k,z,i} < K$ do
    while $\sum_{z=1}^{Z}\sum_{k=1}^{K}\sum_{i=1}^{F} x_{k,z,i} \le F\sum_{z=1}^{Z}\alpha_z$ do
        $x_{k,z,i} \leftarrow 1$    for $k, i$ solving (7)
    end while
    if $\sum_{z=1}^{Z}\sum_{k=1}^{K}\sum_{i=1}^{F} x_{k,z,i} < K$ then
        $(k, z, i) \leftarrow$ solve (7) neglecting constraint (4e)
        $\alpha_z \leftarrow \alpha_z + 1$
        $x_{k,z,i} \leftarrow 1$
    end if
end while
```

Since $Z \le \frac{K}{F}$ in general, the UEs are firstly allocated considering only F RBs per cluster, i.e., one-time slot per IRS configuration, by setting $\alpha_z = 1$ for all z. Once the first ZF UEs are allocated, the algorithm proceeds by solving problem (7), considering the allocation of new time slots in the resource grid (i.e. increasing α_z by one).

We remark that neither the exact nor the estimate of matrices $\boldsymbol{G}_k(f)$ and $\boldsymbol{H}(f)$, for all f, k are required for the scheduling task itself. Indeed, GMAX relies on iterative computations of (2), which only depends on the cascade product $\boldsymbol{G}_k(f)\boldsymbol{\Phi}\boldsymbol{H}(f)$, for all f, k and each codebook entry $\boldsymbol{\Phi}$.

At the end of the procedure, each UE is assigned to a specific RB, satisfying all the constraints of problem (4). As per (3), sets $\mathcal{U}_1, \ldots, \mathcal{U}_Z$, and $\mathcal{S}_i$ are uniquely determined by variables $x_{k,z,i}$, for $z = 1\ldots, Z$, $i = 1,\ldots,F$, $k = 1,\ldots,K$.

C. *Codebook Design And Control Overhead*

To obtain the cell-specific codebook of IRS configurations $\mathcal{C}_{\boldsymbol{\Phi}}$, a clustering-based approach is employed. Similar to the distance-based clustering proposed in [15], the points to cluster are the IRS phase shifters (with b_{I}-bits quantization) that maximize the achievable rate (2), at each $f \in \mathcal{F}$, of M UEs deployed at random positions in the cell, with $M \gg K$. Such configurations are grouped into $|\mathcal{C}_{\boldsymbol{\Phi}}| \ll M$ clusters, according to the well-known K-means (KM) clustering [21], and the codebook entries are the resulting cluster centroids.

The codebook allows a substantial reduction of the IRS control link overhead. Indeed, for each IRS reconfiguration, the gNB transmits $b_{\mathrm{q}} = \log_2|\mathcal{C}_{\boldsymbol{\Phi}}|$ bits, instead of the $b_{\mathrm{I}}N_{\mathrm{I}}$ bits needed to configure each phase shifter individually. Moreover, by further limiting the number of reconfigurations per time frame to Z, the total number of control bits is reduced by a factor $\frac{ZF}{K} \le 1$.

D. *Computational Complexity*

The computational bottleneck of GMAX is the maximum rate evaluation for the initial choice of the Z IRS configurations to solve (5). Specifically, the cascade channel matrix product $\boldsymbol{G}_k\boldsymbol{\Phi}_k\boldsymbol{H}$ dominates computations with a complexity $O\big(N_{\mathrm{g}}N_{\mathrm{I}}^2 + N_{\mathrm{g}}N_{\mathrm{I}}N_{\mathrm{U}}\big)$, and the procedure must be done for all UEs, carrier frequencies, and IRS configurations in the codebook. Similarly, the second loop computes R_k in the same fashion, but the search is restricted to set $\mathcal{P}$, and typically $|\mathcal{P}| = Z \ll 2^{b_{\mathrm{q}}}$. As a result, the overall complexity of GMAX is $O\big(ZF(2^{b_{\mathrm{q}}}+1)(N_{\mathrm{g}}N_{\mathrm{I}}^2 + N_{\mathrm{g}}N_{\mathrm{I}}N_{\mathrm{U}})\big)$. Note that, in the first step the complexity grows exponentially with the codebook overhead, penalizing codebooks of large resolution. This suggests the adoption of a cell-specific codebook to maximize the rate with low overhead.

TABLE I
AVERAGE SUM RATE FOR DIFFERENT IRS SIZES [bit/s/Hz]

	$b_{\mathrm{q}} = 12$	$b_{\mathrm{q}} = 14$	$b_{\mathrm{q}} = 16$	$b_{\mathrm{q}} = N_{\mathrm{I}}$
10H×20V	12.89	19.06	20.92	21.36
20H×40V	44.65	69.97	78.33	92.16
30H×60V	86.90	108.33	131.65	162.19

However, a further complexity reduction can be achieved by observing that, in the first loop, only the optimal IRS configuration of each UE, i.e., the one maximizing its transmission rate, is needed. A possible approach is to derive the optimal IRS configuration $\mathbf{\Phi}_k^{'*}(f_i)$ for all k, i in the continuous phase domain. $\mathbf{\Phi}_k^{'*}(f_i)$ is then mapped to the closest (in the sense of circular distance [15]) codeword in the codebook $\mathbf{\Phi}_k^* \in \mathcal{C}_{\mathbf{\Phi}}$. While the time complexity of deriving $\mathbf{\Phi}_k^{'*}(f_i)$ for each k, i is $O\big(N_{\mathrm{g}}N_{\mathrm{I}}^2 + N_{\mathrm{g}}N_{\mathrm{I}}N_{\mathrm{U}}\big)$ [15], its approximation requires $O(2^{b_{\mathrm{q}}}N_{\mathrm{I}})$ operations. Thus, the total complexity can be reduced to $O\big(K(2^{b_{\mathrm{q}}} + (N_{\mathrm{g}}N_{\mathrm{I}}^2 + N_{\mathrm{g}}N_{\mathrm{I}}N_{\mathrm{U}}N_{\mathrm{I}})) + ZF(N_{\mathrm{g}}N_{\mathrm{I}}^2 + N_{\mathrm{g}}N_{\mathrm{I}}N_{\mathrm{U}})\big)$.

IV. NUMERICAL RESULTS

We consider a urban micro-cell (UMi) 3GPP scenario [22], with all devices lying in the 2-D plane with the gNB placed at the center. According to the 3GPP specifications, the coverage area of the gNB is characterized by an average radius of 167 m and is assumed to lie in the positive x-axis region. We consider $K = 90$ UEs are randomly deployed according to a uniform distribution within the cell area, to be served in downlink by the gNB, assisted by an IRS at coordinates $(75, 100)$ m. The gNB and the UEs are equipped with uniform linear array (ULA) with $N_{\mathrm{g}} = 32$ and $N_{\mathrm{U}} = 4$ antennas, while for the IRS, if not otherwise specified, we adopt a 20H×40V reflective panel ($N_{\mathrm{I}} = 800$), $b_{\mathrm{I}} = 1$ phase shift quantization bits, and $b_{\mathrm{q}} = 14$ bits for the codebook overhead.

Channel. The system operates at a central frequency $f_{\mathrm{c}} = 28$ GHz, the gNB transmission power is 33 dBm, and the noise power density at the receivers is -174 dBm/Hz. Carriers are uniformly spaced in (f_{c}−10 MHz, f_{c}+10 MHz). We employ the 3GPP TR 38.901 spatial channel model [22], wherein channel matrices are computed based on the superposition of different clusters, each consisting of multiple rays that arrive (depart) to (from) the antenna arrays with specific angles and powers. The link between gNB and UEs experiences deep blockage, while we consider a LoS link between the gNB and the IRS; the channels between the IRS and the UEs may exhibit a LoS component depending on the distance, according to [22].

The performance is evaluated in terms of *average sum rate*

$$\bar{R} = \mathbb{E}\left[\sum_{z=1}^{Z}\sum_{k=1}^{K}\sum_{i=1}^{F} R_k(\mathbf{\Phi}_z, f_i)x_{k,z,i}\right], \tag{8}$$

where we average over multiple channel realizations and randomly generated UEs positions.

A. Compared Solutions

In the following, we compare GMAX with different resource allocation policies, under different codebook sizes.

Deterministic allocation (DA). A baseline scheduling where each UE is directly assigned to an RB in cluster z and, upon the assignment, the IRS configuration $\mathbf{\Phi}_z \in \mathcal{C}_{\mathbf{\Phi}}$ maximizes the cluster sum rate.

Unconstrained capacity-based clustering (UOSCBC). This is an extension to OFDMA scheduling of the one-shot capacity-based clustering (OSCBC) proposed in [15] for TDMA. The unique assignment to a particular RB, i.e., constraint (4d) is violated, as there is no limitation imposed in the number of UEs associated with each RB.

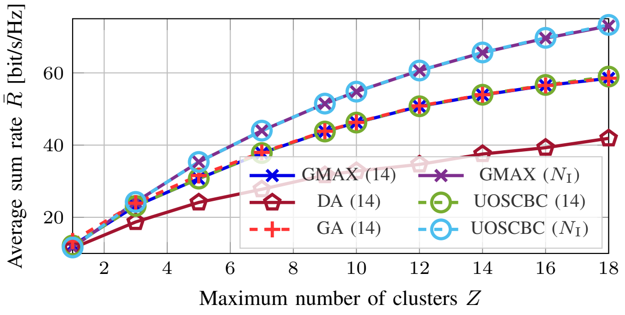


Fig. 3. Average sum rate versus the number of clusters, for $K = 90$ UEs, $F = 5$ carriers. Between brackets is the number b_{q} of bits in the codebook.

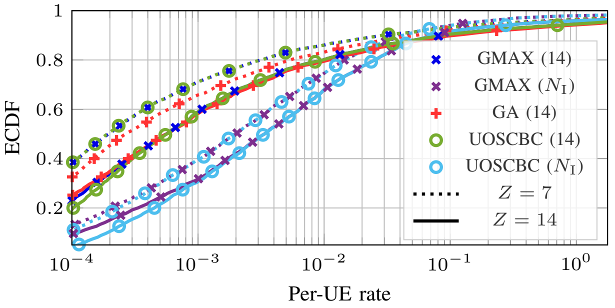


Fig. 4. ECDF of the per-user rate, for $K = 90$ UEs, $F = 5$ carrier frequencies, and $Z \in \{7, 14\}$. Between brackets is the number b_{q} of bits in the codebook.

Genetic algorithm (GA). This is a GA [23] with fitness function (4a), whose initial population includes the GMAX solution. In such GA approach the population generation, crossover, and mutation functions are customized such that all the constraints (4b)-(4e) are always satisfied. This provides the (almost) optimal solution of problem (4).

B. Performance Results

Firstly, Table I shows the relationship between codebook size and system sum rate in the ideal case with F=1 and each UE scheduled with its optimal IRS configuration $\mathbf{\Phi}_k^*$. The results reveal the need for a large codebook to approximate the continuous case (i.e., $b_{\mathrm{q}} = N_{\mathrm{I}}$ for $b_{\mathrm{I}} = 1$). Also, larger IRS panels are more sensitive to the codebook size, as a result of the higher degree of freedom provided by the independent control of each phase shifter. E.g., a 10H×20V-element IRS achieves around 60% of the sum rate achievable with the continuous codebook with only $b_{\mathrm{q}} = 12$, and 98% for $b_{\mathrm{q}} = 16$. Instead, a 30H×60V-element IRS requires $b_{\mathrm{q}} = 16$ to reach 81% of the sum rate achievable in the continuous case.

Fig. 3 depicts the average sum rate as a function of the number Z of clusters, comparing the different clustering strategies. Since each UE must be scheduled once in the resource grid, Z is bounded by K/F. The results show a huge performance gap between the proposal and the DA baseline and highlight the huge performance degradation due to the codebook resolution compared to the slight impact of the frequency assignment constraints (4d)-(4e). In particular, GMAX and UOSCBC with high-resolution codebook ($b_{\mathrm{q}} = N_{\mathrm{I}}$) show a substantial sum rate gap with their respective low-resolution codebook case ($b_{\mathrm{q}} = 14$), while the negligence of constraints (4d)-(4e) with UOSCBC does not provide any substantial benefit on the performance. The proposed GMAX scheduler performs very close to the GA, which is shown to deviate very slightly from the greedy

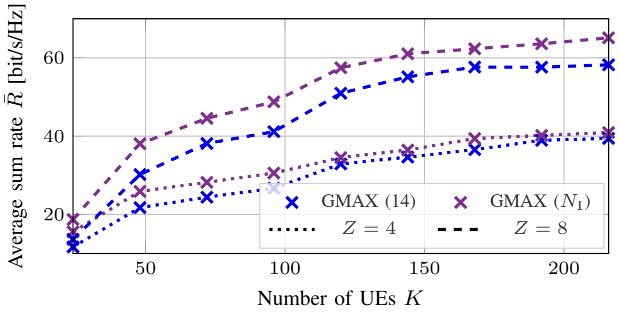


Fig. 5. Average sum rate versus the number of UEs, for $F = 3$ carrier frequencies, and $Z \in \{4, 8\}$. Between brackets is the number b_q of bits in the codebook.

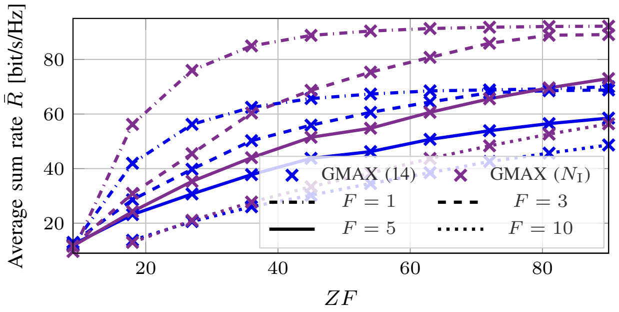


Fig. 6. Average sum rate versus ZF, for $K = 90$ UEs, $F \in \{1, 3, 5, 10\}$ carrier frequencies. Between brackets is the number b_q of bits in the codebook.

solution. Even though the GA approach is not always optimal, such a negligible gap is representative of the validity of GMAX in this context. To emphasize the performance gap between the compared schemes, Fig. 4 shows the empirical cumulative distribution function (ECDF) of the per-UE rate for fixed numbers of clusters $Z \in \{7, 14\}$. While the performance hierarchy remains invariant for almost all compared schemes, the adoption of the sum rate as the fitness function of GA may result in a different rate distribution than GMAX, promoting the UEs experiencing the best channel conditions while penalizing the others.

Fig. 5 shows the average sum rate vs K, for $F = 3$ carrier frequencies. While the sum rate increases with K, for low numbers of clusters ($Z = 4$) the performance gap between GMAX with the low-resolution codebook and GMAX with $b_\mathrm{q} = N_\mathrm{I}$ becomes negligible for a large number of UEs, as the configurations associated to each cluster are sub-optimal in maximizing the sum rate in both cases.

Finally, to analyze the impact of the number of carrier frequencies, Fig. 6 shows the sum rate as a function of ZF. Since $1 \leq Z \leq KF$, the best performance is achievable for fewer carriers, allowing for more frequent reconfigurations. Moreover, it is shown that for large Z the cases $F = 1$ and $F = 3$ exhibit very similar performance. This peculiar behavior is a direct consequence of the considered UMi cell, as $\sim 33\%$ of the UEs on average exhibit a LoS channel component. The channel gain experienced by such UEs is significantly larger than the gains of the UE in non-line-of-sight (NLoS). For $Z = K/3$, such UEs are allocated in different clusters and their optimal configurations are therefore chosen to serve their respective clusters. Thus, $Z = K/3$ is enough to obtain a high sum rate performance.

V. Conclusions

We have discussed the OFDMA downlink scheduling in an IRS-assisted multi-user MIMO system, considering a limited number of IRS reconfigurations per time frame and a discrete codebook of possible configurations. We have tackled the sum rate maximization as an NP-complete generalized multi-knapsack problem, proposing a heuristic solution for the joint IRS configuration and resource allocation and showing its effectiveness in guaranteeing close-to-maximum sum rate compared to a GA-based approach.

References

[1] C. Pan *et al.*, "An overview of signal processing techniques for RIS/IRS-aided wireless systems," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 883–917, May 2022.

[2] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 760–764, Jun. 2020.

[3] J. Lee, J. Choi, and J. Kang, "Harmony search-based optimization for multi-RISs MU-MISO OFDMA systems," *IEEE Wireless Commun. Lett.*, vol. 12, no. 2, pp. 257–261, Feb. 2023.

[4] Y. Guo, Z. Qin, Y. Liu, and N. Al-Dhahir, "Intelligent reflecting surface aided multiple access over fading channels," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 2015–2027, Mar. 2021.

[5] A. Bansal, K. Singh, B. Clerckx, C.-P. Li, and M.-S. Alouini, "Rate-splitting multiple access for intelligent reflecting surface aided multi-user communications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9217–9229, Sep. 2021.

[6] B. Zhuo *et al.*, "Partial non-orthogonal multiple access: A new perspective for RIS-aided downlink," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2395–2399, Nov. 2022.

[7] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz, "Realizing wireless communication through software-defined hypersurface environments," in *Proc. IEEE WoWMoM*, 2018.

[8] R. Flamini *et al.*, "Towards a heterogeneous smart electromagnetic environment for millimeter-wave communications: An industrial viewpoint," *IEEE Trans. Antennas Propag.*, vol. 70, no. 10, pp. 8898–8910, Oct. 2022.

[9] M. Rossanese, P. Mursia, A. Garcia-Saavedra, V. Sciancalepore, A. Asadi, and X. Costa-Perez, "Designing, building, and characterizing RF switch-based reconfigurable intelligent surfaces," in *Proc. ACM WiNTECH*, 2022.

[10] L. Yezhen, R. Yongli, Y. Fan, X. Shenheng, and Z. Jiannian, "A novel 28 GHz phased array antenna for 5G mobile communications," *ZTE Communications*, vol. 18, no. 3, pp. 20–25, 2020.

[11] Q.-U.-A. Nadeem, A. Kammoun, A. Chaaban, M. Debbah, and M.-S. Alouini, "Asymptotic max-min SINR analysis of reconfigurable intelligent surface assisted MISO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7748–7764, Dec. 2020.

[12] V. Jamali, G. C. Alexandropoulos, R. Schober, and H. V. Poor, "Low-to-zero-overhead IRS reconfiguration: Decoupling illumination and channel estimation," *IEEE Commun. Lett.*, vol. 26, no. 4, pp. 932–936, Apr. 2022.

[13] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Capacity and optimal resource allocation for IRS-assisted multi-user communication systems," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3771–3786, Jun. 2021.

[14] 3GPP, "5G; NR; Physical channels and modulation," TS 38.211 (Rel. 16), 2020.

[15] A. Rech *et al.*, "Downlink TDMA scheduling for IRS-aided communications with block-static constraints," in *Proc. IEEE WCNC WKSHPS*, 2023.

[16] F. Guidolin, L. Badia, and M. Zorzi, "A distributed clustering algorithm for coordinated multipoint in LTE networks," *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 517–520, Oct. 2014.

[17] W. R. Ghanem, V. Jamali, M. Schellmann, H. Cao, J. Eichinger, and R. Schober, "Optimization-based phase-shift codebook design for large IRSs," *IEEE Commun. Lett.*, vol. 27, no. 2, pp. 635–639, Feb. 2023.

[18] 3GPP, "5G; NR; Physical layer procedures for data," TS 38.214 (Rel. 16), 2020.

[19] Z.-Q. He and X. Yuan, "Cascaded channel estimation for large intelligent metasurface assisted massive MIMO," *IEEE Commun. Lett.*, vol. 9, no. 2, pp. 210–214, Feb. 2020.

[20] J. Rains, A. Tukmanov, Q. Abbasi, and M. Imran, "RIS-enhanced MIMO channels in urban environments: Experimental insights," in *Proc. EuCAP*, 2024.

[21] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

[22] 3GPP, "5G; Study on channel model for frequencies from 0.5 to 100 GHz," TS 38.901 (Rel. 16), 2020.

[23] J. H. Holland, "Genetic algorithms," *Scientific american*, vol. 267, no. 1, pp. 66–73, 1992.